Reduktion und Elimination in Philosophie und den Wissenschaften

Reduction and Elimination in Philosophy and the Sciences

Alexander Hieke
Hannes Leitgeb
Hrsg.

Beiträge
Papers

31. Internationales Wittgenstein Symposium
31st International Wittgenstein Symposium

Kirchberg am Wechsel
10. - 16. August 2008

31

Reduktion und Elimination in Philosophie und den Wissenschaften
Reduction and Elimination in Philosophy and the Sciences

Beiträge der Österreichischen Ludwig Wittgenstein Gesellschaft
Contributions of the Austrian Ludwig Wittgenstein Society

Band XVI
Volume XVI

Reduktion und Elimination in Philosophie und den Wissenschaften

Beiträge des 31. Internationalen
Wittgenstein Symposiums
10. – 16. August 2008
Kirchberg am Wechsel

Band XVI

Herausgeber
Alexander Hieke
Hannes Leitgeb

Gedruckt mit Unterstützung der
Abteilung Kultur und Wissenschaft
des Amtes der NÖ Landesregierung

Kirchberg am Wechsel, 2008
Österreichische Ludwig Wittgenstein Gesellschaft

Reduction and Elimination in Philosophy and the Sciences

Papers of the 31st International
Wittgenstein Symposium
August 10 – 16, 2008
Kirchberg am Wechsel

Volume XVI

Editors
Alexander Hieke
Hannes Leitgeb

Printed in cooperation with the
Department for Culture and Science
of the Province of Lower Austria

Kirchberg am Wechsel, 2008
Austrian Ludwig Wittgenstein Society

Distributors

Die Österreichische Ludwig Wittgenstein Gesellschaft
The Austrian Ludwig Wittgenstein Society

Markt 63, A-2880 Kirchberg am Wechsel
Österreich/Austria

ISSN 1022-3398

Visuelle Gestaltung: Sascha Windholz
Druck: Eigner Druck, A-3040 Neulengbach

Inhalt / Contents

Formal Mechanisms for Reduction in Science

Terje Aaberge, Sogndal, Norway

1. Introduction

There is a well known story about Victor Hugo who after having submitted *Les miserables* to his editor, went on holiday. He was anxious to know about its reception however, and sent the editor a telegram with the single sign "?". Shortly thereafter he received the response "!" from the editor (Gion 1989). Clearly both telegrams carried a meaning for the receivers. The reason was the existence of the common context determined by the particular situation in which the messages could be interpreted.

The story exemplifies the difference between data and information and how sufficient background knowledge makes it possible to interpret data and turn them into information. The background knowledge defines a context in which to interpret the data. There are two mechanisms for this, either the condition of coherence imposes an interpretation or the context already contain definitions of the data. In any case, the story indicates that if the context is rich then the amount of data needed to describe a state of affairs is smaller than if the context is poor. It thus gives a clue to a preliminary definition of *reduction* with respect to context: a reduction of a context is an enrichment of the context.

In a formal linguistic setting a context is represented by an ontology, *i.e.* a set of implicit definitions of the words of the vocabulary used to describe the domain in question. The ontology provides the formal language with a semantic structure that pictures structural properties of its domain of application. The ontology in itself does not furnish the language with a full semantic. It must be supplemented by an interpretation that relates some of terms of the ontology to external 'objects', *i.e.* objects of its domain of application. The other terms are then given meaning by the definitions. A choice of terms whose interpretation is a sufficient basis for the semantic of a language are said to be primary. All the other terms are defined by the primary terms by means of the definitions. The definitions that only contain primary terms are called axioms (Blanché 1999). An ontology can thus be considered to be constituted by an axiom system or axiomatic core providing implicit definitions of the primary terms and a set of terminological definitions of the additional vocabulary.

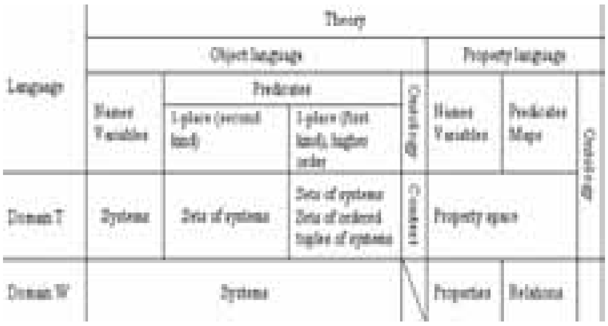
An axiom system for the ontology resumes the syntactic and semantic information in the ontology. It is minimal with respect to both. The syntactic structure represented in the axiom system permits the deduction of all the theorems of the theory and the interpretation of the primary terms gives meaning to the terms introduced by the terminological definitions.

The language is used to describe objects or systems of the domain. The data necessary for a complete description of a system depends on the information content of the axiom system of the ontology. An extension of the system and thus of the ontology provides more information. Accordingly, an extension of the axiomatic system is a formal expression for reduction.

There are two kinds of reductions, ontological and theoretical reduction. Examples of both will be discussed in the following, however, limited to the case of formal scientific languages. By formal I will mean a language whose syntax is provided by first order predicate logic.

2. Structure of a formal scientific language

Any exposition of the structure of scientific theories is based on a number of distinctions representing ontological commitments. Those I have chosen are partly exhibited in the following figure:



<table>
<tr><td rowspan="4">Language</td><td colspan="7">Theory</td></tr>
<tr><td colspan="4">Object language</td><td colspan="3">Property language</td></tr>
<tr><td rowspan="2">Names Variables</td><td colspan="2">Predicates</td><td rowspan="2">Ontology</td><td rowspan="2">Names Variables</td><td rowspan="2">Predicates Maps</td><td rowspan="3">Ontology</td></tr>
<tr><td>1-place (second kind)</td><td>1-place (first kind), higher order</td></tr>
<tr><td>Domain T</td><td>Systems</td><td>Sets of systems</td><td>Sets of systems
Sets of ordered tuples of systems</td><td>Context</td><td colspan="2">Property space</td></tr>
<tr><td>Domain W</td><td colspan="3">Systems</td><td></td><td>Properties</td><td>Relations</td><td></td></tr>
</table>

Figure 1

Here Domain W and Domain T stand for two different perceptions of reality; the Domain W corresponds to logical atomism and Domain T to the more elaborate set theoretical conception. The Figure 1 does not fully represent the relations between Language and Domain. It must complemented by the following diagram,

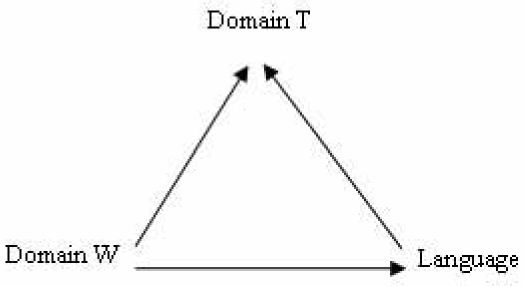


Figure 2

expressing the two interpretations of the correspondence between the structure of language and the reality: that the structure of reality is projected onto language or that the structure of language is projected onto reality. These interpretations are reflected in Wittgenstein's (Wittgenstein 1961) and Tarski's (Tarski 1944, 1985) semantic theories respectively: Wittgenstein's semantic is represented by maps from the domain to language, while the Tarskian semantic is defined by a map from the language to the domain.

In a science there is a need to quantify over systems and properties, however, not both at the same time. Thus, the *a priori* second order language is naturally represented by a juxtaposition of two first order languages, the Object Language (OL) and the Property Language (PL). OL serves to give empirical descriptions of the systems of the domain and PL serves to describe the properties of the systems and to formulate models of systems.

They are both endowed with semantic structures defined by ontologies. Their vocabularies consist of the logical constants and three kinds of terms, the names, variables and predicates, each kind having a particular syntactic role. A name refers to a unique system or property, a predicate to a property (predicate of the first kind) or a category of systems or properties (predicate of the second kind), or a relation between systems or properties. A variable refers to any of the elements in a given category. There is no syntactic difference between predicates of the first and second kind; the distinction is semantic. It is based on the ontological distinction between system and properties. A system is observed and thus conceived as a bundle of properties possessed by the system (bundle theory of substance).

The distinction between the two languages captures scientific practises. In OL the systems are directly referred to, while in PL the reference is indirect; it is given by means of identifying properties that are possessed by the system. Thus, while in OL Newton's second law is expressed by

> *the acceleration of a body equals the net force acting on the body divided by its mass*

in PL the same law is represented by the mathematical formula

$$a = F/m$$

which is without any explicit reference to the *body*. “Body” is not a term in PL. The *body* in question is implicitly referred to by the mass *m* that denotes a property of the body (system).

Figure 1 indicates that the set of properties/relations is represented by an abstract property space in the PL. In this language the relation between the property space and the names of the properties are also included. They are represented by maps that simulate the observation of properties. For example, the set of possible locations in real space is represented by the points of abstract three dimensional Euclidean space and the names of the points by their co-ordinates. This relation is formally represented by a map that relates the points of the abstract space with their co-ordinates. The ontology of the property language incorporates these relations. In the property language it is thus also possible to simulate the act of observation.

A model of a system is a representation of the system in the property language. From the model we can extract a description of the system modelled. The degree of correspondence between the empirical description in the object language and the theoretical description in the property language determines the correctness of the model.

3. Object language and ontological reduction

A domain consists of a set of (physical) systems that possess properties and relations. A system is uniquely identified and described by the properties it possesses. This is done by means of the atomic sentences that attach properties to the system, *i.e.* they are concatenations of the name of the system and the predicates that refer to the properties of the system. The basis for such a description is logical atomism. Each atomic sentence stands for an atomic fact. The conjunction of atomic sentences that applies to a system provides a description or *picture* of the system and serves to distinguish it from the descriptions of other systems.

Some properties are mutually exclusive in the sense that they cannot simultaneously be possessed by a given system; for example, a system cannot at the same time be red and green. This relation of exclusiveness of properties serves to categorise the predicates of the first kind. Each such category is then the range of a map from the set of systems of the domain to the predicates of the first kind. The map, called an observable, relates systems to the predicates denoting properties. Colour is thus an observable. Other examples of observables are form, temperature, position in space, mass, velocity etc.

One distinguishes between two kinds of observables referring to two kinds of properties, properties that do not change in time and thus serves to identify the system, and properties that change. The corresponding observables are identification and state observables respectively. The state properties form a space called the state space of the systems.

The systems can be classified with respect to the identification observables. One starts with one of the observables and uses its values to distinguish between the systems to construct classes. Thus, one gets a class for each value of the observable, the class of systems that possess the particular property, *e.g.* the class of all red systems, the class of all green systems *etc*. The procedure can be continued recursively until the set of identification observables is exhausted. The result is a hierarchy of classes with respect to the set inclusion relation. The basic entities of the classification are the elements of the leaf classes. The discovery of new independent observables will then lead to a refined classification and create new leaf classes and thus new classes of basic entities.

The classes are referred to by predicates of the second kind which thus are ordered naturally in a taxonomy that constitute a linguistic representation of the classification. The taxonomy together with the definitions of the classes is an ontology for the object language. The class definitions impose a semantic structure that mirrors the class inclusion relations and create semantic relations between the predicates. An extension of an ontology due to a refined classification is thus an example of an *ontological reduction*. Moreover, the domain of application of the new language is extended to incorporate the new systems to which some properties of the old systems can be referred. The axiom system for the ontology is given by the definitions of the leaf classes.

An example of a classification is that of material substances. They can be classified in terms of their chemical properties. In particular, the pure chemical elements are given by the periodic table. Taking into account the physical properties however, we get a refined classification distinguishing between isotopes of the same kinds of atoms.

The classification hierarchy can be given a mereological interpretation, *i.e.* the elements of the different classes may be identified by their composition in terms of elementary constituents (Smith et al. 1994). The passage from one level of granularity in terms of elementary constituents to a finer one which in the example above going from the atoms of the periodic table to the constituents of atoms (electrons, protons and neutrons) is an example of ontological reduction.

4. Property language and theoretical reduction

Physics offers many examples of theoretical reduction. We will consider one from classical mechanics. It has several equivalent formulations of which we will discuss two, the Newtonian and Hamiltonian mechanics.

The structure of Newtonian mechanics is defined by a set of axioms covering

Euclidean space and time (abstract)

Action of the Galilei group

Operational definitions of velocity, length and time measures determining coordinatisations

Calculus

Newton's second and third laws

The set of axioms supplemented with terminological definitions constitute an ontology for the property language of Newtonian mechanics.

A model is defined by the specification of a set of equations, the equations of motion. The equations of motion implement Newton's second law and include quantities representing the identification properties of the system modelled and empirical constants, *i.e.* the masses of the objects and the gravitational constant. The solutions, moreover, depend on another set of empirical quantities defining initial conditions.

Hamiltonian mechanics is a formulation of classical mechanics that is a more restrictive way of looking at classical mechanics. It is based on the following elements

Phase space and time as a differential manifold

Action of Galilei group

Operational definitions of momentum, length and time determining coordinatisations

Hamilton's principle of least action

The set of axioms supplemented with terminological definitions constitute an ontology for the property language of Hamiltonian mechanics.

A model of a system is defined by a function on phase space, the Hamiltonian, which includes reference to identification properties of the system modelled. Given the Hamiltonian, the equations of motion are derived from the hypothesis that the dynamics satisfies Hamilton's principle.

The passage from Newtonian mechanics to Hamiltonian mechanics is a theoretical reduction; the axioms of Hamiltonian mechanics impose more structure than those of Newtonian mechanics but at the same time they define a more restrictive theory. The definition of a model is thus more compressed in Hamiltonian mechanics than in Newtonian mechanics. In fact, while the definition of a model of a simple system needs the specification of three functions, the force, in Newtonian mechanics, it is defined by only one function, the energy, in Hamiltonian mechanics. The domain of application of Hamiltonian mechanics is however, smaller than that of Newtonian mechanics. In fact, while Newtonian mechanics can model dissipative systems, Hamiltonian mechanics can only handle conservative systems.

It should be noticed that the terms reduction is also used to denote the limit of physical theories for parameters going to zero.

Literature

Blanché, Robert 1999: L'axiomatique. Paris: Presses Universitaires de France

Gion, Emmanuel 1989 *Invitation à la theorie de l'informatique*, Paris: Éditions du Seuil

Smith, Barry and Casati, Roberto 1994 Naive Physics: An Essay in Ontology, *Philosophical Psychology*, 7/2, pp. 225-244.

Tarski, Alfred 1985 *Logic, Semantic, Metamatematics* (second edition), Indianapolis: Hackett Publishing Company

Tarski, Alfred 1944 The Semantic Conception of Truth and the Foundations of Semantic. *Philosophy and Phenomenological Research* **4**, pp. 341-375

Wittgenstein, Ludwig 1961: Tractatus logico-philosophicus, London: Routledge and Kegan Paul

Wittgenstein on Counting in Political Economy

Sonja M. Amadae, Columbus, Ohio, USA

This paper follows Ludwig Wittgenstein's *Remarks on the Foundations of Mathematics* to investigate the source of the purported necessity delineated in mathematical statements and proofs. It suggests that this "normativity" has a similar structure to that underlying promising, contracting, and political obligation. Whereas many philosophers have abdicated the project of defending that empirical science can yield necessary truths or universal laws,[1] still it is typical that mathematical truths are conceived to be necessary. Therefore the philosopher W.V.O. Quine, although a thorough-going empiricist who attempted to defend mathematics on the grounds of sensory perception, still faced the burden of explaining "why mathematics was (and is) *thought to be* necessary, certain, and knowable a priori."[2] If we understand "normativity" to convey some sort of structural indispensability that may guide judgment and action, then mathematical knowledge represents perhaps the paradigmatic case of a codified, law-like system that embodies non-negotiable relations and claims, that may be intuited by the human intellect.

There is an arresting debate at the foundations of mathematics over whether mathematical objects, or numbers, have an objective existence independent from the mind. To simplify various positions on this question into two varieties, on the one hand are the "realists," who hold that the truth of mathematical statements is externally determinate, even if its status is undecidable within a set theoretic or formal system: "We employ such a conception if we hold that the statement may be determinate in truth-value irrespective of whether we can recognize what its truth-value is."[3]

A second school of mathematics, referred to as anti-realism or intuitionism, accepts that mathematical truths exist only in the mind of mathematicians: they are constructed. Such an acceptance of the imaginative work done by mathematicians would seem to be on par with Wittgenstein's emphasis of the social character of the normativities of counting, calculating, and proving. "Wittgenstein's general treatment of the topic of rule-following entails that the status of a proof, or calculation, is always in need of *ratification*."[4] By this account, human counting practices retain their shape, or consistent patterns, over time not because they are laid down by iron-clad procedural rules, but because we commit ourselves to interpreting and acting on the rules as consistently as our contingent intersubjective context makes possible.

This lack of agreement about the foundation of mathematics, over whether the objects of its investigation actually exist or not, stands in parallel to debates over whether moral systems represent truths independent from the cultures in which they are expressed. There is a symmetry between the assertion of the existence of deontological moral truths, such as the Kantian categorical imperative, and the claim of independent validity of mathematical truths; either case, so far as we know, cannot in principle confirm its verification-transcendent authority. Even if this parallel is striking, it is further apparent that whereas deontology in morals is a position marginalized by mainstream scientific approaches to human behavior, [5] realism in mathematics is the more widely accepted status quo in philosophies of science and math.[6] This realism essentially accepts that humans have "the capacity to grasp a verification-transcendent notion of truth"[7] in matters of mathematics, but doubts the same in matters of morals or ethics. We routinely accept verification-transcendence in mathematics but not in ethics.

Granted this general privileging of the normativity of mathematics as evincing necessary, a priori, yet verification independent, truths, a philosophy of mathematics is called upon to "account for the at least apparent necessity and priority of mathematic[al knowledge]."[8] Indeed, it seems that much of the present-day celebration of scientific naturalism, that casts doubt on the reality of moral and ethical judgment, strives to present a position on mathematics that navigates the notoriously unbridgeable chasm between a priori and a posteriori knowledge. Quine, Hilary Putnam and P. Maddy are leading philosophers who have attempted this line of argumentation, ultimately seeking to preserve the nonnegotiable quality of math while grounding it on knowledge derivable from empirical observation.[9] However, this line of inquiry consistently concedes both that empiricism is irrelevant for the actual practice of mathematics, and that mathematical truth is independent from our procedures of knowing it.[10] Rather, it suggests that mathematics will finally be vindicated in scientific application.[11] Conveniently, Wittgenstein presents an anti-realist philosophy of math, consistent with intuitionism in many of its details and implications, but with the added benefit of not advocating any need to revise mathematical practice.

In exploring the character of mathematics as a language game that perhaps best represents our paradigmatic case of "rule-following," Wittgenstein suggests that the laws of mathematics stand as imperatives and commands, and not as objectively verifiable truth claims: "Mathematical discourse is not fact-stating; its role is rather to regulate forms of linguistic practice."[12] If we distance our understanding of the source of mathematical normativity as flowing from objective objects and relations that exist outside our minds and practices, then we may understand that mathematical statements have the character of declarations,

[1] For example, W.V.O. Quine, for discussion see Shapiro, Thinking About Mathematics, 218,

[2] Shapiro, Thinking About Mathematics, 218.

[3] Crispin Wright, Wittgenstein on the Foundations of Mathematics (Cambridge: Harvard University Press, 1980), 7; even philosophers of mathematics who hold a naturalistic position that ultimately mathematics should be verifiable through scientific (empirical) means, endorses numeric realism: "As a realist [P.] Maddy (1990: cha. 4, ss 5) agrees with Gödel that every unambiguous sentence of set theory has an objective truth-value even if the sentence is not decided by the accepted set theories" (Shapiro, 224).

[4] Wright, Wittgenstein, 128.

[5] Jean Hampton, The Authority of Reason (Cambridge University Press, 1998).

[6] Shapiro, Thinking about Mathematics, "Numbers Exist," 201-225.

[7] Wright, Wittgenstein, 10.

[8] Shapiro, Thinking About Mathematics, 23.

[9] See Shapiro, Thinking About Mathematics, "Numbers Exist," 201-225.

[10] Shapiro, 220, 224.

[11] Shapiro, 220.

[12] Wright, Wittgenstein, 157.

imperatives, or commands in the form of admonishing adherence to rules that we assent to follow. The intuitionist Dummett, whose position Wittgenstein's resembles, refers to mathematical statements as quasi-assertions:

> Quasi-assertions are declarative sentences which are not associated with determinate conditions of truth and falsity but share with assertions properly so-called the feature that there is such a thing as *assenting* to them; where such assent is communally understood as a commitment to some definite type of linguistic or non-linguistic conduct, and receives explicit expression precisely by the making of the quasi-assertion.[13]

The subtle aspect of understanding the distinction between mathematical statements as in principle verifiable against an objective reality, versus having the character of being ratified by voluntarily acceptance, is that although we seek to preserve some sense of non-arbitrary structure, we must locate its apparent "necessity" in our discretionary compliance rather than in some facet of extra-mental reality. This necessity has the form of willingly binding ourselves to a normative correctness that we enact in our practice. Hence we have the sufficient leverage to not only ask "[o]f someone who is trained [in a specific type of rule-following] 'How *will* he interpret the rule in this case?'", but further to raise the question, "How *ought* he to interpret the rule for this case"?[14]

This view of mathematics as having a humanly devised command structure instead of a structure insured by objective reality alters our picture of the type of normative guidance underlying mathematical judgment. Instead of being guided in making mathematical statements by facts, we consider that "all mathematical propositions [are] expressed in the imperative, e.g., 'Let 10 x 10 be 100.'"[15] The significance is that this depiction of mathematics makes the consistency of its structure dependent on our voluntary commitment to uphold conceptual relations in specific ways:

> Such an account is exactly what we should intuitively propose for sentences expressing the making of a promise. No one would ordinarily suppose that the use of sentences of the form, 'I promise to …' is best understood as the making of a statement, true or false; though their being prefixed by 'it is true that …' is grammatical sense.[16]

The promissory quality, then, of mathematical normativity is that mathematical rules suggest what we "ought to conclude," and in participating in these rule-following exercises we accede to draw the conclusion implied by the rule. It is not that some feature of an objective world of numbers intercedes to form the basis of our judgment in a necessary fashion. Rather, in mathematical rule-following, we agree to abide by the rules as prefiguring or commanding our judgment. If we consider the role proofs play in mathematics, "it marks not a discovery of certain objective liaisons between concepts, but something more like a resolution on our part so to involve them in the future."[17]

If our understanding of the normativity structuring apparently necessary truths in mathematics rests on our commitment to follow the rules of mathematics, then it is possible to see that the rule-following nature of math is little different from other rule-following institutions throughout our society. This opens the possibility of considering that social-norms that stand as a system of rules have as much sanctity as do the rules of mathematics. Typically, social norms are regarded as subject to preference; either an individual prefers to follow a social norm or not; if she chooses to follow a social norm, this is because she prefers to do so. However, in the case of mathematical judgment, preference is seldom invoked as a source of decision over the result of a calculation or proof.

This recasting of the foundation, as it were, of mathematics from fact and objective truth to socially constructed and ratified laws suggests the possibility for drawing a parallel between legal systems of rule-following and mathematical systems. In his essay, "The Groundless Normativity of Instrumental Rationality," Donald Hubin argues that neo-Humean instrumentalists "must engage in the same 'lowering of expectations' [of the source of normativity of instrumental rationality to the same level] that the legal positivist must."[18] For Hubin, practical rationality, of which instrumentality is part, is not an objective matter. In making his point, he draws on legal positivism's retreat from natural law theory, and draws on H.L.A. Hart to expand on this view. [19] Hubin is making the point that even though a legal system provides a normative basis for action, it cannot ground its ultimate principles. I am reworking Hubin's parallel between positive law and instrumental reason to contrast a realist account of math with an alternative declarative understanding. In an anti-realist mathematics, the binding quality of rules only holds insofar as we assent to them.

It has traditionally been the case the social and political normativity has been viewed as of a lesser pedigree than instrumental and mathematical normativity insofar as the former is conditional, and the latter is non-negotiable. For example, Phillip Pettit provides an explanation for how social norms may be derived from instrumental agency as the former is conditional on individual rational self interest.[20] In his *Theory of Justice*, John Rawls was widely criticized from within rational choice theory for placing action according the "the reasonable," which included the political theoretic concept of fair play, on par with agency conforming to the dictates of expected utility theory.[21] It was not automatically obvious from within rational choice theory that agents had a duty to uphold the rules of government if they did not further an agent's ends in each and every circumstance of action.[22] Therefore, without some sanctioning device that alters payoffs, the rule of law does not in and of itself provide a reason for action that trumps agents' preferences over end states. Rawls concludes of his contrasting approach to justice as fairness, "There is no thought of trying to derive the content of justice within a

[13] Wright, Wittgenstein, 155.

[14] Ludwig Wittgenstein, Remarks on the Foundations of Mathematics, ed. by G.H. von Wright, R. Rhees, and G.E.M. Anscombe, trans. By G.E.M. Anscombe (Cambridge, MA: MIT Press, 1996) (RFM), V-9, p. 267.

[15] Wittgenstein, RFM, 155.

[15] Ludwig Wittgenstein, RFM, V-17, p. 276.

[16] Wright, Wittgenstein, 157.

[17] Wright, Wittgenstein, 135.

[18] Donald Hubin, "The Groundless Normativity of Instrumental Rationality", The Journal of Philosophy 98:9(2001), 445-468, 466.

[19] Hubin, "Groundless Normativity," 463.

[20] Philip Pettit, "Virtus normativa: Rational Choice Perspectives," in his Rules, Reasons, and Norms (Oxford University Press, 2002), 308-343.

[21] John Rawls, A Theory of Justice (Harvard University Press, 1971); John Rawls, "Justice as Fairness: Political not Metaphysical," Philosophy and Public Affairs, 14:3 (summer, 1985), 223-51.

[22] This is the problem David Gauthier faces in Morals by Agreement (Oxford University Press, 1985).

framework that uses an idea of the rational as the sole normative idea."[23]

I am suggesting that mathematics, in any form, but even more specifically as it is harnessed to anchor all manners of institutions in political economy that depend on "accurate counting" for their functioning, embodies the normativity of Rawls' "reasonable" as opposed to the rational.[24] By Rawls' description, "if the participants in a practice accept its rules as fair, and so have no complaint to ledge against it, there arises a prima facie duty…of the parties to each other to act in accordance with the practice when it falls upon them to comply."[25] Most of us accept the normativity of mathematical rule-following automatically out of habit or a sense of duty. We do not at first perceive that this virtually innate compliance cuts across the grain of the competing, and supposedly more basic, normativity of instrumental agency which recommends counting in one's favor when one can get away with it. In fact, considerations of expected utility do interrupt counting practices in cases of embezzlement, fraud, bribery, and ballot box stuffing. The normativity of counting and calculating represents the logic of appropriateness and not the logic of consequences. Adherence to mathematical rules confines judgment; judgment is not a function of preferences over outcomes.

Counting practices throughout political economy resemble the rule of law insofar as they do not have an independent object or autonomous truth-value separate from the rules constituting them. Although most of us do not actually determine, or even consent to, the rules governing these procedures in banking, insurance, taxation, inheritance, or elections, still there is an evident presumption that one counts in accordance to the rules free from considerations of our obvious interest in the outcomes. Much like Rawls' formulation of "the Reasonable," most of us have been conditioned to accept, or even to reflexively consent to, an inherent necessity of counting in accordance with the rules directing the activity.

[23] Rawls, "Justice as Fairness," 237.

[24] For a discussion of the distinction between the rational and the reasonable in Rawls, see Rawls' "Justice as Fairness," and S.M. Amadae, Rationalizing Capitalist Democracy (Chicago University Press, 2003), 271-3.

[25] Rawls, "Justice as Fairness," 60.

Referential Practice and the Lure of Augustinianism

Michael Ashcroft, Melbourne, Australia

This paper is an examination and defence of Wittgenstein's thesis that language itself promotes an Augustinian picture of its workings. Let us define Augustinianism as the thesis that the meaning of an expression is its referent, and distinguish a strong variant that restricts the referents of expressions to ostensively indicatable material objects. In this paper I will argue that if Wittgenstein is correct about reference talk, linguistic practice tempts us to (incorrectly) adopt both positions. I shall begin by describing a naïve notion of reference. Then I will examine the role of reference in contemporary meaning theories and draw parallels with Wittgenstein's own account in order to elucidate the latter. Finally I will explain why the resulting practices can lead us to accept both forms of Augustinianism, and why these positions are mistaken.

At first blush, Wittgenstein's ‘meaning is use’ thesis seems to offer a simple account of reference. As he noted at *PI* 10:

> What is supposed to shew what [words] signify, if not the kind of use they have?

I take Wittgenstein to accept that, in one sense of ‘refers’ or ‘signifies’, the referential link between a sign and its referent lies in the fact that the rules for some signs use are such that their correct use intimately involves (a) particular ostensively indicatable material entity/entities which are thereby the referent(s) of the sign. It is this sense that captures what I shall term ‘naïve referential practice’.

But, Wittgenstein points out, it is not this sense of reference that motivates the question of what the expressions of his simple language refer to. Since he had explained the use of the expressions he was at that point dealing with, in its naïve sense the question is already answered. Thus, Wittgenstein continues, the question must be a request ‘for the expression “This word signifies *this*” to be made part of the description’ of the expressions use. There must, alongside our naïve referential talk, be a sophisticated variant wherein the uses of expressions are explicated via referential claims. Certainly, even in ordinary language, ‘refers’ has a much broader role than the naïve practice allows. We talk of our expressions referring to abstract objects like numbers, fictional objects like Sherlock Holmes, properties like blue, and many other things besides. The only hypothesis here seems to be that this broader use of ‘refers’ is involved in elucidating the use of expressions. For the purposes of this paper I shall assume this is correct. For what I wish to argue is that it is the way Wittgenstein believed that expressions such as ‘This word signifies *this*’ and ‘This word refers to *this*’ are made part of the description of words’ uses that leads to the conclusion that language itself tempts us to understand it in an Augustinian fashion.

To explain this, let us begin by turning to the role of reference in formal meaning theories. Presuming a Fregean syntax and ignoring complications required to deal with quantifiers, a typical meaning theory attributes semantic values to names and treats predicates as functions from names to the semantic value of sentences – where an expression’s semantic value is that which indicates the contribution the expressions make to the meanings of the sentences it can be part of, whilst a sentence’s semantic value is its meaning. The theory then gives a functional account of the logical connectives which permits the production of semantic values for complex sentences, and lastly (and most problematically) provides a theory for how the use of sentences can be deduced from the semantic values the meaning theory attributes to them. In attributing semantic values to (the sub-sentential expressions the theory parses as) names, the names are said to refer to objects, which, in a deliberately set-theoretic construal of what is going on, we can take to be grouped in the meaning-theory’s domain. The theoretical relation of reference thus introduced can be expanded such that one might also say that definite descriptions and predicates refer to the objects that satisfy them and (possibly empty) sets of objects respectively. The latter case looks very akin to saying that predicates refer to properties, and to assist this exposition let us explicitly accept that properties are sets. In this case, a set-theoretic construal of the quantifiers permits us to understand them as referring to properties (sets) of sets – taking ‘all’ to refer to the property of being identical to the universal set and ‘some’ the property of not being identical to the empty set. Importantly, the single criterion for a successful meaning theory (as a descriptive account of the meanings we do attribute to others) lies in its getting its theorems correct. In the rarefied air of theoretical semiotics, it makes no sense, Davidson pointed out, to complain that a meaning theory comes up with the right theorems time after time, but has the logical form (or deep structure) wrong. [Davidson; 1977] The objects to which an expression refers are therefore not something that can be examined directly, but are determined by the legitimacy of the theorems the referential axioms produce.

One might object that referential axioms are not so thoroughly unconstrained, for they relate singular terms to objects. Therefore only those things that actually exist are kosher referents in the theory. So, for example, since there is no object *Atlantis*, a meaning theory ought not to accept the axiom ‘‘Atlantis’ refers to Atlantis’. One might reply that by the criterion given above what is important is merely that the meaning theory produces the correct theorems. So whilst one could, there is neither need nor justification in restricting the axioms of a meaning theory such that one ought to include as referents only objects one is ontologically committed to. But this reply is too quick. For the objection’s motivation is likely not the given criterion for determining a correct meaning theory, but Quine’s thought that accepting any theory requires ontological commitment to the objects it quantifies over (or, since a theory may be satisfied by models with different domains, it requires existential ontological commitment to there being one such domain). Insofar as, for any singular term of a theory, *t*, the theory implies $(\exists x)(x{=}t)$, a theory’s singular terms refer to objects of its domain of quantification – to objects which we therefore ought to be ontological committed.

There are reasons to object to this claim. But I shall not pursue them here. Let us accept that a theory requires ontological commitment to the objects its quantifiers range over. In the case of a meaning theory, these objects are the semantic values of (expressions parsed as) names. But these objects have not been shown to be the middle-sized dry goods we would, in the aforementioned naïve reference talk, say are the referents of most of the mentioned expressions in the referential axioms. On the contrary, formal semantics is a mathematical discipline: First order set theory. Given the possibility of a set-theoretic construal of formal meaning theories, as well as their historical development from Tarskian model theory, we might think the same is true in their case; or more weakly, we might think it possible to interpret them in this way. If so, then although we owe ontological commitment to the members of a meaning theory's domain, these would, or at least could, be urelements. In which case the axiom ''Atlantis' refers to Atlantis' demands ontological commitment to nothing more than an (existent) urelement, not a (non-existent) continent.

This foray into formal meaning theories casts light on how the expression "This word signifies *this*" can be made part of the description' of the word's use. As in formal meaning theories, so in folk practice: It occurs through reference talk coming to be used to indicate at least certain aspects of the expression's semantic role. This indication can be wider or narrower. We have, for example, numerals in our language that are characterised as referring to natural numbers. They are characterised this way both *en masse*, in that referring to natural numbers is what numerals do, and individually, in that each numeral has a specific natural number it refers to. Presuming the practice does not also describe complex arithmetical equations as referring to numbers (or numbers alone), to say that a person uses a particular expression to refer to a natural number is to indicate that they mean it as a numeral. To indicate that they use it to refer to a particular natural number is to indicate that they give it the same meaning as a particular numeral. Let us assume, as seems plausible, that natural language has the semantic vocabulary – expressions denoting the categories of objects, properties, relations, truth functions, properties of properties, etc, and the means to provide indefinitely many names of the individuals entities of the various categories – to allow us to think of every sub-sentential expression (as parsed in the canonical syntax, which we can assume to be Fregean) as referring to particular referents of a particular category. Let us call these the canonical referents of the language's sub-sentential expressions. This permits information about the meaning a person gives a sub-sentential expression to be expressed by the class of entity that the expression is said to refer to: to learn that someone uses a sub-sentential expression to refer to an object is to discover that they mean it as a name (or definite description), whilst to learn they use it to refer to a property is to find they mean it as a predicate, etc. Referents, via the referential relation, provide a model for language on the basis of referential claims of the form ''*a*' refers to *b*' and 'There is some *x* such that '*A*' refers to an *x*'. To those familiar with the practice, this model categorises the correct use of expressions. Explaining the model a person utilises helps explain the meaning they provide their expressions. Telling others the model they ought to use helps to teach them to use language as we wish them too. Since such a model provides referents that suffice, within the practice, to entirely represent the contribution the expressions make to the meanings of the sentences they can be part of, then knowledge about what a person refers to by an expression will provide knowledge of what the person means by the expression.

It is but a short step from believing our language and canonical syntax permits such a referential practice to thinking we possess the same. Such a sophisticated referential practice would not be redundant. As well as facilitating learning, it permits translations from one language to another; indeed it permits extremely subtle translations that can elucidate the similarities and differences in structures between the two languages (*cf PI* 10). But when applied to one's own language in the presence of competent users the practice idles, it produces trivial substitution instances of the schema ''A' refers to A', or ''A' refers to the property (of) A', etc (perhaps with small amounts of declination or conjugation to produce appropriately reified canonical referents). This is harmless enough, but note that reference is simultaneously important in elucidating meaning and every expression is (given a recursive categorisation system and an ability to provide names for previously undiscussed members of categories) tautologically provided with a referent, and this referent is (also tautologically) the meaning (semantic role) of the expression.

Thus, as in formal meaning theories, saying that an expression possesses a particular referent, or possesses a referent of a particular type, provides information about the expressions' semantic value. (And certainly, Wittgenstein exorcises any concern about the legitimacy of the used expressions on the right of reference claims. We can think of this sophisticated reference talk as a *sui generis* linguistic practice whose utility lies in its creation of this referential model. The objects of this model, which we arguably need to be ontologically committed to, are nothing more than other expressions of the language.) Such, I think, is Wittgenstein's understanding of how expressions such as 'This word refers to *this*' are made part of the description of the use of words.

It is clear how such a linguistic practice lures us towards Augustinianism. For in the sophisticated practice every expression possesses a referent which is, in some sense, the expressions meaning (semantic role). Two points elucidate the lure and problems of the weak and strong Augustinian accounts respectively:

(i) Within sophisticated referential practice, reference talk provides a *model* of the semantic role of expressions in that referential claims *represent*, to those familiar with the practice, the semantic role of expressions. It is a mistake to think that the possession of a referent in this sense *causes* an expression to have a semantic role.

(ii) Within sophisticated referential practice, all expressions possess (their canonical) referents which represent their semantic role. But, as noted, we also naïvely talk about expressions referring in the sense that their correct use intimately involve (a) particular material entity/ entities which are thereby their referent(s). It is a mistake to think that the fact that all expressions possess referents in the sophisticated sense entails that they possess referents in the naïve sense. It is likewise a mistake to think that the referents expressions may possess in the naive sense represent the semantic role of the expression.

The Augustinian account confuses modelling with explaining and, in its strong variety, conflates the naïve concept of reference with the sophisticated. But the ease of these mistakes is why Wittgenstein felt that, given a sophisticated referential practice, our language itself attempts to foist an Augustinian understanding upon us. In searching for what Wittgenstein described as the 'life' of our expressions we immediately confront a picture of meaning provided by a practice wherein the semantic role of expressions is given by their referents. To paraphrase his characterisation, this picture holds us captive. We cannot get outside it, for it lies in our language and languages repeats it to us inexorably. But we can equally see why the Augustinian accounts are mistaken, confusing modelling with explanation and, in the strong case, trading on ambiguity.

Literature

Davidson, Donald, “Reality without reference”, (1977) in his *Inquiries into Truth and Interpretation*, Oxford University Press, 1984, p. 223

Wittgenstein, Ludwig, *Philosophical Investigations*, Basil Blackwell, Oxford, 1963

The Date of Tractatus Beginning

Luciano Bazzocchi, Pisa, Italy

1. Tractatus and Prototractatus

I suggest considering the so-called "Prototractatus" notebook (MS104) not as "an early version of the Tractatus Logico-Philosophicus"[1], but as the effective manuscript of Wittgenstein's book. We know that the ultimate typescript was dictated in August 1918; it's considered a final writing, despite a dozen of later inserted propositions[2]. Well, the MS104 notebook (if we look at the whole of it and not only at its published part) contains in its turn all the material of that typescript, including title, dedication, motto and Preface, with the exception of only five propositions.[3] In particular, the first fifty remarks of the manuscript, the backbone of the work, passed almost unaltered into the final book, and 41 of them maintained also the same decimal number: so we can consider the date of composition of these first pages as the real starting date of the Tractatus itself. Unfortunately, there isn't any agreement on Prototractatus' composition date.

The content similarity between Prototractatus and Tractatus was just what led von Wright in error when he advanced "the conjecture […] that work on the 'Prototractatus' immediately preceded the final composition of the book in summer of 1918" (Wittgenstein 1971, p.9). This may perhaps be true for the last part of the notebook, i.e. pp. 103-120, not edited and not considered "Prototractatus" by von Wright; but most of MS104 was written a long time before. The decisive philological proof was found by McGuinness in 1989, when he published a list taken from the correspondence of Hermine Wittgenstein and dated January 1917. It mentions some of Wittgenstein's manuscripts; the fifth entry of the list ("a large chancery volume, containing the revision of [the first three notebooks] for publication") seems to refer precisely to the Prototractatus notebook. McGuinness argues that at the end of 1916 the notebook was filled at least until page 71, in correspondence with proposition 7 insertion, or perhaps until the end of the successive layer of text at page 78 (McGuinness 2002).

2. Prototractatus first 70 pages

While we can accept his conclusion, it's not so much clear in which circumstances these pages were composed. McGuinness expects that from Hermine's list we can deduce the non-existence of an eventually lost diary connecting the three we have (MS101, 102 and 103 of von Wright's catalog); the period between MS102 and MS103, from June 1915 to March 1916, would instead be dedicated to the Prototractatus compilation, until the line traced at page 70. This hypothesis is very uncertain. The work around the Prototractatus is utterly unlike the work on the diaries. Compared with the structured and formal aim of the *Abhandlung*, their date-ordered arrangement (which is identical before and after the interleaving period: left pages with encoded personal notes, right pages with philosophical free entries) answers to very different needs. Besides, it's likely that an intermediate lost diary existed, as Geschkowski arguments in his book[4]. Finally, it's probable that the third entry of the Hermine's list just refers to this (now lost) notebook, and not to the successive MS103, as McGuinness thought.[5] But from his objections to McGuinness, without any cogent reason, Geschkowski concludes that the first 70 pages of the Prototractatus were filled only in the autumn of 1916, on the basis of a gathering of material on loose sheets.

3. Prototractatus first 28 pages

On the contrary, as I elsewhere discussed (Bazzocchi 2007a and 2007b), I think that the method of composition of the notebook's first layer, until p. 28, reveals a typical first writing, where the decimal numbers play the role of heuristic guide. It seems improbable that the proposition numbers were added later, as instead Geschkowski is forced to assert (if the decimals were present at the moment of the supposed copy from the loose sheets, the notebook wouldn't be in such disorder as it is). We can prove, indeed, that the decimal numbers were in use from the beginning. In fact, the proposition 2.23 in second page was deleted by pencil and transferred to the fourth page under the new decimal 2.181: at the moment of deletion, it already had its number, perfectly coherent with all the others.

In short, I think that the first 28 pages were filled before the letter to Russell of October 22nd 1915, because from the letter we can deduce that some time before *a copy of* the notebook first stratum *into* a "last summary on loose sheets" was made (see Bazzocchi 2006). So we can date this first layer compatibly with McGuinness' dating (although by different reasons), i.e. around 1915 summer.

4. Prototractatus first 12 pages

Now I want to introduce a new argument, that until now critics haven't noticed.

In the diary entry of June 18th 1915, one can find a very baffling passage. In the middle of a long discussion on generality and particularity, illustrated by an example of a picture and its dots, there is an incongruous reflection: "Not: a propositions follows from another one, but: the truth of a proposition follows from the truth of the other". Then the text continues about pictures and dots. What is the sense of this mention? Why such sudden inspiration? One idea is that Wittgenstein, incidentally remembering *some other remark*, decided on a new way of expression and

[1] See the subtitle of (Wittgenstein 1971).

[2] These were added by hand to the typescript, generally on the overleaf of its sheets, during Wittgenstein's permanence in the Montecassino camp.

[3] They are the remarks 3.251 (derived from a note of June 19th 1915), 4.0311 (taken from Nov 4th 1914) and the second paragraph of 4.01 (from Oct 27th 1914): perhaps these were already in a supposed parallel version of MS104 notebook, requested by the so-called 'Korrektur' . Instead, 3.22 (from Dec 29th 1914) and 3.221 (from May 26th and 27th 1915) were added directly during the process of dictation. On the other end, one of the later twelve insertions, the proposition 5.2523, had already appeared as last entry of MS104.

[4] See (Geschkowski 2001), chapter "2.1 Reasons for the existence of a further diary". Passing, I add that the lost diary can be even partially reconstructed. Presumably, Prototractatus pages 79-81 contain selected propositions from its second part in perfect continuity with pages 81-86, that systematically contain all the good propositions of the consecutive MS103 diary.

[5] McGuinness himself argues that at the time the notebook in question had been "in part" reversed in the Prototractatus; but MS103 notebook was exploited only later, starting from page 81 of the manuscript.

hurried to fix it on the page. The remark to be modified is not in the diary. But if we look at page 12 of the Prototractatus notebook (a page that as in McGuinness' as in my hypothesis takes place *around* that period), we find exactly the contested expression: "5.041 In particular *a proposition follows from another one* if all the truth-grounds of the first are truth-grounds of the second"[6]. Well, the remark is emended with the precise insertion of "the truth of": "the truth of a proposition follows from the truth of another one".[7]

Here we have two indubitable facts: on June 18th 1915 Wittgenstein fixed a correction, and at Prototractatus page 12 the same correction took place. There are only two possibilities: *or* first Wittgenstein stated the amendment in abstract, and then the case took place and he corrected it exactly as stated some time before, *or* first he wrote the previous form on the Prototractatus, and then reviewed it and remarked the adjustment on the diary. The first case is very unlikely. It's hard to believe that Wittgenstein decided in abstract such a particular (and indeed not so clear) correction of his thought; that then (a few weeks later, in McGuinness' hypothesis) twice[8] he made just the "mistake" he had already criticized; and that finally he corrected it following a previous such foresighted purely theoretical amendment. The only effective possibility is that the compilation of Prototractatus page 12 *precedes* the discovery of the inaccuracy and its record on the diary. Note that the question does not concern only the wording of propositions 5.04 and 5.041 – that at the time, one may think, could have been recorded on some other slip of paper – but properly Prototractatus page 12, because the correction is unquestionably on it.

Hence we can conclude that the Prototractatus notebook started *before* (and not *after*) the end of the MS102 diary, that in fact contains a reference to its page 12. McGuinness' hypothesis seems to fall off anyway, but onto the opposite side compared to what Geschkowski argued.

5. Prototractatus first page

The Prototractatus compilation was indeed a very slow process, at an average speed of three or four pages a month: the total 120 pages of August 1918 were already 71 as the end of 1916, at least 28 in October 1915, and 12 in June[9]. So we can presume that the starting point was in April or May 1915. In this case, the letter to Russell of May 22nd 1915 assumes a definite sense. In the previous communication to Russell, in November 1914, Wittgenstein said: "If I should not survive the present war, the manuscript of mine that I showed to Moore at the time will be sent to you, along with another one which I have written now, during the war" [Wittgenstein 1974, p. 62]. The second manuscript is evidently the 1914 diary, whose first notebook was completed in October 30th. But in May the reference is quite different: "I'm extremely sorry that you weren't able to understand Moore's note – Wittgenstein writes – Now, what I've written recently will be, I'm afraid, still more incomprehensible. […] If I don't live to see the end of this war, […] you must get my manuscript printed whether anyone understand it or not".

Here Wittgenstein refers to only one coherent manuscript ["*mein Manuskript*"], started in the last period ["*in der letzten Zeit*"], very different and more incomprehensible than the one showed to Moore. This recent writing can hardly be identified with the two wartime notebooks MS101 and MS102, already cited in the previous letter and presented as similar to the pre-war notebook. Besides, this is the first time, despite Russell's frequent solicitations, that Wittgenstein speaks about printing some work of his – or rather, insists it "must" be printed. After his reluctance to publish anything that is less than perfect, his diaries seem the less indicated works for publication.[10] But the most puzzling reference is the final clause: "The problems are becoming more and more lapidary and general and the method has changed drastically. –"[11]. Wittgenstein wasn't in the habit of telling something without a good reason. Such a relevant change of method is not detectable in the diary entries, nor in the passage from MS101 to MS102. The method here remains discursive and dubitative, without any increasing "lapidarity". On the contrary, everyone would say that with the first pages of the Prototractatus "problems become more and more lapidary and general". Here Wittgenstein cannot refer to the diaries, but to new records (may be also in other sheets or notebooks) which in brief will converge (or are in the process of converging) into the Prototractatus notebook. No doubt that starting from its first page the method does "change drastically", adopting Tractatus' top-down numerical structure. So we aren't far from the truth if we think that the first page of the notebook, the proper *Abhandlung* starting point, was filled between April and May 1915.[12]

This conclusion is not without consequences. If in general the 1915-16 notebooks do not precede the definition of the *Abhandlung* propositions on the Prototractatus register, nor are they independent and alternative, but accompany it, as a counter-song that discusses its apodictic statements, it's useful to read the two documents in parallel. It's essential to hypothesize a definite date scansion of the Prototractatus notebook, and above all to follow the sequence of its itinerary, which – it's convenient to repeat here – doesn't have anything in common with Tractatus' arrangement in sequential order of decimal number. The notebook privileges a top-down process, from high-level sequences to ever deeper reflections; all the skeleton of the arguments is stated before the successive waves of specific comments.[13] In particular, the first twenty-eight pages of the Prototractatus

[6] As I discuss in (Bazzocchi 2005), this proposition is surprising recurrent in Tractatus' story. It is quoted, in a double allusive manner, in a note at Prototractatus' head; besides, it maintains an embarrassing logical error, whose correction involved a correspondence between Ramsey and Wittgenstein, and determined an unsatisfactory adjustment of the entire pass.

[7] In German, from: "Insbesondere folgt ein Satz aus einem anderen…" into: "Insbesondere folgt die Wahrheit eines Satzes aus der Wahrheit eines anderen…".

[8] The same correction appears also in the previous statement, 5.04, whose ending ("so sagen wir dieser Satz folge aus der Gesamtheit jener anderen") becomes: "so sagen wir die Wahrheit dieses Satzes folge aus der Wahrheit der Gesamtheit jener anderen". The four insertions "die/der Wahrheit" are very evident on the page.

[9] I refer to Wittgenstein's page numeration. Note that the first page of text, with the first fifteen propositions, is numbered as page 3.

[10] Compare with Hermine's list, where not the diaries, but only Prototractatus notebook is marked: "for publication".

[11] "Die Probleme werden immer lapidarer und allgemeiner und die Methode hat sich durchgreifend geändert. –". Surprising, in the "Historical introduction" to the Prototractatus von Wright quotes almost the whole letter, except this revealing conclusion. So von Wright can argue: "What he here calls 'my manuscript' is, I conjecture, the manuscript he had shown to Moore and the first two wartime notebooks" (Wittgenstein 1971, p.6).

[12] After a consistent period of non-productivity and depression, until April 15th ("Es fällt mir nichts Neues mehr ein! […]Ich kann auf nichts mehr Neues denken"), the encoded journal shows a turn in April 16th ("Ich arbeite") and 17th ("Arbeite"). A period "of grace" is testified with unusual emphasis at the end of the month: "Ich arbeite" (April 24th), "Arbeite" (26th), "Arbeite! In der Fabrik muß ich jetzt meine Zeit verplempern!!!" (27th), "Arbeite wieder!" (28th), "Die Gnade der Arbeit!" (May 1st).

[13] So the Prototractatus structure is very alike the Tractatus hypertext arrangement, in the sense illustrated in [Bazzocchi 2008].

do not correspond to recorded propositions on the diaries, but are in general their structural ancestors. So, it becomes clear how could the 1915-16 notebooks contain so many propositions of detail which will find place, without corrections, in the final work, since at the moment of their first conceiving, the entire structure of reference was already fixed on the contemporary *Abhandlung*. The Prototractatus stratification sets a series of nuclear prototypes, in some way discussed and commented in the diary, whose inspection can improve the comprehension of the whole enterprise.

Literature

Bazzocchi, Luciano 2005, "The Strange Case of the Prototractatus Note", in *Time and History – Papers of 28. International Wittgenstein Symposium*, Kirchberg am Wechsel, 2005, pp. 24-26.

Bazzocchi, Luciano 2006, "About «die letzte Zusammenfassung»" in *Cultures: Conflict-Analysis-Dialogue – Papers of the 29th International Wittgenstein Symp*osium, Kirchberg am Wechsel, pp. 36-38.

Bazzocchi, Luciano 2007a, "Hypertextual interpretation of the decimals and architectonic hermeneutics of Wittgenstein's Tractatus", in *The Labyrinth of Language*, G.P.Gàlvez ed., Castilla-La Manche, Cuenca, pp. 95-103

Bazzocchi, Luciano 2007b, "A database for a Prototractatus Structural Analysys", in *Philosophy of The Information Society – Papers of the 30th International Wittgenstein Symposium*, Kirchberg am Wechsel, pp. 18-20.

Bazzocchi, Luciano 2008, "On butterfly feelers. Some examples of surfing on Wittgenstein's Tractatus", in *Philosophy of the Information Society*, Vol. 1, Alois Pichler & Herbert Hrachovec eds., Ontos-Verlag, Frankfurt a.M., 2008, pp. 129-144.

Geschkowski, Andreas 2001, Die Entstehung von Wittgensteins Prototractatus, Bern.

McGuinness, Brian 2002, "Wittgenstein's 1916 «Abhandlung»", in *Wittgenstein and the Future of Philosophy*, R.Haller, K.Puhl eds., Wien.

Wittgenstein, Ludwig 1971, ^{2}1996, *Prototractatus*, B.F McGuinness, T.Nyberg and G.H. von Wright eds., Routledge & Kegan Paul, London.

Wittgenstein, Ludwig 1974, *Letters to Russell, Keynes and Moore*, B.F McGuinness and G.H. von Wright eds., Basil Blackwell, Oxford.

The Essence (?) of Color, According to Wittgenstein

Ondřej Beran, Prague, Czech Republic

Wittgenstein's treatise of the topic of colors can be seen as an interesting development of the view on the nature or essence of color (colors), but such development that ends with a considerable weakening (not to say deconstruction) of the conception of any essence.

Wittgenstein was attracted to the question of colors in *Tractatus* (Wittgenstein 1993) where he deals the first time with the color exclusion problem. His conception of elementary propositions is such that any elementary proposition is true or false independently on any other elementary proposition (or all of them). This independence can be seen from the fact that a conjunction of two elementary propositions can be neither tautology, nor contradiction. This is not the case of the conjunction of two ascriptions of color to the same point in space and time (to say of some point that it is green and it is red, is a contradiction – see 6.3751). Hence ascriptions of color seem not to be elementary. Does it mean that the essence of color is to be found somewhere deeper that in what shows to us as "color"? Wittgenstein provides no clear answer. What is confusing here is the fact that color ascriptions serve to many empiricist philosophers (including Vienna Circle (but, presumably, not including Wittgenstein)) as notorious examples of a primitive observation.

This problem becomes clearer and more insistent in the later texts, beginning with "Some Remarks on Logical Form" (1929). Wittgenstein discusses here possible ways of the analysis of color ascriptions. What is it that is ascribed when we say that something is "red"? The concept "red" seems to be not primitive, reducible. Where can one find the elementary propositions constituting the allegedly complex color ascription? Wittgenstein proposes an analysis into mathematized elements – that in a color ascription we ascribe *n* (certain number of) elements (quantities) of color (so that what we usually call "color" is a complex of such elements). However, there is a problem: since in mathematics any *n* includes also *n*-1, and *n*-2, then when we say (as an "elementary" proposition) that something possesses *n* elements (quantities) of "red", it implies that it possesses also any lower number of these elements (and so all the lighter (or darker?) tones of the ascribed "color"). Which is counterintuitive – the essence of color thus cannot be analyzed this way, going under the surface of what we see as "color". In this sense, and in opposition to what Wittgenstein says in *Tractatus*, color ascriptions are elementary. But on the other hand, there is the problem with their interdependence (any ascription of color excludes ascriptions of any other color). It seems that there are some types of elementary propositions that are interdependent. The logical form of our language is thus not uniform, it must respect the diversified shape of worldly phenomena.

This quite strong phenomenological sketch (that the structure of phenomena influences and grounds the logical form of language) is quickly revoked in *Philososophical Remarks* (Wittgenstein 1964). But not so that language, previously seen as "realistically" based on worldly phenomena becomes now "arbitrary" (this is what Austin (1980) suggests). That language cannot be straightforwardly compared with the world, doesn't mean that it doesn't or needn't respect its conditions (the world is still an environment whose claims and needs must language cope with, though it cannot be treated independently on language – compare Lance 1998 and his conception of language as a sport). Phenomenology now becomes identical with grammar. That is to say: the regular structure of the possibilities of experience (phenomenology) cannot be distinguished from the regular structure of what can be meaningfully said (grammar). How does this concern colors and their essence, if any? As for their essence, nothing changed much. Colors are still primary, elementary, irreducible, and their ascriptions are still interdependent (exclude each other). What is substantial for colors (for what "colors" are), the constitutive, normative relations among them, in this sense their "essence", can be demonstrated by means of certain schemes.

Wittgenstein introduces here the scheme of color-octagon, or two octagonal pyramids joint in their bases. The points of the octagon are red, violet, blue, blue-green, green, yellow-green, yellow and orange, the vertices of the pyramids are black and white. This scheme encloses the phenomenology, i.e. grammar of colors. It is normative, since the relations between concepts of colors (the laws of experience) are not liable to a subjective licence. Of course, the shape of the particular language is contingent, but for its respective speakers it is *a priori*. A contingent *a priori* (see Rorty 1991), pragmatically well-functioning.

A bit later *The Big Typescript* (Wittgenstein 2000) Wittgenstein makes the scheme a little more complicated. He tries to distinguish between so called basic (primary) colors – red, blue, green, yellow, and the other four, that are "mixed" colors. The octagon (or the double pyramid) is replaced by the color-circle, where the basic colors are fundamental (within their continuum the "pure" color is identifiable as a point), whereas mixed colors are not identifiable as points and represent only a continuum. Wittgenstein is led to this distinction by the different status of color mixtures. As he shows, the mixture of red and yellow is not a mixture in the same sense, as the one of violet and orange. The latter one just doesn't produce the color which stands in the circle between the constituents (i.e. red). That is to say: all colors are not of the same kind (or the relations among colors are not always the same or symmetric). What is even more disquieting is Wittgenstein's consideration about the exclusive ascriptions: of course, to say that something is red and that it is green doesn't make sense (in a sense), but an average speaker needn't necessarily feel it this way. What is decisive for the conclusion whether something makes sense or not, is whether any speaker can (feel that she/he can) use the "sentence" meaningfully in some situation. If she/he can, then philosophy cannot forbid it to her/him. It is linguistic practice, not philosophical generalization that decides what does make sense and what doesn't. The essence of color is expressed in grammar, i.e. meaningful use, and even if it includes that ascriptions of colors exclude each other, it doesn't mean that anyone cannot make the exclusive conjunction meaningful. In this sense, the essence of colors can seem "illogical" (in the usual, everyday sense of the word "logic").

This gap becomes much wider in *The Brown Book*. Whereas previously the four-polar color-circle was the ultimate authority; for example for the conclusion that any red-green combination doesn't make sense, here Wittgenstein presents another perspective: If for example some social class ("patrician") is characterized by red and green clothes, the combination red-green will be perfectly meaningful, in the sense of "patrician". An analogous example is: if some culture doesn't have a common name for our "blue" and calls dark blue "Oxford" and light blue "Cambridge", these people's answer to the question what Cambridge and Oxford have in common will be: Nothing (see Wittgenstein 2005, p. 134f). Of course, this sense of color combination is quite different from the problematic idea of one point in space time having two different colors, or the one of "reddish-green" color (which is such "in itself", so to speak). Hence, the purpose of these counterexamples of "patrician" colors or the distinction Cambridge/Oxford is not to refute the older statements about the color exclusion. The notion of what are the constitutive relations among (i.e. phenomenology of possibilities of) colors, hence, what is the essence of colors, is only broadened this way. It is not easy just exclude anything from the essence of color (from what is meaningful to say about colors and relations among them, in whatever sense – all this belongs to their "essence", as Wittgenstein sees).

These examples, though fictitious, introduce relativistic questions: is it possible that various people or rather various cultures have various systems of colors? And can we decide which system is "true"? For now, Wittgenstein answers nothing. Later, he will admit the possibility, but with certain (to so speak Davidsonian) limitations; but the decision, if any, will have to be done otherwise than by a straightforward comparison of the color concepts with colors "in reality".

Wittgenstein then had left the topic of colors for more than ten years, and returned to it in *Remarks on Colors* (Wittgenstein 1992), his response to Goethe's *Farbenlehre* which incited his great interest. The main purpose of Goethe's analysis of colors is to provide a criticism and alternative of Newton's optical experiments. For Goethe, the nature of colors in general cannot be conceived by one optical experiment, unjustly generalized. White doesn't consist of all the rainbow-colors, except of the context of light fraction. A color-theorist, claims Goethe, must respect the variety of color laws and relations among them, which differ from context to context. If there is any medium within which what is essential for colors is available, it is the medium of our experiencing (*Erleben*) – which includes the regular impact of colors and their combinations on the perceiver, as well as all the conventional (allegoric, symbolic etc.) constituents of the meaning of colors (Goethe 2003).

Wittgenstein's late return to colors, inspired by Goethe, proves his slight weakness with respect to the temptation of phenomenology (for the problematics of Wittgenstein's "phenomenology" see Gier 1981 or Kienzler 1997). However, he is well aware of the disparate character of the "essence" of colors. Either "phenomenologically", or "grammatically", one cannot find a simple, unite "essence".

The central question he asks – and the central problem he sees – here is the one of the "sameness" of color. He discusses several problematic examples: 1) We call "red" both the autumn leaves and some red clothes – however, "in a sense", it is not the same color. Actually, all the things we call "red" can seem quite different (and the difference is not only the one of light/dark). 2) One can paint both "white" things and "illumined grey" things (things usually conceived and seen as such), using the same palette color. 3) When one paints a dark room in the full light, how can she/he then compare the colors of the painture painted and seen in the full light with the colors of the room seen in the dark?

All these examples show that it is not at all easy to state how can two things have the same color, how to compare it, and what does this "sameness" mean. The universal, *unum versus alia,* seems here to be nothing more than one word standing against all the disparate phenomena. But it would be a philosophical error to search for some *one* thing (in whatever sense of "thing") hidden behind the *one* word ("craving for generality" – Wittgenstein 2005, p. 17ff). In this sense, Wittgenstein seems to be a kind of nominalist – the universal shared by all the particular things is a word, *nomen*. But there is no further analysis of what this universal word capturing the "essence" is. The universality of the word means nothing more and nothing less than the universality of use (just the fact we use the one word in all the different contexts). And that we know that something is red, cannot be further explained (the only possible explanation is that we have learned English – see Wittgenstein 1958, § 381).

The relations among colors become still more diversified. In one context (optical) colors differ: some can be seen-through, and some cannot (white, black, brown); in another context (colors of a paper) all the colors are of the same sort. However, philosophy shouldn't try to explain away these differences and reduce them on a simple essence and simple essential relations among colors, but on the contrary to try to conceive as many such differences as possible. The essence of colors lies in the meaning of the words for colors; there is no better (in fact no other) way how to conceive the "essence" of colors than by a description of this *variety*.

As for the relativistic problems with alternative systems of colors, Wittgenstein introduces two types of anti-relativistic argument. One of them is so speak Davidsonian (cf. Davidson 1974): in order that we are able to state that something is a concept of color, though differing from our concepts and not quite understandable for us, it must be somehow akin to our concepts. We must always have some auxiliary evidence to discover whether something is a concept of color (a bit like the evidence of whether someone is a good tennis or chess player which doesn't require that the author of the judgment is himself/herself a good tennis or chess player). After all, we have no better criterion for being a color than that it is one of our colors. The other argument is: if we are to decide between two different conceptions (lists) of the primary colors (one of them includes green among them, one of them considers it as a mixture of blue and yellow), we must look at which one of them works better in practice. I.e. which one of them enables us to fulfill more tasks (or more complicated tasks). Wittgenstein thinks, which is not without problems, that the conception of four basic colors is better in this sense. But whatever is the answer here, the only acceptable relativism is the relativism of systems that are akin and that function equally well in practice.

Literature

Austin, James 1980 “Wittgenstein's Solutions to the Color Exclusion Problem”, *Philosophy and Phenomenological Research* 41, 142-149.

Davidson, Donald 1974 “On the Very Idea of a Conceptual Scheme”, *Proceedings and Addresses of the American Philosophical Association* 47, 5-20.

Gier, Nicholas F. 1981 *Wittgenstein and Phenomenology*, Albany.

Goethe, J.W. 2003 *Farbenlehre*, Stuttgart.

Kienzler, Wolfgang 1997 Wittgensteins Wende zu seiner Spätphilosophie: 1930-32, Frankfurt a.M.

Lance, Mark 1998 “Some Reflections on the Sport of Language”, in James Tomberlin (ed.), *Philosophical Perspectives, 12: Language, Mind, and Ontology*, Oxford.

Rorty, Richard 1991 Contingency, Irony and Solidarity, Cambridge.

Wittgenstein, Ludwig 1929 “Some Remarks on Logical Form”, *Proceedings of the Aristotelian Society*, *Suppl. vol.* 9, 162-171.

Wittgenstein, Ludwig 1964 *Philosophische Bemerkungen*, Oxford.

Wittgenstein, Ludwig 1958 *Philosophische Untersuchungen*, Oxford.

Wittgenstein, Ludwig 1992 *Werkausgabe Band 8*, Frankfurt a.M.

Wittgenstein, Ludwig 1993 *Tractatus logico-philosophicus*, Praha.

Wittgenstein, Ludwig 2000 *The Big Typescript*, Wien.

Wittgenstein, Ludwig 2005 *The Blue and Brown Books*, Oxford.

Wittgenstein’s Externalism – Getting Semantic Externalism through the Private Language Argument and the Rule-Following Considerations

Cristina Borgoni, Granada, Spain

I.

Since Kripke has defended that “the real ‘private language argument’ [P.L.A.] is to be found in the sections preceding § 243” (Kripke, 1982, p. 3) of *Philosophical Investigation* [PI], it has become an imperative – for those who want to enter the discussion - to figure out its relation to the rule-following argument [R.F.A].

In this paper, I will maintain that both arguments are connected to each other, but not in the Kripkean sense. By doing this, I will be able to offer a double externalist interpretation to them. On the one side, the P.L.A., when considered as independent from the R.F.A, will lead us to a negative formulation of the externalist thesis, through a *reductio ad absurdum* of the internalist conception of the mental. On the other side, when both arguments are considered as concerning to the same question, they will lead us to a positive defence of the externalism.

I will take *externalism* as the position that defends that mental contents are individuated with reference to external factors to the mind.

II.

A great part of the discussion about the P.L.A. is centred in the case proposed by § 258. A case where we are asked to imagine ourselves writing in a diary the occurrence of a certain “private” sensation. In this diary, we should write the sign “S” every time we had that sensation. Wittgenstein warns us with respect to the traits of this exercise: “(…) The individual words of this language are to refer to what can only be known to the person speaking; to his immediate private sensations. So another person cannot understand the language (Wittgenstein, 1953, § 243).

The notion of private language criticized by Wittgenstein involves several questions; the question about completely private experiences (in the sense that no one could have access to them but its owner), the question about the development of a language able to describe such experiences, and, the question about the possibility of a language understood only by its creator. When Wittgenstein argues against the idea of a private language, he is arguing against such notions. Furthermore, he is arguing against a specific theory of language, that one which supposes that an ostensive connection between a word and a sensation (or between a word and an object) is sufficient to establish a meaning. § 258 leads us to the ultimate consequences of thinking in those terms:

> (…) A definition surely serves to establish the meaning of a sign. —Well, that is done precisely by the concentrating of my attention; for in this way I impress on myself the connexion between the sign and the sensation. —But "I impress it on myself" can only mean: this process brings it about that I remember the connexion *right* in the future. But in the present case I have no criterion of correctness. One would like to say: whatever is going to seem right to me is right. And that only means that here we can't talk about 'right' (Wittgenstein, PI, § 258).

There are those who have interpreted such an argument as dealing with a skeptical problem about memory. Such an interpretation says that, although an ostensive definition can be made plausible, the problem is how to warrant the future connection between the sensation “S” to its name. However, it seems that this kind of skeptical problem is not the core of Wittgenstein’s argument (Gert, 1986, p. 429). In the case proposed by § 258, the problem is not to apply the same word I am using now in the future, nor it is about how to remember the way I have used it in the past; more than that, the problem is that even in the current case we are not allowed to say that any meaning was established at all.

Another interpretation of the P.L.A. is the known defence by Kripke, that P.L.A. is not but a particular case of the R.F.A., an argument that leads us to another skeptical paradox.

The R.F.A. can be exemplified with the case proposed in § 185. In such a case, a pupil is taught to write down the series of cardinal numbers of the form 0, n, 2n, 3n, etc, at an order of the form “+n”. “So at the order ‘+ 1’ he writes down the series of natural numbers” (Wittgenstein, PI § 185). We are asked to suppose that the pupil has been tested up to 1000. Then, the pupil is asked to follow the series beyond 1000 and following the order “+2”. He writes 1000, 1004, 1008, 1012.

> We say to him: "Look what you've done!"—He doesn't understand. We say: "You were meant to add *two*: look how you began the series!"—He answers: "Yes, isn't it right? I thought that was how I was *meant* to do it."—Or suppose he pointed to the series and said: "But I went on in the same way."—It would now be no use to say: "But can't you see....?" —and repeat the old examples and explanations (Wittgenstein, PI § 185).

Kripke indicates that the core of the R.F.A. is to demonstrate that “[a]dequate reflection on what it is for an expression to possess a meaning would betray (…) that that fact could not be constituted by any of *those*”; by any “available facts potentially relevant to fixing the meaning of a symbol in a given speaker’s repertoire” (Boghossian, 1989, p. 508). Under this interpretation, § 185 proposes a skeptical paradox in similar terms to what seems to be suggested in the following aphorism:

> This was our paradox: no course of action could be determined by a rule, because every course of action can be made out to accord with the rule. The answer was: if everything can be made out to accord with the rule, then it can also be made out to conflict with it. And so there would be neither accord nor conflict here (…) (Wittgenstein, PI § 201).

Although this aphorism continues saying that “It can be seen that there is a misunderstanding here from the mere fact that in the course of our argument we give one interpretation after another” (Wittgenstein, PI § 201), Kripke insists on the skeptical scenario. A scenario that spreads to the P.L.A.: nothing could fix the meaning of the sign "S", as well as nothing could fix the meaning of the sign "+2" in the pupil's case.

The solution found by Kripke to the supposed skeptical paradox is the communitarism; if there is nothing as a "semantic fact" to determinate the difference between looking right and being right, to decide about this difference is something that belongs to the community.

McDowell (1984), however, who disagrees with Kripke's interpretation, offers us not just an important criticism to that interpretation, he also shows us another way of understanding Wittgenstein's position. What McDowell does is to stress the conditions to the very perception of the skeptical paradox, insisting on the continuation of the § 201:

> (…) What this shows is that there is a way of grasping a rule which is *not* an *interpretation*, but which is exhibited in what we call "obeying the rule" and "going against it" in actual cases (Wittgenstein, PI § 201).

McDowell maintains that "Kripke's paradox" occurs only if we keep considering meaning as an interpretation. The necessary step, therefore, would be to change the idea that understanding always supposes offering an interpretation. That would be Wittgenstein's lesson. If the R.F.A. does not concern the desperation of how to establish the difference between right and wrong, the Kripkean conclusion is not maintained either. If McDowell is right in his diagnosis, it is not the case that the P.L.A. is just another instance where we can verify the skeptical paradox. In the case of the sign "S", we are not allowed to say that we have established any meaning at all, but this is not the case with the sign "+2". In a sense, both arguments are connected because they both dismiss the idea of meaning as being the univocal relation between a sign and an object, or between a sign and a mental image. However, they set apart in the sense that, the case of "+2" has a correction criterion, thought not established by a semantic fact, while in the case of "S" it has not. In this sense, we could say that the P.L.A. establishes a specific criticism to the idea of mental entities giving meaning to our language. So, I propose to reformulate the P.L.A. in the following terms:

(i) Possessing a correction criterion is a condition of possibility to a language;
(ii) A private language lacks correction criteria;
(iii) A private language is impossible. There is no such a thing as a private language because it is not a language.

"Having a meaning is essentially a matter of possessing a correctness condition" (Boghossian, 1989, p. 515). The first premise seems to be widely accepted. A statement is meaningful if it can be true or false.

The second premise appears clearly at the end of § 258. The attempt to point privately to a certain sensation, to a private one, leaves us without a correction criterion. The very sensation can not itself give me such a criterion, as it seems to be supposed by an ostensive definition between the sensation and the name I give to it. Wittgenstein rejects this image, not only here, but in most parts of his work. The R.F.A. is an example of this rejection, but it appears also in the earlier aphorisms of PI, when Wittgenstein criticizes the Augustinian image of the language.

Given the two first premises, the immediate conclusion of such an argument is that the "concept of a private language is one that cannot be defended, at best, and is incoherent, at worst" (Preti, 2002, 56).

The P.L.A. has a deep externalist character. The notion of private language could indeed be elaborated in opposition to an externalist position: the components of such a "language" are not identified by external factors to the mind, but purely by internal ones. Because of that, to argue for the incoherency of such a notion opens the way to reach externalism through a *reductio ad absurdum*. And the conclusion is that it becomes unintelligible to talk, at the same time, about instances of language (it does not matter if we are talking about the world or about our subjective experiences) and about private correction criteria.

If, by arguing the P.L.A., we show the incoherency of internalism, we could consider this path as a kind of motivation to reach externalism, though a negative one. It is possible, however, to also find a positive motivation in Wittgenstein's arguments, but taking both P.L.A. and R.F.A. as working together. And this is possible if we think that, more than a criticism, they offer us an alternative option to think about meaning which does not need the idea of semantic facts.

Kripke defends that the Wittgensteinian argument leads us to communitarism. We could understand him as saying that the premise (ii) is true because any correction criterion is to be established by a community. In this sense, one could find in Kripke's interpretation some externalist appearance if we could retain the idea that individuating mental contents belongs to the community and never to oneself privately. However, the Kripkean position is much stronger than that; the community is provided with full powers to the very establishment of meanings. While this position could sound as an externalism, it would also sound as the complete isolation of the community inside itself. At this moment, "[o]ne would like to say: whatever is going to seem right to us is right. And that only means that here we can't talk about 'right'" (McDowell, 1984, p. 49, n. 12)

As I have tried to defend, not only the Kripkean interpretation does not seem to be the most satisfactory one, but his solution also causes a discomfort to which McDowell calls our attention. If in an internalist position we could be isolated from the community, now we could, all together, be isolated from the world. And this does not seem to be Wittgenstein's position, as Preti warns:

> From the fact that our fellows in the community play a constitutive role in determining content it will not follow that content is *not* the "queer", inner mental process that Wittgenstein is concerned to deny. (…) Perhaps, that is, it is true that what determines meaning or content must be partly constituted by the minds of others – but it won't follow from this that the content in *other* minds in the community isn't determined by *their* inner mental processes. Merely being *other* is not enough to thwart the inner state conception of meaning, and it may be that Wittgenstein appreciated this (Preti, 2002, p. 60).

There is, however, another way of making plausible the idea that correction criterion can only belong to the public sphere without the commitment to the communitarism. And that is possible when we realize that the institution and the application of meanings are not distinct activities. If the moments of application of meanings are so important in Wittgenstein approach, this is so because they are not separated from the moments of institution of meanings. The externalism here would follow a more positive way than the one that was reached with the accusation of incoherence of the notion of private language. Here the meanings would be established with relation to external factors to one's mind, but also, with relation to external factors to any mind.

The positive character of Wittgenstein's argumentation is, without doubt, which brings with itself the dispute about the interpretation of his arguments. The dispute, for example, about which notion of meaning Wittgenstein defends at all. I believe, however, that it is important to point to the sense of "internal" Wittgenstein is rejecting. As Preti points well, one could understand the notion of "private" only as in opposition to "social", as Kripke does. But such a notion does not exhaust in fact all that is being rejected by Wittgenstein: "the hidden, the inner, the introspectively accessible, the mentalistic (Preti, 2002 p. 60). It seems that the externalism reached through Wittgenstein's arguments involves the rejection of all this set of notions.

Literature

Boghossian, Paul 1989 "The Rule-Following Considerations", *Mind* 98, 507-549.

Gert, Bernard 1986 "Wittgenstein's Private Language Argument", *Synthese* 68, 409-439.

Kripke, Saul 1982 Wittgenstein on Rules and Private Language, Oxford: Blackwell.

McDowell, John 1984 "Wittgenstein on Following a Rule", in: Alexander Miller and Crispin Wright (eds.) 2002, *Rule Following & Meaning*. Chesham: Acumen, 45-80.

Preti, Consuelo 2002 "Normativity and Meaning: Kripke's Skeptical Paradox Reconsidered", *The Philosophical Forum* 33, 39-62.

Wittgenstein, Ludwig 1953 *Philosophical Investigations* (translated by G. E. M. Anscombe 1979) Oxford: Basil Blackwell.

Wittgenstein, Ludwig 1969 *On Certainty* (translated by G. E. M. Anscombe and Denis Paul 1979) Oxford: Basil Blackwell.

Informal Reduction

E.P. Brandon, Cave Hill, Barbados

This paper was provoked by Horst's recent book (*Beyond Reduction: Philosophy of Mind and Post-Reductionist Philosophy of Science*) that argues for a metaphysical pluralism largely on the basis of claims about the status of what one might call the reductionist programme in philosophy of science.

Horst's position is that the idea that the mature physical sciences display extensive, metaphysically significant reductions is an illusion that philosophers of science have exposed but which too many of us in other areas of philosophy mistakenly cling to. Whatever the formal obstacles that Horst points to, I am not convinced that they constitute a refutation of a metaphysically important reductionism, so my aim is to try to clarify, informally, what that kind of reductionism is concerned with, and to suggest that our best bet in identifying successful, and unsuccessful, reductions of that type remains with the scientists themselves, rather than with the models we have created of what ideal reduction should involve.

Ernest Nagel provides, for many, the standard account of what the reductionist programme aspired to. My strategy is to briefly review what Nagel actually claimed for reduction, and then to consider the type of issue that more recent specialists have urged against it.

Very roughly, the formal ideal Nagel set out involves the laws and predicates of two theories. Theory one is reduced to theory two just in case there are bridging principles linking the predicates of theory one with those of theory two, and the laws of theory one can be deduced from those of theory two with the help of such bridging principles. While writing of deduction here, Nagel is clear that these deductions should embody explanations of theory one in terms of theory two, so the relationship is deduction plus whatever other constraints explanation requires.

Rather than focussing on Nagel's formal account itself, I want to stress that methodologically he was a naturalised epistemologist *avant la lettre*: his discussion derives from the positions of practising scientists on whether reduction has been successful; he takes reduction as a fact accepted by most scientists in the relevant fields and aims to characterise it. It is important for him also to stress the many failures of reduction, perhaps most notably the impossibility of a mechanical reduction of electromagnetism, as well. His is not the style of argument that shows that a putative reduction fails to fit a formal model and so is shown not to be a case of genuine reduction. While formal, his model is recognised to be a model, an ideal case, and so inherits the messiness of models throughout the sciences.

So, for instance, Nagel supposes that the Boyle-Charles' law were the only thing derivable from the kinetic theory of gases and says "it is unlikely that this result would be counted by most physicists as weighty evidence for the theory … For prior to its deduction, so they might maintain, this law was known to be in good agreement with the behavior of only "ideal" gases, … Moreover, physicists would doubtless call attention to the telling point that even the deduction of this law can be effected only with the help of a special postulate connecting temperature with the energy of the gas molecules—a postulate that, under the circumstances envisaged, has the status of an *ad hoc* assumption, … In actual fact, however, the reduction of thermodynamics to the kinetic theory of gases achieves much more than the deduction of the Boyle-Charles' law. There is available other evidence that counts heavily with most physicists as support for the theory and that removes from the special postulate connecting temperature and molecular energy even the appearance of arbitrariness" (Nagel 1961, 359-60). He notes that the special assumptions can be replaced by others known to be closer to reality, and that the reducing theory can "augment or correct" the currently accepted body of laws of the reduced theory.

The same methodological point can be derived from an article quoted by Horst (from Silberstein), though neither author goes on to elaborate on it, that most scientists would think that the errors of philosophers show they have a bad model of reduction rather than that there are no reductions. "Focus on actual scientific practice suggests that either there really are not many cases of successful epistemological (intertheoretic) reduction or that most philosophical accounts of reduction bear little relevance to the way reduction in science actually works. Most working scientists would probably opt for the latter claim" (Silberstein 2002, 94).

I suggest that the main point[1] of the reductionist programme is to claim that some particular area of interest can be comprehended, explained, by certain entities, features, ways of working, and *no others are needed*. The area of interest is then nothing but the entities, features and their ways of working used in these explanations. These claims are rough, they constitute a scientist's informal patter, rather than the technical, strict derivations she might offer within a theory. They indicate the wider significance of the theory, and of course they may prove to be as fallible as anything else, but they are not groundless. If this idea is on the right lines, it casts considerable doubt on the salience of the points that have been made against Nagel, and relied upon by Horst in his application to the philosophy of mind.

This account can be defended by looking at Scerri's discussion of the relation between chemistry and quantum mechanics, in particular the role of the latter in accounting for the periodic table. Scerri insists that we cannot rely on the quantum mechanical theory and the approximations it allows *ab initio* but rather we accept them when they yield what are known empirically to be the right results and complicate them when we have empirical evidence that the first approximation is mistaken. I am in no position to deny that this is what happens, nor I think is his commentator, Friedrich, but our point is that there is absolutely no suggestion that these cases where the theoretically derived structure is known to be mistaken involve inexplicable emergent properties or new theoretical notions. They simply show that our approximations have left out something important which we already knew about. So, for instance,

[1] As with virtually every philosophical position, there are a number of variants. The historically informed account of emergent properties (Timothy O'Connor and Hong Yu Wong 2006, http://plato.stanford.edu/entries/properties-emergent/) in the Stanford Encyclopedia of Philosophy, for example, distinguish between ontological/metaphysical and epistemological construals of the same issue, and then further subdivides those approaches.

Scerri's says about the case of chromium: "It appears that both non-relativistic and relativistic calculations fail to predict the experimentally observed ground state which is the $4s^1 3d^5$ configuration" (2004, 101), but he immediately goes on to admit "Of course I do not deny that if one goes far enough in a more elaborate calculation then eventually the correct ground state will be recovered. But in doing so one knows what one is driving at, namely the experimentally observed result. This is not the same as strictly predicting the configuration in the absence of experimental information." Right, it isn't; but that failure has not revealed anything new at work.

As Galison says of a different case, "The reductionist physicists reply that it is true that you might not guess these collective behaviors; but if you ask *why* very cold copper superconducts, the answer includes nothing other than electrons and electrodynamics-there's no magical supplementary thing over and above these" (2008, 121).

We can see the point at issue in a case even Horst acknowledges is not decisive: the mathematical intractability of the three-body problem for Newtonian theory is no reason to suppose that anything new has entered a Newtonian system when a third mass is introduced. The cases Horst thinks are more significant than this do not seem to involve anything very different, however. I appeal to Azzouni's authority in agreeing with Nagel that the idealising assumptions required in the thermodynamics case are not a barrier to the worthwhile kind of reduction she calls "scientific". She says "Imagine, contrary to fact, that a genuine derivational reduction is available, but only if constraints are placed on gas states that are—given the physics of micro-particles—quite probabilistically low. In such a case … physicalism fails: emergent phenomena, indicated by physically inexplicable constraints on the probability space of micro-particles, show this" (2000, 45). She comments in a footnote: "Garfinkel (1981:70-1) seems aware of this possibility, but seems also, falsely, to think that the actual derivation of the Boyle-Charles law from the statistical behavior of the ensemble of molecules illustrates it just because of the use of the conservation of energy (in closed systems) and the assumption of a normal distribution of velocities."

There are, of course, contentious issues in what scientists consider successful or unsuccessful reductions (e.g. the temporal isotropy of statistical mechanics as against the directedness of phenomenological thermodynamics). But the types of idealising assumption that Horst is appealing to are not usually something that should undermine our confidence in a reduction.

While I have been happy to call upon Azzouni for support, and indeed find her account of scientific reduction very close to what I have been urging myself,[2] I will close with one quibble. If we can legitimately see scientific reduction as not requiring formal derivations but simply "*all* that is desired is an extension of the scope of an underlying science in a way illuminating both to that science and the special science above it" (2000, 40), I would like to suggest that we can extend the same sympathy to Nagel's own derivational model of reduction. As I indicated at the start, Nagel saw the relation as explanatory. Working with the tools available to him he took that to involve formal derivation, though actual cases, of explanation and of reduction, might well only exhibit the elements that Hempel called an "explanation sketch" (Hempel 1965 [1942], 238,). Allowing for Nagel's clear recognition that his account is indeed an idealisation of the reductions we actually find, we may resist agreeing with Azzouni that the concern for derivation has been altogether a mistaken "obsession with words". It provides a picture that we can adjust to get closer to the realities we are interested in, as indeed her own explorations reveal.

Literature

Azzouni, Jodi 2000 Knowledge and Reference in Empirical Science, London: Routledge.

Friedrich, Bretislav 2004 "…Hasn't it?", Foundations of Chemistry 6, 117–132.

Galison, Peter 2008 "Ten Problems in History and Philosophy of Science", Isis 99, 111–124.

Hempel, Carl G. 1965 [1942] "The Function of General Laws in History", in: Aspects of Scientific Explanation, New York: The Free Press.

Horst, Steven 2007 Beyond Reduction: Philosophy of Mind and Post-Reductionist Philosophy of Science, Oxford: Oxford University Press.

Nagel, Ernest 1961 The Structure of Science, London: Routledge and Kegan Paul.

O'Connor, Timothy, and Hong Yu Wong 2006 "Emergent Properties", in Stanford Encyclopedia of Philosophy (http://plato.stanford.edu/entries/properties-emergent/).

Scerri, Eric 2004 "Just How Ab Initio is Ab Initio Quantum Chemistry?" Foundations of Chemistry 6, 93-116.

Silberstein, M., 2002. Reduction, Emergence and Explanation. In The Blackwell Guide to the Philosophy of Science, edited by P. Machamer and M. Silberstein. Oxford: Blackwell.

[2] She says, for instance, "talk of there being a scientific reduction in this sort of situation is still legitimate because we really do take As, and what is going on with them, to be nothing more than Bs, and what is going on with them; we recognize and expect that if, in certain cases, we overcome (particular) tractability problems (as we sometimes do) in treating As as Bs, we will not discover recalcitrant emergent phenomena. Scientific reduction is a project with methodological depth: the idealized model is one where deviations from what is actually going on are deviations we can study directly, extract information from, and, when we're lucky, minimize. This is the full content of the claim that As, and what is going on with them, are really just Bs, and what is going on with them" (43-44).

An Anti-Reductionist Argument Based on Spinoza's Naturalism

Nancy Brenner-Golomb, Bilthoven, The Netherlands

In this paper I wish to concentrate on one aspect of an anti-reductionist view, namely on the central idea underlying the so called 'bottom-up' principle of the structure of science. This idea says that although the behaviour of any structured entity is governed by laws which apply to this kind of structure alone, these laws are the result of, or emerge from, the properties of its basic elements. The most important aspect of this view is the relationship it establishes between the unity of science and the unity of nature. Feynman, for example, argued that the greatest success of the quantum theory is in increasing the unity of science. He claimed that the advantage of the possibility to explain the whole of chemistry in terms of quantum mechanics is weighed against the previously accepted empirical principle, that in order to accept a theory, a detailed understanding is required of what goes on in every experiment. This advantage of quantum mechanics, he says, shows that we are on the right track. And he adds that this advantage is accentuated by the fact that if chemistry can be so reduced to physics, then the whole of life can be reduced to it as well. According to him, the most important hypothesis in biology is that there is nothing that living things do, that cannot be understood by seeing them as made of atoms acting according to the laws of physics [Feynman 1989, 3-3 and 3-6].

In other words, Feynman's conception of science is that of physicalism, understood as everything that can be explained by physics, including non-material things, like laws of nature, the geometry of space or abstract concepts like energy. He emphasizes that we do not know *what energy is* (the emphasis is his). All we know is that this abstract quantity has many forms; that it can be calculated in each of them, and that their sum total is constant, which is The Law of Conservation of Energy [Feynman 1989, 4-1]. And 'explained by physics' means 'explained by a hierarchy of natural sciences which are ultimately reducible to physics.' The 'bottom-up' principle says that this hierarchy reflects the evolution of the structure of the universe.

My first claim in this paper is that although Spinoza argued against Descartes' conception of science, his arguments apply also to physicalism. This is because the unity of science has remained the same as Descartes claimed in the seventeenth century,, namely that all that science can do is to explain the physical world, in spite of the fact that most scientists do not accept Cartesian dualism.

My second claim is that starting from Spinoza's view of nature, the 'bottom-up' principle cannot be sustained as a universal law. This is because by the 'bottom-up' principle the properties of a structure which emerges from the properties of its basic elements have no effect on the structure of its elements. For example, the machinery of a cell includes a process for the production of proteins. The first step in this process is performed by an RNA molecule which selects that part of the DNA which prescribes its production. The 'bottom up' principle in this case says that, although this selection depends on the shape of this molecule, its biological function in the cell has no role in determining this shape. Its shape is exclusively determined by the laws of chemistry. In order to disprove the rival hypothesis, that it was a vital force of the cell that was responsible for determining the shape of this molecule, molecular biologists who adhered to the 'bottom-up' principle removed the RNA molecule into a test tube, heated it so that it lost its shape, and allowed it to cool down outside the cell. As a result, the molecule regained its 3-dimensional shape, proving that there was nothing in the structure of the cell that contributed to its formation [Cairns 1997. pp. 101and 94]. However, according to Spinoza's naturalism this independence cannot be maintained if the scientific hierarchy includes the structure of society emerging from the properties of individual people as its elements.

Spinoza's naturalism does not reject the idea that biology underlies a theory of mind. On the contrary. He explains that in order to recognize Peter the mind must abstract some essence of his by which he appears to us as the same person every time we see him. Yet, it is only by reflection on our factual recognition that we know that this must be the case. In fact, our brain derives this essence while we remain ignorant of it and of the process by which it is derived [Spinoza 1979 p.237]. In general, he says "no one has yet been taught by experience what the body can do merely by the laws of nature in so far as nature is considered merely as corporeal or extended, and what it cannot do save when determined by the mind."And he explains further that "*the body* can do many things by the laws of its nature alone at which the mind is amazed... when men say that this or that action arises from the mind which has power over the body, they know not what they say..." [Spinoza 1979 p.87].

Spinoza agreed with the empirical scientists of his time that *whenever possible* we must seek evidence for a theory of mind as much as we must do so for knowledge of the physical world. An argument to this effect we find in his comment on the idea that a person cannot judge something to be bad for him and yet want it. This, he says, is *contrary to experience.* As philosophers, we should acknowledge the fact that a person can very well want what is bad for him, and look for a natural explanation for it [Spinoza 1998, p.138].

I emphasize the phrase 'whenever possible' because Spinoza agreed with Descartes that we have some knowledge for which we cannot find evidence in the sense acceptable to empirical scientists. In fact, his own claim that there is nothing outside nature is not provable in this way. But according to him, this assumption is essential for creating a correct science. It is essential because it serves the best guide for research and the best standard of truth for its judgements [Spinoza 1979 p.241]. Of course, physicalism is also held to be the best guide and a standard of truth for research. The question is whether biology, which takes the theory of evolution as its guide and standard of truth can accept the 'bottom-up' principle as advocated by physicalism, or whether its inclusion of humanity in the evolving animal world is better explained by accepting Spinoza's conception of the human mind as part of natural evolution.

According to Spinoza, Descartes' assumed distinction between Thought and Extension is in fact a distinction between two ways by which the world can be understood. Either according to its conceived abstract laws or by its causal relations as they are observed in space.[Spinoza 1979, p.7 (note to proposition X)]. The distinction, he explains, must be made only because none of these ways of understanding can be derived from the other. Taking an example from physics, instead of his own [Spinoza 1966, p.7], the abstract law of gravitation cannot be derived from observed movements alone, and knowledge of this law is not sufficient for explaining a particular movement in space. But the world they explain is clearly the same.

Again we should note that although not many scientists or philosophers adhere to Cartesian dualism, Spinoza's argument is still relevant because this dualism has been replaced by a new one, namely of culture versus nature. Being beyond the permitted length of this paper, I can only point out that in spite of the influence of Darwin, his followers only included the human body in their study of evolution. And an influential scientist like Richard Dawkins, or philosophers like Charles Peirce, Quine, Wittgenstein and Daniel Dennett, among many others, see in rational thinking a cultural invention, where a culture is largely independent of nature. But by Spinoza's view a culture cannot be independent of nature. Anything which can affect human behaviour must be explained in natural terms because there is nothing outside nature.

Spinoza's conception of *substance* is his conception of Nature as a whole. Its definition says that substance is its own cause and is to be conceived through itself, namely by nothing outside itself [Spinoza 1979 p.1, definitions I and III], implying that the laws of nature are not imposed by God on inert matter, as Spinoza's contemporaries, and even Newton, believed. These definitions say that the laws of nature express the internal dynamic force of material existence – which is the meaning of his equating God to Nature, and that every thing which comes into existence is a modification of substance, and its own internal forces must be understood in terms of the internal forces of Nature.

In his Metaphysical Thoughts [Spinoza 1998 p.120] Spinoza argues that the essence of life should be understood as "the force through which things persevere in their own being." It is because this force can be *conceptually* distinguished from the things themselves, he explains, that the idea arose that things *have* life, namely souls, as if life was distinct from the living things themselves. In the *Ethics* he generalizes the idea to all structured things. *All things*, he says, behave so as to sustain their own survival [Spinoza, 1979 p.91 (proposition VI)].

Commenting on Descartes' "*I think therefore I am*" Spinoza says that Descartes indeed discovered an essence of man. But this essential feature is part of the internal forces by which people persevere in their natural existence [Spinoza 1998 pp.9-10]. Spinoza explained the function of reason, as a corrective mechanism by which ideas are accepted or rejected by a balance of reasons, akin to the balance of forces in the body [Spinoza 1979 p.255]. He explains the necessary inclusion of this mechanism in human nature as a result of his other explanation that the more a body can perceive and respond to many things at the same time, the more it depends on understanding [Spinoza 1979 p.48].

This explanation is given in a note to proposition xiii in part II of the *Ethics*, which in a slightly different formulation says that an idea always reflects either an objective state of the human body or a certain mode of existence outside the body, and nothing else [Spinoza 1979 p.47]. In order to understand this proposition we may start by noting that 'ideas' should be understood as including everything of which we are conscious. For example, feeling hungry is also an idea. The objective state, or as he says, the object of this idea, in the body is comparable to a biologist saying that this feeling *is* the set of processes in the body which produce it. A feeling is clearly not the same as these processes. But if it is what we are conscious of when certain changes occurs in the body, in terms of which feeling hungry is fully explained. In this sense we may talk of a reduction of this mental state to a physical one. However, according to Spinoza, this explanation is not complete because a feeling is categorized as a kind of pain – a general term describing transitional states of the body by which its power of action is reduced [Spinoza 1979, p.128 (definition III and the explanatory note)]. And it follows from his conception of life, that this feeling must be combined with a desire to restore the body to its natural capacities, which in this case means a desire to assuage the pain of hunger.

While the objective state in the body underlying feeling hungry is a universal state reducible to biology, the actual behaviour for restoring the body to its natural capacities depends on the knowledge how to do it. Hence, the objects of the ideas constituting this knowledge are 'certain modes of extension actually existing' outside the body. This knowledge cannot be universal. If it were universal to our species, it would have meant that perception of these objects outside the body together with, as he says, the amazing laws of nature that move the body without the mind's interference, would have been sufficient for survival. And a theory of mind would be reducible to biology, even if environmental influence includes learning by imitating other animals of the same species. In this case, the 'bottom-up' principle might have been saved. The reason why this is not so for human beings is that the objects outside the body which affect behaviour are the behaviours of other people whose desire is to live according to *their* natural drives.

Again, Spinoza's naturalistic approach does not reject the assumption that the laws which govern a social structure emerge out of the properties of its elements, namely the properties of individual human beings. In the first chapter of his Political Treatise he says that his intention is to demonstrate that a sound political science can and ought to be based on what is known both of human nature and of political practice. This, according to him, agrees with other branches of science which verify or reject their theories by available evidence. What his study of human nature taught him is that passions are stronger motives of behaviour than reason. It follows that when people in power design rules for preserving the integrity of their community, they can never be free from the influence of their passions. Yet, he also learned that *all* people know that if they want to pursue their own plans of life they must surrender a great part of their power to the state [Spinoza 1951 pp.296-297 (15-16)]. This knowledge, according to him, is not a result of using reason – as Hobbes argued at the time – but is an intuition, which stated in modern terms means innate knowledge, that we need each other's help. People could not have discovered this essence of political life if they were not already living in societies [Spinoza 1966 p.269]. This he says, applies to all knowledge of a true essence of a thing, even to mathematics. We would not be able to know the essential equation of a parabola, for example, without first knowing parabolas. And we know parabolas because they exist [Spinoza 1998, p.99].

Spinoza explains that the basic political problem is not the imposition of law and order but the tendency of people in power to suppress the tendency of other people to use reason, so that they passively accept these leaders' ideas, as if they necessarily provide the best way to satisfy everybody's desire to live according to their nature in peace and security [Spinoza 1951 pp.215-216 and 313-315]. Spinoza's intention with developing his political science was to show that the best way to satisfy this basic desire was to design civil laws which would encourage rationality and thereby prevent this behaviour of leaders. But my purpose in this paper is only to show that, at least when the study of the human mind is included in the scientific project, it is impossible to maintain that a structure has no effect on the structure of its elements. This is because, as Spinoza maintained, to say that something is natural does not mean that it cannot be distorted [Spinoza 1979 pp.139-140]. For example by the influence of the natural behaviour of leaders.

Literature

[the year of Spinoza's books refer to the editions I used].

Cairns, J.: 1997. *Matters of Life and Death*, Princeton Univ. Press, N.J.

Feynman R.: 1989. *Lectures on Physics* 1989. A Commemorative Issue, edited by Leighton and Sands.

Spinosa: *1951, A Political Treatise* (PT), Dover Publications Inc. New York.

- *A Theologico-Political Treatise* (TPT), published together with PT.

- 1966. *Correspondence of Spinoza*, translated and edited by A Wolf, Frank Cass & Co.

- 1979. *Ethics*, Everyman's Library, Dutton: New York.

- *Metaphysical Thoughts* (MT), published with PCP.

- 1998. *Principles of Cartesian Philosophy* (PCP). Translation into English by Samuel Shirley. Hackett Publishing Company, Indianapolis/Cambridge

- Treatise on the Correction of the Understanding, published together with Ethics.

Did I Do It? – Yeah, You Did! Wittgenstein & Libet On Free Will

René J. Campis C. / Carlos M. Muñoz S., Cali, Colombia

1. Libet

RP is a concept developed by neuroscience to give an account of intentional action. It is basically 'brain electrical activity found to start increasing about 0,8 seconds before voluntary movement' (*Cf.*: Kornhuber and Deecke 1965, Deecke *et al.* 1969 and Libet *et al.* 1983). Libet involves the concept in an experiment (fig. 1) attempting to establish a temporal distinction between the onset of RP and "conscious wish".

Libet's main presupposition is: "If the moment of conscious intention preceded the onset of the RP, then the concept of conscious free will would be tenable: the early conscious mental state could initiate the subsequent neural preparation of movement." (Haggard & Libet 2001, p. 48). Since motor act is not a direct effect of conscious intention (CInt), but of an indirect one of cerebral potential for unconscious initiation of the action (RP) -he concludes, free will (FW) should be revised.

On Libet's viewpoint, intentional actions begin with RP followed by conscious intention. Libet did not register electrophysiological evidence of brain states associated with the content of W-judgments (verbal reports just at the moment of awareness of a choice –W-j) or, according to his analysis, with the "*first awareness of wish to act*" (Libet, 1999, p. 49) –Libet registered the onset of CInt when W-j's was reported.

Two types of data were used by Libet to arrive to his hypothesis, namely, *introspective* and *electrophysiological*; the former was constituted by W-j and M-judgments (verbal reports just at the moment we think that our motor act begins), and the latter by EEG and EMG evidence (fig. 1). His conclusions both combine and depend on these sources of evidence.

The study of FW from Libet''s perspective requires to track causal estimations between two types of data: 'if the moment of conscious intention followed the onset of the RP, then conscious FW cannot exist: a conscious mental state must be a consequence of brain activity, rather than the cause of it' (Haggard & Libet 2001, p. 48). We reject this approach to the explanation of human intentional actions and FW.

Libet's findings have led to a new model (fig. 2) that emerges from a *causal approach* in opposition to the classic model, where intentional action was supposed to be an indirect effect of CInt.

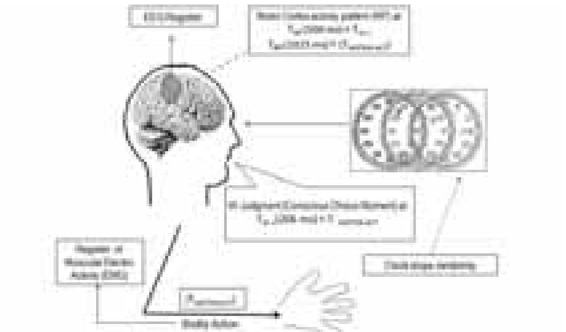

EEG Register
Motor Cortex activity pattern (RP) at
W-judgment (Conscious Choice Moment) at
Register of Muscular Electric Activity (EMG)
Clock stops randomly
Bodily Action

Fig 1. Libet's Experiment for Self-initiated (preplanned) acts (e.g. vid.: Libet, 1983)

[illegible]

After Libet's rejection of the classic concept of FW, he posits that there is a "free won´t" (FWN), since an individual can stop the motor act before its completion –overriding the RP and blocking the triggering of its associated action (*Cf.*: Libet 1985 and 2003). He claims that FW still stands since the subject's intentions are involved in his act of FWN as an act of intentional control.

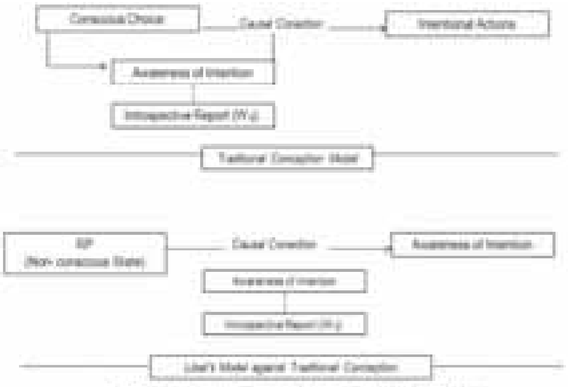


Fig 2. Traditional Conception Model and Libet's New Model

2. Wittgenstein

It is hard to state what Wittgenstein would say about the above mentioned issues – it is difficult enough to summarize what one could consider to be his actual stance on FW. The multiple opinions proposed by him in different occasions in respect to FW make it virtually impossible to draw clear conclusions, but there is some previous work in relation to this concept (remarkably, Hacker 1996, Vol. 4, part V). What then, comes out clear about will? Our first claim is that Wittgenstein –though being obscure on will himself- wasn´t all that wrong compared to the trap in which Libet falls into by rejecting the classic concept of FW based upon the temporal precedence of RP over the motor act.

Two concepts can be appreciated in his early works: «will as an act» and «will as a content of thought» (i.e. an idea). Such concepts reflect the terms of traditional discussion in philosophy: "The will seems always to have to relate to an idea" (8/11/1916; also 11/6/1916) and "The act of the will is not the cause of the action but is the action itself" (*id.*).

Wittgenstein claims that intention (after e.g., flexing your wrist) is properly the *act of the will* in itself, not merely a propositional attitude[1]. This analysis goes from behavior to thought (not inversely). However, Wittgenstein seems to accept that will begins with our desires and with our thought in general (*Cf*.: 21/7/1916); thus, will is not merely a cognitive condition for intentional actions, but also represents the possibility to assign specific contents to thoughts. In Wittgenstein´s words: "this is clear: [...] One cannot will without acting. If the will has to have an object in the world, the object can be the intended action itself. And the will does have to have an object." (Wittgenstein, 08.11.16). In this way, a human being lacking of will seems impossible (see *Id*.): "The will is an attitude of the subject to the world. The subject is the willing subject." (4/11/1916).

Traditionally, one is a free agent if one has intentional actions -if one's actions depend on one's will. Two concepts are problematic here: 'agent' and 'will'. We reject Libet's conclusions because they imply to mistakenly identify subjective choices as being equal to beliefs; for Libet, beliefs are not the cause of intentional actions, since the actual cause is the RP (a state over which the agent has not conscious control of). We claim that the concept of 'agent' in Libet's study is inadequate. For us, RP could mainly be related to prior fixation of the reference for our intentional actions and 'agent' to the relevant domain in the scrutiny of what we call 'efficient causal agent' (an agent that could be accurately accounted for as an actual causal relation avoiding domain confusions).

3. *RP* Revisited

3.1. Content Approach and Cognitive Path

FW debate differs from that of free actions (*vid*. Tugendhat 2006). The latter is about conditions of conscious intentions and choices as a particular aspect of volition, while the former is about conditions of intentional actions i.e., actions made and consciously controlled by an agent (someone doing something). We shall focus now on *cognitive* conditions of conscious intentions; in §4 we will focus on *domain* conditions of intentional actions.

In the *square-in-the-mirror* example Wittgenstein posits that FW might be intrinsically related to the focus of attention (*Cf*.: 4/11/1916). Picking potential stimuli intentionally plays a role in the individuation of an act of the will.

This conception seems to derive from an *intensionality-centred-perspective* (ICP) for intentional actions –for which "What is the relevant mental content to perform intentional actions?" is the main question. An ICP standpoint leads to a question: «What is the relevant mental content controlled by an agent while performing intentional actions?»

From a naturalized view of cognition, we propose that focusing attention is a neurocognitive-process depending on an agent's intentions. Agents have control of this process; FW depends on our dispositions to selectively choose contents of thought and to fixate intentions. Temporal precedence of RP over motor acts leads not to conclude that RP does not depend on attentional fixation; otherwise, RP is *content-dependent* and, therefore (in optimal conditions –excluding, say, hallucinations), *context-dependent*.

Once we have falsely discarded classic FW, we still would need to explain why we think about our actions as effects of our beliefs (why we fall in the "illusion of FW". See Fig. 4). *The resulting analysis is not that our intentions are completely isolated epiphenomenal facts, but our attentional processes precede our intentions, and plausibly, our RPs*. The contrary would depend on evidence of RP associated with the fixation of attention.

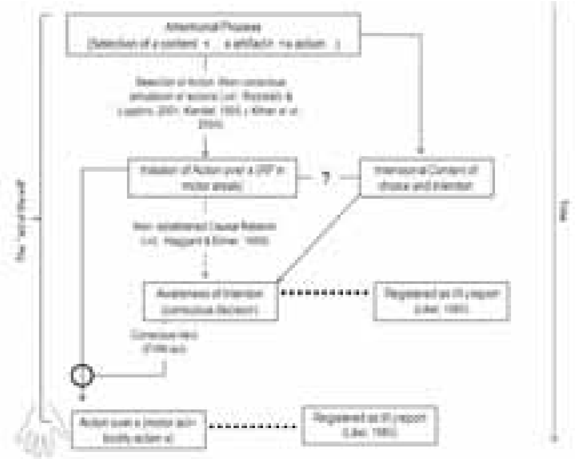

Attentional Process
?
Awareness of Intention (conscious decision)
Registered as W report (Libet 1983)
Registered as M report (Libet 1983)

Fig 3. ICP (a content approach - not a causal eclectic approach as Libet's). Attentional content as core factor in fixing the content of choices and intentions.

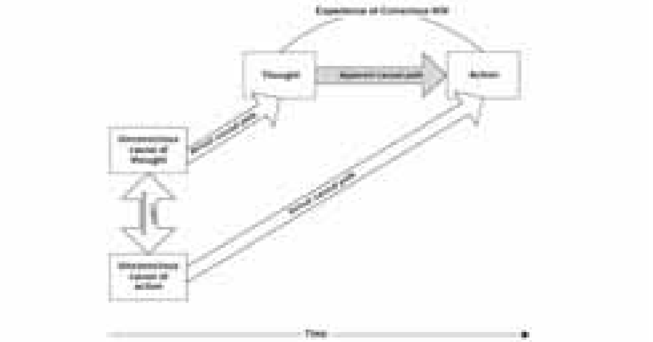

Experience of Conscious Will
Thought
Apparent causal path
Action
Unconscious cause of thought
Actual causal path
Unconscious cause of action
Actual causal path
Time

Fig. 4 "The experience of conscious will arises when a person infers an apparent causal path from thought to action. The actual causal paths are not present in the person's consciousness. The thought is caused by unconscious mental events, and the action is caused by unconscious mental events, and these unconscious mental events may also be linked to each other directly or through yet other mental or brain processes. The will is experienced as a result of what is apparent, though, not what is real." [Fig. 3.1 in Wegner 2002. Reproduced under permission of the author].

[1] For will as a thought, see 14/7/1916.

3.2 Ourselves: Agents

You arrive to your neighbor´s house, knock on the door, he opens and welcomes you. *Who* do you think it was the one that opened the door? His brain? Is your neighbor a brain or a *bunch-of-RPs*? Do you actually greet his brain or, rather, a person? There is an apparent confusion between common understanding of FW and that of neuroscientific approaches.

Paradox: for a radical monist –accepting physical world's causal closure-, brain processes are not unconscious *per se,* but rather are part of a neurobiological flow that generates a physical event called conscious awareness; for a phenomenist or an anti-reductionist, the type of relevant objects that give content to intentional actions are those that you know as a *person* –not as a brain: the door, the doorbell, your friend. Libet’s analysis is somewhere between these two domains.

RP is not an agent, but a factor involved in motor acts of an agent. The tension arises when an apparently monist stance is mixed with the domain in which our concept of will makes sense.

An obstacle is the fact of the vagueness of traditional use of concepts such as 'will' and 'wish' and similar in German (for instance, 'wollen', 'möchten') and Spanish ('querer', 'pretender'). Hacker 1996 speaks of “ambiguities that have characterized the efforts of philosophers to illuminate the nature of the will and of human action” and Bennett & Hacker 2005 draw a similar diagnosis in the case of some neuroscientific explanatory efforts.

Hacker also points out that “philosophers have invented a new use for the words 'will', 'want' and 'volition'.” Following Wittgenstein: “How is "will" actually used? In philosophy one is unaware of having invented a quite new use of the word, by assimilating its use to that of, e.g., the word "wish". It is interesting that one constructs certain uses of words especially for philosophy, wanting to claim a more elaborated use than they have, for words that seem important to us.” (RPP I §51).

To bring meanings of terms from natural language to technical domains is a common habit. Such concepts begin to lose their initial meanings and uses and start to be wrapped by presuppositions of the new domains. Although common, it has not been proven as the best strategy since it seems to be a result of 'traditional anxiety for generality'.

We do not need to track causal connections between a partial state of an agent (e.g. a belief) and his intentional action to destroy the concept of FW; what we need is to undo the causal connection between the agent –be it a whole of neurobiological states or a subject- and his intentional actions. Adopting Libet's approach, the conscious agent seems an epiphenomenal factor reduced to beliefs (registered as W-j) in the causal flow that generates motor act (see Hacker 1996, Id. §2).

There are a lot of processes that biologically compose an agent. The agent does not have control over most of them, but they are causally involved in its actions. One standpoint against FW lies in identifying an agent’s state isolated from the rest of the agent's mental states. This is not Libet’s path: neither he, nor others have demonstrated yet that RP is isolated from other brain states involving conscious content.

In 1963 Walter turned electric brain states (EBS, perhaps RPs) into agents: he connected EBS recorders to the brains of subjects and these to a slide-viewer. Slides were changed by this efficient, but bizarre-electric-agent. In this experiment the efficient causal agent was not human and the subjects' conscious states seemed to be mere epiphenomenal facts, but we are not epiphenomenal states placed somewhere between electric-agents and actions.

4. Conclusions

Libet´s conclusions on FW represent an instance of *mereological fallacy* (*vid*. Bennett & Hacker 2005). The notion of agent is not the same in his works as the one relevant in the dispute for FW. Our *(neuro)cognitive conjecture* is that the processes that lead to fixating our attention are prior to the appearance of RP (Kornhuber & Deecke 1965); fixating our attention is an intentional activity, whereas RP is not such by definition –at least, further research is necessary to settle the dispute (e.g., Kilner *et al.* 2004). Reducing conscious intentions to W-j reports is also inappropriate. Subjective conscious choices and intentional *cognitive processes* are not to be reduced to beliefs -though beliefs, intentions and desires have classically been considered as propositional attitudes with the same logical form-. Finally, a causal account based upon tracking temporal precedence between events pertaining to two sources of evidence is wrong; thus, an ICP seems to bring us to prudent conclusions –for empirical reference on a similar direction see Haggard & Eimer 1999.

Again, we are not epiphenomenal states. Neither Libet, nor others have demonstrated that RP is isolated from other brain states involving conscious content. Philosophers such as Wittgenstein have contributed with elements that neuroscientists are compelled to consider. Philosophical hypothesis seem to give meta-theoretical feedback to scientific theories of mind and brain, despite the associated despise for them and the frantic and systematic ignorance derived from 'traditional anxiety for generality'.

Acknowledgments

We thank Gonzalo Munévar and Peter Hacker.

Literature

Bennett, Max, Hacker, Peter 2005 *Philosophical Foundations of Neuroscience*, Blackwell.

Deecke, Lüder, Scheid, P. and Kornhuber, Hans 1969 "Distribution of readiness potential, pre-motion positivity, and motor potential of the human cerebral cortex preceding voluntary finger movements", *Experimental Brain Research* 7, 2, 158-168.

Hacker, Peter 1996c Wittgenstein: Mind and Will, An Analytical Commentary to Philosophical Investigations 4, Oxford, Blackwell.

Haggard, Patrick and Eimer, Martin 1999 "On the relation between brain potentials and conscious awareness", *Experimental Brain Research*, 126, 128–133.

Haggard, Patrick and Libet, Benjamin 2001 "Conscious Intention and Brain Activity", *Journal of Consciousness Studies*, 8, # 11, 47-63.

Kandel, Erick Schwartz, James & Jessell, Thomas 1995 *Essentials of Neural Science and Behavior*, Hertfordshire, Prentice Hall.

Kilner James, Vargas, Claudia, Duval, Sylvie, Blakemore, Sarah-Jayne and Sirigu, Angela 2004 "Motor activation prior to observation of a predicted movement", *Nature Neuroscience*, Vol.: 7, # 2, 1299-1301.

Kornhuber, Hans and Deecke, Lüder 1965, "Hirnpotentialänderungen bei Willkürbewegungen und passiven Bewegungen des Menschen: Bereitschaftspotential und Reafferente Potentiale", *Pflügers Archiv*, 284, 1-17.

Libet, Benjamin 1985 "Unconscious cerebral initiative and the role of conscious will in voluntary action", *Behavioral and Brain Sciences*, 8, 529-566.

Libet, Benjamin 1999 "Do We Have FW?", *Journal of Consciousness Studies*, 6, 8- 9, 47-57.

Libet, Benjamin 2003 "Can Conscious Experience Affect Brain Activity", *Journal of Consciousness Studies*, 10, 2, 24-28.

Libet, Benjamin, Gleason, Curtis, Wright Elwood and Dennis Pearl 1983 "Time of conscious intention to act in relation to onset of cerebral activity (readiness potential): The unconscious initiation of a freely voluntary act", *Brain*, 102, 623–42.

Rizzolatti, Giacomo and Luppino, Giuseppe 2001 "The Cortical Motor System", *Neuron* (Sept.), 31, 889-901.

Tugendhat, Ernst 2006 "Libre albedrío y determinismo", *El Hombre y la Máquina*, 26, 80-87.

Wegner, Daniel 2002 *The Illusion of Conscious Will*, Cambridge Mass and London, England, Bradford Books and MIT Press.

Wittgenstein, Ludwig 1916 Notebooks 1914- 1916, *The Collected Works of Wittgenstein*, Wright, George and Anscombe, Gertude, (eds.), Oxford, Basil Blackwell.

Wittgenstein, Ludwig 1980 *Bemerkungen über die Philosophie der Psychologie* (Tss 229, 232, 244-245).

Mental Causation and Physical Causation

Lorenzo Casini, Canterbury, Kent, England, UK

1. Introduction

The recent debate between Antony and Kim on the nature of mental causation offers the possibility to evaluate the reason underlying two up-to-date physicalist positions: Kim's Reductive Physicalism and Antony's Non-Reductive Physicalism. Despite differences, both share a common metaphysical task. They look for a *systematic account* of the relations between the physical and the mental, which is needed, so they say, because higher-level properties can enter into 'genuine' laws only if they inherit the causal power of ontologically prior lower-level entities. This means that there could not be regularities concerning mental states without underlying physical mechanisms. In particular, only the physical mechanisms at work at the microlevel can secure and explain the truth of psychological generalisations. Only at the microlevel, Antony and Kim argue, we find the entities involved in 'genuinely' causal phenomena.

I show that if the paradigmatic feature which the microphysical is to display is that it conforms to a particular model of causal *production*, as Kim explicitly suggests, this prevents Reductive- and Non-Reductive Physicalism to achieve their tasks. In fact, certain quantum mechanics' phenomena cannot be described in terms of causal production. If we accept a statistical-regularist approach to describe phenomena in the quantum domain, then quantum mechanics and psychological phenomena are on a par with respect to their causal features. The physicalists, who claim the necessity to account for the mental in physical terms, is to clarify what peculiar feature microphysical mechanisms possess, and the mental is to inherit, for psychological generalisations to be secured and explained.

2. The Metaphysical Picture: Physicalism and Reductionism

Both Antony and Kim conceive the world as *layered*, i.e., made of different levels organised in a hierarchical structure. The determining level, the physical bottom level, is *ontologically prior* to all the other higher levels, because its entities stand with those of the higher levels in a part-whole relation, such as that which occurs between one oxygen and two hydrogen atoms and a H_2O molecule. However, at each level there are properties which make their first appearance at that level. For instance, properties like density or viscosity of H_2O molecules were not present at the lower level of their atomic constituents. In particular, (i) the entities of psychology, such as sensations and propositional attitudes, are nothing over physical complexes, such as patterns of neurons; (ii) each mental property (e.g.: a toothache, the belief that 'the water is wet', etc.) is a property of some physical entity or system of physical entities (e.g.: an underlying pattern of neurons). The question, then, is: How can mental properties have the causal power they have in a world ultimately constituted by physical entities and mechanisms?

Antony's and Kim's answers, however, are different. The former does, whereas the latter does not, accept that systems of lower-level entities can acquire *mental* properties, i.e., *mental* causal powers, which are *emergent* from the lower base, and *non-ontologically-reducible* to it. For Kim and Antony, a property at a given level is emergent, iff non-ontologically-reducible to the lower-level property it emerges from. For Antony, the properties of the psychological domain are like H_2O with respect to its atomic constituents. This would legitimate the autonomy of psychology, whose properties must inherit their causal power from lower-level physical entities but are not to be reducible to the properties of these entities, on pain of identifying psychology with branches of physics. In order to meet these desiderata, she wants reductive explanation without ontological reduction. In contrast, Kim claims that psychological properties do not constitute a proper scientific domain, given that they can be ontologically reduced to physical properties. In fact, so he reasons, they are *not*, strictly speaking, emergent as H_2O is, insofar as they apply to precisely the same objects as do their realiser properties—i.e., mental properties and their realisations are properties of entities at the *same* level, and have the same causal powers (Kim 1998, 82-3). The difference between their positions depends on whether or not multiple realisability (MR) holds. According to MR, each mental property can be realised by many distinct physical properties—therefore is not identical to any of them. If MR is true, then no reduction is possible, and psychology is autonomous (Antony 2007, 154-5).

Kim does not accept the idea that there is *in principle* an indefinite number of realisers among the individuals of the same species or structure type. That is, he challenges the *truth* of MR, and claims that a structure-specific reduction is—*in principle*—possible, granted that the physical realisers the psychological properties are reduced to are sufficiently similar to one another (Kim 1993, 89, 313). Mental properties can be identified with physical properties which play the same causal role—for each mental property there is also a physical property which is *necessary and sufficient* for the mental property to arise, given that they are, in fact, one and the same (Kim 2006, 280). Mental terms which stand for disjunctions of different physical properties, instead, have no ontological correlates—therefore no scientific value (Kim 1993, 334-5). As a consequence of reduction, psychology loses its proper subject matter, and together its autonomy.

For Antony, in contrast, ontological reduction is impossible because of MR. In order to vindicate mental causation *as mental* and justify the autonomy of psychology, whilst *consistently* holding that the physical is ontologically prior, she advocates the possibility—*in principle*—of a reductive explanation of every mental property in terms of physical properties, such that some physical property is *sufficient* but *not necessary* for a mental property to emerge. Mental properties are properties of some physical system or other—therefore ontologically acceptable, *and* proper scientific kinds, because they enter into realisation-independent regularities, i.e., regularities which do not depend necessarily on one specific physical property (Antony, Levine 1997, 92-4).

3. Back to Basics

I do not explain the differences between Kim's model for ontological reduction and Antony's model for reductive explanation. Instead, I stress a fundamental similarity between the two approaches. The task of both Antony's and Kim's metaphysical projects is "back to basics".

For Antony, reductive explanation of psychological regularities in terms of basic physical entities and mechanisms constitutes the *necessary* metaphysical condition for an explanation to be true. We need 'a *systematic* account of mental phenomena in terms of physical microstructures' (Antony, Levine 1997, 94-ff.). Although—for Antony—there are regularities that cannot be apprehended at more basic levels of descriptions, such as those of psychology, these 'entail the existence of some ultimately physical mechanism', 'a pattern of lower-level events that guarantees, contingent on features of the background, the emergence of some higher level regularity' (Antony 1995, 441).

The same holds for Kim. In fact, whether or not reduction succeeds depends on the possibility to identify *at least* the sufficient condition for the higher-level property to inherit its causal power from its lower-level realiser, in order for higher-level generalisations to be linked with "real" entities and mechanisms: 'The psychological capacities and mechanisms posited by a true psychological theory must be *real* [italics mine], and the only reality to which we can appeal in this context seems to be physical reality' (Kim 2006, 161). *Macro*causation, i.e. causation at any higher level, can be proved to be "real" only if systematically linked to *micro*causation, i.e., the causation at work at the bottom physical level, out of which it emerges (Kim 1993, 100).

For both Antony and Kim the *real causal job* is only done by "real" entities, i.e., entities which belong to the ultimate ontology of the layered world. Any higher-level observed regularity is maintained by some "genuinely causal" interaction between ontologically prior physical entities. Thus, mental laws describing these regularities have explanatory force only if linked to microphysical causal mechanisms. Some interesting questions arise. *First*: What is the "genuine" feature of the mechanisms at the microlevel which guarantees the truth of explanations at higher levels? *Secondly*: What are the "real" entities involved in these mechanisms?

Antony claims that a 'physical model of causation' is to be applied to mental events (Antony, Levine 1997, 102), but, regrettably, she does not go much further. It is clear, however, that this model of physical causation is neither regularity-based nor counterfactuals-based, insofar as these are exactly the kinds of causation—holding for the mental—that she is not satisfied with.

Kim shares the same perplexities but he is much more explicit (Kim 1998, 45, 71; Kim 2007, 230-5). The problem of mental causation cannot be resolved by invoking a regularist-nomological or a counterfactual-dependence approach to causation, *real* causation being "production", or generation. What Kim means by model of causal production is something close to a Salmon-Dowe *conservative quantity theory* of physical causation (CQ) (Kim 2007, 240 n.13; Dowe 1992; Salmon 1994). Kim's reasons for preferring this kind of causation are that only the notion of causal production (i) permits the distinction between real causal processes and pseudo-processes—i.e., processes generating accidental, non-lawlike, regularities, rendering dispensable the use of nomological- and counterfactuals-based regularities (Kim 1993, 93-ff.; Kim 1998, 45; Kim 2007, 231) and (ii) has the characteristic of *locality*, for which 'causes are connected to their effects via spatiotemporally continuous sequences of intermediaries'—i.e., generate their effects via processes which propagate in spacetime along a continuous trajectory (Hall 2004, 225; Kim 2007, 235). As Kim puts it, human agency, i.e., the capacity to perform actions in the physical world on account of beliefs, desires, etc, 'requires the productive/generative conception of causation' (Kim 2007, 236). Thus, mental causation can be secured and explained only by backing psychological regularities to causally productive mechanisms.

4. The 'reality' of microcausation

Unfortunately, there are strong reasons to doubt that production can do the job. In fact, this model does not apply to those phenomena where action-at-a-distance seems to occur (Hall 2004, 226, Salmon 1984, 210, 242-59; Salmon 1998, 23, 224, ch. 16). In fact, there are quantistic phenomena, where no continuous spatiotemporal process can be identified, such as the well-known problem of EPR causal anomaly—it takes the name of Einstein, Podolsky, and Rosen, who formulated it in 1935, charging quantum mechanics of incompleteness. Consider a quantum system consisting in an atom of positronium—a positron (positive electron) and a negative electron orbiting about one another. The system's total intrinsic angular momentum, or spin, is zero. Let the particles be separated from one another without affecting the angular momentum of the total system or of either parts. The EPR problem is that a measurement performed upon the positron seems to influence the physical state of the electron, even if there is no physical interaction between the two at the time of the measurement. The enigma is how the *remote* parts of the system can react *instantaneously*, i.e., without the medium of a causal process in spacetime, to a *local* interaction with one of the parts. This is the ground for Einstein's opposition to quantum mechanics: either quantum mechanics is incomplete—i.e., there are "hidden variables" explaining the phenomenon, or the relationship between momentum and position is "non-real" (Mehra 1974, 70-1). However, no later studies have discovered the presence of hidden variables and dissolved the problem. As Salmon himself admits, a single consistent description that explains what happens in terms of spatiotemporally continuous causal processes and local causal interactions cannot be given for the quantum domain (Salmon 1984, 245).

Quantistic phenomena are currently considered *genuinely* and *irreducibly* stochastic. Obviously, this does not exclude that quantistic laws are incorrect, or that quanta are not the ultimate microparticles, but I do not see how it can *suggest* that causal production is at work at the microlevel. Interestingly enough, Kim's desideratum of locality, as a continuous sequence of causal intermediaries in spacetime, cannot be met exactly with reference to mechanisms involving the physical entities of the microlevel. Appeals to regularities or counterfactuals are *not dispensable* at the microlevel. Does this mean that we have to deny the 'reality' of the phenomena of quantum mechanisms and treat them as pseudo-processes? Or can we be content with a causal explanation in terms of statistical correlations, i.e., a regularist approach?

Notice that Kim (2007, 232) concedes that only regularities and Humean "constant conjunctions" may be present at the microlevel. For him, this means either that (i) 'it makes no sense to speak of "underlying" mechanisms, or "real" causal processes at a lower level', or that (ii), 'although only "constant conjunctions", but no causation, exist at the fundamental level […], causal relations can, and do, exist (or "emerge") at higher levels'.

But the problem is not evaded: (i) if we accept microlevel regularities as having a real—yet non-productive—causal role, *why* should we still hold that the possibility of mental causation rests *necessarily* on that of reduction? In fact, reducing and reduced phenomena do not differ with respect to the "genuinity" of their causal features; (ii) if we claim that real causation exists or emerges only at higher levels, *why* is reduction of mental phenomena to microphysical phenomena *necessary*, given that the microworld lacks the essential feature which secures and explains mental causation? To say that causal relations emerge at higher levels does not help, once entities and features of the microworld are taken as paradigmatic and ontologically prior.

Far from regarding only Kim, the problem regards Antony too. As mentioned, she fails to specify what she means by physical causation. Nonetheless, she claims that a physical model of causation is to be applied to mental events. If by physical causation she means production, the same objections against Kim hold. If she means something different, mental causation is not different from, and no more genuine than, microphysical causation.

5. Conclusion

The reason underlying Kim's Reductive Physicalism and Antony's Non-Reductive Physicalism is that only the mechanisms at work at the microlevel can secure and explain the truth of psychological generalisations. I have shown that, if the supposed feature that these mechanisms should have, and that mental ones inherit from them, is that they conform to a CQ model of causal production, then psychological and microphysical laws are on a par with respect to their causal features. In fact, there are microphenomena not explainable in terms of continuous sequences of causal intermediaries in spacetime, as the CQ model requires.

Antony and Kim have gone to great effort to convince us that—the possibility *in principle* of—a systematic link between mental properties and regularities and physical entities and mechanisms is *necessary*, because only the latter, so they argue, can secure and explain the truth of the explanations given by means of the former. But they do not put the same effort in telling us *on what* features of the microlevel, which the mental level lacks, the truth of our psychological generalisations depends. I would urge them to specify what it is that distinguishes physical mechanisms from mental regularities, and renders the former the secure basis for the latter, in order to convince us about the necessity of their enterprises, whether explanatorily or ontologically reductive.

Literature

Antony, Louise 1995 "Law and Order in Psychology", *Philosophical Perspectives*, 9, 429-446.

Antony, Louise 2007 "Everybody Has Got It: A Defense of Non-Reductive Materialism", in: McLaughlin, Brian L. and Cohen, Jonathan (eds.), *Contemporary Debates in Philosophy of Mind*, Oxford: Blackwell, 143-59.

Antony, Louise, and Levine, Joseph 1997 "Reduction With Autonomy", *Philosophical Perspectives*, 11, 83-105.

Dowe, Phil 1992 "Wesley Salmon's Process Theory of Causality and the Conserved Quantity Theory", *Philosophy of Science* 59, 195-216.

Hall, Ned 2004 "Two Concepts of Causation", in: Collins, John, Hall, Ned, and Paul, Laurie A. (eds.), *Causation and Counterfactuals*, Cambridge, MA: MIT Press, 225-76.

Kim, Jaegwon 1993 *Supervenience and Mind*, Cambridge University Press.

Kim, Jaegwon 1998 *Mind in a Physical World*, Cambridge, MA: MIT Press.

Kim, Jaegwon 2006 *Philosophy of Mind*, Cambridge, MA: Westview.

Kim, Jaegwon 2007 "Causation and Mental Causation", in: McLaughlin, Brian L. and Cohen, Jonathan (eds.), *Contemporary Debates in Philosophy of Mind*, Oxford: Blackwell, 227-42.

Mehra, J. 1974 *The Quantum Principle: Its Interpretation and Epistemology*, Dordrecht-Holland, Boston-U.S.A.: D. Reidel Publishing Company.

Salmon, Wesley C. 1984 *Scientific Explanation and the Causal Structure of the World*, Princeton: Princeton University Press.

Salmon, Wesley C. 1994 "Causality Without Counterfactuals", *Philosophy of Science*, 61, 297-312.

Salmon, Wesley C. 1998 *Causality and Explanation*, Oxford University Press.

On Two Recent Defenses of The Simple Conditional Analysis of Disposition-Ascriptions

Kai-Yuan Cheng, Chia-Yi, Taiwan

I. Introduction

A wide variety of reductionist projects in philosophy appeals to dispositions to do the work. Dispositional analyzes can be found in the areas of inquiry on mental states (Ryle, 1949), meaning (Kripke, 1982; Quine 1960), colors (McGinn, 1983), values (Lewis, 1989), goodness (Smith, 1994), properties (Shoemaker, 1980), and so on. That the dispositional explanatory strategy is broadly adopted by reductionists is not hard to explain. A traditional view, which is rooted in empiricism (see Bricke, 1975) and continues to be shared by contemporary philosophers, such as Carnap (1936), Goodman (1955), Quine (1960), Mackie (1973), Prior (1985), and many others, analyzes a disposition-ascription “x has D” in terms of a simple counterfactual conditional “If x were p, x would q”, which mentions only a pair of possible events. If this analysis were correct, dispositional properties would be themselves reduced to mere possibilities of events, and thus rendered ideal to figure in reductive accounts of other properties regarded as captivating and problematic.

Things are not so straightforward, however. Counterexamples to the simple conditional analysis have been offered by Martin (1994), Smith (1977), Johnston (1992), and Bird (1998), and are extensively considered as decisive in refuting the analysis in question. The nature of dispositions is consequently not as simple as the conditional analysis seems to suggest. Viewing a disposition as a robust property and not merely as possible events is an expected result. However, exactly how to characterize it has become a major challenge and focus of heated debates for contemporary metaphysicians (e.g., Armstrong, Martin, & Place, 1996; Mumford, 1998; etc.).

Against this realist trend, recently two philosophers stand out—Choi (2006) and Gundersen (2002)—in defending the simple conditional analysis of dispositions (see Fara, 2006). They make a glaring claim that various counterexamples fail to refute the simple conditional analysis. Their attempts to reduce disposition-ascriptions to conditionals, if successful, would lead to “the ontological consequence that there are no dispositions *qua* properties” (Mumford, 1998). Given the significance of this issue, the aim of this paper is to examine whether these two philosophers succeed in their attempts. I shall argue that they do not, and show that each founders on a similar ground. Below I will begin with a brief review of the counterexamples raised by Martin and Bird, to which Choi and Gundersen have aimed at responding.

II. Counterexamples to A Simple Conditional Analysis by Martin and Bird

According to a simple conditional analysis, a disposition-ascription is analyzed into a counterfactual conditional. Take fragility for example. A simple conditional analysis has it that DA iff CC:

> DA. Something x is fragile at time t.
> CC. If x were to be struck at t, then it would break.

Martin (1994) considers a pair of cases with an example to show that this bi-conditional analysis fails in both directions. To use a variant of Martin’s (1994) electro-fink example, imagine that a sorcerer brings about an effect on a glass in the following two ways (this case is due to Lewis, 1997): i) as soon as a fragile glass is about to be struck, the sorcerer protects the glass from breaking by instantaneously casting a spell that renders the glass no longer fragile; ii) as soon as a non-fragile glass is about to be struck, the sorcerer renders it fragile and causes it to break when struck. In case i), DA is true, but CC is false. This means that CC is not necessary for DA. In case ii), DA is false while CC is true. This means that CC is not sufficient for DA. As a result, disposition-ascription is not logically equivalent to a conditional. Martin infers from this result that dispositions *qua* real properties cannot be reductively explained by conditionals.

Lewis (1997) takes Martin’s (1994) refutation of SCA as decisive, but maintains that a conditional analysis can be remedied by refining it as follows:

> RCA. Something x is disposed at time t to give response r to stimulus s iff, for *some intrinsic property* B that x has at t, for some time $t^{\prime}$ after t, if x were to undergo stimulus s at time t and retain property B until $t^{\prime}$, s and x’s having of B would jointly be an x-complete *cause* of x’s giving response r.

where an x-complete cause is “a cause complete in so far as havings of properties intrinsic to x are concerned, though perhaps omitting some events extrinsic to x” (Lewis, 1997, p. 149). Lewis’s proposal consists of two main ideas: 1) to have a disposition is to have *some intrinsic property* that serves as the causal basis of giving response r upon receiving stimulus s; 2) the clause of retaining the intrinsic property B during the time lag between t and t’ can deal with Martin’s counterexample.

It is worth noting that Lewis does not seem to apply RCA directly to deal with Martin’s counterexample. Choi (2006, p. 370) brings our attention to Lewis’s taking two different steps in coming up with an analysis of a disposition-ascription (1997, p. 142-146). The first step is to put an ordinary disposition-ascription such as DA into an “overly dispositional locution” by specifying the stimulus and the response of fragility as follows:

> ODL. Something x is fragile at time t iff x has the disposition at t to give the response of breaking to the stimulus of being struck.

The second step is to apply RCA to ODL to yield the following analysis of fragility:

> RCA*. Something x is fragile at time t iff, for *some intrinsic property* B that x has at t, for some time $t^{\prime}$ after t, if x were to be struck at time t and retain property B until $t^{\prime}$, x’s being struck and x’s having of B would jointly be an x-complete *cause* of x’s giving response r. (c.f. Choi, 2006, p. 371)

Noting this two-step procedure inherent in Lewis’s analysis is crucial to our subsequent discussion and evaluation of Choi’s own position.

RCA* handles Martin's counterexample nicely. It correctly dictates that a glass would be attributed with fragility, *if it were to retain the intrinsic property* when struck. The analysis also justly predicts that a glass would not be ascribed fragility in the second case. This is because if the glass were to retain the intrinsic property between t and t', that property would be causally irrelevant to breaking the glass; what causes the glass to break in this case is some extrinsic factor, i.e., the sorcerer.

Bird (1998) argues, however, that Lewis's analysis remains a failure, given the cases of antidotes. An antidote is defined by Bird as "something which, when applied before t', has the effect of breaking the causal chain leading to r, so that r does not in fact occur" (1998: p. 228). An example of an antidote is a physical device that absorbs the shock waves of a glass when struck. Consequently, the glass retains its fragility at t' but does not break when dropped, thanks to the device. In this case, the *analysandum* on the left-hand-side of RCA* is met, but the *analysans* on the right-hand-side of RCA* is not fulfilled. This means that a conditional is not necessary for a disposition-ascription. Another counterexample that works in a converse order is offered by Lewis himself (1997, p. 145-146). A styrofoam S is not fragile. But as soon as the Hater of Styrofoam hears the distinctive sound made by S when struck comes and tears S apart by brute force. In this case, the *analysans* is true: it is clear that if S were to be struck and retained its intrinsic property B, the striking and B would jointly be an S-complete cause of S's breaking. However, the *analysandum* is false: S is plainly not fragile. This is a case of mimickers. It shows that a conditional is not sufficient for a disposition-ascription. Lewis's RCA* thus has to be rejected by the two counterexamples (see Johnston, 1992, for making similar points).

III. Choi's Two-Step Approach

Choi's (2006) defense of the simple conditional analysis of disposition-ascriptions is taken through an indirect route. He first argues that Lewis's two step procedure can be suitably exploited to restore Lewis's own reformed conditional analysis from the antidotes and mimickers counterexamples. He then shows that the same approach can be adopted to develop a plausible simple conditional analysis which can equally defeat all the relevant counterexamples including Martin's fink cases. Consequently, Lewis's original motivation for advocating a reformed conditional analysis is invalidated. Moreover, given that the simple conditional analysis is ontologically more economic, with no commitment to construing a disposition as an intrinsic property, the simple version should be preferred to the reformed version. I shall argue that despite Choi's illuminating discussion and intriguing suggestion, the two step approach does not escape a basic problem which Martin raises for the simple conditional analysis.

To see how the two step approach works, first consider how Lewis himself deals with the Hater of Styrofoam case. Lewis maintains that S obviously does not qualify as a fragile object, because its breaking does not go through *a certain direct and standard process* (1997, p. 145). Lewis suggests that ODL be revised by adding this constraint to the specification of the manifestation of S, which is the first step of the analysis. The second step is to apply RCA, which is kept intact, to this revised form of ODL. The result will be a new analysis which dictates that S is not fragile, since S goes through an indirect and non-standard process of manifestation which renders the conditional on the right-hand-side of the bi-conditional analysis false.

Choi's innovating idea is to adopt a similar method to treat the presence of fragility-antidotes as a non-standard stimulus condition, which a plausible ODL had better exclude in its formulation. Generalizing these two counterexamples, the Styrofoam and antidote cases, Choi (2006, p. 373) proposes that the following two steps be taken. The first step is to revise ODL:

> ODL'. Something x is fragile at time t iff x has the disposition at t to exhibit a *fragility-specific manifestation* in response to a *fragility-specific stimulus*,

where a fragility-specific stimulus includes x's being struck in the absence of antidotes to fragility, and a fragility-specific manifestation includes x's breaking through a certain direct and standard process. The second step is to apply RCA to ODL' to produce a new analysis of fragility:

> RCA**. Something x is fragile at time t iff, for *some intrinsic property* B that x has at t, for some time t' after t, if x were to undergo a fragility-specific stimulus at time t and retain property B until t', s and x's having of B would jointly be an x-complete *cause* of x's exhibiting a fragility-specific manifestation.

RCA** can thus well handle the Styroform and antidote counterexamples.

Choi (2006, p. 374) then makes a crucial claim that the same two-step strategy can be adopted to restore the simple conditional analysis of the following form:

> SCA. Something x has the disposition at time t to give response r to stimulus s iff, if x were to undergo s at time t, it would give response r.

The procedure is to take the first step of adopting the revised ODL' instead of ODL, and then take the second step of applying SCA to ODL' to imply a new analysis of fragility:

> SCA*. Something x is fragile at time t iff, if x were to undergo a fragility-specific stimulus at t, it would exhibit a fragility-specific manifestation.

SCA* can overcome the Styrofoam and antidote cases. For the Styrofoam S would not break when struck *in the absence of fragility-mimickers*, and hence would be correctly classified as non-fragile. The glass would break when struck *in the absence of fragility-antidotes,* and hence would be qualified as fragile. Choi also quite compellingly shows that SCA* can handle Martin's fink cases, if the specification of ODL' in the first step of the analysis suitably includes *the absence of finks* like the sorcerer (2006, p. 375-376). Given that SCA* can counteract all the counterexamples as well as RCA** does, without having to introduce an intrinsic property B as x's causal basis in its formulation, Choi concludes that the simple conditional analysis is superior to Lewis's reformed conditional analysis, under the framework of the two step approach.

The problem that Choi's two step approach to restoring the simple conditional analysis faces seems to be this. The key to dealing with counterexamples in this analysis is to focus on the first step, by formulating an ordinary disposition-ascription into an overtly disposition locution in such a way that it excludes certain factors which might causally interfere with the typical manifesting process in response to a typical stimulus. For example, when specifying a fragility-specific stimulus, the analysis includes the absence of fragility-finks, fragility-antidotes, fragility-mimickers, and relevant others. For this formulation to work, however, it has to specify a full list of factors which are relevant to bringing about counterexamples to the analysis. How to provide such a list is, as Choi himself acknowledges, "a nontrivial and indeed hard problem" (2006, p. 377). What seems to be worse is that it is hard to see how this task could be done without having to presuppose the very dispositional concept *fragility*, or even invoking the concept itself. Doesn't the concept of fragility, when put into an overtly dispositional locution, simply becomes one "which nothing prevents it from being fragile"? This would be strikingly circular.

The difficulty involved here is, in my view, not different from the problem for proponents of the original simple conditional analysis who try to handle the fink cases by adding a *ceteris paribus* clause to the antecedent of the conditional. The trick is to enable us to treat the presence of finks as a condition where other things are not being equal, and thus allow us to legitimately exclude the fink counterexamples to the conditional analysis. As Martin (1994, p. 5-6) convincingly points out, however, the idea of introducing the *ceteris paribus* clause is to include the set of all the events which would *bring about the same effects* as finks, and this simply amounts to stating that *nothing happens to make it false that the disposition in question is in place*. This modified simple conditional analysis is blatantly circular. It seems to me that the simple conditional analysis in Choi's two-step approach merely transfers the circularity problem from the level of a conditional (in the second step) to the level of formulating an overtly dispositional locution (in the first step), without making a genuine progress over the original version discussed by Martin.

IV. Gundersen's Appeal to Standard Conditions in Subjunctive Conditionals

The basic objection to the simple conditional analysis SCA relies on an intuitive and gripping picture of the world, which is nicely expressed by Bird (2000, p. 229) as follows:

> Some object might possess a disposition, and continue to have it, and also receive the appropriate stimulus, yet fail to yield the manifestation.

Bird's explanation of this widespread phenomenon is also a natural one: antidotes (might) exist and interfere with the causal process leading to the manifestation of a disposition. Gundersen (2002) examines several ways of construing and defending Bird's antidote counterexamples to SCA, and argues that none of them works. Below I shall focus on one of these lines of argument, and show why I think Gundersen does not make a compelling case for the defense of SCA.

Gundersen first points out that Bird's antidote counterexamples can be given a modalized reading, as suggested by Bird's own expressions:

> The state of the world we are interested in is one described, albeit incompletely, in my illustrative story. It is one that includes among other things the context of the boron rods being lowered and the presence of the relevant stimulus for [the pile's disposition to chain react]. I shall call this state *w.* It is sufficient for a counter-example to the conditional analysis to show that *w* is possible, where it is the case that in *w*, [Fx] is true and *m* is false. It is agreed that in w, [Fx] and [- m if the boron rods are lowered]. Since, as just remarked, w includes the context [of the boron rods being lowered], it follows that in w, [- m]. (Bird, 2000, p. 232; c.f. Gundersen, 2002, p. 400)

In Gundersen's understanding, Bird regards a disposition as an intrinsic property, which renders the *analysandum* (a disposition-ascription) of SCA true in whatever context the disposition is (or might be) in, and is also simultaneously committed to an *ultra-contextualism*, according to which the *mere possibility* of a world state *w* renders the *analysans* (a subjunctive conditional) of SCA false.

Gundersen then maintains that an ultra-contextualism regarding subjunctive conditionals is untenable. The reason is that it amounts to the thesis that a super-causal link exists between stimulation and manifestation; put differently, it gives us an understanding of subjunctive conditional in terms of strict entailment where the consequent is true in every possible world in which the antecedent is true. Gundersen contends that this is a thesis too strong and unreasonable to be accepted, stating that "no one believes an object has a certain dispositional property if and only if the characteristic manifestation *must* be displayed whenever stimuli conditions obtain" (2002, p. 401). Gundersen claims that SCA is as good as it stands, and what needs to be discarded is the following modalized version of SCA:

> SCAm. Necessarily, something x has the disposition at time t to give response r to stimulus s Iff, if x were to undergo s at time t, it would give response r.
> (c.f. Gundersen, 2002, p. 401)

Gundersen thus seems to suggest that SCA holds, even given counterexamples such as those raised by Bird. This means that Gundersen must think that there are certain cases, cases that do not include counterexamples, in which a subjunctive conditional in SCA is rendered true. What then are those cases?

Gundersen has an answer to the above query. It goes as follows (2002, p. 402):

> … subjunctive claims only require for their truth a causal link which typically associates them in standard, or better, sufficiently nearby environments.

We may continue to ask: What are those environments, which are deemed *standard*, or *sufficiently nearby*, in which subjunctive claims are rendered true? To this question, Gundersen admits that "that surely is a hard question", but insists that subjunctive semantics depends on an implicit acknowledgement of such standard conditions" (2002, p. 402). Gundersen claims that the standard in question is objective, which serves as the ground for our making subjunctive claims. Nonetheless, Gundersen appears to leave such a standard unspecified.

This is highly unsatisfactory. In a simple conditional analysis, we rely on a subjunctive conditional to inform us whether a disposition-ascription is true. In the version recommended by Gundersen, it is a subjunctive conditional under standard conditions that fulfills this task. However, we are not provided with any explicit specification of what those standard conditions are or any method of how to identify them. We are then on no sound ground to determine whether a disposition-ascription is true or not. In other words, the simple conditional analysis faces a dilemma. On one horn, it lacks a clear specification of the standard conditions in question, and hence renders a subjunctive conditional of SCA vague and undetermined in its truth-value. On another horn, to specify it would risk presupposing the disposition under inquiry, and hence renders the analysis circular. Either horn of the dilemma seems to render Gundersen's defense of the simple conditional analysis futile.

V. Conclusion

The simple conditional analysis of disposition-ascriptions is well motivated, given its implication for shedding light on the ontology of dispositions and for the prospects of reductionist projects in a wide variety of philosophical inquiries. However, some basic difficulties seem to persistently plague any attempts to advocate such an analysis. The difficulties in question have to do with how the analysis handles counterexamples to it. Either some phrase like the *ceteris paribus* clause has to be added to the antecedent of a conditional in the analysis, which is notoriously vague, or the phrase has to be specified clearly, which ends up unavoidably circular.

Choi and Gundersen seem to run into similar difficulties in each of their sophisticated defenses of the simple conditional analysis. Choi's two-step approach separates the task of formulating a disposition-ascription into an overtly dispositional locution from that of giving the dispositional locution a conditional analysis. The hope is to keep the conditional analysis intact, while let the formulation in the first step do the trick of dealing with counterexamples. It turns out that the formulation is either incomplete, or circular when further specified. This leaves the analysis as a whole deeply problematic. Gundersen, on the other hand, holds that counterexamples do not refute a subjunctive conditional, because there is an objective standard which determines when the causal link between manifestation and stimulus specified by the conditional obtains. Such a standard, however, is merely left unspecified. It remains a daunting challenge to give a substantial specification of the standard in question without rendering the analysis circular. In conclusion, it appears that the prospects of restoring the simple conditional analysis are dim.

Literature

Armstrong, D. Martin, C. B. & Place, U. T. 1996: *Dispositions: A Debate*, London: Routledge.

Bird, A. 1988: “Dispositions and Antidotes”, *The Philosophical Quarterly* 48: 227-234.

---2000: “Further Antidotes: A Response to Gundersen”, *The Philosophical Quarterly* 50: 229-33.

Bricke, J. 1975: “Hume's Theory of Dispositional Properties”, *American Philosophical Quarterly* 10, 15-23.

Carnap, R. 1936: “Testability and Meaning”, *Philosophy of Science* 3: 420-468.

Choi, S. 2006: “The Simple vs. Reformed Conditional Analysis of Dispositions”, *Synthese* 148: 369-379.

Fara, M. 2006: “Dispositions”, Standford Encyclopedia of Philosophy.

Goodman, N. 1954: *Fact, Fiction and Forecast*, Cambridge, Mass.: Harvard University Press.

Gundersen, L. 2002: “In Defense of the Conditional Account of Dispositions”, *Synthese* 130: 389-411.

Johnston, M. 1992: “How to Speak of the Colors”, *Philosophical Studies* 68: 221-263.

Kripke, S. 1982: Wittgenstein on Rules and Private Language, Cambridge, Mass.: Harvard University Press.

Lewis, D. 1989: “Dispositional Theories of Value”, *The Proceedings of the Aristotelian Society*, Supplementary Volume 63: 113-137.

---1997: “Finkish Dispositions”, *The Philosophical Quarterly* 47, 143-158.

Mackie, J. L. 1973: *Truth, Probability and Paradox*, Oxford: Oxford University Press.

Martin, C. B. 1994: “Dispositions and Conditionals”, *The Philosophical Quarterly* 44:

1-8.

McGinn, C. 1983: The Subjective View: Secondary Qualities and Indexical Thoughts, Oxford: Clarendon Press.

Mumford, S. 1998: *Dispositions*, Oxford: Oxford University Press.

Prior, E. W. 1985: *Dispositions*, Aberdeen, Aberdeen University Press.

Quine, W. V. 1960: *Word and Object*, Cambridge, Mass.: MIT Press.

Ryle, G. 1949: *The Concept of Mind*, London: Hutchinson.

Shoemaker, S. 1980: “Causality and Properties”, reprinted in Shoemaker, 1984.

---1984: *Identity, Cause and Mind*, Cambridge: Cambridge University Press.

Smith, A. D. 1977: “Dispositional Properties”, *Mind* 86 (343): 439-445.

Smith, M. 1994: The Moral Problem, Oxford: Blackwell.

Queen Victoria's Dying Thoughts

Timothy William Child, Oxford, England, UK

In a number of passages, Wittgenstein suggests that we can make perfectly good sense of ascriptions of thoughts that we have no means of verifying: thoughts that not only *are not* but *could not be* manifested in behaviour. For example:

> Lytton Strachey writes that as Queen Victoria lay dying she 'may have thought of', say, her mother's youth, her own youth, Prince Albert in a Grenadier's uniform (LPP 274. See also LPP 32-3, 99, 152, 229; RPP i 366).

We clearly understand Strachey's speculation. But it seems perfectly possible not only that Queen Victoria *did not* report her dying thoughts but that she *could not* have done so. And in that case, we cannot make sense of claims about her dying thoughts in terms of what she was disposed to report thinking; she had no such disposition. So how do we understand what Strachey says?

One idea would be to appeal to counterfactuals: if Queen Victoria *had been* able and willing to report what she was thinking, she *would have* reported thinking such-and-such. But Wittgenstein takes a different line. We learn 'She thought X', he thinks, in cases where people say what they thought, and where the question what they thought has some practical importance. But with our understanding secure in those basic cases, we can go on to apply the same words to cases where there is no possibility of verification, and where no practical consequences attach to someone's having thought one thing or another. Thus:

> We understand 'He thought X but would not admit it', but we get the use of 'He thought X' from 'He admits X', i.e. says X, writes in his diary X, acts in an X-like way . . . Thinking and not admitting comes in only after thinking and admitting. It's an exception-concept. You'd have to explain to someone who did not know what 'thinking and not admitting' was in terms of thinking and admitting (LPP 329).

In Wittgenstein's view, then, a central role is played, in determining the content of the concept of thought, by cases in which someone's thoughts are manifest in their words or actions. That is a particular case of a more general principle: that a central role is played in determining the content of a concept by cases in which the concept is manifestly instantiated. That principle does not apply to every concept. The content of a highly theoretical concept, for instance, is determined by the theory in which it appears, not by cases where it is manifestly instantiated. Similarly for concepts that can be analyzed in terms of descriptive conditions. But it is very plausible that there are some cases where the principle does apply.

Colour concepts are an obvious example. Cases where something is manifestly red, where it is observed to be red, have a crucial role in determining the content of the concept *red*. But the concept *red* also applies to things that are not observed to be red, and to things that in some reasonably strong sense *could not* be observed to be red: things that can only exist in conditions where human life is impossible, and so on. How should we understand the application of the concept in those cases? One idea is to appeal to counterfactuals: for an unobserved object to be red is for it to be true that, if it *were* observed by a suitable observer in suitable conditions, it *would* look red. That proposal might work in explaining how we understand applications of the concept *red* to objects that merely *are not* observed. But it is hard to see how it could work for the case of an object that *could not* be observed to be red. Yet we do seem able to make sense of the thought that such an object is red. So we need a different idea. An obvious proposal is this: cases in which objects are manifestly red play an essential role in determining the content of the concept *red*. What it is for an unobserved object to be red is then explained by relation to what it is for an observed object to be red: an unobserved object is red just in case it is the same colour as an object that is observed to be red.[1]

Now Wittgenstein might complain that such a view would be question-begging. If we are trying to explain what it is for an unobserved object to be red, we cannot simply help ourselves to the idea of the object's being *the same colour* as an observed red object. For (adapting what he says about a different case): I know well enough that one can call an observed red thing and an unobserved red thing 'the same colour', but what I do not know is in what cases one is to speak of an observed and an unobserved thing being the same colour' (cf. PI §350). But how far would Wittgenstein push this objection? He would certainly insist that what it takes for one thing to be the same colour as another cannot just be taken for granted: it must be understood by reference to a humanly-created concept of colour; and the existence of the concept depends on a whole practice of sorting and classifying things according to their colours, of agreeing and disagreeing about which things are the same colours, and so on. But once that point is accepted, does Wittgenstein think there is a further problem about extending the concept *red* from things that are observed to things that are not, and could not be, observed? It seems plausible that, for the case of objects that are unobserved but *could be* observed, he would accept the dispositional view mentioned in the previous paragraph: what it is for an unobserved table to be brown is for it to be disposed to appear brown to the normal sighted under certain circumstances (see RC §97). But how would he understand the application of the concept *red* to things that *could not be* observed? I know no passage where Wittgenstein explicitly considers that question.[2] Perhaps he would regard such an application as unintelligible. But if that is his view, it needs further argument. For, on the face of it, there is no obvious reason why the concept of colour that we develop in connection with practices involving observed things should not be straightforwardly applicable to things whose colours we could not observe.

What about the concept of thinking? Two points about Wittgenstein's view seem clear. First that, as I have said, a central role is played in determining the content of the concept by cases in which what someone is thinking is

[1] My formulation of this proposal draws heavily on Peacocke's account of 'identity-involving explanations of concept possession' (see Peacocke 2008, especially chapter 5). But I have not attempted to represent Peacocke's own view.

[2] PI §§514-15 considers the question whether a rose is red in the dark, in the context of a discussion of forms of words that look like intelligible sentences but are not. But Wittgenstein's point seems not to be that the sentence 'a rose is red in the dark' is unintelligible but, rather, that it is not the possibility (or not) of imagining a rose being red in the dark that shows the sentence to be intelligible (or not).

manifest because she says or otherwise manifests what she is thinking. Second, that our grasp of what it is for someone to think so-and-so in a case where her thoughts cannot be manifested is dependent on our grasp of what it is for someone to think so-and-so in a case where her thoughts are manifested. But exactly what is the relation between the content of the concept in the two kinds of case? We can distinguish three quite different models, each of which is consistent with the two points just made.

On the first model, the relation between the case where someone says what she is thinking and the Queen Victoria case is like the relation between the cases of observed colour and unobservable colour suggested above. The concept of thinking cannot be explained without making use of examples of thinking; we acquire the concept of thought, in part, in connection with cases where we can tell what someone is thinking. But, having explained the concept of thinking as it applies in cases where we can tell what someone is thinking, we can apply the same concept without further explanation to cases where people's thoughts are not and could not be manifested. At one point, Wittgenstein presses the question, '*what* we can do with' a sentence about Queen Victoria's dying thoughts – '*how* we use it' (RPP i 366). On the current model, that question has a straightforward answer. We use the sentence 'Queen Victoria saw so-and-so before her mind's eye' to speculate about Queen Victoria's dying thoughts. We engage in such speculation because we are interested in what she was thinking about immediately before her death. And we are interested in that question for its own sake – not because we think it has any practical implications.

Maybe Wittgenstein would accept that answer. But some of what he says suggests a quite different model. On this second model, the content of the concept of thought as applied in the Queen Victoria case cannot simply be read off the content of the concept in the more basic cases; it must be understood by giving a direct account of the nature and point of the practice of describing and speculating about thoughts whose ascription cannot possibly be verified. We find it natural to take the word 'thought' from the basic cases, where we can tell what someone is thinking, and apply it in Queen Victoria cases. The meaning of the word in these new applications is parasitic on its meaning in the basic cases, but it is not fully determined by that use; it depends also on the actual use of the word in the new applications. And that use is a matter of our shared interest in developing narratives about the inner lives of others: narratives that have no practical purpose, and for which there is no standard of correctness other than what people agree in regarding as plausible or appropriate. On this view, the practice of discussing Queen Victoria's dying thoughts comes closer to the practice of discussing fiction than to that of ascribing thoughts in more basic cases.

A third model is suggested by the following passage:

> What is the purpose of a sentence saying: perhaps N had the experience E but never gave any sign of it? Well, it is at any rate possible to think of an application for the sentence. Suppose, for example, that a trace of the experience were to be found in the brain, and then we say it has turned out that before his death he had thought or seen such and such etc. Such an application might be held to be artificial or far-fetched; but it is important that it is *possible* (RPP i 157).

On this view, the sentence 'perhaps N had the experience E but never gave any sign of it' has an application, a meaning, because there is in principle some way of verifying whether or not N did have the experience E. If we apply this line to the Queen Victoria case, we will say that we understand the ascription of thoughts in such a case by supposing that there is, after all, a method of verifying such ascriptions, albeit a method that looks not to the subject's actual or potential words and actions, but to physical traces of her thoughts.

If Wittgenstein accepts the first model of our understanding of the ascription of thoughts in the Queen Victoria case, his treatment will be decisively non-verificationist. If he accepts the second model, his account of the meanings of such ascriptions will, again, avoid verificationism; but it will nonetheless be a form of anti-realism. For it will explain the meanings of such ascriptions in a way that gives up the idea that there is any independent fact of the matter about what Queen Victoria was thinking in her dying moments. If he accepts the third model, his account of the Queen Victoria case will, after all, be a form of verificationism. For on this view, the meaningfulness of ascriptions of thought in the Queen Victoria case depends on the supposition that those ascriptions are not, after all, inaccessible to *every* form of verification.

Which of the three models would Wittgenstein accept? I think his position is unclear. The first model is consistent with much that he wants to say. But there is some evidence that he would reject that model; that he would insist that an account of the meaning of the word 'think' as applied in Queen Victoria cases must say something more substantive about our practice of using the word in such cases. The very fact that he presses the question, what we *do* with the sentence 'Queen Victoria may have thought . . .' suggests that, even when we have explained the meaning of ascriptions of thought in cases where a subject's thoughts are manifested, there is a further question, how we understand ascriptions of thoughts that lie beyond our normal methods of verification. That, in turn, suggests that when we apply the concept of thought in Queen Victoria cases, we are in some way developing or extending the concept, or using it in a secondary sense. A view of that sort seems right for the application of the adjectives 'fat' and 'lean' to days of the week. Perhaps it is right for the application of the concept *calculating* to cases in which there is no overt process of calculation. But it is hard to believe that it is right for the application of the concept *thinking* to Queen Victoria cases. If Wittgenstein was tempted by such a view, it is a temptation he should have resisted.

Literature

Peacocke, Christopher 2008 *Truly Understood*, Oxford: Oxford University Press.

Wittgenstein, L. LPP Wittgenstein's Lectures on Philosophical Psychology 1946-47,

London: Harvester, 1988.

Wittgenstein, L. PI *Philosophical Investigations*, 2nd edition, Oxford: Blackwell, 1958.

Wittgenstein, L. RC *Remarks on Colour*, Oxford: Blackwell, 1977.

Wittgenstein, L. RPP i Remarks on the Philosophy of Psychology vol i, Oxford: Blackwell, 1980.

Diagonalization. The Liar Paradox, and the Appendix to *Grundgesetze*: *Volume II*

Roy T Cook, Minneapolis, Minnesota, USA & St Andrews, Scotland, UK

1. Diagonalization in the Grundgesetze

The standard story regarding Frege's Grundgesetze is as follows: Frege's system amounts to nothing more than higher-order logic plus the inconsistent Basic Law V:

BLV: $(\forall X)(\forall Y)[\S(X) = \S(Y) = (\forall z)(Xz = Yz)]$[1]

There are a number of aspects of Frege's logic that differentiate it from standard higher-order systems, however.

The first of these is that Frege treats statements (or, more carefully, what we would think of as statements) as names of truth values. Thus, the connectives are, quite literally, truth-functions, and quantification into sentential position is allowed. (These are first-order quantifiers distinguishing Frege's approach from higher-order logics which allow for second-order quantification into sentential position, interpreting such quantifiers as ranging over 'concepts' of zero arity). For example, the Grundgesetze analogue of:

$(\exists x)(\sim x)$

is both well-formed and a theorem in Frege's formalism.

Once we realize that the quantifiers of the Grundgesetze range over not just value ranges and other mathematical (and perhaps non-mathematical) objects, but also over truth values, the second aspect of Frege's system which will be of interest becomes apparent. Frege's language contains a falsity predicate:

$x = \sim(\forall y)(y = y)$

In other words, an object is the false if and only if it is identical with the truth value denoted by:

$\sim(\forall y)(y = y)$

Thus, within the *Grundgesetze*, we can quantify over statements and we can construct a falsity predicate. The next question to ask is whether the *Liar Paradox* can be constructed within Frege's system. The answer is "Yes". We define our diagonalization relation as follows:

$\mathrm{Diag}(x, y) \;=\; (\exists Z)(y = \S Z \wedge x = Z(y))$

"Diag" holds between x and y if and only if y is the value-range of some concept Z and x is the truth value obtained by applying Z to the value-range of Z. We can now prove the following version of diagonalization:

Theorem 1: In the Grundgesetze, for any predicate $\Phi(x)$, there is a sentence G such that:

$\Phi(G) = G$

is a theorem.

Proof: Given $\Phi(x)$, let:

$F(y) = (\exists x)(\mathrm{Diag}(x, y) \wedge \Phi(x))$

$G = F(\S F)$

The following are provably equivalent in the *Grundgesetze*:

(1) $\Phi(G)$

(2) $\Phi(F(\S F))$

(3) $(\forall x)(F(x) = F(x)) \wedge F(\S F) = F(\S F) \wedge \Phi(F(\S F))$

(4) $(\exists Z)((\forall x)(F(x) = Z(x)) \wedge Z(\S F) = Z(\S F) \wedge \Phi(Z(\S F)))$

(5) $(\exists Z)(\S F = \S Z \wedge Z(\S F) = Z(\S F) \wedge \Phi(Z(\S F))$

(6) $(\exists x)(\exists Z)(\S F = \S Z \wedge x = Z(\S F) \wedge \Phi(x))$

(7) $F(\S F)$

(8) G

[(1) and (2) are equivalent by the definition of G, (2) and (3) by logic, (3) and (4) by logic, (4) and (5) by BLV, (5) and (6) by logic, (6) and (7) by the definition of F, and (7) and (8) by the definition of G.]

The basic idea of the proof is that we can 'fake' the standard proof of diagonalization (see e.g., Boolos and Jeffrey [1989], Chapter 15) by using the value ranges of concepts as 'names' of those concepts, and quantification over truth values in lieu of names of statements, thereby sidestepping the need for Gödel numbers or analogous coding devices.

We can immediately generate the Liar paradox. Applying Theorem 1 to our falsity predicate results in a sentence Λ such that:

$\Lambda = (\Lambda = \sim(\forall y)(y = y))$

is a theorem. But this entails:

$\sim(\forall y)(y = y)$

Note that we can derive (8) from (1) without the use of BLV. In other words, letting Grundgesetze – BLV denote the system obtained by removing BLV from thes Grundgesetze, we have:

Corollary 2: In the *Grundgesetze* – BLV, for any predicate $\Phi(x)$, there is a sentence G such that:
$\Phi(G) \rightarrow G$
is a theorem.

This does not lead to contradiction, however. Applying Corollary 2 to the falsity predicate we obtain:

$(\Gamma = \sim(\forall y)(y = y)) \rightarrow \Gamma$

which entails merely:

Γ

This is not surprising, since the consistency of the BLV-free fragment of the *Grundgesetze* is relatively easy to demonstrate.

[1] Here, and below, I use modern symbolism instead of Frege's two-dimensional notation, primarily for typographical convenience. All proofs, etc., can be straightforwardly translated into Frege's original formalism. Particular attention should be paid to the use of identity, since in Frege's system identity holding between two statements (i.e. names of truth values) is roughly equivalent to our biconditional.

It is worth noting that we can also prove:

Corollary 3: In the *Grundgesetze* – BLV, for any predicate $\Phi(x)$, there is a sentence G such that:
$G \rightarrow \Phi(G)$
is a theorem.

This result is obtained by replacing our definition of "Diag" above with:

$$\text{Diag}(x, y) \;= (\forall Z)(y = \S Z \rightarrow x = Z(y))$$

The trick is that without BLV we cannot prove that there is a *single* sentence G such that both:

$$\Phi(G) \rightarrow G$$

and:

$$G \rightarrow \Phi(G)$$

Thus, we can prove an analogue of Gödel's diagonalization lemma within the *Grundgesetze*, and restricted versions of diagonalization hold in the consistent sub-system not containing BLV. The reader might wonder why we have made so much of these results. After all, we already knew that the *Grundgesetze* (including BLV) was inconsistent, so the news that one can construct the *Liar paradox* as well as *Russell's paradox* within Frege's system is not exactly earth-shattering (although the 'naturalness' of the construction of the *Liar paradox* in the *Grundgesetze* is somewhat surprising, at least to the author). In addition, the corollaries that follow for the consistent subsystem *Grundgesetze*–BLV are trivial in any system of sufficient expressive strength – just let G be any tautology in Corollary 2, and any contradiction in Corollary 3.

The interest of these results lies in their connection to Frege's attempted fix of the *Grundgestze* in the appendix to Volume II, to which we now turn.

2. Diagonalization and the Appendix to Grundgesetze

A quick examination of Theorem 1 reveals that the full strength of BLV is not required in order to prove the full, biconditional form of diagonalization. Instead, we merely need the resources to infer line (2):

$$\Phi(F(\S F))$$

from line (5):

$$(\exists Z)(\S F = \S Z \wedge Z(\S F) = Z(\S F) \wedge \Phi(Z(\S F))$$

In order to get from (5) to (2), we do not need it to be the case that concepts with the same value-range are always co-extensive. Instead, we merely need concepts to agree on their shared value-range. Thus, we can recapture Theorem 1 by replacing BLV with the (prima facie weaker) Fixed-Point Principle for value-ranges:

$$\text{FPP:}\;\; (\forall X)(\forall Y)(\S(X) = \S(Y) \rightarrow (X(\S X) = Y(\S X)))$$

If FPP holds, then we can move from:

$$\S F = \S Z$$

to:

$$F(\S F) = Z(\S F)$$

and thus from:

$$\Phi(Z(\S F))$$

to:

$$\Phi(F(\S F))$$

Thus, any principle meant to replace BLV and provide identity conditions for value ranges cannot, on pain of *Liar*-induced contradiction, imply FPP.

Surprisingly, in response to the detection of *Russell's paradox*, and without any (apparent) knowledge that the *Liar paradox* could also be derived within the *Grundgesetze*, Frege isolated FPP as exactly the problematic consequence of BLV.

In the appendix of Volume II of the *Grundgesetze*, Frege begins his discussion of *Russell's paradox* by distinguishing between the two 'directions' of BLV:

$$\text{BLVa:}\; (\forall X)(\forall Y)((\forall z)(X(z) = Y(z)) \rightarrow \S X = \S Y)$$

$$\text{BLVb:}\, (\forall X)(\forall Y)(\S X = \S Y \rightarrow (\forall z)(X(z) = Y(z)))$$

He notes that, if we are to individuate concepts extensionally (an assumption he is unwilling to give up), then BLVa cannot be the problem – after all, *any* function f from concepts to objects will satisfy:

$$(\forall X)(\forall Y)((\forall z)(X(z) = Y(z)) \rightarrow fX = fY)$$

So BLVb must be where the problem lies, and Frege sets out to discover exactly what goes wrong with this principle. He outlines his strategy as follows:

> We shall now try to complete our inquiry by reaching the falsity of (Vb) as the final result of a deduction, instead of starting from (Vb) and thus running into a contradiction. (1893, p. 288 in the Frege Reader)

Thus, in order to understand exactly what it is about BLVb that causes the problem, we need to find a direct proof of its negation, and not rely merely on a *reductio* of it via Russell's construction. In other words, Frege requires a direct proof of:

$$(\exists X)(\exists Y)(\S X = \S Y \wedge (\exists z)(X(z) \wedge \neg Y(z)))$$

In searching for such a proof, Frege discovers that he can obtain a stronger result, which I have elsewhere (Cook [in progress]) called:

Frege's Little Theorem: For any function f from concepts to objects one can prove:

$$(\exists X)(\exists Y)(f(X) = f(Y) \wedge X(f(X)) \wedge \neg Y(f(X)))$$

So, given any function from concepts to objects, there exist two concepts such that the function maps both concepts to the same object, yet the concepts differ on that very object.

Here is the rub: The instance of Frege's Little Theorem obtained by substituting the the value range operator "§" for "f" is the negation of FPP! In other words, the principle that Frege identifies as causing *Russell's paradox* is exactly the principle that is needed to turn the proofs of our corollaries into proofs of the diagonalization.

The proof runs as follows (see Frege 1893, pp. 285 – 288 in the *Frege Reader*, for Frege's original proof):

Proof:

Given a function f from concepts to objects, let:

$$R(x) = (\exists Y)(x = f(Y) \wedge \neg Y(x))$$

Then:

(1) $\neg R(f(R))$	Assump for *Reductio*
(2) $\neg(\exists Y)(f(R) = f(Y) \wedge \neg Y(f(R)))$	(1), Df. of R
(3) $(\forall Y)(f(R) = f(Y) \rightarrow Y(f(R)))$	(2), Logic
(4) $R(f(R))$	(3), Logic
(5) $R(f(R))$	(1) – (4), *Reductio*
(6) $(\exists Y)(f(R) = f(Y) \wedge \neg Y(f(R)))$	(5), Df. of R
(7) $(\exists Y)(f(R) = f(Y) \wedge R(f(R)) \wedge \neg Y(f(R)))$	(5), (6), Logic
(8) $(\exists X)(\exists Y)(f(X) = f(Y) \wedge X(f(X)) \wedge \neg Y(f(X)))$	(7), Logic

Frege concludes that such 'fixed points' are the root of Russell's paradox:

> We can see that the exceptional case is constituted by the extension itself, in that it falls under only one of the two concepts whose extension it is; and we see that the occurrence of this exception in no way can be avoided. Accordingly the following suggests itself as the criterion for equality in extension: The extension of one concept coincides with that of another when every object that falls under the first concept, except the extension of the first concept, falls under the extension of the second concept likewise, and when every object that falls under the second concept, except the extension of the second concept, falls under the first concept likewise. (1893, p. 288 in *The Frege Reader*)

As a result, Frege suggests a modification of BLV:

BLV* $(\forall X)(\forall Y)(\S X = \S Y = (\forall z)((z \neq \S X \wedge z \neq \S Y) \rightarrow (X(z) = Y(z))))$

According to the amended principle two concepts receive the same value range if and only if they hold of exactly the same objects other than their value ranges.

The inadequacy of Frege's BLV* is well-known, although the reasons commonly given for its failure are mistaken. The well-known works addressing the formal aspects of BLV*, Frege's so-called 'way out', such as Quine (1955) and Geach (1956), report that Frege's amended principle is consistent, but inadequate for his purposes, since it implies that at most one object exists. What they fail to appreciate, however, is that since Frege's *Grundgesetze* allows for quantification into sentential position, one can (without any version of BLV, amended or not) prove the existence of at least two objects (the true and the false). As a result, from the perspective of Frege's *Grundgesetze*, BLV* is just as inconsistent as was BLV (Landini (2006) comes closest to this, as he proves that BLV* is inconsistent if the truth values are their own singletons, as Frege intended, and also proves that BLV* is inconsistent if the truth values are not value-ranges at all).

3. Lessons Learned

The ultimate failure of Frege's attempt to salvage his life's work does not imply that it contains nothing of value. I will conclude by identifying two lessons that can, and should, be drawn from all of this.

The first is that we should take care in attributing the inadequacies of BLV* to some sort of panicked, half-hearted attempt by Frege to amend his. Quine describes this common attitude to the appendix:

> It is scarcely to Frege's discredit that the explicitly speculative appendix now under discussion, written against time in a crisis, should turn out to possess less scientific value than biographical interest. Over the past half century the piece has perhaps had dozens of sympathetic readers who, after a certain amount of tinkering, have dismissed it as the wrong guess of a man in a hurry. (1955, p. 152)

While the 'fix' might have been written in a hurry, and BLV* is inconsistent, the discussion leading up to it has much to teach us about the mathematics of abstraction principles in general and the roots of *Russell's paradox* and related phenomenon in particular. In this respect, Frege's Little Theorem is not the incorrect guess of a man in a hurry, but rather a deep insight into the puzzling nature of abstraction and the paradoxes that can arise from its unfettered application.

This brings us to the second lesson. Connections are often drawn between the *Liar paradox* and *Russell's paradox* (and between the semantic and set-theoretic paradoxes more generally), but these connections tend to be quite loose, relying on the intuition that circularity of some vicious sort is at the root of both phenomena (for a project that draws the connections much more tightly, however, the reader is urged to consult Cook 2007!). The construction of the *Liar paradox* within Frege's system, and his identification of the exact principle that is the root of both this paradox and the one communicated to him by Russell, suggests that further study of Frege's system (or modern variants that retain object-level quantification into sentential position, such as that provided in Landini 2006) hold promise for a deeper understanding of these paradoxes individually and of the links that bind them together as distinct aspects of a single problem.[2]

Literature

Boolos, G. & R. Jeffrey, 1989 Computability & Logic, 3rd Ed., Cambridge: Cambridge University Press.

Cook, Roy T 2007 "Embracing Revenge: On the Indefinite Extensibility of Language", in Revenge of the Liar, JC Beall (ed.), 2007 Oxford: Oxford University Press.

Cook, Roy T (in progress) Frege, Numbers, and Sets (book manuscript).

Frege, Gottlob 1893, 1903 Grundgezetze der Arithmetik I & II, Hildesheim: Olms.

Frege, Gottlob 1997 The Frege Reader, M. Beaney (ed.), Oxford: Blackwell.

Geach, Peter. 1956 "On Frege's Way Out", Mind 65, 408 – 409.

Gödel, Kurt 1992 On Formally Undecidable Propositions. New York: Dover.

Landini, Gregory 2006 "The Ins and Outs of Frege's Way Out", Philosophia Mathematica 14, 1 – 25.

Quine, W.V.O. 1955 "On Frege's Way Out", Mind 64, 145 – 159.

[2] Thanks go to Greg Taylor for helpful discussion of earlier versions of this paper and suggestions for improvement. In addition, the present paper has benefited greatly from feedback recieved at Arché: The Philosophical Research Centre for Logic, Language, Metaphysics, and Epistemology at the University of St Andrews when I presented portions of Cook [in progress] to members of the Centre.

Exorcizing Gettier

Claudio F. Costa, Natal, Brazil

> Knowledge is not simply justified true belief,
> but it is justified true belief, justifiably arrived at.
> *Robert J. Fogelin*

Gettier's problem[1] seems to be a daunting treat to our belief in the rationality of the human knowledge. In what follows I intend to show with some formal precision the natural way out of the trap.

Using the symbol *a* to a person, *K* to knowledge, *B* to the belief, *E* to a reasonable justifying evidence (justification), and *p* to the proposition, we might symbolize the tripartite definition of knowledge as follows:

$$(Df.1)\quad aKp \;=\; \overset{(i)}{p}\ \&\ \overset{(ii)}{aBp}\ \&\ \overset{(iii)}{aEBp}$$

According to this definition, *a* knows that *p* (*aKp*) means the same as the conjunction of these three conditions, namely (i) that *p* is true, (ii) that *a* believes that *p* is true, and (iii) that *a* has a reasonable justification for her belief in the truth of *p*. As it is well-known, Gettier's problem arises from the discovery of counterexamples to this definition, namely, from cases where the person *a* fails to attain knowledge though satisfying these three conditions.

To remember Gettier's counterexamples, consider the following[2]. Suppose that professor Stone said to Mary yesterday that he would come to the university this night to give a lecture. Since Mary knows that Stone is a highly responsible person, she can claim that she knows that he came to the university this night. However, unknown to her, one of Stone's sons suffered an accident and he needed to drop the lecture. However, it is true that he came to the university, since he was momentarily in his room to take some documents. Mary's claim to know that Stone came to the University this night seems to satisfy the conditions to the traditional definition: it is a true belief and the justification presented by her is reasonable enough. Nevertheless, its truth is only accidentally achieved and nobody would say that Mary really knows that Stone was at the university tonight.

As it was sometimes noted, there is a straightforward and effective way to answer the problem, which seems to be nearly buried under the considerable amount of alternative answers explored in the literature[3]. It consists simply in the request that *a* sound *epistemic justification must belong to what we are able to accept as making the proposition* p *true*[4]*.* So, Mary's justification for her belief that professor Stone came to the University this night, based on the evidence given by his statement that he would give a lecture, might be reasonable, but is epistemically unsound, since this information is no part of what we – as the epistemic evaluators of Mary's knowledge claim – are prepared to accept as making true the belief that Stone came to the university this night. Reasonability is not enough. A justification must also be epistemically sound, by making itself acceptable to the epistemic evaluators of a knowledge claimer *a* as making a proposition *p* true[5]. In the case of the gettierian counterexamples, these epistemic evaluators have always some information that overrides the epistemic soundness of the reasonable justification given by the knowledge claimer.[6]

My aim here is to improve the tripartite definition of knowledge by stating more formaly this intuitive solution. This can be done by making explicit the internal link between the condition of justification and the condition of truth. In order to do it, we shall review the formulation of the conditions (i) and (iii) of (*Df.1*).

We begin with the condition of truth. As it appears in the traditional definition, it is surely a simplification. For it seems like the truth-value of the proposition when it is contemplated by God. Since God doesn't need to verify anything in order to know the truth, he does not need to consider whether any truth-condition is satisfied. So, for him "*p*" or "*p* is true" is enough. However, if we intend to make justice to the condition of the truth of *p* as it is known by us (that is, by the knowledge-evaluators of knowledge-claimers), we need to consider whether the truth-conditions were satisfied. Now, how to do it? We need first to see that, when an evidence *E* for the ascent of *p* is found, it must be seen by us as *sufficient* to make the proposition *p* true. The meaning of 'sufficient' here can be made precise as follows:

> An evidence *E* is sufficient for the assent of *p* as true iff *E* makes *p* either (i) necessarily true (when *p* is a non-empirical, deductively grounded truth) or (ii) probable in a very high level (for the cases of empirical, inductively grounded truths)[7].

We can introduce the symbol '~>' (to be read as "is sufficient to") in order to express this conditional. Thus, given the evidence *E* for the ascent of *p*, this means that *E* ~> *p*, in other words, that for us either *E* makes *p* necessarily true or very probably true.

With this in mind we can introduce the symbol *E** to designate the set of all justifying evidences that we consider *individually sufficient* for the truth or falsity of *p* in the already specified sense. To give an example: suppose

[1] E. L. Gettier: "Is Justified Belief Knowledge?" *Analysis* 23, 6, 1963, 121-23.

[2] I take this example (with slight changes) from D. J. O'Connor and Brian Carr, *Introduction to the Theory of Knowledge* (The Harverster Press: Brighton 1982).

[3] Similar considerations can be found in D. J. O'Connor and B. Carr, *Introduction to the Theory of Knowledge*, p. 82. The origin of this view seems to be due to Robert F. Almeder, particularly in the paper "Truth and Evidence", *The Philosophical Quarterly* 24, 1974, 365-68. The most original and compelling defense of a similar view can be found in Robert Fogelin's book, *Pyrrhonian Reflections on Knowledge and Justification* (Oxford University Press: Oxford 1994), chapter 1.

[4] This requirement was stated by D. J. O'Connor and by Brian Carr, who also say that "the reason why the proposition is true must not be independent on the facts asserted in the proposition constituting the grounds for the belief", claiming for elaboration (*Introduction to the Theory of Knowledge,* p. 81). Robert Fogelin stated the same point more concisely his definition of knowledge: "S knows that P iff S justifiably came to believe that P on grounds that establish the truth of P" (*Pyrrhonian Reflections on Knowledge and Justification*, p. 28)

[5] My distinction between a *reasonable* justification and an *epistemically sound* justification is equivalent to the distinction between a personal justification (epistemically responsible) and a justification given on the basis of adequate grounds. See Michael Williams, *Problems of Knowledge: a Critical Introduction to Philosophy* (Oxford University Press: Oxford 2001) pp. 22-23.

[6] The words 'we' and 'us' point usually to the knowledge-evaluators, with their usually wider informational set. However, this does not precludes the possibility that the knowledge-evaluator is the knowledge-claimer herself, by making a self-evaluation of her own past knowledge claims.

[7] I am not considering Kripkian cases like that of necessary *a posteriori* beliefs derived from *E* (they are also controversial).

that I am sure that it is true that Stone came to the university tonight because of *E1*: “I saw him parking outside”, and/or because of *E2*: “he called me on phone, saying that he was coming here”. Since I take these evidences as true, and I see each of them as sufficient to make me accept the truth of the proposition *p*, I can say that $E1 \rightsquigarrow p$, that $E2 \rightsquigarrow p$, and that $E^* = \{E1, E2\}$. An important characteristic of E^* is that, under the assumption that we are rational evaluators, either all its members are sufficient to make *p* true or they are all sufficient to make *p* false, otherwise they would cancel one another[8]. With these concepts we can redefine the condition of truth by making explicit the role of the evidential truth-conditions to our acceptance of *p* as true. Here is the formulation for the condition of truth:

(i’) $(E^* \;\&\; (E^* \rightsquigarrow p))$

This is the same as saying that *p* is true, since given our acceptance of E^* as true (that is, the truth of at least one evidence *E* such that $E \rightsquigarrow p$), our acceptance of the truth of *p* follows (by *modus ponens* or inductively). The difference is that now the satisfied evidential truth-conditions can be made explicit as the members of the set *E*.* I claim that this is what we, fallible truth-searchers, ultimately mean with the condition (i).

The second improvemente concerns the reformulation of the condition of justification in the definition of knowledge, linking this justification with the set of evidences that make the proposition *p* true. What we need to do is only to require, additionally, that the evidential justification *E* given by *a* might be seen by us as belonging to our accepted **E*, namely, to the set of evidences that we (as the evaluators of knowledge-claims) are prepared to accept as the satisfied truth-conditions which are individually sufficient to make *p* true (cases in which $E^* \rightsquigarrow p$). Here is our reformulation of the third condition:

(iii’) $aEBp \;\&\; (E \in E^*)$

The condition (iii’) says that, additionally to the condition that *a* has a reasonable evidence justifying the truth of *p*, it is required that this evidence, for being sound, must be able to be accepted by us as belonging to the set of evidences that we are prepared to accept as individually making *p* true.

With this in mind we are prepared to reformulate the tripartite definition of knowledge in a way that makes explicit the internal relation between the condition of justification (iii) and the condition of truth (i). Here it goes:

$$(Df.2)\;\; aKp = \;\overset{\text{(i')}}{(E^* \;\&\; (E^* \rightsquigarrow p))} \;\;\&\;\; \overset{\text{(ii)}}{aBp} \;\;\&\;\; \overset{\text{(iii')}}{(aEBp \;\&(E \in E^*))}$$

Dropping the condition (ii) as redundant, since it is repeated in the first conjunct of (iii), we get the following version:

$$(Df.3)\;\; aKp = \;\overset{\text{(i')}}{(E^* \;\&\; (E^* \rightsquigarrow p))} \;\;\&\; \overset{\text{(iii')}}{(aEBp \;\&(E \in E^*))}$$

What these definitions tells us is that the justifying evidence *E* given by *a* must belong to the set of evidences (of fulfilled truth-conditions) that might be hold by us (the knowledge-evaluators) as individually sufficient to make *p* true. If the evidence *E* given by *a* belongs to E^*, and E^* is so that its individual members lead to the necessary or at least highly probable truth of the proposition *p*, so that $E^* \rightsquigarrow p$, than *E* is epistemically sound, for it assures us the truth of *p* either as necessary or as practically certain.

Now, consider again our gettierian counterexample. Mary’s evidence *E* (“Stone said to me he would give a lecture today”) would *not* be accepted by us (since we are better informed, and also know about the accident with his son etc.) as belonging to our E^*, even if we know that Stone was (by different reasons) at the university this night. So we conclude that, according with our definition of knowledge, she really does not know. And this result can be generalized in order to exorcize any conceivable gettierian counterexample. Since in no counterexample of Gettier kind the justifying evidence *E* belongs to the set E^*, none of these counterexamples satisfies the proposed reformulation of the tripartite definition of knowledge.

[8] For example: evidences for the roundness of the earth are *E1* (photos from the all) and *E2* (the circumnavigation of the globe). Each one is a member of E^*, sufficient for the truth the proposition *p* saying that the earth is round. But if $\sim E2$ were an element of E^*, *E1* would loose its force and would not be a sufficient condition, do not belonging to E^* anymore.

A Wittgensteinian Approach to Ethical Supervenience

Soroush Dabbagh, Tehran, Iran

Introduction

What can we say with regard to the extent of the patternability of the reason-giving behaviour of a morally relevant feature in different ethical contexts? The main issue between generality and particularity in moral reasoning concerns the existence of patterns in use of moral vocabulary that would permit the formulation of general statements governing the applicability of that vocabulary. Particularism challenges an intuitive notion of generalism. There are general patterns to which the reason-giving behaviour of a morally relevant non-moral property in different contexts is responsive and this is the main issue in evaluating arguments of particularism and generalism. It concerns the way in which a morally relevant feature contributes to the moral evaluation of different cases. The subject can be formulated using the idea of supervenience, according to which if two concrete ethical situations are relevantly similar with respect to their non-moral (descriptive) properties, their moral (evaluative) properties would be the same. Suppose we are confronted with a concrete ethical situation, in which a moral property F supervenes on non-moral properties G and H. According to the generalist, should we come across a similar ethical situation in which G and H are combined together, the ultimate moral evaluation of the case would be the same —F would apply. So, subscribing to the existence of supervenience leads to approving the existence of general patterns to which the reason-giving behaviour of a morally relevant non-moral property can fit. In other words, with the aid of such patterns, we can see how a morally relevant non-moral property contributes to the moral evaluation of different cases.

According to generalists who subscribe to the notion of supervenience, the reason-giving behaviour of a morally relevant feature in different cases is generalisable in the sense that its reason-giving behaviour is answerable to patterns of word use. But a particularist like Dancy prefers to talk about the idea of resultance with regard to the way in which non-moral properties are related to moral properties in ethical contexts. According to him:

> Resultance is a relation between a property of an object and the features that ‘give’ it that property. Not all properties are resultant; that is, not all properties depend on others in the appropriate way. But everyone agrees that moral properties are resultant. A resultant property is one which ‘depends’ on other properties in a certain way. As we might say, nothing is just wrong; a wrong action is wrong because of other features that it has…Supervenience, as a relation, is incapable of picking out the features that make the action wrong; it is too indiscriminate to be able to achieve such an interesting and important task (2004, 85-88).

According to this view, there is no such thing as a general pattern which summarises the reason-giving behaviour of a morally relevant feature and we cannot see how a morally relevant feature contributes to the moral evaluation of different cases by appealing to supervenience. Supervenience deals with the behaviour of a morally relevant feature in different ethical contexts, the way in which moral properties supervene upon the *class* of non-moral properties. In contrast, resultance concerns the way in which a moral property results from non-moral properties in a specific ethical situation. So, a particularist who claims there is no metaphysical account available of generality in moral reasoning, emphasises that the reason-giving behaviour of a morally relevant feature and its contribution to moral evaluation can vary from case to case as a result of combining with other features in many different ways. So, the reason-giving behaviour of a morally relevant feature is not generalisable to say, its relevance for reasoning in different cases is not answerable to general patterns of word use. Rather, the reason-giving behaviour results from the way in which different morally relevant features are combined together in a specific moral situation. Therefore, according to Dancy, the idea of resultance, unlike supervenience, can better systematise our common sensical intuitions with regard to the way in which several morally relevant features are combined together in different ethical contexts.[1]

Now I outline the particularist’s answer with regard to the extent of the patternability of the reason-giving behaviour of morally relevant features in different contexts which is associated with resultance while undermining superveneince.

1. The Particularists’ Answer

According to the particularists’ standpoint, moral principles are strongly context-dependent in the sense that the reason-giving behaviour of a morally relevant feature is not answerable to general patterns.

The main argument in support of particularism draws on the idea of *holism* about reasons for action. According to holism, morally relevant nonmoral properties are highly contextual, and may change their reason-giving behaviours from case to case where they are compounded with other morally relevant non-moral properties, so that what makes an action wrong in one case may make it right in another case. In other words, the deontic valence of a moral consideration (such as one's duty to fulfil his promise to someone else) is not constant, and may vary from case to case.

Dancy’s argument in favour of holism about reasons for action is an application of holism about normative reasons in general. Dancy claims that normative reasons for belief are obviously and non-controversially holistic (highly contextual), and that it is very odd to account for reasons for action as non-holistic. But how could normative reasons for belief be holistic? Dancy’s argument for this claim is as follows: suppose that something is in front of me, and I experience it as a red pencil. Experiencing something as a red pencil is a justified reason for me to believe that a red pencil is in front of me. Again suppose that, as a thought experiment, I have taken a pill which makes blue things seem red to me. In this case,

[1] For more on the distinction between resultance and supervenience, see Dancy, J (1981) ‘On Moral Properties’, Mind, 90, pp, 367-385, 380-382 & (1993) Moral Reasons (Oxford: Blackwell), pp. 73-79. See also R∅nnow-Rasmussen, T. (1999) ‘Particularism and Principles’, Theoria, 65, pp.114-126, 115-119. See also Sinnott-Armstrong, W. (1999) ‘Some Varieties of Particularism’, Metaphilosophy, 30, pp. 1-12, 2-5.

experiencing something as a red pencil is a reason that justifies me in believing that a blue pencil is in front of me. Therefore, it is not the case that experiencing something as red always justifies me in believing that there is something red is in front of me. Conversely, it can justify me in believing that there is something blue is in front of me. Dancy says:

> It is not as if it is some reason for me to believe that there is something red before me, though that reason is overwhelmed by contrary reasons. It is no longer *any reason at all* to believe that there is something red before me; indeed, it is a reason for believing the opposite (2004, p.74).

This means that reasons for belief behave holistically, and the way in which they are combined together and contribute to ultimate justification can vary from context to context. In other words, they have no intrinsic and invariant valence outside context, for their valence can change as a result of reacting to other reasons.

2.Criticising the Particularistic Position: Wittgensteinian account of normativity

In order to criticise Dancy's constitutive and metaphysical claim concerning the way a morally relevant feature contributes to the moral evaluation of different contexts, I draw on the account from Wittgenstein with regard to the nature of concepts[2].

Suppose we want to articulate and define the concept 'game'. On the face of it, it seems that in order to do this we need to state common properties of games with which we have been confronted, such as: basketball, handball, snooker, chess, boxing, wrestling etc. On the basis of the common properties obtained, we would say that:

If x meets the condition g_1, g_2, $g_{3,}$ … g_n, x is a 'game'.

This view supposes that there is something in common which needs to be articulated and categorised to arrive at the definition of the concept 'game'. It suggests that there is something in common among different kinds of games. By utilising the obtained general rule, we can say whether or not a new phenomenon can be regarded as a game. In this model, the general pattern acts as the normative standard of the rightness and wrongness of the use of words.

However, Wittgenstein rejects the existence of such a common property in different kinds of games; something which can be articulated as an essence of the concept 'game'. The whole idea of 'family resemblance' in *Philosophical Investigations* is concerned with the denial of such an approach to defining a concept like game. There is nothing in common among different games which can be articulated. For instance, if someone says that losing and winning can be regarded as a common feature of different games, we can *show* him other games in which there is no such thing as losing and winning like the child who builds a house using Lego. Moreover, if we want to consider equipment such as a ball, goal, net, racket etc. as a common feature or features of different games, one can *show* other games such as: boxing, wrestling etc. in which these items not used. So, it seems that there is an open-ended list of game-making features which forms the different games with which we are familiar. So, it seems that we cannot arrive at what the concept 'game' is through articulating a feature common to different games. Nevertheless, we, as language-users use the word 'game' in our communication meaningfully. In other words, although there is an open-ended list of game-making features, we cannot regard anything we like as an example of the concept 'game'. It seems that there is a normative constraint that requires us to see whether or not the phenomenon with which we are dealing can be regarded as a game. Wittgenstein attempts to show that the normative constraint that we are talking about cannot be put into words. Rather, it can only be grasped through ongoing practice of *seeing* the similarities and dissimilarities. There is nothing beyond seeing the similarities which can do this job. He states:

> What does it mean to know what a game is? What does it mean, to know it and not be able to say it?… Isn't my knowledge, my concept of a game, completely expressed in the explanations that I could give? That is, in my describing examples of various kinds of games; showing how all sorts of other games can be constructed on the analogy of these (1953, §75).

According to Wittgenstein, it is not the case that I know what the concept 'game' is before being engaged in the practice of seeing the similarities. Rather, what we see within practice is all we have about the concept 'game'. This results in the denial of the pre-existing concept of game. However, the more we are engaged in the practice of using the word, the more clearly we see what a game is. This is an open-ended process. To grasp the meaning of a concept such as game, all we have is seeing the similarities: this is a game, that is a game, this is not a game etc. and this is not ignorance. Being engaged in practice is not a halfway and second hand explanation of what a game is. This is all we have at hand and it does not mean that any phenomenon can be regarded as an example of the concept 'game'. Rather, there is a normative constraint which lies in the way in which we are engaged in seeing things as similar. In other words, it is not the case that regarding a new phenomenon as a game is a matter of taste and can be done arbitrarily or at random. Rather, there is a normative constraint which can be seen within practice. There is an account which can be given with regard to whether or not the new phenomenon is a game. The account becomes clearer to the extent that we are engaged in the practice of seeing things as similar. There is no such thing as a pre-existing and abstract pattern which can be utilised in order to see whether or not the new phenomenon is a game. Rather, there is an account with regard to the normative standard of the rightness and wrongness of the use of words which is associated with the way in which we are engaged in seeing the similarities. The crucial thing at this stage is that there is an account with regard to a normative constraint which can be given. In fact, in place of the notion of the pre-existing source of normativity, there is a normative constraint which can be seen merely within practice.

To the extent that we are engaged in the activity of seeing things as similar, we can see what the concept 'game' is. We have a role in shaping the concept. In other words, the concept 'game' *emerges* following our ongoing practice of seeing the similarities. Moreover, the concept 'game' is extendable. The more we are engaged in the practice of seeing similar games, the more the concept is extended. Practice has an indispensable role in the extendibility of the concept 'game'. So, we can say that there

[2] Note that, at this stage, I shall apply the Wittgensteinian account with regard to the nature of concepts to repudiate Dancy's constitutive claim regarding the reason-giving behaviour of a morally relevant feature in different contexts. The justification of the Wittgensteinian account of the nature of concepts is another issue and can be evaluated separately and on its own.

is some generality in the concept ‘game’, albeit one that emerges.

What follows from the Wittgensteinian story is that the reason-giving behaviour of the word ‘game’ in different contexts is answerable to general patterns of word use. This is the constitutive and metaphysical claim with regard to the existence of patterns of word use.

Considering Wittgensteinian account of patternability and the way in which the reason-giving behaviour of a morally relevant feature is answerable and responsive to patterns of word use, it seems that Dancy’s claim about the very idea of supervenience is implausible. According to Dancy - as there is no such thing as an exactly similar ethical situation - to say that the reason-giving behaviour of a morally relevant feature would be answerable to general patterns in other ethical contexts is useless.

But as we saw in the example of the concept ‘game’, although several game-making features are combined together in different ways, they are not responsive to general patterns of word use: Answerability to general patterns is not necessarily associated with the existence of *exactly* similar situations. As far as an emerging pattern is concerned, there is no such thing as a finite list of features which make the pattern. Nevertheless, there is such a thing as a normative constraint which can be seen to the extent that we are engaged in practice. So, we can subscribe to the idea of supervenience, according to which moral properties supervene upon non-moral properties in the sense that the reason-giving behaviour of a morally relevant feature in different context is answerable to patterns without resorting to phrases like ‘exactly similar situation’. In other words, the modest-generalist can agree with a particularist like Dancy in criticising the idea of a pre-existing and fixed pattern according to which a new phenomenon has to be subsumed under a determined and rigid pattern. Such an account of pattern requires the new phenomenon to be *exactly similar* to the components of the pattern. But the modest-generalist can appeal to the idea of open-endedness to give a constitutive account of patternability without appealing to pre-existing and determined pattern.

To summarise, Dancy’s claim with regard to the way in which the reason-giving behaviour of a non-moral feature contributes to the moral evaluation of different cases can be reconciled with the generalistic Wittgensteinian position which deploys the idea of patternability and answerability. It follows from this that still we can stick to the very idea of supervenience, as far as the reason-giving behaviour of a morally relevant feature in different contexts is concerned.

Literature

Dancy, J (1981) ‘On Moral Properties’, *Mind*, 90, pp, 367-385, 380-382 & (1993) *Moral Reasons* (Oxford: Blackwell), pp. 73-79.

Dancy, J. (2004) Ethics Without Principles (Oxford: Oxfors University Press).

Rønnow-Rasmussen, T. (1999) ‘Particularism and Principles’, *Theoria*, 65, pp.114-126.

Sinnott-Armstrong, W. (1999) ‘Some Varieties of Particularism’, *Metaphilosophy*, 30, pp. 1-12, 2-5.

Wittgenstein, L.(1953) *Philosophical Investigations* (Oxford: Blackwell).

There can be Causal without Ontological Reducibility of Consciousness? Troubles with Searle's Account of Reduction

Tárik de Athayde Prata, Fortaleza, Brazil

I. Introduction

In his writings about the philosophy of mind John R. Searle often deals with the question of *reduction*, because the main question in this field can be defined in these terms: do the mental phenomena have a special mode of existence or are they *reducible* to physical phenomena? (see SEARLE, 1992, p. 2). But it is not clear whether his account of reduction is really coherent. Searle distinguishes different types of reduction (see SEARLE, 1992, p. 113-114), but when he speaks about consciousness, he makes incompatible claims. The two types that are relevant here are *causal* and *ontological* reduction. The main problem is that he thinks of consciousness as a special case, in which these two types of reduction are not equivalent: consciousness can be *causally* but *not ontologically* reduced, and that seems to commit him with the contradictory claims that consciousness *is* and *is not* identical to brain behavior. In the present paper, Searle's conception of causal reduction and its relations with ontological reduction will be examined (section II), as well as his argument for the ontological irreducibility of consciousness (section III), which seems to be in contradiction with this conception of causal reduction. After that, Searle's arguments for the thesis that ontological irreducibility does not force us to dualism are going to be discussed (section IV). My conclusion is that this last argument fails, so that ontological irreducibility entails a kind of dualism, and Searle states and denies (in contradictory way) the identity of consciousness and brain processes.

II. Causal and Ontological Reduction

Searle defines the causal reducibility of consciousness as follows:

> "Consciousness is causally reducible to brain processes, because *all features* of consciousness are accounted for causally by neurobiological processes going on in the brain, and consciousness has no *causal powers* of its own in addition to the causal powers of the underlying neurobiology." (SEARLE, 2002b, p. 60, my emphasis)

A causal reduction of consciousness consists of the *causal explanability* of its surface features by brain processes at the microlevel and the *identity of causal powers* of both. These two aspects are closely related to an identity thesis concerning consciousness and brain behavior.

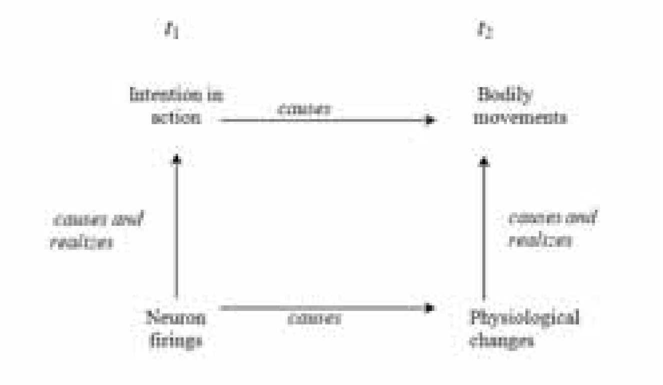
Firstly, causal explanability entails that the surface features of the phenomenon are *caused by* the behavior of the system's microstructure in which the phenomenon is realized in. But this causation does not mean that we have to do with two different things. In *Intentionality* the author mentions:

> "there can be causal relations between phenomena at different levels *in the very same underlying stuff* (…) to generalize at this point, we might say that two phenomena can be related by both causation and realization provided that they are so at *different levels of description*." (SEARLE, 1983, p. 266, my emphasis)

Searle's conception of the causation of surface features by the system's microstructure behavior does not concern an *event* which causes another, but a *sufficient condition* without temporal connotations (see SCHRÖDER, 1992, p. 100).

Secondly, the identity of causal powers is presented by Searle as a consequence of an identity relation between both phenomena. In one of his first writings on the philosophy of mind Searle defended the *causal efficacy* of mental phenomena and thought the description of its causal powers as possible at different levels: "Mental states are no more epiphenomenal than the elasticity and puncture resistance of an inflated tire are, and interactions *can be described both at the higher and lower levels*, just as in the analogous case of the tire." (SEARLE, 1980, p. 455, my emphasis)

Furthermore, it is clear that the identity of causal powers follows from the fact that both phenomena are *the same thing* described at different levels. These two points (the connection of *causal explanability* and of the *identity of causal powers* with the identity of both phenomena) become more understandable if we consider Searle's scheme for the representation of the causal functioning of mental states. In *Intentionality* he draws the following picture:



Searle asserts explicitly that "the phenomena at t_1 and t_2 respectively are *the same phenomena* described at different levels of description" (SEARLE, 1983, p. 269, my emphasis), what entails that the "cross level" causation between neuron firings and the intention in action is causation *with identity*, and the simultaneous relation of realization between them determines this identity.[1] This is, the causal explanability of the features of a conscious mental phenomenon is made possible by causal relations without time gap, by causal relations between different levels of the *same system*. And once that the phenomena at t_1 and t_2 are identical, he says that there are also "diagonal" causal relations between the phenomena at t_1 and t_2:

[1] Explaining the realization relation in the case of liquidity Searle writes: "the liquidity of a bucket of water is not some extra juice secreted by the H2O molecules. When we describe the stuff as liquid we are just describing those very molecules at a higher level of description than that of the individual molecule." (SEARLE, 1983, p. 266, my emphasis)

> "Notice that on this model (…) we could also draw diagonal arrows which in this case would show that the intention in action causes physiological changes and that the neuron firings cause bodily movements. Notice also that on such a model the mental phenomena are no more epiphenomenal than the rise in temperature of the firing of a spark plug."(SEARLE, 1983, p. 270)[2]

I think that these "diagonal" causal relations correspond to the identity of causal powers of conscious mental phenomena and brain processes, the second aspect of Searle's conception of causal reduction. And if it is really so, then this identity of *causal powers* is grounded on the identity of the *phenomena themselves*. The connection of (a) causal explanability and (b) identity of causal powers with (c) the identity of the phenomena is a strong evidence for the connection of causal and ontological reduction, because ontological reduction yields the conclusion that entities of certain types "consist in nothing but" (SEARLE, 1992, p. 113) entities of other types, what is for him a peculiar form of identity relation that exists also by properties (as liquidity, solidity and *consciousness*). Moreover, Searle himself acknowledges that, in general, successful causal reductions lead to ontological reductions: "where we have a successful causal reduction, we simply redefine the expression that denotes the reduced phenomena in such a way that the phenomenon in question can now be identified with their causes." (SEARLE, 1992, p. 115) It seems to me that the causal reduction makes such a possible *redefinition* because the causal explicability and the identity of causal powers allow an *identity statement* concerning both phenomena (for example liquidity and molecular behavior). But in Searle's opinion *there is an exception*, there is at least a phenomenon whose causal reduction does not lead to an ontological reduction: consciousness.

III. The Argument for Ontological Irreducibility

Ontological irreducibility leads to a situation that is in my opinion very strange, namely that "Consciousness is entirely causally explained by neuronal behavior but it is not thereby shown to be nothing but neuronal behavior." (SEARLE, 2004, p. 119) We saw above that causal explicability (and identity of causal powers) entails in Searle's view an identity relation between the phenomena in question, but if it is not the case that consciousness is *nothing but* neuronal behavior, then consciousness is *something else* as neuronal behavior, so that it is not clear *how* consciousness could be causally reducible. Appealing to Thomas Nagel's, Frank Jackson's and Saul Kripke's conceptions, which (in his opinion) have articulated the same argument in different ways (see SEARLE, 1992, p. 116-117), Searle offers the following formulation:

> "Suppose we tried to say the pain is really 'nothing but' the patterns of neuron firings. Well, if we tried such an ontological reduction, the essential features of the pain would be left out. No description of the third-person, objective, physiological facts would convey the subjective, first-person character of the pain, simply because the first person features are different from the third-person features." (SEARLE, 1992, 117)

He says explicitly that subjective and objective features *are different*, what is in my opinion incompatible with his conception of causal reduction presented above (section II). A redefinition of consciousness in terms of objective entities (as brain processes) is *impossible*, and it seems to me to *undermine* the possibility of a causal reduction in Searle's model. Such a causal reduction *requires* an ontological reduction. But now we need to examine his argument for the claim that ontological reduction has no deep consequences and to evaluate if it can make causal reduction compatible with ontological irreducibility.

IV. Is Ontological Irreducibility Harmless?

Searle refuses the general opinion that an ontological irreducibility of consciousness is a challenge to our scientific world view, and tries to prove that this irreducibility does not force us to a property dualism. He believes that ontological irreducibility is in this sense harmless because it is a consequence of our *interests* about consciousness, and not a consequence of the structure (or essence) of the phenomenon itself (see SEARLE, 1992, p. 123). According to him, an ontological reduction consists to carve off the surface features of a phenomenon and *to redefine* it in terms of the microlevel's causes of these surface features. We make this when our interest is to know about the microcauses. The only difference between *subjective* states of consciousness and *objectives* system features (as liquidity or solidity) is that in the case of consciousness our interest are the surface features, so that we cannot carve off them.

But what draw my attention is that Searle compares subjective with certain objective phenomena (as mud and music, see SEARLE, 2004, p. 120) – because, when we use the expressions "mud" and "music", we are interested on the surface features of these phenomena – and, moreover, says that we *could* make the redefinition if we want. These statements suggests (a) that consciousness *is* identical to brain processes and (b) that we are not interested in this identity when we use the expressions "consciousness", "pain", etc. – as we are not interested in the identity of music and air movements when we speak, for example, about Beethoven's ninth symphony. But these two claims seem problematic to me. Searle himself says that subjective and objective features are *different* – what becomes clear when we note that the description of molecular behavior *can* convey the surface features of mud, while the description of brain processes *cannot* convey the surface features of consciousness – and the fact that we are not interested in microcauses when we speak about surface features is *trivial* and cannot explain ontological irreducibility. If objective descriptions never would convey the subjective character of conscious states, because they are *different*, then ontological irreducibility *does not* follows of our pragmatics interests in the surface features.[3]

[2] In my presentation of Searle's view of causal reducibility I refer to his remarks about intentional states, while the subject of this paper is his account of the reduction of consciousness. But it seems not problematic for me, because Searle thinks consciousness and intentionality as connected, and makes similar remarks about the causal efficacy of conscious sensations (see SEARLE, 1995, p. 219). Moreover, he suggests that consciousness is identical to brain behavior (although consciousness is caused by it – see SEARLE, 2002a, p. 9)

[3] A further strategy to defend Searle's view would be to say that he takes consciousness not for identical but for supervenient to brain processes. But his position about supervenience is ambiguous. On one hand he says: "It is certainly true that consciousness is supervenient on the brain" (SEARLE, 2004, p. 148). On the other hand he finds this concept not helpful and thinks that his own theory of cross-level causation (that implies identity) is more interesting: "the concept of supervenience adds nothing to the concepts the we already have, such concepts as causation, including bottom-up causation, higher and lower levels of description, and higher order features being realized in the system composed of the lower level elements." (SEARLE, 2004, p. 149). The

V. Concluding Remarks

I think that causal and ontological reduction – in Searle's conception – is essentially linked and that causal reducibility is incompatible with ontological irreducibility. Because of this, Searle's theory implies *contradictory claims*: in some moments he asserts that consciousness and brain processes are *identical*, in other moments he says that they are *different*. It seems to me that Jaegwon Kim realizes this inconsistence when he comments Searle's claim that causal interactions between mental and physical phenomena can be rediscribed at different levels: "Obviously, the redescription strategy is available only to those who accept 'M=P', namely reducionist physicalists (Searle of course does not count himself among them)." (KIM, 2005, p. 48). Moreover Searle's strategy to show that ontological irreducibility is harmless seems to *repeat* the same mistake, then he suggests that consciousness and brain processes are *identical*, what is incompatible with his claims about ontological irreducibility (difference). This irreducibility is for me the most troublesome thesis of biological naturalism, and it would be very helpful for the credibility of the theory if this thesis was *eliminated*. Perhaps Searle should conceive the difference between consciousness and brain processes in another way which is not ontological.[4]

Acknowledgments

Financed by the DCR-Program in accord of CNPq and FUNCAP foundations (Brazil)

Literature

Kim, J. (2005) *Physicalism, or something near enough*. Princeton; Oxford: Princeton University Press.

SCHRÖDER, J. (1992) "Searles Auffassung des Verhältnisses von Geist und Körper und ihre Beziehung zur Identitätstheorie" In: *Conceptus* XXVI, nr. 66, pp. 97-109

Searle, J. (1980) "Intrinsic Intentionality" In: *Behavioral and Brain Sciences* 3, pp. 450-6.

__________ (1983) *Intentionality: an essay in the philosophy of mind*. Cambridge: Cambridge University Press.

__________ (1992) *The Rediscovery of the Mind*. Cambridge Mass., London: MIT Press.

__________ (1995) "Conciousness, the Brain and the Connection Principle: A Reply" In: *Philosophy and phenomenological Research* 55(1) pp. 217- 32.

__________ (2002a) *Consciousness and Language*. Cambridge (UK): Cambridge University Press.

__________ (2002b) "Why I Am Not a Property Dualist" In: *Journal of Consciousness Studies*, 9, No 12, pp. 57-64

__________ (2004) *Mind: a brief introduction*. Oxford: Oxford University Press.

concepts of levels of description and of realization implies identity (about realization see footnote 1 above).

[4] I am very grateful to Guido Imaguire and Noa Latham for many helpful comments and to Ananda Badaró for the correction of the English.

Algorithms and Ontology

Walter Dean, New York, USA

This purpose of this note is to advertise — but not answer — a question which is of significant foundational importance to both mathematics and computer science but which has been largely overlooked within philosophy of mathematics. Succinctly stated it is as follows:

> (A) Are the mathematical procedures conventionally termed *algorithms* themselves mathematical objects?[1]

I will assume that some general notion of algorithm — i.e. of a practical method for solving a mathematical problem — is already implicit in mathematical practice. This seems reasonable since algorithms for simple arithmetic operations (e.g. the "grade school" long division algorithm) have long been commonplaces of our informal computational practice. Specific algorithms (e.g. Euclid's algorithm) are well known not only because of their antiquity but also because of the ways in which they have contributed to modern mathematics (e.g. in the definition of Euclidean domain or in the proof of Sturm's theorem). Finally, a great many other algorithms have been developed in conjunction with specific subfields of mathematics — e.g. Brent's method in numerical analysis, Gosper's algorithm in combinatorics, Strassen's algorithm in matrix algebra — and will thus be known to specialists in these fields.

Mathematicians have traditionally been most interested in applying algorithms to solve mathematical problems such as determining whether a given number is prime or that a function has a root in a given interval. This flags two important observations about the role of algorithms in contemporary mathematics: 1) that many of the individual mathematical statements which we now take ourselves to *know* (e.g. that certain numbers are prime) have been derived by the application of specific algorithms; and 2) that mathematical interest has generally been focused on the results of applying these methods rather than on the computational properties of the methods themselves.[2]

The situation is quite different in contemporary computer science. In this context, algorithms are regarded as abstract objects in their own right whose properties may be directly studied and compared. This is evident from the sort of language used to describe individual algorithms, of which the following observations are typical:

> I) Individual algorithms are referred to by *proper names* — e.g. "Euclid's algorithm", MERGESORT, HEAPSORT, etc.
>
> II) Such names are used to predicate computational properties directly of individual algorithms — e.g. "MERGESORT" has running-time $O(n\log_2(n))$."
>
> III) General results are stated using quantifiers ranging over algorithms — e.g. "There is a polynomial time algorithm for primality", "If $P \neq NP$, then there is no polynomial time algorithm for deciding propositional satisfiability", "There is no comparison sorting algorithm with running-time less than $O(n\log_2(n))$."

If we apply conventional standards of ontological commitment to I)-III), we are led to the conclusion that computer science is committed to regarding algorithms realistically — i.e. as forming a class of objects to which algorithmic names (such as those in I)) refer, and over which quantifiers (such as those in III)) range.

To get an impression of what is at stake in our interpretation of such claims, it will be useful to consider the developments which led to the adoption of the idiom exemplified by I)-III). Statements of this sort are characteristic of a field known as *algorithmic analysis* which was established in the late 1950s by (Knuth 1973). Knuth proposed a means of measuring and comparing the efficiency of algorithms in terms of their so-called *big*-O *running-time complexity* — i.e. the asymptotic rate of growth of the number of steps $O(t_A(|x|))$ it takes algorithm A to return a value on an input x as a function of its size. The development of this theory was motivated by the dual observations that i) there exist intuitively distinct algorithms A_1 and A_2 which compute the same function but which differ in their asymptotic running-time and ii) the relative efficiency of A_1 and A_2 in practice is often invariant with respect to how they are implemented relative to a particular formal model of computation $\boldsymbol{M}$ (e.g. as RAM machines).

The first of these observations illustrates that within computer science, algorithms are treated *intensionally*. This is to say that an algorithm A is generally not identified with the function f_A it computes, but rather with a procedure or method whose operation induces this function. That algorithms are indeed individuated in this manner may be illustrated by observing that a computational predicate like "A has running-time $O(t(|x|))$" creates a context in which the substitution of names for algorithms which compute the same function need not preserve truth value.[3] The second observation records the fact that it is conventional to treat algorithms as the intrinsic bearers of asymptotic complexity theoretic properties. For not only is it often difficult to reason mathematically about the complexity of an algorithm if we are forced to work with a particular mathematical representation (e.g. RAM machine), but the combinatorial features of such representations often turn out to be irrelevant for comparing the behavior of different algorithms in practice.

These observations shed some light on why it is useful to adopt an idiom which treats algorithms as

[1] Although it appears that this question has not been systematically investigated by philosophers, the technical proposals of (Moschovakis 1998) and (Gurevich 1999) both seek to establish positive solutions. However, both of these approaches arguably fall victim to the problem of "computational artifacts" which is discussed below.

[2] This is not, of course, to say that properties of algorithms are completely ignored by mathematicians. For in particular, it is acknowledged that prior to claiming that the fact that the application of an algorithm A to a value a yields b as output is a proof that the value of a function f at a is equal to b, A must be proven *correct* with respect to f -- i.e. it must be shown that $\forall x[f(x) = A(x)]$. Such proofs generally proceed by constructing a mathematical model M of A -- i.e. a purely mathematical representation of its mode of operation. The availability of such representations might be taken to suggest that mathematical practice is already committed to some version of (A). However, since mathematicians are not generally interested in the computational properties of individual algorithms (e.g. their running time), they will generally accept correctness proofs based on models M which only weakly reflect the operation of the algorithms which they are introduced to represent.

[3] For instance, it does *not* follow from the fact that 1) MERGESORT has running-time $O(|x|log2(|x|))$, and 2) that MERGESORT and SELECTIONSORT compute the same function that 3) SELECTIONSORT has running-time $O(|x|log2(|x|))$.

objects. But they also leave unanswered a variety of foundational questions concerning what it means to regard algorithms in this manner. For note that if algorithms are indeed treated as intensional entities within computer science, then we might fear that will be forced to posit a novel class of non-extensional (and perforce non-mathematical) abstract objects in order to account for the truth conditions of statements of types II) and III).[4] This concern serves to illustrate the importance of establishing a positive answer to (A).

Some hope that such an answer may be given comes from reflecting on the origins of computer science within computability theory. The origins of the latter subject can be traced to the call to provide a mathematical definition of the class *C* of *effectively computable functions* as it arose within the Hilbert programme. It this context, there was general agreement that a function *f:* $N^k \rightarrow N$ is effectively computable just in case there exists an algorithm for computing its values. But in order to show that a given function is *not* effectively computable required that ***C*** be given a precise definition.

The developments leading to the consensus that ***C*** should be identified with the class of partial recursive functions — i.e. to the claim now known as *Church's Thesis* [CT] — are sufficiently familiar that they need not be repeated here. What is less well recognized is that the original arguments for CT did *not* proceed by first giving a mathematical definition of a class ***A*** which could plausibly be taken to consist of objects corresponding to algorithms and then defining

$$\text{(C)} \quad \mathbf{C} =_{\mathrm{df}} \{f : \mathrm{N}^k \rightarrow \mathrm{N} : f(x_1, ..., x_n) = A(x_1, ..., x_n) \; \& \; A \in \boldsymbol{A}\}$$

Rather, Church, Turing, Gödel, and Post all proceeded by defining a class of formal models ***M*** (which I will refer to somewhat inaccurately as *machines*) which formalize different notions of what it means to be a finitary procedure. For instance, for Gödel, ***M*** consisted of the class of general recursive definitions. Such definitions can be taken to formalize a variety of ways in which functions can be introduced so that their values can be explicitly computed (e.g. by course of values recursion). But if we define $\boldsymbol{F_M}$ to be the class of functions computable by members of ***M***, for each choice of ***M***, the question remains as to whether the corresponding class $\boldsymbol{F_M}$ exhausts all effectively computable functions.

In order to demonstrate that we ought to accept ***C*** = $\boldsymbol{F_M}$ thus requires an additional argument that for any informally characterized algorithm *A*, there exists a machine $M \in \boldsymbol{M}$ such that *A* and *M* determine the same function. This appears to be an extensional claim about the relationship between two classes of functions. But note that any argument in its favor must apparently proceed by the following intensional route: i) given any algorithm *A*, there is an $M \in \boldsymbol{M}$ such that each step in the informally characterized operation of *A* can be correlated with one or more steps in the operation of *M*; ii) hence the function induced by the complete operation of *M* coincides with that induced by that of *A*. Incipient arguments to this effect may be found in the original papers of Church, Turing, and Post from 1936. Better fleshed out versions appear in the writings of more recent commentators such as (Rogers 1967), (Gandy 1980), and (Sieg & Brynes 2000).

Inasmuch as any sound argument for CT must proceed in the manner just suggested, one might reasonably conclude that at least certain choices for ***M*** will include a formal representation of every algorithm.[5] And on this basis, one might conclude that it is allowable to take ***A*** = ***M*** in (C). But the members of ***M*** will generally be finite combinatorial objects, and thus mathematical objects *par excellence*. Thus one also might conclude that not only should (A) be answered in the positive, but such an answer is already implicit in our acceptance of CT.

The fact that such a conclusion is *not* warranted follows by reflecting further on some basic results which have emerged from algorithmic analysis. As noted above, for instance, algorithms are individuated at least as finely as their big-O running-times. But for certain choices of ***M***, we can find examples of algorithms *A* computing certain functions *f* to which we assign running-time $O(t_A(|x|))$ but for which it may be shown that there is no $M \in \boldsymbol{M}$ computing *f* with running-time $O(t_M(|x|)) \leq O(t_A(|x|))$.[6] If we take the property of having running-time $O(t_A(|x|))$ to be a property of *A* itself, then results like this suggest that we cannot take algorithms to be *identical* to members of any specific class ***M***. For if we were to do so, there would be no guarantee that there is a member of ***M*** which faithfully represents *A*'s computational properties.

This situation highlights the kind of conceptual and technical problems which arise when we attempt to settle (A) directly by identifying algorithms with machines. For on the one hand, the argument for CT sketched above promises to show that for every informally presented algorithm *A*, there will exist a machine $M_A \in \boldsymbol{M}$ which mimics its step-by-step operation. But on the other hand, the question of determining when the existence of a particular form of step-by-step correlation is sufficient to allow us to conclude that M_A is identical to *A* appears to require that we have a prior characterization of the properties of *A* itself. This is to say that before we can be in a position to assess whether a given argument for (A) of this form might be successful, we must first agree on how our computational practices fix the properties of individual algorithms.

At this point, a number of analogies between (A) and various reductive proposals in the philosophy of mathematics can be drawn. For note that if we agree that algorithms are regarded as intensional objects in computer science, (A) amounts to the claim that reference to such entities can be eliminated in favor of extensional mathematical ones. The desire to demonstrate such a claim can thus be compared to the traditional nominalist desire to show how reference to mathematical entities can be eliminated in favor of reference to concrete ones.

This observation suggests that other strategies are available in attempting to demonstrate (A) than simply attempting to identify a mathematical object to correlate with each individual algorithm — i.e. what (Burgess &

[4] The gravity of this concern will ultimately depend upon how tightly the practice of computer science pins down the identity conditions which must be imposed on algorithms. It follows from the example of the previous note that extensionally equivalent algorithms cannot be identified when they differ with respect to a definite computational property such as asymptotic running-time complexity, But at the same time, there do not appear to be cases in which statements of algorithmic *non-identity* -- i.e. of the form *A*1 ≠ *A*2 -- are accepted in computer science when no such property serves to distinguish *A*1 and *A*2.

[5] This point is put by (Rogers 1967) (p. 19) as follows: "[T]here is a sense in which each of the standard formal characterizations appears to include all possible *algorithms* ... For given a formal characterization..., there is a uniform effective way to 'translate' any set of instructions (i.e. algorithm) of that characterization into a set of instructions of one of the standard formal characterizations."

[6] A number of examples of lower-bound results of this nature are known for the single-tape, single-head Turing machine model ***T*** which is most often referenced in Rogers-style translation arguments. For instance, while there is a trivial O(*n*) algorithm for determining whether a binary string is a palindrome, no machine $T \in \boldsymbol{T}$ can solve this problem in time faster than O(*n2*) (cf. Hopcort and Ulman 1979).

Rosen 1997) call *objectual reduction*. Rather, we might start out by treating our computational practice as constituting a term-introducing "procedural" theory T_p. Such a theory would contain not only standard mathematical terms and quantifiers, but also terms (A_1, A_2, ...) naming algorithms and quantifiers ($\forall X_1$, $\forall X_2$, ...) ranging over such entities.

The question as to whether T_p commits us to the existence of a non-mathematical class of entities corresponding to the range of the procedural quantifiers can accordingly be formalized by asking whether it is possible to interpret T_p over a purely mathematical theory $T_\mathrm{m} \subseteq T_\mathrm{p}$. In particular, we can ask whether T_p is conservative over T_m for purely mathematical statements and also whether it is possible to formulate T_m in a manner such that it is able to derive appropriate interpretations of results of types II) and III).

Demonstrating the former fact is likely to be straightforward as it requires only that T_m is able to prove the correctness of various algorithms (in the sense of note 2) relative to some means of representing them which need not reflect their intensional properties such as running-time. However, constructing an interpretation which is also capable of accounting for results in algorithmic analysis will most likely require that we attend to the details of how we make reference to individual algorithms in practice. Reflection on this topic suggests that: 1) the only linguistic means we have of referring to individual algorithms is via expressions of the form "the algorithm implemented by machine *m*"[7]; 2) we generally take it to be possible to refer to *the same* algorithm by referring to *different* machines. As a consequence, T_p is likely to contain many statements of the form $imp(m_1) = imp(m_2)$, and $imp(m_1) \neq imp(m_2)$ (where $imp(\cdot)$) is intended to formalize "the algorithm implemented by *m*").

This latter observation illustrates why it is unlikely that the interpretation of an algorithmic name like MERGESORT can be taken to be identical to any *particular* machine.[8] If T_p is to reflect the grammatical structure of statements like those in II) and III), this suggests that we must take the values of $imp(\cdot)$ to be equivalences classes of machines under a definition of *computational equivalence* ≈ defined over a suitable class of machines ***M***.[9] Such a definition would ideally serve to analyze the meaning of statements of the form

(M) machines m_1 and m_2 implement the same algorithm

in a manner which additionally satisfied all statements of algorithmic identity and non-identity contained in T_p. On this proposal, the sustainability of (A) will rest on the availability of such a definition of equivalence. If such a definition could be given, we would have shown how it was possible to *contextually reduce* procedural discourse to mathematical discourse (again in the sense of Burgess & Rosen). The ontological status of algorithms could accordingly be taken to be that of (neo)-Fregean abstracts over ***M*** relative to **≈**.

Literature

Burgess, John & Gideon Rosen 1997 A subject with no object, Oxford: Clarendon Press.

Gandy, Robin 1980 "Church's thesis and principles for mechanisms" in Jon Barwise,
H. J. Keisler, and K. Kunen (eds.) The Kleene Symposium, Amsterdam: North-
Holland, 123–148.

Gurevich, Yuri 1999 "The sequential ASM thesis." Bulletin of the EATCS, 67, 93-125.

Hopcroft, John & Jeffrey Ullman 1979 Introduction to Automata Theory, Languages, and Computation Boston: Addison-Wesley.

Knuth, Donald 1973 The art of computer programming, volumes I-I I I. Boston: Addison Wesley.

Moschovakis, Yiannis 1998 "On the founding of a theory of algorithms" H. G. Dales & G. Oliveri, (eds.), Truth in mathematics, 71–104. Oxford: Clarendon Press.

Rogers, Hartley 1967 Theory of Recursive Functions and Effective Computability.

[7] The other option is to treat algorithms as corresponding to the denotations of *programs* -- i.e. linguistic descriptions of procedures given over a formal programming language. However, reference to algorithms via this route arguably collapses into reference via machines as each program will be interpretable as a machine via an appropriate form of operational semantics.

[8] For if we take $A = M$ for a fixed M there will generally be no way of defining *imp* so that these identity and non-identity statements are satisfied. More generally, such a proposal will entail that the computational properties of *A* are identical to those of *M*. But this will generally be unacceptable since machines possess a variety of "artifactual" properties which we generally do not attribute to algorithms -- e.g. have a fixed number of states, having exact (as opposed to asymptotic) running-time, etc.

[9] This fact is recognized by Moschovakis (who identifies algorithms as equivalence classes of computational models known as *recursors*) but it is ultimately denied by Gurevich. Even for Moschovakis, however, the question remains whether his chosen notion of equivalence either 1) serves to analyze the meaning of statements of the form (M) and 2) induces identity questions on algorithms which are consistent with those reflected by *T*p.

The Knower Paradox and the Quantified Logic of Proofs

Walter Dean / Hidenori Kurokawa, New York, USA

The Knower paradox was originally introduced by (Montague and Kaplan 1960) [M&K]. We will begin by recording a simple version of the paradox adapted from (Egré 2005). Suppose that T extends Q and let $K(x)$ be a (possibly complex) predicate in L_T. It follows that T proves a fixed point theorem of the following form:

(**FP**) For every open formula $\varphi(x)$ in L_T, there exists a sentence δ such that

(*) $T \vdash \varphi(\underline{\delta}) \leftrightarrow \delta$.

Now suppose $K(x)$ additionally satisfies

(**T**) $T \vdash K(\underline{\varphi}) \rightarrow \varphi$
(**Nec**) if $T \vdash \varphi$, then $T \vdash K(\underline{\varphi})$

Then it may be shown that T is inconsistent by letting δ be such that

1) $T \vdash \neg K(\underline{\delta}) \rightarrow \delta$
2) $T \vdash K(\underline{\delta}) \rightarrow \neg\delta$

via (**FP**) and then arguing as follows

3) $K(\underline{\delta}) \rightarrow \delta$ **T**
4) $\neg K(\underline{\delta})$ 2), 3)
5) δ 1), 4)
6) $K(\underline{\delta})$ 5), **Nec**
7) $\perp$

The foregoing presentation of the Knower departs from that of M&K in two respects. The first of these is that rather than using a sentence δ satisfying 1), 2), they use one satisfying $K(\underline{\neg\delta}) \leftrightarrow \delta$. The second is that we have employed the rule **Nec,** as opposed to assuming that $K(x)$ satisfies the axioms

(**U**) $K(\underline{K\ \varphi \rightarrow \varphi})$
(**I**) $K(\underline{\varphi})\ \&\ I(\underline{\varphi},\underline{\psi}) \rightarrow K(\underline{\psi})$

wherein $I(\underline{\varphi},\underline{\psi})$ expresses that ψ is derivable from φ. It may reasonably be claimed that the original derivation of M&K rests on a set of principles which more precisely isolates the source of the paradox than those we have employed. We have elected to base our treatment on 1)-7) because the resolution we suggest below will also be applicable to the choice of fixed point and weaker principles employed by M&K.

It is also notable that the Knower was originally formulated in an arithmetic language as opposed to one with a propositional operator. This reflects the fact that M&K assume that such a setting is required in order to ensure the existence of self-referential statements and argues that they took the paradox to weaken Quine's argument that modal operators must be conceived as predicates of sentences. As Érgé convincingly argues, however, the availability of self-reference in a language with modal operators is essentially independent of whether we think of these operators as taking sentences or propositions as arguments.

This observation suggests that by viewing the foregoing derivation in a modal setting, it may be possible to isolate the principles which lead to paradox in a manner that does not depend on the mechanism by which self-reference is achieved. It is an easy observation that this derivation remains valid when we reinterpret $K(x)$ as a propositional operator and treat the arithmetic sentence δ as a denoting a fixed proposition D of which

8) $\neg\Box D \leftrightarrow D$

is provable. When recast in this light, the derivation can be taken to show that there is a general conflict between the modal reflection axiom T (which is the analogue of **T**) and any modal principles which would imply 8).

One means by which this can be demonstrated is to note that the logic S4 (which includes T) is incompatible with self-reference in the sense that not only is it incapable of proving any instance of 8) but also

9) S4 + $\Box$ ($\neg\Box D \leftrightarrow D$) is inconsistent[1]

This result might be taken to bear on the Knower not only because its proof essentially recapitulates 1)-7), but also because there is a well-known interpretation of S4 whereby $\Box$ is assigned the reading

10) $\Box F$ iff F is informally provable

Such an interpretation was first proposed by (Gödel 1933) in an attempt to provide a provability semantics for intuitionistic logic. The details of what follows do not, however, rely specifically on the relationship between S4 and intuitionism. Rather, they depend on the availability of so-called *explicit* refinements of S4 which can be employed to reason about knowledge qua provability.

For present purposes, an explicit modal logic can be taken to be one that possesses an infinite family of modalities of the form $t{:}F$. As opposed to 10), wherein $\Box$ expresses a notion of provability in which proofs are kept implicit, this notation is conventionally assigned the interpretation

11) $t{:}\varphi$ iff t verifies φ

Here t may be a structured term which, in the paradigmatic case, is taken to denote an explicit mathematical proof. A system employing this notation was envisioned by (Gödel 1938). However, a complete formalization of a logic of explicit proof was first provided by (Artemov 2001) under the name LP (the *Logic of Proofs*).

LP itself is not sufficient to express the versions of **FP** and **T** which are required to formulate the Knower. For note that on their intended interpretations, both the arithmetic knowledge predicate $K(x)$ and the informal provability predicate $\Box$ contain implicit quantifiers over proofs or other evidentiary entities. This is clearest in the case of $K(x)$, which is standardly taken to extend an arithmetic provability predicate $Bew(y)$ which itself abbreviates a statement of the form $\exists x Proof(x,y)$. Although as Gödel already observed, the $\Box$ of S4 cannot be interpreted as expressing provability within formal

[1] This tension also surfaces with respect to the provability logic GL in which statements like 8) are provable. However, it may easily be shown that GL + T is inconsistent.

arithmetic, 10) is already suggestive of quantification over a domain of informal proofs — cf. (Tait 2001).

In order to reconstruct the Knower in a system of explicit modal logic, we need a version of LP which contains quantifiers ranging over proofs. Such a system is presented in (Fitting 2004) under the name QLP. The language of QLP is given by first specifying a class of *proof terms*

$$\mathrm{Term}_{QLP} = c \mid x \mid !t \mid t_1 \cdot t_2 \mid t_1 + t_2 \mid \langle t\forall x\rangle$$

The class of formulas of LP is then specified as follows:

$$Form_{LP} = P \mid t{:}\,\varphi \mid \neg\varphi \mid \varphi \rightarrow \psi \mid (\forall x)\varphi \mid (\exists x)\varphi$$

The axioms of QLP are as follows:

LP1	all tautologies of classical propositional logic
LP2	$t{:}(\varphi \rightarrow \psi) \rightarrow (s{:}\,\varphi \rightarrow t{\cdot}s{:}\psi)$
LP3	$t{:}\,\varphi \rightarrow \varphi$
LP4	$t{:}\,\varphi \rightarrow !t{:}t{:}\,\varphi$
LP5	$t{:}\,\varphi \rightarrow t{+}s{:}\,\varphi$ and $s{:}\,\varphi \rightarrow t{+}s{:}\,\varphi$
QLP1	$(\forall x)\,\varphi(x) \rightarrow \varphi(t)$
QLP2	$(\forall x)(\psi \rightarrow \varphi(x)) \rightarrow (\psi \rightarrow (\forall x)\,\varphi(x))$
QLP3	$\varphi(t) \rightarrow (\exists x)\varphi(x)$
QLP4	$(\forall x)(\varphi(x) \rightarrow \psi) \rightarrow ((\exists x)\varphi(x) \rightarrow \psi)$
UBF	$(\forall x)t{:}\varphi(x) \rightarrow \langle t\forall x\rangle{:}(\forall x)\varphi(x),\ x \notin FV(t)$

Axioms LP1-LP5 correspond to versions of the S4 axioms wherein instances of □ have been "realized" by proof terms. Axioms QLP1-QLP4 correspond to a set of axioms adequate for classical predicate calculus and to which the usual free variable restrictions apply. UBF is an explicit form of the Barcan formula and is justified on the basis of the observation that if we possess a proof term t which serves to *uniformly* verify $\varphi(x)$ for all x, then there should be a proof (denoted by the complex proof term $\langle t\forall x\rangle$) which serves to justify $(\forall x)\,\varphi(x)$. The rules of QLP consist of modus ponens and universal generalization together with a rule known as *axiom necessitation*. This rule says that if φ is an axiom of QLP, then we may introduce $c{:}\varphi$ where c is a so-called *proof constant* — i.e. an unstructured proof term introduced as an atomic justification for φ.

Before reconstructing the derivation of the Knower in QLP, it will be useful to record the following technical result:

Theorem (Lifting) [Artemov/Fitting]

If QLP $\vdash \varphi$, then for some proof term t, QLP $\vdash t{:}\varphi$.

The Lifting Theorem reports that if a statement φ is derivable in QLP, its derivation may be internalized within the system so as to yield a proof term t which exhibits its structure. As such, the Lifting Theorem serves as a sort of explicit counterpart to the S4 necessitation rule (i.e. $\vdash F \,/ \vdash \Box F$) which itself is an implicit form of the rule **Nec** used to justify the step 5)-6).

The final step which we must undertaken before reconstructing the Knower is to introduce some means of introducing an explicit analog of a self referential statement which mirrors (*). The most straightforward way to proceed is to simply consider the result of adjoining a statement of the form

12) $d{:}(\neg(\exists x)x{:}D \leftrightarrow D)$

which formalizes "d is a proof of 'there does not exist a proof of D iff D'." Reasoning in QLP from 12) as a premise, we may now derive a contradiction as follows:

13)	$\neg(\exists x)x{:}D \rightarrow D$	left to right direction of 11)
14)	$(\exists x)x{:}D \rightarrow \neg D$	right to left direction of 11)
15)	$(\exists x)x{:}D \rightarrow D$	derivable in QLP
16)	$\neg(\exists x)x{:}D$	propositional logic
17)	D	
18)	$t : D$	for some term t obtainable via Lifting
19)	$t : D \rightarrow (\exists x)x{:}D$	QLP3
20)	$(\exists x)x{:}D$	
21)	$\perp$	

The step 17)-18) is analogous to the step 5)-6) in the original derivation. In the case of QLP, however, this step is elliptical in the sense that although we know a term t exists via the Lifting Theorem, such a term must be explicitly constructed by internalizing steps 13)-18). Constructing t requires not only constants d_1 and d_2 such that

21) $d_1{:}(\neg(\exists x)x{:}D \rightarrow D)$
22) $d_2{:}((\exists x)x{:}D \rightarrow \neg D)$

(which may be constructed from d in 12)) but also a proof term which serves as a verification of 16). Note that while this statement is a explicit analog of an instance of the reflection axiom T, it is *not* an axiom of QLP. Not only must this statement be derived in QLP, but to construct t, its proof must also be lifted. This may be done as follows:

24)	$x{:}D \rightarrow D$	LP3
25)	$r{:}(x{:}D \rightarrow D)$	axiom necessitation
26)	$(\forall x)r{:}(x{:}D \rightarrow D)$	universal generalization
27)	$(\forall x)r{:}(x{:}D \rightarrow D) \rightarrow \langle r\forall x\rangle{:}\forall x{:}(x{:}D \rightarrow D)$	UBF
28)	$\langle r\forall x\rangle{:}(\forall x){:}(x{:}D \rightarrow D)$	
29)	$(\forall x)r{:}(x{:}D \rightarrow D) \rightarrow ((\exists x){:}D \rightarrow D)$	QLP4
30)	$q{:}[(\forall x)r : (x{:}D \rightarrow D) \rightarrow ((\exists x){:}D \rightarrow D)]$	axiom necessitation
31)	$q \cdot \langle r\forall x\rangle{:}((\exists x){:}D \rightarrow D)$	LP2

With this derivation in hand, it is then easy to see that we may take

31) $t \equiv d_1 \cdot ((a \cdot (q \cdot \langle r\forall x\rangle))\cdot d_2)$

where a is a proof constant for the tautology $(\varphi \rightarrow \psi) \rightarrow ((\varphi \rightarrow \neg\,\psi) \rightarrow \neg\,\varphi)$.

The insight which we think QLP provides into the Knower can now be framed by considering the role which UBF plays in the foregoing derivation. For note that QLP includes neither a general necessitation rule analogous to **Nec**, nor even a local instance of this principle akin to **U**. As we have just seen, however, the explicit forms of both principles — i.e.

32) $(\exists x){:}D \rightarrow D$ and
33) $q \cdot \langle r\forall x\rangle{:}((\exists x){:}D \rightarrow D)$

— are derivable in QLP. Both of these principles are required in order for the derivation 13)-21) to go through. However UBF turns out to essential to the derivation of 33) as it may be shown , without UBF, no statement of the form $t{:}((\exists x){:}D \rightarrow D)$ is derivable in QLP.[2]

This is significant for diagnosing how the principles involved in the original derivation of the Knower conflict. Several recent commentators have proposed that the paradox should be resolved by rejecting **U**.[3] However, the

[2] This follows from the fact that UBF is not conservative over the QLP-UBF for statements not containing terms of the of $\langle r\forall x\rangle$. In particular, it may be shown that for no φ, do we have QLP−UBF $\vdash (\exists y)y{:}((\exists x){:}\,\varphi \rightarrow \varphi)$.

[3] More specifically, among the three principles employed in the original M&K derivation (i.e. **T**, **U** and **I**), the consensus among recent commentators has

original motivation for adopting this principle over the **Nec** rule seems to be mainly to reduce the strength of the assumptions required to develop the paradox. (Note in particular that the epistemic rational which is commonly given for adopting **U** seems to be a special case of that which is given for **Nec**.) If we develop the Knower in the context of QLP, not only is neither principle accorded elementary status, but the foregoing observations demonstrate that if we think of knowledge in terms of proof existence, that there is an implicit interaction between the knowledge modality and proof quantification implicit in the original derivation. It is precisely this interaction which is exposed by the role of UBF in QLP derivation.

This observation prompts a reconsideration of UBF itself. The original motivation for its inclusion in QLP was to preserve the Lifting Theorem (a version of which also holds for LP). However, in light of the original setting of the Knower, one might also inquire into its arithmetic significance. In this regard, a parallel may be drawn between UBF and implicit form of the Barcan formula — i.e.

34) $(\forall x)\Box\varphi(x) \rightarrow \Box(\forall x)\varphi(x)$

— in the context of Quantified Provability Logic. As (Boolos 1993) observes, if we take $\Phi(x) \equiv \neg Proof_{PA}(x,\bot)$, then it may readily be seen that 34) is not arithmetically valid. For note that on this interpretation, the antecedent expresses the fact that no natural number is provably a proof of $\bot$, while the consequent expresses the fact that it is provable that there is no proof of $\bot$. But of course the former statement is true (in the standard model), whereas, per the second incompleteness theorem, the latter is false (assuming that PA is consistent).

Now define an arithmetic interpretation of QLP to be a mapping $(\cdot)^*$ which i) replaces every propositional letter P with an arithmetic sentence $(P)^*$ and every proof term t with a natural number or variable according to its type, ii) is such that $(x{:}\varphi)^* = Proof_{PA}(x,\varphi^*)$, iii) commutes with connectives, and iv) is such that $((\forall x)\varphi)^* = (\forall x)[Pf(x) \rightarrow \varphi^*]$ (where $Pf(x)$ expresses that x is a code of a proof). On the basis of such an interpretation, it may similarly be shown that UBF is not arithmetically valid. In particular, the interpretation of UBF for $\Phi(x) = \neg x{:}\bot$ corresponds to the claim that if for all natural numbers x, $(b)^*$ is a proof that if x codes a proof, then $\neg Proof_{PA}(x,\bot)$, then $(<b\forall x>)^*$ is a proof that there is no proof of $\bot$. On the assumption that $(<b\forall x>)^*$ denotes a standard natural number (and that PA is consistent), this conclusion also violates the second incompleteness theorem. And from this it follows that there can be no uniform means of arithmetically interpreting proof terms of the form $<t\forall x>$.

We take the foregoing observations to highlight the applicability of explicit modal logic to the Knower, but also point to a more precise diagnosis of the root of the paradox. For not only does the use of constructive necessitation in the derivation allow us to see logical structure which is hidden by the use of principles like **U** or **Nec** in the original derivation, but it also appears that there are good reasons to be suspicious of at least one of the principles which is suppressed in the original derivation — i.e. UBF — at least if we wish to assign it an arithmetic interpretation.

This desire may be reasonable if we look to arithmetic for the source of self reference required to develop the Knower. However, if our aim is merely to reason about justified knowledge more generally, there may also be good reasons to retain UBF. For not only does it arises naturally out of reflection on the notion of informal proof and provability, it also allows us to prove the provable consistency of our reasoning about these concepts. Both facts appear to have been foreseen by (Gödel 1938, p. 101-103). Much more can be said about these issues, but doing so is outside the scope of the current paper.

Literature

Artemov, Sergei 2001 "Explicit Provability and Constructive Semantics", *The Bulletin of Symbolic Logic* 7(1), 1-36.

Boolos, George 1993 *The Logic of Provability*, New York: Cambridge University Press.

Cross, Charles 2001 "The Paradox of the Knower without Epistemic Closure", *Mind* 113, 109-114.

Egre, Paul 2005 "The Knower Paradox in the Light of Provability Interpretation of Modal Logic", *Journal of Logic, Language and Information* 14, 13-48.

Fitting, Melvin 2004 "Quantified LP", Technical report, CUNY Ph. D. Program in Computer Science Technical Report TR2004019.

Gödel, Kurt 1933 "An Interpretation of the Intuitionistic Propositional Calculus", in: Solomon Feferman et al. (eds.), *Collected Works*, Vol. 1 K. Gödel, New York: Oxford University Press.

Gödel, Kurt 1938 "Lecture at Zilsel's", in: Solomon Feferman et al. (eds.), *Collected Works*, Vol. 3, K. Godel, New York: Oxford University Press.

Kaplan, David and Montague, Richard 1960 "A Paradox Regained", *Notre Dame Journal of Formal Logic* 1, 79-90.

Maitzen, Stephen 1998 "The Knower Paradox and Epistemic Closure", *Synthese* 114, 337-54.

Tait, William 2006 “ Gödel’s interpretation of intuitionism”, *Philosophia Mathematica* 14, 208-228.

been to blame the paradox on either **U** or **I.** (Maitzen 1998) argues that the paradox may be resolved by rejecting the assumption that knowledge is closed under deductive consequence as embodied by **I**. However, (Cross, 2001) shows that a version of the Knower may be developed by using a modified knowledge predicate which is not assumed to be deductively closed. This observation appears to lay the blame squarely on the principle **U** -- a point of view which is adopted by both Cross and Érgé. We take our explicit reconstruction of the Knower to deepen the motivation for adopting this position.

Quine on the Reduction of Meanings

Lieven Decock, Amsterdam, The Netherlands

Quine's semantic nihilism is well-known. From his earliest work onwards, he expelled meanings from his ontology. One of the important innovations in his doctoral thesis, *The Logic of* Sequences, which is a reworked version of Whitehead and Russell's *Principia Mathematica*, was extensionalism. Quine replaced the intensional propositional functions by extensional classes. During his trip to Europe in 1933, he discovered that this had become standard practice in Europe, and has defended extensionalism ever since, even in his latest writings. The only universals one should accept are classes. He regarded classes or sets as *bona fide* objects, because there is a clear criterion of identity, viz. classes can be identified through their members. For intensions no such criterion is available, so they cannot be hypostasised (1960:244). For some years, Quine also tried to get rid of classes (Quine 1936; Goodman & Quine 1947), but he came to recognize the necessity of positing sets, thus giving up strict nominalism.

Quine's extensionalism has determined his views on meaning and semantics. Attributes or meanings, the intensional components of universals, are only acceptable if they can be given a clear criterion of identity. In practice, this meant that meanings are only acceptable insofar they can be reduced to clearly identifiable objects, i.e. classes of classes, … , of physical objects. In Quine's work, one can find two concrete proposals for such a reduction. In the first proposal empirical meanings are characterised as stimulus meanings, i.e. classes of physical stimuli, in the second they are classes of linguistic expressions. Quine judges both proposals unsuccessful.

Quine defines a *stimulus meaning* as the ordered pair of the affirmative stimulus meaning and the negative stimulus meaning (1960:31-35). The affirmative stimulus meaning is the class of all stimulations to which a given speaker at a date would assent; the negative stimulus meaning the class of stimulation to which she would dissent. The proposal can be sharpened by using reaction time to measure doubt, or by introducing a modulus, i.e. a maximum time duration for stimulations. So far, stimulus meaning is reduced to an ordered pair of classes of stimulations. An ordered pair can be reduced by means of Kuratowski's or Wiener's reduction method to sets (1960 §53). Hence, a stimulus meaning is clearly identifiable if stimulations can be reduced to simpler entities. The stimulations can be ocular irradiation patterns together with "the various barrages of other senses, separately and in all synchronous combinations" (1960:33). Ocular irradiation patterns are types of evolving chromatic irradiation patterns of all durations up to some modulus. An alternative is defining an external momentary stimulation as "the set of [a person's] triggered receptors." (1981:50) Quine's notion of stimulus meaning is unproblematic from a reductionist point of view. Empirical meanings are reduced by means of reduction strategies that are acceptable for Quine to entities that Quine finds unproblematic, namely physical objects and classes.

The reduction strategy is clearly inspired by Carnap's reduction programme in *Der logische Aufbau der Welt*. Quine's worries concerning stimulus meaning accord with his objections to Carnap's early reduction programme and his later verification theory of meaning (Carnap 1936). Quine believes that stimulus meaning is restricted to observation sentences, whereas most sentences are not immediately linked to sensory stimulations. Only sentences at the boundary of the web of belief have stimulus meaning, but this is only a small fraction of all sentences. Hence, stimulus meaning is not a viable basis for semantics, because the meaning of most sentences cannot be explained as stimulus meaning. In brief, Quine has given an impeccable reduction strategy, and at the same time pointed out its severe limitations.

In Quine's second reduction strategy, "we could define the meaning of an expression as the class of all expressions like it in meaning." (1992:52; see also 1960:201; 1979:140; 1981:46). The reduction of expressions is unproblematic, either to classes (via Gödel numbering and the reduction of numbers to sets) or to classes of physical objects (inscriptions). More noteworthy, the class of meaningful expressions can be precisely delineated in grammar (see Decock 2002:86). For the reduction strategy to work, the only further requirement is that a precise characterisation of the dyadic predicate "*x* is alike in meaning with *y*" or "*x* is synonymous with *y*" is elaborated. In his early work, Quine is extremely sceptical about this notion of synonymy, especially for standing sentences (1960:201). Of course, one can use stimulus meaning to define synonymy, and even with the help of first order logic extend this notion to 'cognitive synonymy' (1979), but this will not help for standing sentences. The only alternative is to characterise synonymy by means of the notion of analyticity: "Once we have analyticity, cognitive equivalence is forthcoming; for two sentences are cognitive equivalent if and only if their truth-functional biconditional is analytic." (1992:54*f*) In view of Quine's well-known demise of the analytic-synthetic distinction, this looks like a dead end.

However, in an interview on the occasion of the Rolf Schock prize in November 1993, Quine said:

> Yes so, on this score I think of the truths of logic as analytic in the traditional sense of the word, that is to say true by virtue of the meanings of the words. Or as I would prefer to put it: they are learned or can be learned in the process of learning to use the words themselves, and involve nothing more. They are analytic in the same sense in which the standard example such as "No bachelor is married", is analytic: something that's learned in the process of learning to use the word "bachelor" itself.
> (Bergström & Føllesdal 1994, 199*f*)

This passage and other more covert passages (1974:79; 1992:55) look like a recantation of one of Quine's most famous claims. Quine admits that theorems of first order logic can be analytic, and that sentences such as "No bachelor is married" can be analytic. It is arguable that Quine has regarded first order logic as analytic since the end of the 1940s, or even propositional logic as analytic since 'Truth by convention' (1936). Nevertheless, the claim that "No bachelor is married" can be analytic, is a radical departure from earlier claims. Quine here offers a clear behaviouristic characterisation of analyticity. This further implies that synonymy and meaning become unproblematic, at least for expressions that are analytically equivalent. It also implies that an inventory of analytic statements can be made up, and that with the help of first order logic semantical rules or meaning postulates can be distilled.

In other words, the section about 'meaning postulates' in Quine's 'Two dogmas of empiricism' becomes unconvincing. Section 4 of 'Two dogmas of empiricism' is a long critical discussion of semantical rules. Quine writes:

> Now the notion of semantical rule is as sensible and meaningful as that of postulate, if conceived in a similarly relative spirit – relative in time, to one or another particular enterprise of schooling unconversant persons in sufficient conditions for truth of statements of some natural or artificial language *L*. But from this point of view no one signalization of a subclass of truths of *L* is intrinsically more a semantical rule than another; and, if 'analytic' means 'true by semantical rules', no one truth of *L* is analytic to the exclusion of another. (1953:34)

We see that Quine argues that the characterisation of analyticity is in the end circular. However, in view of Quine's extreme antifoundationalism this is hardly an objection. Second, he argues that the distinction of analytic and synthetic statements on the basis of semantical rules is arbitrary. But this is hardly a serious objection to Carnap. Already in 1934, in *The Logical Syntax of Language*, Carnap had formulated his principle of tolerance, claiming that there are no morals for setting up linguistic frameworks. The arbirtrariness of the choice of semantic rules, and thus of L-determinate statements is an essential ingredient of Carnap's later philosophy. In the early thirties, Carnap still believed that a formal characterisation of analyticity could be found, but he was soon convinced by Gödel, Tarski, and McLane that the construction was flawed. Quine must have known that Carnap was only interested in language-relative notions of analyticity. Moreover, there is no reason to believe that the verification theory of meaning is still an essential part of Carnap's notion of meaning. Hence, it is hard to see how Quine's critique relates to the position Carnap endorsed in *Meaning and Necessity*. It seemingly took several decades before Quine realised that analyticity, synonymy, meaning, and semantical rules can rather innocuously be grounded in behavioural practice. Quine did have serious arguments against Carnap's various proposals of an analytic-synthetic distinction, and certainly with regard to the distinction between factual and mathematical truths, but it is ironical that Quine's most famous argument is least firmly grounded, and even later to a large degree withdrawn.

In *The Roots of Reference*, some of the critical remarks of 'Two dogmas' still find an echo though. Quine gives the following behaviourist definition of analyticity:

> If the samples first acquired qualify as analytic, still they gain thereby no distinctive status with respect to the language or the community; for each of us will have derived his universal categorical powers from different first samples. Language is social, and analyticity, being truth that is grounded in language, should be social as well. Here then we may at last have a line on a concept of analyticity: a sentence is analytic if *everybody* learns that it is true by learning its words. Analyticity, like observationality, hinges on social uniformity. (1974:79)

The complaint about the complete arbitrariness of choosing meaning postulates is here replaced by the observation that different people learn the domestic language in a different way, so that everyone could have an idiosyncratic notion of analyticity. The list of analytic truths is thus drastically reduced through the requirement that everyone must have learned the truth of an analytic sentence through learning the language. This important qualification notwithstanding, it is noteworthy that already in 1974 Quine gave a precise definition of analyticity by means of which in principle a reduction of meanings was possible.

A further step in Carnap's direction can be taken. If analyticity hinges on social uniformity, it becomes possible to impose the uniformity through social linguistic engineering. For Carnap, this is entirely unproblematic. He was actively engaged in the promotion of artificial languages such as Esperanto. In the Vienna Circle, an artificial pictorial language, ISOTYPE, had been constructed by Otto Neurath and his wife. Carnap's construction of artificial linguistic frameworks in his major semantical works ties in with his engineering approach towards natural language. On this view, it is possible to regard semantical rules not as arbitrary formal postulates, but as social imperatives concerning the use of certain expressions. Social uniformity need not be the result of every individual's learning the language, but may be effectuated through teaching the language in standardised ways. The normative force of schoolbooks, dictionaries, etc. can thus significantly broaden the class of analytic expressions. As a result, language can be transformed and streamlined.

For Quine, however, this line of reasoning is problematic. Quine regularly stresses that the formal frameworks must be interpreted, and often gives the impression that he believes that this is only possible by borrowing their meaning from the natural language in which they are embedded. Both in 'Two dogmas of empiricism' and 'Carnap on logical truth', he demands that the notion of analyticity be clear in the natural language before application of the notion to artificial languages be feasible (1953:36; 1976:127). On the other hand, in the introduction to the chapter on 'regimentation', i.e. the transformation of a scientific theory expressed in natural language into a theory expressed in first order logic, in *Word and Object*, Quine writes:

> Opportunistic departure from ordinary language in a narrow sense is part of ordinary linguistic behavior. Some departures, if the need that prompts them persists, may be adhered to, thus becoming ordinary language in the narrow sense; and herein lies one factor in the evolution of language. Others are reserved for use as needed. (1960:157).

The passage illustrates Quine's ambivalence towards artificial languages and linguistic engineering. Sentences can be meaningful in natural language, but meaning postulates in artificial languages are usually parasitic on the meaningfulness of natural languages, or at best, as the quoted passage illustrates, can become meaningful in the long run. Moreover, dictionaries do not stipulate meanings, but are merely an inventory of a variety of uses:

> Though the word 'meaning' is ubiquitous in lexicography, no capital is made of a relation of sameness of meaning. An entry gets broken down into several "meanings" or "senses," so called, but only ad hoc to explain how to use a word in various dissimilar situations. When a word is partly explained by paraphrasing a sample context, as is so often the way, the paraphrase is meant only for typical circumstances, or for specified ones; there is no thought of sameness of meaning in any theoretical sense. (1995:83)

In conclusion, Quine has two reduction methods for meanings. The first, interpreting empirical meaning as stimulus meaning has limited applicability. In the second method, meanings are regarded as sets of synonymous expressions, whereby synonymy is characterised through analyticity, which is in a behaviourist way explained as true as a result of learning the language. This could be explained as resulting from socially imposed semantical rules, but Quine refrains from taking this last step. It is as if only his aversion of transforming the natural language prevents him from taking it.

Literature

Bergström, L. and Føllesdal, D. 1994 “Interview with Willard Van Orman Quine in November 1993” *Theoria* 40, 193-206.

Carnap, R. 1936 “Testability and Meaning” *Philosophy of Science* 3, 419-471; 4, 1-40.

Decock, L. 2002 *Trading Ontology for Ideology*, Dordrecht: Kluwer.

Goodman, N. and Quine, W.V. 1947 “Steps toward a Constructive Nominalism” *Journal of Symbolic Logic* 12, 105-122.

Quine, W.V.O. 1936 “A Theory of Classes Presupposing No Canons of Type” *PNAS* 40, 320-326.

Quine, W.V.O. 1953 *From a Logical Point of View*, Cambridge MA: Harvard.

Quine, W.V.O. 1960 *Word and Object*, Cambridge MA: MIT Press.

Quine, W.V.O. 1974 *The Roots of Reference*, La Salle: Open Court.

Quine, W.V.O. 1976 *The Ways of Paradox*, Cambridge MA: Harvard.

Quine, W.V.O. 1979 “Cognitive Meaning” *Monist* 62, 129-142.

Quine, W.V.O. 1981 *Theories and Things*, Cambridge MA: Harvard.

Quine, W.V.O. 1992 *Pursuit of Truth*, Cambridge MA: Harvard.

Quine, W.V.O. 1995 *From Stimulus to Science*, Cambridge MA: Harvard.

The Scapegoat Theory of Causality

Marcello di Paola, Rome, Italy

1.

In the Tractatus, Wittgenstein's position was radically anti-factualist. Hume's influence was evident: the cause-effect relation cannot be observed: belief in the causal nexus is superstition.

But Wittgenstein also embraced the Kantian insight: though there are no causal facts, the logical structure of the world/language is causal, i.e. causality is the only form in which our descriptive systems can be conceived. Natural laws, whether they exist or not, are the grammar of our thoughts and language. Causality is the grammar of science.

At the end of the Philosophical Investigations, however, Wittgenstein throws in a totally original viewpoint, questioning the primacy of grammar in general:

> "If the formation of concepts can be explained in reference to natural facts, then, rather than on grammar, should we not perhaps involve ourselves with what, in nature, grounds it?" PI, XII
>
> "Compare a concept with a style of painting. Can we just choose it or not? Are we here simply talking of what's pretty and what's ugly?" PI, XII

Indeed, in On Certainty knowledge would finally be characterized as a decision:

> "We do not learn the praxis of empirical judgement by learning rules; we are taught judgements, and their connections to other judgements. We are presented with the plausibility of a totality of judgements". OC, 140
>
> "My 'state of mind', the "knowing", is for me not a guarantee of what happened. It consists of this: that I would not be able to see where a doubt could arise, where supervision would be possible". OC, 356
>
> "But here, is it then not shown that knowledge resembles a decision?" OC, 362

The roots of this view are to be sought in Cause and Effect. There Wittgenstein does the background work for his final conception of what "knowing" is. Before ramifying into the world, logical structures germinate from the seeds of action. The way we think matches the morphology of the way we act. Action is decision. To know is to judge. To know with certainty is:

> "When a guy says that he will not recognize any experience as evidence for the contrary; this is no doubt a decision". OC, 368

To know is to pass a verdict. This fits with popular characterizations of reason as a tribunal. What reason does is investigating; but, pace Kant and Tractatus, this is not a logical enterprise. The grammar of the world/language evolves from the practical facts of society. We may, for our convenience, invent many alternative natural histories in order to study concepts: but to know with certainty we must decide and elect only one among them, and not doubt our decision thereafter. A concept is like a style of painting, but we do not choose it on aesthetic grounds: it embodies the evolution of social judgements. This interpretation of PI, XII sees reason as a tribunal, and human practice as the jury.

2.

In CE, Wittgenstein rejects Russell's thesis on causality. To explain why we describe the world as causally structured there is no need to postulate any direct intuitions of causal relations: it is enough to point out that certain statements, describing a first event as the cause of a second, are simply never subjected to criticism. The linguistic game of causality does not start with a doubt. However, to consider causal statements as "beyond doubt" does not amount to their being transcendentally grounded (contra Kant); nor (contra Russell) to their being intuitions, as when I am hit with a stick, experience pain, and intuitively know that the blow caused the pain.

The experience of pain is one we may genuinely call "experience of a cause", says Wittgenstein. But not because we are directly and unmistakably made aware of a specific cause. There could be endless possible alternative causes for the pain: while the blow may only have the function of giving me the impression of touch, pain could actually be exploding inside me (a micro-bomb, previously inoculated).

Causal propositions are beyond doubt not because they are solidly grounded on a priori categories or intuitions, but because their being grounded at all is not even in question. I cannot be certain about any specific cause: but I must (I want to) be certain about there being a cause in general. Not to question certain things is a practical methodology.

In CE, Wittgenstein constructs an elegant Gedankexperiment to show how we come to speak of causes:

> two plants, a rose and a poppy. I am led to think that the macro-differences I see between them correspond to micro-differences in their seeds' biological compositions. Different seeds cause different plants: I doubt not that fine-grained genetic inspections would find the seeds to differ in some respect. This is the medieval doctrine that all the "perfections" of the effect already be present in the cause: the "pipeline" conception of causality (Martin, 2008).

Wittgenstein proposes to block the pipeline: suppose the seeds are found to be identical. How to explain the rose and the poppy being two different plants? We would not know what to think, quite literally. Now suppose we finally do find a difference, perhaps at the quark level. Wittgenstein still asks us to prove that such micro-difference is the pertinent one, so that the macro-difference between the two plants does not merely correspond to, but is causally determined by, the micro-one. We cannot be certain of that, and neither Kant nor Russell can help, at this point. We may keep on searching desperately (CE, App.1); or we may simply proclaim a cause.

> "… We also speak of 'tracking the cause': in a simple case we follow, so to speak, the rope, to see

> who's pulling it. When we find such individual – how do we know that it is him, his pulling, the cause of the fact that the rope is moving? Do we establish that through a series of experiments?" CE, p. 15.10

We don't. The main point of the causality issue is that, when something happens we look for (what we call) the cause of it. At the root of the grammar of causality are not scientific facts, logical categories, or direct intuitions. There is action: there are acts of investigation. Investigation is not modelled on science, but vice-versa. The search for causes is a non-scientific, eminently practical activity. We react to the cause, our eyes running from one thing to another:

> "… to call something a 'cause' is like pointing to someone and say "He did it!" CE, 24.9

> "He who follows the rope and finds who's pulling can take a further step, and conclude: so this was the cause, - or rather, is it not the case that all he wanted to find was whether someone was pulling, and who?"CE, 16.10

3.

The practice of scapegoating is anthropologically ubiquitous. The individuation of scapegoats is not an experimental, much less a logical enterprise. The chain between the scapegoat and the misfortune it is said to have caused does not need to be spelled out scientifically. All that matters is that someone did it: if that is the case, then something can be done back.

> "In one case 'he is the cause' simply means: he pulled the rope. In other cases it means something like: these are the facts that I must change in order to eliminate this phenomenon … But how do I get to the idea of changing a circumstance in order to eliminate a phenomenon? ... Yes, it may be said that this presupposes that I am looking for a cause, that from a phenomenon I go look for another".
> CE, p.20

The search for a cause is a human reaction to the social facts of existence. We do not observe causal relations, we do not project causality onto the world, nor do we experience it intuitively. These are chit-chats (CE, 22.10). We proclaim it.

> "… In alternative to what? Certainly to never pull the strings, always remaining uncertain about what really is the cause of the phenomenon; as if it made sense to say: strictly speaking it is impossible to know with certainty, so that what would come closer to the truth would be to leave the question open. This idea is based on a total misunderstanding of the roles that pertain to exactness and doubt"
> CE, 21.10

> "The simple form (and this is the primitive form) of the game of cause and effect is the determination of the cause, not the doubt" CE, 21.10

The primitive form of the causality game is the hunt for a scapegoat, guilty of all bad, even and especially when the trajectory of emergence of such bad is un-reconstruct-able. The proclamation needs not be substantiated scientifically – all is needed is that the general mechanism not be questioned.

In CE, the genealogical argument starts with an inspection of the grammar of doubt: linguistic games in which we doubt (that something is the cause of something else) originate as complications of simpler games, in which there is no doubt.

I now submit that Wittgenstein's position is best made sense of by an evolutionary interpretation.

4.

The evolutionary position has it that some functions of our mind, which philosophers, struck by their pervasiveness, have hypostasized as transcendental categories, or direct intuitions, are indeed specializations that have evolved in response to social situations humans have found themselves in during their history as a species.

Such hypothesis was explored by Cosmides and Tooby (1992), who maintained that problems we find confusing when expressed in naked logical terms become very clear when coated in social ones — we score high at logical inference if the latter refers to contexts of interaction: and those are the contexts faced by our ancestors when establishing patterns of socio-economic connection. Our mind has evolved a specialized capacity to tackle socially significant problems, such as individuating those who defect from covenants.

When confronted with social problems, a specialized mental mechanism moves our eyes from one thing to another. Thousands of years of social negotiation have equipped us with a somewhat automatic drive to look for, and ability to find, who's pulling the rope.

Now, keeping all that in mind, as well as our brief discussion on scapegoating and Wittgenstein's Gedankexperiment, consider the following statement:

> "… If I say: history cannot be the cause of development, that does not mean that I cannot foresee development starting from history, for this is precisely what I do; but it means that I do not call this a 'causal connection', that this is not about predicting the effect from the cause.

> To say: 'There must be a difference in the seeds, even if we cannot find it, plainly displays how powerful it is within us the impulse to see everything through the scheme of cause and effect … 'there must be', that is: we want to use this image in any case". CE, 26.9

Causality in the scientific sense means predicting the effect from its cause. In evolutionary, genealogical, Wittgensteinian sense, it means tracking the cause from its effects. This is the scapegoat theory of causality.

When the group is hit by misfortune, the linguistic game of explanation is enacted in causal terms, with reference to a violation of social trust, which in turn implies a violation of the group's covenant with its natural context, which explains the misfortune. The mysterious cause of nature's operations is thus searched for and individuated within the group. The elimination of the guilty scapegoat is a necessary and sufficient condition for the continuation of social life. But what is important is that the causal chain linking the scapegoat to misfortune actually runs the other way: from misfortune to scapegoat. The cause can only be genealogically reconstructed: before they break the social covenant, community members are, as members, indistinguishable, just like the two seeds in the Gedankexperiment. In both cases, the inability to predict effects is ubiquitous: the grammar of a genuinely causal explanation in the scientific sense has no application.

We may have evolved a specialized capacity to detect defectors from covenants, which has later been adapted by our minds to other kinds of operations, such as scientific investigation. The seed of the causality game is not in the world, in our speculative intellect, or in our intuitions: it is in the realm of social action. Investigation is not modelled on science, but vice-versa.

5.

The scapegoat theory of causality implies, contra Hume, that effects (misfortunes; different plants) are in the past: from past facts we extrapolate causes, and it is thus causes that, properly speaking, follow effects. In his critique, Hume chronologically ordains effects and causes the other way, himself operating a first, and crucial, rationalization, which misleads him into considering causality a theoretical, not a practical, problem.

Kant does not question Hume's formulation. Transcendentalism imputes the pervasiveness of causal extrapolations to a priori, immutable categories of the intellect. Wittgenstein does not abandon the Kantian idea of world-descriptions being only conceivable in causal terms, but he rejects the claim that this is so because there are immutable logical categories underlying the world/language. While Kant sees causality as a universal category of our descriptions, Wittgenstein sees it as a fact about our descriptions, genealogically traceable to the practice of linguistic games more akin to scapegoating than to science. To verbalize such games in cognitive terms conceals their origins as social activities. Pace Kant, causality is not a cognitive lamp with which rational beings illuminate the world. It is an unspoken presupposition that circumscribes the linguistic activity of men within circumstances that are primarily social. Such presupposition is not transcendental: indeed it is not conceptual at all, it is eminently practical (reactive + adaptive).

> "Knowledge is interesting only within a game".
> CE, 18.10

Finally, the directness of Russell's intuition finds no expression in a linguistic game:

> "To 'intuitively recognize the cause' means: to know it in some way (to experience it in a non-usual way) ... Is he not then in a situation no different from that of one who correctly guesses the cause?"
> CE, 18.10

> "We can of course imagine someone saying, in the bliss of inspiration, that he now knows the cause: but that does not preclude us from checking whether he knows the right thing." CE, 18.10

Checking from within the linguistic game of causality we play.

Intuitionism misleads us out of this game. The latter is the not-primarily-scientific one of social adaptation: a way to know causes that has no role within such game is "not interesting". Much more interesting are the words of a medicine man pointing at the scapegoat to explain the mysteries of nature.

We have no intuition of causality as if it existed apart from the use we make of it in linguistic games. The scapegoat theory describes the game of causality as that of finding a cause in any case. This implies an active search for it, accompanied by a non-scientific trust in its existence.

6.

In line with Wittgenstein, an evolutionary interpretation suggests that the use of the causality relation within linguistic games responds to adaptive requirements, primarily social, so strong as to account for both the dimensions of "universality" and "instinctive-ness" that transcendentalism and intuitionism, respectively, wished to capture.

We trust it that there is a cause for every effect, much like primitive groups trust it that there is a scapegoat for every misfortune. The game played is similar, and does not involve "knowing".

> "They tell me that in these circumstances this thing happens. They discovered it by checking a few times ... In the end, I trust those experiences, or their reports, and in conformity with those I orient, unscrupulously, my actions. But this trust, has it not performed well? For all I can see – yes". OC, 603

Logic Must Take Care of Itself

Tamara Dobler, Norwich, East Anglia, England, UK

1. A fundamental tension[1] in Wittgenstein's early conception of logic, which he became aware of at the time he started with *Notebooks 1914-1916*[2], surfaces in the question stated in the second entry: "How is it reconcilable with the task of philosophy, that logic should take care of itself?" (NB, 2). 'The task of philosophy', I take it, refers to the idea of *complete analysis* that is central to both Frege's and Russell's projects.

The following brief reconstruction will outline several basic assumptions that underlie the concept of logically perfect language and logical analysis within Frege's and Russell's frameworks. Firstly, this conception entails a sharp divide between thoughts and expressions of thoughts in language – thoughts are what logic is interested in, not its expressions in everyday language, which is a matter for psychology. Consequently, we have a separation between logical form – which logic is exclusively interested in – and grammatical form[3] – which has no importance for the 'science of logic' except as the source of impurity and confusion ("Instead of following grammar blindly, the logician ought rather to see his task as that of freeing us from the fetters of language" Frege 1997 [1897] 244). Logic deals with propositions – that is, with proper expressions of thoughts – not with sentences of ordinary language. Symbolism or logically perfect language (modelled on the example of maths) should be able, in contrast with the sentences of ordinary language, to present clearly the logical form of our thought ("A language of that sort would be completely analytic, and will show at glance the logical structure of the facts asserted or denied" Russell 1956 [1918], 197-98). Every (assertoric) sentence of our language should be translatable into symbolism – that means that a sentence is subjected to analysis. The idea behind symbolism is "one word and no more for every simple object" as Russell put it, or in Frege's words: "every expression constructed as a proper name… in fact designate an object". A combination of these simple words or names (in a proposition) is assumed to refer to a fact, or a complex made of simple objects ("In a logically perfect language the words in a proposition would correspond one by one with the components of the corresponding fact" Russell 1956 [1918], 197). Analysis is completed when we dissect the proposition so that simple objects that constitute a fact are shown to be clearly represented by simple names that stand for them. This also means that the logical form of a proposition is rendered perspicuous, and that the task of philosophy, as far as the proposition in question is concerned, is fulfilled.

Note that this is a rather oversimplified version of the story. We have to bear in mind that many fine differences become visible if we focus more closely on the relation between Russell's and Frege's conceptions of logical analysis. One conception is given in the *Tractatus* as well. Here I merely sketch how the goal of 'complete analysis' relates to the task of philosophy and the shared basic assumptions of such a goal.

2. Now we need to flesh out a rationale behind the "extremely profound and important insight" that "logic must take care of itself" (NB, 2). The significant portion of Wittgenstein's early philosophy is condensed in the first few entries of the *Notebooks*. The account presented in this section is largely based on these opening passages and on several earlier remarks from *Dictations to Moore* (1914)[4].

Firstly, we are invited to consider the idea of something like the self-sufficiency of (logical) syntax – we must be able to set the rules of syntax by looking at the symbols alone. That is, every mention of the meaning of a sign is an empty move, as it were; it is absolutely unneeded as nothing is being said which was not already *seen* ("If syntactical rules for functions can be set up at all, then the whole theory of things, properties, etc., is superfluous" (NB, 2). This pertains especially to the theory of types: any such theory is superfluous, tautological and senseless – for it tries to do something that is always already done in a more trivial way in our language ("Even if there *were* propositions of [the] form "M is a thing" they would be superfluous (tautologous) because what this tries to say is something which is already *seen* when you see "M"" (DM, 110). What is given by ordinary sentences is enough for us to have a pretty good idea of what makes sense, i.e., that which we *understand* ("It is *obvious* that, e.g., with a subject-predicate proposition, *if* it has any sense at all, you *see* the form, so soon as you *understand* the proposition, in spite of not knowing whether it is true or false" (DM, 110). The same moral is expressed in the thought that whatever is possible is also legitimate ("A *possible* sign must also be capable of signifying" (NB, 2), *viz*. logic that governs the formation of any possible sign in our language makes it legitimate, puts it in traffic. I.e. "Every possible sentence is well-formed" (NB, 2).

Secondly, we are faced with a question of nonsense: it is not as if the signs were responsible for the breakdown of sense – the responsibility is completely on our part ("Let us remember the explanation why "Socrates is Plato" is nonsense. That is, because *we* have not made an arbitrary specification, NOT because a sign is, shall we say, illegitimate in itself!" (NB, 2). We are free to utter whatever gibberish we like, only that does not entail that whatever it is that brings sense to our utterances will automatically lose its significance. Even though Wittgenstein does not spell out at this point what those conditions of sense might be, it is certain, given the quotations above, that every possible linguistic construction is designed legitimately i.e. to make the sense possible.

A month after Wittgenstein wrote the above remarks, he envisaged conditions of sense in terms of the (extended) picture metaphor as the agreement between our thoughts, our language, and how our world is. But that discussion falls out of the scope of this paper. The crux of our examination herein is to point out an important contrast: in Frege/Russell's case, it is ordinary language that is on trial, "for very many of the mistakes that occur in

[1] See the abstract
[2] Hereafter NB
[3] This general view is also highlighted in 4.0031 of the Tractatus

[4] Hereafter DM

reasoning have their source in the logical imperfection of language" (Frege 1997 [1897] 244), whereas, for Wittgenstein, it is *we* who have not "*given* any meaning to certain of its [sentence's] parts. Even when we believe we have done so" (NB, 2).

The first move towards making the picture metaphor applicable across the board has thereby been made: all sentences are to be treated equally in their capacity to display logical features; they are all able to express sense. Unanalysed sentences are not to be treated as logical failures. If there is something for analysis to determine, it is not sentences' logical integrity, for they possess such integrity inherently ("Remember that even an unanalysed subject-predicate proposition is a clear statement of something *quite definite*" (NB 4). If a difference between the analysed and unanalysed form of a sentence is to be made, it should not depend on its capacity to express sense but perhaps on something additional.

3. We can now go back to Wittgenstein's question: "How is it reconcilable with the task of philosophy, that logic should take care of itself?"

Here the idea that logic is already at work in any possible sentence clashes with the task of philosophy conceived in terms of complete analysis. Wittgenstein becomes acutely aware of this tension when he asks himself "Does such a complete analysis exist? *And if not*: then what is the task of philosophy?!!?" (NB, 2) That is, if everything that we need of logic is always already there in our language, is "shown by the existence of subject-predicate SENTENCES", then why should we need analysis *at all*? The question is all the more pressing, for analysis is conceived as *the* task of philosophy, as our real *need* ("Then: if *everything* that needs to be shewn is shewn by the existence of subject-predicate SENTENCES etc., the task of philosophy is different from what I originally supposed" (NB, 3).

With this astonishingly important question, Wittgenstein for the first time touched the heart of the matter – Frege's and Russell's expectations about what philosophy should accomplish, i.e., the ultimate clarity of logical form via the complete analysis of propositions, wherein analysis is taken as a *necessary* route towards such clarity, just did not fit the idea that "logic must take care of itself" whose main features I outlined above.

It is hard to overstate the significance of this acknowledgement. Being stated in the form of a question it also suggests that the deepest difficulties related with what he took as a given from his teachers, in contrast with where his own investigations had brought him by this point, are yet to be met. If everything is already in 'perfect order' in our language, as his new picture of logic implies, then in what sense do we really need analysis? Is analysis a necessary precondition of clarity about the logic of our language such that in the absence of analysis we would not be able to know what we think, or what does and does not make sense?

Wittgenstein was obviously not ready to reach any final verdict at this point, so likewise I could offer merely preliminary suggestions regarding his removal from Frege and Russell. Again, the thing to keep in mind is that the *Tractatus* does contain an account of 'complete analysis' and accordingly we should be wary of being overly or prematurely dismissive with regard to a possible role for analysis in achieving a certain level of perspicuity of the linguistic expressions. Equally, given that Wittgenstein's 'fundamental insight' also appears in the *Tractatus*, it seems plausible to at least wonder if the reasons for having such a need for logical analysis are somewhat different than in Frege/Russell's case as the following passage, for instance, suggests:

> Can't we say: It all depends, not on our dealing with unanalysable subject-predicate sentences, but on the fact that our subject-predicate sentences behave in the same way as such sentences in *every* respect, i.e. that the logic of our subject-predicate sentences is the same as the logic of those. The point for us is simply to complete logic, and our objection-in-chief against unanalysed subject-predicate sentences was that we cannot construct their syntax so long as we do not know their analysis. But must not the logic of an apparent subject-predicate sentence be the same as the logic of an actual one? If a definition giving the proposition the subject-predicate form is possible at all...? (NB, 4)

4. As a result, I suggest that one way to gain a better perspective on the role of the 'picture metaphor' in Wittgenstein's early work is to focus on his urge to reconcile what struck him as two conflicting lines of thought. On the one hand, he was partially committed to the idea that the task of philosophy, as Frege and Russell held, ought to address imperfections of ordinary language by a means of analysis ("a considerable part of what one would have to do to justify the sort of philosophy I wish to advocate would consist in justifying the process of analysis" Russell 1956 [1918], 178), and, on the contrary, he was seriously engaged with the idea that logic always takes care of itself and that ordinary language sentences are perfectly fine as they are.

Hence, the 'need' that the picture metaphor attempted to fulfil could hardly have arisen from the conception that "logic must take care of itself", as this view entails that, in principle, logic does not have needs that a logician is invited to discover and satisfy. It was actually one of 'Russell's needs' i.e. the need to answer the question *when the analysis should be considered complete* that sought the fulfilment (or as Wittgenstein put it "when those signs [signs that behave like signs of the subject-predicate form] are completely analysed?" (NB, 2) In order to account for the problem of 'completeness' in the above mentioned sense, the analysis' advocate needs the 'world' as an ontological excuse. I.e. he needs to assume '*simple objects*' in the world which would, when reached, give him a 'wink' that the analysis is completed and, therefore, the logical form of a sentence rendered clear (logical atoms as "the last residue in the analysis" Russell, 1956 [1918], 178). Secondly, he needs to bridge the world and propositions so that the simple names arrived at in the process of analysis correspond to simple objects.

Note, however, that the need is not to seek answers from the world, but to tune the metaphysics of the world/reality in such a way to serve the 'analyst' with the desired targets ("The demand for simple things *is* the demand for definiteness of sense" NB, 62).

The metaphor of picturing was introduced as an account of the agreement between our sentences/thoughts and pieces of the world that allegedly dictate their analysis.

The trouble is, I fear, that at least *initially* Wittgenstein adhered to 'Russell's need' somewhat dogmatically, and thus the metaphor of picturing, which was to offer the fulfilment, turned out to be dangerously oversimplified. By the time of the *Tractatus*, however, Wittgenstein's thoughts might have already gone in another direction, as the famous proposition 6.54 suggests – only this must stay a topic for some different occasion.

Literature

Frege, Gottlob 1997 [1897] 'Logic', in: M. Beaney (ed.) *The Frege Reader,* Oxford: Blackwell, 227-250

Russell, Bertrand 1956 [1918] 'The Philosophy of Logical Atomism', in: *Logic and Knowledge*, London: Allen and Unwin, 177-281

Wittgenstein, Ludwig 1998 *Notebooks 1914-1916*, Oxford: Blackwell Publishers

Wittgenstein, Ludwig 1998 [1914] Notes Dictated to G. E. Moore in Norway in: Notebooks 1914-1916 Appendix II, Oxford: Blackwell Publishers

Wittgenstein on Frazer and Explanation

Keith Dromm, Natchitoches, Lousiana, USA

In his “Remarks on Frazer’s *Golden Bough*,” Wittgenstein identifies at least two problems with Frazer’s explanations for religious and magical practices. First, Frazer’s explanations are implausible. Frazer regards them as nascent forms of contemporary science that reflect “faulty views” about physics, medicine, or technology (Wittgenstein 1993, p. 129). According to Wittgenstein, this is to treat these practices as “pieces of stupidity”: “But it will never be plausible to say that mankind does all that out of sheer stupidity” (Wittgenstein 1993, p. 119). Wittgenstein’s second criticism would seem to have priority. He writes: “the very idea of wanting to explain a practice . . . seems wrong to me” (Wittgenstein 1993, p. 119). However, some commentators have focused on the first criticism, and they find in Wittgenstein’s remarks a more plausible account of religious and magical practices. Rather than the antecedents of contemporary science or technology, the practices examined by Frazer are elaborations on either expressive or instinctive behaviors. As expressive behaviors, magical practices, for example, do not attempt to effect some change in the natural world; they are expressions of wishes, desires, or other attitudes toward the world. Wittgenstein seems to be suggesting this view of magic when he writes that “magic brings a wish to representation; it expresses a wish” (Wittgenstein 1993, p. 125; see, e.g., Hacker 1992, p. 286).[1] Other commentators have focused more on Wittgenstein’s references to instinctive behavior within these remarks (e.g., Clack 1999; De Lara 2003). For example, Wittgenstein refers to “Instinct-actions” within an observation about the non-instrumental character of ritualistic actions (Wittgenstein 1993, p. 137). Elsewhere, he associates a ritual with an instinctive behavior (Wittgenstein 1993, p. 141). Wittgenstein seems to be suggesting in these places biological origins for religious and magical practices. Some supporters of the instinct reading have vigorously opposed the expressivist reading (e.g., Clack 1999 and 2003). However, both readings agree that, according to Wittgenstein, ritualistic actions are performed without regard to their utility. As such, they are misleadingly compared to modern technology or medicine. These readings also take Wittgenstein to be opposed to the view that these practices are manifestations of a primitive science, since—as Wittgenstein insists in several places—they should be not characterized in terms of the beliefs of their participants. He writes: “the characteristic feature of ritualistic action is not at all a view, an opinion” (Wittgenstein 1993, p. 129; see also p. 123 and 129). As such, they do not represent beliefs, whether true or false, about nature.

According to these interpretations, Wittgenstein’s second criticism of Frazer amounts to the claim that the kind of explanation that Frazer offers is not appropriate for these practices. Since magical and religious practices are not based on beliefs about the world or anything else, they should not be explained in terms of their participants’ beliefs. However, this is still to attribute to Wittgenstein an explanation for these practices. The explanation is importantly different than the one Frazer offers; we can characterize it as a causal explanation as opposed to Frazer’s intellectualist explanation. The causes that Wittgenstein is supposed to have identified for these practices preclude the interpolation of participants’ beliefs in an explanation for their performance. The practices arise naturally out of certain instinctive or expressive behaviors of humans without the mediation of beliefs. But Wittgenstein’s second criticism does not challenge the *type* of explanation that Frazer offers for these practices. Again, Wittgenstein says that there is something wrong with the “very idea of wanting to explain a practice.” If we are to reconcile these two criticisms, some other purpose for Wittgenstein’s discussions of expressive and instinctive behaviors needs to be found. This purpose must be something other than explaining religious and magical practices. Identifying this purpose will be my task in what follows.

P. M. S. Hacker offers some correct advice in dealing with Wittgenstein’s remarks on Frazer: “If one wants to learn from them, they should not be squeezed too hard” (Hacker 1992, p. 278). They were only slightly revised after their initial composition. Only the first part of them (MS 110) was preserved in a transcript (TS 221), and those remarks were subsequently dropped from a later version of that transcript (TS 213). The second part of the remarks comes from scraps of paper that were probably inserted by Wittgenstein into his copy of the abridged version of *The Golden Bough* (MS 143).[2] But while the remarks were not worked over like those collected in the *Philosophical Investigations*, they deserve some attention. They are about a book in which Wittgenstein had a serious interest (Drury 1981, pp. 134-5) and, if read properly, they can illuminate not only their subject but other areas of Wittgenstein’s thought. The best strategy for approaching them is to read them in light of the more reliable records of Wittgenstein’s thought. This strategy will warn us away from taking Wittgenstein to be offering in them his own explanation for religious and magical practices.

Wittgenstein famously asserts in the *Philosophical Investigations* that in philosophy “We must do away with all *explanation*, and description alone must take its place” (Wittgenstein 2001, §109). Explanations cannot remedy the confusions that generate philosophical problems. Instead of the novel information that an explanation provides, we require a better understanding of language or other practices in order to be relieved of our confusions. Wittgenstein’s second criticism of Frazer seems to extend this admonition to our efforts to understand ancient and otherwise unfamiliar practices. But how can mere descriptions improve our understanding of alien practices? This depends on the type of deficiency in our understanding that we are trying to rectify. Wittgenstein understands Frazer’s central problem to be the strangeness and unfamiliarity of certain religious and magical practices. Frazer is attempting to make sense of these practices. So, his question is less about *where* they came from, and more about *why* they are performed. The former can be answered without answering the latter. And

[1] While Hacker (1992) seems to endorse, at least in part, the expressivist interpretation, his understanding of Wittgenstein’s use of “perspicuous representations” and developmental hypotheses in his remarks on Frazer is very close to mine. Paul Redding (1987) also provides a similar interpretation.

[2] See the editors’ introduction to the “Remarks on Frazer’s Golden Bough” for more information on their sources (pp. 115-117).

whereas the former question can be answered by uncovering new facts about the practices, the latter question requires a different kind of solution.

In attempting to explain these practices, by either revealing the beliefs of their practitioners or fitting them within a developmental hypothesis (a method of Frazer's that we will consider later), Frazer is succumbing to what Wittgenstein calls in these remarks the "the foolish superstition of our time" (Wittgenstein 1993, p. 129), which is to believe that every puzzle can be remedied by a scientific explanation. In one of his transcripts, Wittgenstein identifies this as the "scientific way of thinking" and says:

> What is disastrous about the scientific way of thinking (which today possesses the whole world) is that it wants to respond to any disquiet with an explanation. (TS 219, p. 8; author's translation)

The disquiet that Frazer suffers from, that which motivates him to seek an explanation for these practices, is caused by their strangeness and unfamiliarity. However, this cannot be remedied through an explanation. Instead, Wittgenstein says in these remarks, in a variation on his advice to philosophers, that "one can only *describe* and say: this is what human life is like" (Wittgenstein 1993, p. 121).

Wittgenstein uses a concept that plays an important role in his discussions of the treatment of philosophical problems to characterize the sort of description that can provide the desired understanding: "perspicuous representation" (Wittgenstein 1993, p. 133). Such a representation will help us see that "there is also something in us which speaks in favor of those savages' behaviour" (Wittgenstein 1993, p. 131). He provides an example of this in a passage that has been used to support both the expressivist and instinctive interpretations of his "Remarks on Frazer":

> When I am furious about something, I sometimes beat the ground or a tree with my walking stick. But I certainly do not believe that the ground is to blame or that my beating can help anything. "I am venting my anger". And all rites are of this kind. Such actions may be called Instinct-actions.—And an historical explanation, say, that I or any ancestors previously believe that beating the ground does help is shadow-boxing, for it is a superfluous assumption that explains *nothing*. The similarity of the action to an act of punishment is important, but nothing more than this similarity can be asserted.
>
> Once such a phenomenon is brought into connection with an instinct which I myself possess, this is precisely the explanation wished for; that is, the explanation which resolves the particular difficulty. And a further investigation about the history of my instinct moves on another track.
> (Wittgenstein 1993, p. 137)

A description alone can reveal such a connection between an opaque practice and something I do. In doing this, it would satisfy Wittgenstein's criterion for a perspicuous representation:

> This perspicuous representation brings about the understanding which consists precisely in the fact that we "see the connections." Hence the importance of finding *connecting links*.
> (Wittgenstein 1993, p. 133)

That Wittgenstein puts the connection in terms of a shared "instinct" should not be taken as a commitment by him to some biological account of the origins of ritualistic practices. Such an account, as well as any version of the expressivist theory, would be as incapable as Frazer's explanations of making an alien practice seem less strange. Wittgenstein also says that an investigation of the instinct's history "moves on another track," suggesting that an exact characterization of it is irrelevant to the purposes served by its identification.

Instead of revealing the emotional or biological roots of ritualistic actions, Wittgenstein is drawing our attention to what he elsewhere calls the "common spirit" that underlies the practices being compared:

> All these *different* practices show that it is not a question of the derivation of one from the other, but of a common spirit. And one could invent (devise) all these ceremonies oneself. And precisely that spirit from which one invented them would be their common spirit. (Wittgenstein 1993, p. 151)

It is only by recognizing the "common spirit" in which a practice is performed that it can be relieved of its strangeness. The recognition is not a matter of knowing certain facts about the practice, facts that an explanation can provide. Rather, it involves being able to occupy imaginatively the place of a participant in the other practice. Our ability to do this can be facilitated by a description of the practice that highlights a "common spirit" or "connecting link" between the alien practice and one in which we are already a participant. A description that is able to do this will provide the "satisfaction," as Wittgenstein puts it, that Frazer sought through his explanations:

> I believe that the attempt to explain is already therefore wrong, because one must only correctly piece together what one *knows*, without adding anything, and the satisfaction being sought through the explanation follows of itself.
> (Wittgenstein 1993, p. 121)

If we fail to recognize the "common spirit" in which the practices are performed, then no amount of new information provided by an explanation will make the alien practice any less opaque.

Wittgenstein does admit a role for explanations in facilitating our understanding of alien practices. However, in serving this role they are importantly different than the explanations that Frazer offers (as well as those sometimes attributed to Wittgenstein). For example, in order to account for the sinister quality a contemporary spectator would discern in the Beltane Fire Festival, Frazer offers a developmental hypothesis for the ritual that locates its origins in human sacrifice. But this explanation's ability to increase our understanding of the practice does not depend upon the explanation's truth. As Wittgenstein explains:

> The deep, the sinister, do not depend on the history of the practice having been like this, for perhaps it was not like this at all; nor on the fact that it was perhaps or probably like this, but rather on that which gives me grounds for assuming this.
> (Wittgenstein 1993, p. 147)

The explanation can function as a "perspicuous representation" of the practice that is able to highlight those features of it by which we can, as Wittgenstein puts it, discern its "connection with our own feelings and thoughts" (Wittgenstein 1993, p. 143). In order to do this, the hypothesis about the practice's origins need not be true (it need not even be supposed to be true); it only needs to draw our attention to those aspects of the practice that are shared by ones in which we participate.

This is also the case with the developmental hypotheses identified in these remarks by the expressivist and instinctive interpretations. The purpose of these hypotheses is not to inform us about the origins of religious and magical practices, but to facilitate our understanding of these practices. This is the same function served by other hypotheses we find in Wittgenstein's writings, such as those that associate the development and acquisition of language with instinctive or "primitive" reactions (e.g., Wittgenstein 2001, §244). For Wittgenstein's purposes in these writings, the truth of these hypotheses is irrelevant. Instead, as he puts it, "the correct and interesting thing to say is not: this has arisen from that, but: it could have arisen this way" (Wittgenstein 1993, p. 153). While their truth certainly makes a difference in other contexts, it does not make a difference to Wittgenstein's efforts to relieve us of certain confusions.

Literature

Clack, Brian 1999 *Wittgenstein, Frazer and Religion*, New York: St. Martin's Press.

Clack, Brian 2003 "Response to Phillips", *Religious Studies* 38, 203-209.

Drury, M. O'C. 1981 "Conversations with Wittgenstein", in: Rush Rhees (ed.), *Personal Recollections*, Oxford: Blackwell.

De Lara, Philippe 2003 "Wittgenstein as Anthropologist: The Concept of Ritual Instinct", *Philosophical Investigations* 26, 109-124.

Hacker, P.M.S. 1992 "Developmental Hypotheses and Perspicuous Representations: Wittgenstein on Frazer's *Golden Bough*", *Iyyun: The Jerusalem Philosophical Quarterly* 41, 277-299.

Redding, Paul 1987 "Anthropology as Ritual: Wittgenstein's reading of Frazer's *The Golden Bough*", *Metaphilosophy* 18, 253-269.

Wittgenstein, Ludwig 1993 "Remarks on Frazer's *Golden Bough*", in: James Klagge and Alfred Nordmann (eds.), *Philosophical Occasions: 1912-1951*, Indianapolis: Hackett.

Wittgenstein, Ludwig 2001 *Philosophical Investigations*, Oxford: Blackwell.

References to Wittgenstein's unpublished writings follow von Wright's catalogue. The typescript (TS) or manuscript (MS) number is followed by page number(s).

Dummett on the Origins of Analytical Philosophy

George Duke, Melbourne, Australia

Introduction

Michael Dummett's claim that 'the fundamental axiom of analytical philosophy [is] that the only route to the analysis of thought goes through the analysis of language' (1993, 128) has been criticized on the grounds that it excludes seminal figures in the analytical tradition such as GE Moore and Bertrand Russell (for example in Monk and Palmer, 1996). In this paper I begin by suggesting that Dummett's characterization has some validity if restricted to what Alberto Coffa (1991) has called 'the semantic tradition' (that part of the analytical tradition represented by figures such as Frege, the Russell of 'On Denoting', the early Wittgenstein, Carnap, Tarski and Quine), in which the role played by logical analysis based on mathematical techniques is central. The restricted applicability of Dummett's characterization, even when suitably qualified in this way, is instructive because it allows for a clearer view of the extent to which it is possible and/or meaningful to characterize the analytical tradition as a whole and its relation to what Dummett calls 'other schools' (1993, 4).

1. The Linguistic Turn

Stated without further qualification, Dummett's characterization of analytical philosophy raises obvious objections. It is simply not the case that the seminal thinkers of the analytical tradition form a united front around the notion that a philosophical account of thought can only be achieved through a philosophical account of language. Apart from the examples of Moore and Russell already mentioned, Frege is equally problematic, on account not only of his lifelong ambivalent attitude towards imprecise natural language but also his 'realist' view that thoughts unthought by a thinker are still true or false (53, 1900).

Dummett's characterization has the virtue, from his own perspective, of bringing together those components of the thought of Frege and late Wittgenstein to which he is particularly sympathetic. It is hard not to think, however, that he has been led astray by his almost exclusive concern upon the historical relations between Frege and Husserl in *Origins of Analytical Philosophy*, which, given Husserl's commitment to a phenomenology of pure consciousness, could lead to the conclusion that the linguistic turn is distinctive of the analytical school as against other philosophical approaches.[1]

While no one would deny the centrality of linguistic considerations to the analytical tradition, Dummett's formulation is too rough-grained to offer any meaningful characterization of a particular tradition. A better approach would be to focus on the origins of what Alberto Coffa has called 'the semantic tradition', a tradition which includes many of the major thinkers of analytical philosophy. What unifies these figures, however, is not so much an emphasis upon linguistic meaning and rejection of intuition (Russell and Quine are counter-examples to this thesis), as a belief in the capacity of logical analysis to illuminate traditional philosophical problems.

2. Frege's new logic

When we read Dummett's characterization of analytical philosophy in the context of his views on Frege's place in the history of ideas it in fact accords with the privileged place of logical analysis. According to Dummett, 'only with Frege was the proper object of philosophy finally established' (1975, 458). This involves the thesis, 'first, that the goal of philosophy is the analysis of the structure of thought; secondly, that the study of thought is to be sharply distinguished from the study of the psychological process of thinking, and, finally, that the only proper method for analysing thought consists in the analysis of language' (1975a, 458).

For Dummett, therefore, Frege began a revolution in philosophy as overwhelming as that of Descartes (1973, 665-666 and 1975, 437-458). Whereas the Cartesian revolution consisted in giving the theory of knowledge priority over all other areas of philosophy, Frege's primary significance consists in the fact that he made logic the starting point for the whole subject (1973, 666). Dummett here means logic in the broad sense of a theory of meaning or the search for a model for what the understanding of an expression consists in (1973, 669). The thought is that Frege inaugurated an epoch in which 'the theory of meaning is the only part of philosophy whose results do not depend upon those of any part, but which underlies all the rest' (1973, 669).

In appealing to the linguistic turn as decisive for analytical philosophy, Dummett therefore points towards the introduction of semantic considerations that he takes to be embodied in Frege's employment of the context principle in *Die Grundlagen der Arithmetik* (1884). Faced with the Kantian question concerning how it is possible to be given numbers, when we do not have representations or intuitions of them (1993, 5), Frege, Dummett alleges, converted 'an epistemological problem, with ontological overtones' into one about 'the meaning of sentences' (1991, 111).

It is Frege's new predicate logic introduced in *Begriffsschrift*, based on the extension of function-argument analysis from mathematics to logic, which provides the technical means to carry out this strategy. In *The Logical Basis of Metaphysics*, Dummett argues that while the philosophy of thought has always in a sense been regarded as the starting point of the subject 'where modern analytical philosophy differs is that it is founded on a far more penetrating analysis of the general structure of our thoughts than was ever available in past ages, that which lies at the base of modern mathematical logic and was initiated by Frege in 1879' (1991, 2).

Dummett's defence of analytical philosophy against 'the objections of laymen', who lament the abandonment of 'fundamental' questions for technical investigations, sets out from the fact that the analysis of inference carried out in modern logic presupposes an analysis of the structure of propositions. From this point of view, one could see why

[1] Moreover, for leading representatives of the European tradition after Husserl, such as Gadamer and Derrida, linguistic considerations are central. While these thinkers were not concerned with giving an account of thought in the apposite sense, and their approach to language is based on hermeneutic and semiotic considerations respectively rather than semantics and logic, this raises more questions as to the adequacy of Dummett's attempt to distinguish the two dominant philosophical schools of the twentieth century.

an adequate syntactic analysis of our language has priority in philosophical explanation. If we grant the further thesis that Frege's new language of quantifiers and variables represents the most perspicuous means of representing natural language, we can apparently in good conscience justify the privileged role of logical analysis in analytic philosophy.

To privilege the role of Frege's predicate logic is not to understate the importance for the semantic tradition of either the attack on psychologism, which Dummett calls 'the extrusion of thoughts from the mind', or the context principle. This is because these two tenets of analytical philosophy in its classical phase are coeval with the introduction of Frege's new logical symbolism. Frege's notions of concept and object are correlative to the symbolic notions of function and argument; by taking concept as a function of an argument, we can understand the process of concept formation without appeal to extraneous psychological considerations. And the context principle is, as Frege states explicitly, inspired by the rigorisation of the calculus, whereby infinitesimals are banished through an explanation of the meaning of 'contexts' containing expressions such as df(x) or dx rather than seeking to explain them in isolation.

It is generally acknowledged that the introduction of quantifier notation and bound variables was the single most important advance in logic since Aristotle. Frege's way of parsing sentences involving quantifiers offers a tremendous increase in expressive power insofar as it can adequately represent the statements of multiple generality that had troubled traditional syllogistic. Although the significance of Frege's revolution in logic is well-known, however, the original intention informing his development of his new conceptual notation is easily understated in the contemporary context. Dummett's statement that 'the original task which Frege set himself to accomplish, at the outset of his career, was to bring to mathematics the means to achieve absolute rigor in the process of proof' (1973) is obviously accurate, but, informed by an awareness of the incompleteness of second-order proof procedures, also understates the extent of Frege's ambition.

An historically unprejudiced reading of the preface to *Begriffsschrift* cannot avoid the conclusion that Frege conceived of his new formula language as a vital contribution to the realization of the Enlightenment project of a *mathesis universalis*, a universal methodical procedure capable of providing answers to all possible problems. While conceding the slow advance in the development of formalized languages, he notes recent successes in the particular sciences of arithmetic, geometry and chemistry (1879, XI), and also suggests that his own symbolism represents a particularly significant step forward insofar as logic has a central place with respect to all other symbolic languages and can be used to fill in the gaps in their existing proof procedures (1879, XII). On account of its seemingly limitless generality, the new predicate calculus, with its expressive power to represent functions and relations of higher level, is conceived by Frege as the most significant advance yet made on the way towards Leibniz's grandiose goal of a universal characteristic.

3. Transformative Analysis and Semantic Logicism

Recent work by Michael Beaney (2007) and Robert Brandom (2006) further clarifies the distinctive philosophical perspective of the semantic tradition. Brandom's characterization of the notion of 'semantic logicism' is particularly revealing, in that it provides a way of bringing together philosophers for whom logical analysis of language and meaning is the core concern and naturalistic and empiricist approaches which are less easily accommodated by Dummett's fundamental axiom.

Beaney explicates three conceptions of analysis in the Western philosophical tradition, claiming that the third of these - transformative analysis - is characteristic of analytical philosophy in its classical phase as embodied by Frege, Russell, the early Wittgenstein and Carnap. The first form of analysis is the decompositional - the breaking of a concept down into its more simple parts. The decompositional approach is prevalent in early modern philosophy and encapsulated in Descartes' 13th rule for the direction of the mind that if we are to understand a problem we must abstract from it every superfluous conception and by means of enumeration, divide it up into its smallest possible parts. The second kind of analysis is regressive analysis, according to which one works back towards first principles by means of which something can be demonstrated. This conception is predominant in classical Greek thought, for example in Euclidean geometry. Transformative analysis works on the assumption that statements need to be translated into their 'correct' logical form before decomposition and regression can take place. Classic examples are Frege's attempt to reduce mathematics to logic and Russell's theory of definite descriptions. The epistemological and ontological explanatory power of Frege's predicate logic would thus appear to be the major assumption of analytical philosophy in its classical phase.

Robert Brandom introduces the notion of 'semantic logicism' to characterize 'classical' analytical philosophy. According to Brandom, analytical philosophy in its classical phase is concerned with the relations between vocabularies – 'its characteristic form of question is whether and in what way one can make sense of the meanings expressed by *one* kind of locution in terms of the meanings expressed by *another* kind of locution' (2006, 1). So, what is distinctive of analytical philosophy is that '*logical* vocabulary is accorded a privileged role' (2006, 2) in specifying semantic relations that are thought to make the true epistemological and ontological commitments of the former explicit.

In explicating the classical project of analysis as 'semantic logicism', Brandom notes that it involves, to employ Dummettian phraseology, the translation of epistemological and ontological questions into a semantic key. Brandom describes how two core programs of classical analytical philosophy, empiricism and naturalism, were transformed in the twentieth century 'by the application of the newly available logical vocabulary to the self-consciously semantic programs they then became' (2006, 2). The generic challenge posed by such projects is to demonstrate how target vocabularies, for example, statements about the external world, can be reconstructed from 'what is expressed by the base vocabulary when it is elaborated by the use of logical vocabulary' (2006, 3).

Brandom's characterization of semantic logicism is more inclusive than Dummett's fundamental axiom, but nonetheless does not completely cover the range of philosophers who would commonly be considered analytic. Apart from thinkers like Moore and Ryle, to whom it does not seem strictly applicable, more recent analytical thinkers have in fact placed the basic thesis of semantic logicism in question.

Brandom suggests that the main challenge to analytical philosophy in its classical phase came from Wittgenstein's rejection of the assumption that, following a codification of the meanings expressed by one vocabulary, through the use of logical vocabulary, into that of another vocabulary, we can derive properties of use. Emphasising the dynamic character of linguistic practice, Wittgenstein rejects the assumption of classical semantic analysis that vocabularies are stable entities with fixed meanings, replacing this model with a piecemeal account of the uses to which language is put in various language games.

From this perspective, if we accept that semantic logicism is in some way characteristic of analytical philosophy in it classical phase, the pragmatist challenge of Wittgenstein and subsequent thinkers such as Rorty, is best viewed as a response to the original assumptions of the semantic tradition based on a realization of the limits of the application of mathematical techniques to natural language and everyday experience. As has often been noted, these responses in fact share much in common with the thought of major twentieth century continental thinkers, such as Heidegger and Gadamer. The fact that many dominant programs in contemporary analytical philosophy, such as contextualism, no longer have unmitigated faith in the program of logical analysis is also a recognition of the limits of the original aspirations of logical analysis.

As Michael Friedmann has suggested, the Carnap-Heidegger debate is highly instructive here, in that it highlights two radically different philosophical attitudes not only to logic and mathematics but also to the modern natural science built upon their edifice. This explains why the work of thinkers like Davidson, McDowell and Brandom, who have sought to explicate the logical space of reasons and reintroduced hermeneutic considerations, is accurately thought to represent a rapprochement between divergent traditions.

4. Conclusion

In this paper I have argued that Dummett's fundamental axiom of analytical philosophy is inadequate not only because of what it excludes, but also insofar as it risks understating the role of logical analysis for that part of the tradition which he himself privileges. While representative of his own commitment to a position which reconciles semantic logicism with the dictum that meaning is use, Dummett's axiom is at risk of covering over both the true origins of analytical philosophy in its classical phase and the extent to which its original project has been placed in question.

To provide a more complete characterization of analytical philosophy and its relation to 'other schools' one would need to spell out the relation between 'instrumental' and 'reflective' rationality. Arguably, the failure of 'other schools' in the twentieth century, with some notable exceptions, was precisely their inability to present an adequate account of an alternative account of rationality to the instrumental i.e. their critique of instrumental rationality was indiscriminate in the sense that it was often prosecuted against rationality *per se*. This is why the recent 'hermeneutic' turn in analytical philosophy represents a more significant development than the earlier 'pragmatist challenge'.

Literature

Beaney, M. 2007. 'Analysis' (http://plato.stanford.edu/entries/analysis/index.html) in The Stanford Encyclopedia of Philosophy.

Brandom, R. 2006. The 2005-2006 John Locke Lectures. Between Saying and Doing: Towards an Analytic Pragmatism. Trinity Term 2006: Oxford University.

Coffa, A. 1991. *The Semantic Tradition from Kant to Carnap*. Cambridge: Cambridge University Press.

Dummett, M. 1973a. *Frege: Philosophy of Language*. London: Duckworth.

Dummett, M. 1975. 'Can Analytical Philosophy be Systematic and Ought it to Be?' in Dummett, M. 1978. *Truth and Other Enigmas*. Cambridge, Massachusetts: Harvard University Press.

Dummett, M. 1991. *The Logical Basis of Metaphysics*. Cambridge Massachusetts: Harvard University Press.

Dummett, M. 1993. *The Origins of Analytical Philosophy*. Cambridge Massachusetts: Harvard University Press.

Frege, G. 1879. *Begriffsschrift*. 1998. Hildesheim: Georg Olms Verlag.

Frege, G. 1884. *Die Grundlagen der Arithmetik*. 1988. Hamburg: Felix Meiner Verlag.

Frege, G. 1900. ‚Der Gedanke' in *Logische Untersuchungen*. 1976. Patzig, G (ed.). Göttingen: Vandenhoeck & Ruprecht.

Monk, R & Palmer, A. 1996. Bertrand Russell and the Origins of Analytical Philosophy. Bristol: Thoemmes Press.

Wittgenstein meets ÖGS: Wovon man nicht gebärden kann …

Harald Edelbauer/Raphaela Edelbauer, Hinterbrühl, Österreich

0. Das Projekt

Die rezente Studie *Sprache Macht Wissen*, kommt zu dem Ergebnis, „daß das Bildungswesen in Österreich für gehörlose/hörbehinderte SchülerInnen und Studierende reformbedürftig ist, und chancengleiche Bildungsmöglichkeiten für diese Personengruppe nicht immer gegeben sind.“ (Krausnecker/Schalber 2007)

Hier hakt unser Projekt Evaluierung von Wittgensteins Sprachphilosophie(n) anhand der Gebärdensprache ein. Formal zielt es auf die Übertragung des Tractatus logico-philosophicus sowie der Philosophischen Untersuchungen in die Österreichische Gebärdensprache (ÖGS) ab; im zweiten Schritt soll die Übersetzung dieser Werke in die ‚alphabetische’ Gebärdenschrift, wie sie C. Papaspyrou entworfen hat, erprobt werden.

Ziel des Projekts ist einerseits eine Hilfestellung für Gebärdendolmetscher, die hinter dem Katheder philosophische Inhalte an gehörlose Studierende vermitteln sollen; zum andern die Einübung semantischer Kompetenz auf höherem Niveau für Mitglieder der Gebärden-Sprachgemeinschaft. Es soll überprüft werden, inwiefern auch in Wittgensteins Konzepten noch implizite sonozentrische Annahmen stecken.

„Deshalb ist der Vorgang der Übersetzung fast noch wichtiger als ihr Ergebnis“, erläutert der organisatorische Leiter des Projekts, Thomas Nagy.

Jede Übersetzungseinheit, an der gehörlose Student(inn)en und hörende Dolmetscher(innen) mitwirken, wird filmisch dokumentiert. Geplant ist darüber hinaus die anschließende Fixierung der – im Konsens vorläufig akzeptierten - Gebärden mittels einer neuen Notation. (Papaspyrou 1990)

1. Expedition in semantisches Neuland

> Definitionen sind Regeln der Übersetzung von einer Sprache in eine andere. Jede richtige Zeichensprache muß sich in jede andere nach solchen Regeln übersetzen lassen. *Dies* ist, was sie alle gemeinsam haben. (Wittgenstein 1984)

Dieser Satz des *Tractatus* - 3.343 – enthält quasi Wittgensteins frühe Sprachkonzeption ‚in a nutshell’; er birgt sogar in nuce den *Grundgedanken*, daß die logischen Konstanten nicht vertreten.

Auch wenn wir – wie ihr Autor selbst – die Feststellung *3.343* nicht mehr unterschreiben würden, bleibt das Problem der Übersetzung für die philosophische Semantik grundlegend.

Gerade die Übertragung der beiden ‚Zentralwerke’ Wittgensteins – des *Tractatus* (im folgenden ‚T’) sowie der *Philosophischen Untersuchungen* (im folgenden ‚PU’) – offenbart eine faszinierende Selbstreferenz: eben jene philosophisch-semantischen Probleme, von denen der zu übersetzende Text handelt, tauchen in unerwarteter Brisanz *als Probleme der Übersetzung* wieder auf.

Und dieses ‚thematische Feedback’ nimmt enorm zu, wenn die Zielsprache aus Gebärden anstatt aus Lauten besteht; denn, wie schon Wilhelm Wundt am Anfang des 20. Jahrhunderts erkannte, tun sich hier kategoriale Abgründe auf:

> Wie sehr man dabei meist noch geneigt blieb, einfach die der Lautsprache entnommenen Kategorien auf die Gebärden zu übertragen, dafür bildet freilich die noch heute vollständigste Sammlung von Zeichen dieser Art einen Beleg. Sie unterscheidet die Gebärden lediglich in Symbole für Hauptwörter, Eigenschaftswörter und Zeitwörter, ohne darauf Rücksicht zu nehmen, daß diese grammatischen Kategorien in der Form, in der sie die Lautsprache besitzt, für die Gebärde überhaupt nicht existieren. (Wundt 1911)

Diese kategoriale Inkompatibilität – der Wittgensteins Hypothese (‚Definitionen als Regeln der Übersetzung’) nicht standhält – verdankt sich vor allem den völlig distinkten Kommunikationskanälen. Chrissostomos Papaspyrou, ein selbst gehörloser Linguist, unterscheidet hier zwei Sprachfamilien verschiedener *Substanz*:

> Jede natürliche Sprache weist bekanntlich sowohl eine Form, als auch eine Substanz auf, die als materieller Träger der Form dient. … Die Substanz hat unmittelbare Beziehung zu der Aktualisierungsmodalität, bei der sich eine natürliche Sprache auf physiologische Art und Weise manifestiert. … Jedoch, wie ein Blick in die relevante Literatur zeigt, ist die Substanz als Vergleichsfaktor nicht berücksichtigt worden. Die Annahme, daß alle menschlichen natürlichen Sprachen Lautsprachen sind, und somit eine gemeinsame Substanz besitzen, klammerte diese Möglichkeit von vornherein aus. … Doch es gibt die Gebärdensprachen, die die Gültigkeit der oben erwähnten Annahme offensichtlich aufheben. Als visuell-manuelle Zeichensysteme bieten die Gebärdensprachen, anders als die unterschiedlichen Lautsprachen, eine noch tiefer ausgeprägte Kontrastierung: die Kontrastierung der Aktualitätsmodalitäten zueinander. (Papaspyrou 1990)

Um zu prüfen, ob und wieweit Wittgensteins Sprachkonzepte den Übergang von einer ‚Substanz’ zur anderen heil überstehen, haben wir den empirischen Weg gewählt, den Versuch einer Übersetzung von *T* und *PU* in die – in Österreich unter Gehörlosen gebräuchliche – Gebärdensprache: Wittgenstein meets ÖGS.

2. Wittgenstein – der Maßstab auf dem Prüfstand

Warum Wittgenstein? Weil er für uns noch immer die *maßgebliche* Instanz der Philosophie der idealen *und* der normalen Sprache bleibt. Seit er, im T, Bedeutung als Bild und später, in den PU, Bedeutung als Gebrauch charakterisiert hat, ist bis jetzt nichts Neues an vergleichbarer Kraft und Tiefe hinzugekommen.

Daß manche bedeutende Werke der philosophischen Literatur ihrer Ursprungs-Sprache unablösbar eingeschrieben bleiben, ist ein bekanntes und oft diskutiertes Faktum; für manche auch ein Ärgernis. Heideggers Ontologie läßt sich ebensowenig gänzlich vom Deutschen lösen, wie Sartres Ontologie vom Französischen. Man mag zu Quines *Unterbestimmtheit der Übersetzung* stehen, wie man will; sie gilt zumindest für den Großteil der philosophischen Klassiker. Wer sie sozusagen persönlich kennenlernen will – ohne kompromißbelastete Übertragung - muß die Sprache lesen können, in welcher sie verfaßt sind.

Innerhalb der Sprachphilosophie wird die feste Bindung eines Systems an ein bestimmtes Idiom weit weniger tolerabel. Dort, wo es um das Verhältnis von Sprache schlechthin zur Wirklichkeit geht, muß ein Gedankengebäude auch auf fremdem Grund fest stehen können. Eine Theorie der Bedeutung beispielsweise, die sich etwa nur in *Whorfs* SAE (,Standard Average European) - Sprachen vollständig und korrekt formulieren läßt, würde ‚Bedeutung' zu einer Eigentümlichkeit dieser Idiome degradieren. (Whorf 1970)

Wir wollen diese Minimalforderung das *Kopernikanische Prinzip der philosophischen Semantik* nennen: Jede Hypothese (mit allgemeinverbindlichem Anspruch) über die Natur sprachlicher Bedeutung schlechthin sollte sich in sämtliche Sprachen, die über reflexive Potenz verfügen, übersetzen lassen. Unter *reflexiver Potenz* verstehen wir hier die Möglichkeit, innerhalb eines Verständigungssystems Bedeutungs-analyse zu betreiben, d.h. Phänomene wie Intention, Sinn, Begriff zu untersuchen und zu klären.

Cum grano salis ist die Eignung einer Sprache als ihre eigene Metasprache gemeint.

Bei den Gehörlosensprachen ÖGS und DGS handelt es sich zweifellos um zwei (miteinander nahverwandte) Gebärdensprachen mit reflektiver Potenz, z.B. es kann rekursiv und ohne Begrenzung über verwendete Gebärden und ihre Bedeutung gebärdet werden.

Der erste Ertrag unseres Projekts liegt in der quasi objektiven ‚Meßbarkeit' der Berechtigung wittgensteinscher Sichtweisen: Was sich davon als *prinzipiell* nicht in die ÖGS übersetzbar erweist, ist – entsprechend unserem ‚Kopernikanischen Prinzip' - noch nicht allgemein genug für eine *Universalsemantik*.

Freilich muß von Fall zu Fall rigoros untersucht werden, ob wirklich prinzipielle Unübersetzbarkeit vorliegt – nicht etwa ein Nachholbedarf auf dem Gebiet der Gebärdensprache, Inkompetenz der Gebärdensprecher oder tiefliegende Mißverständnisse.

Den zweiten Ertrag bilden sozusagen die Prolegomena zu einer philosophischen Gebärden-fachsprache. Es ist nicht einzusehen, warum sich gehörlose Menschen den Zugang zu tiefen und komplexen Fragestellungen stets nur über ihre Zweitsprache verschaffen können, ohne Möglichkeit, die Inhalte innerhalb ihrer Sprachgemeinschaft an weniger Laut-Schriftkundige weiterzuvermitteln.

3. Erste Erfahrungen: (un)gebärdige Metaphern

Schon im Zuge der ersten Übersetzungsversuche waren formale Schwierigkeiten deutlich von den inhaltlichen zu unterscheiden. Zu den Barrieren formaler Art zählen Eigentümlichkeiten der Gebärdensprache, wie daß z.B. Konjunktionen am Anfang eines Nebensatzes (‚daß'/‚ob') kein Gebärdenzeichen entspricht, oder daß der Konjunktiv durch die Körperhaltung ausgedrückt wird. Das macht es einigermaßen anstrengend, eine Feststellung wie T 2.0211

> Hätte die Welt keine Substanz, so würde, ob ein Satz Sinn hat, davon abhängen, ob ein anderer Satz wahr ist. (Wittgenstein 1984)

zu gebärden.

Diese Art von Hürden sind aber bei einiger Sorgfalt durch Zerlegung und Umformulierung zu umgehen.

Ernstere Hindernisse ergaben sich angesichts der von Metaphorik und Analogienbildung dicht durchzogenen Sprache Wittgensteins. Unsere erste ‚Gewährsfrau', eine gehörlose Übersetzerin mit ÖGS als Erstsprache, konnte mit der Zentralmetapher des *Tractatus* – Gedanken bzw. Sätze als Bilder von Tatsachen - nichts anfangen.

Das widersprach diametral unseren Erwartungen, da rund 40 Prozent der Gebärden der ÖGS ikonischer Natur sind, weiters unsere Gesprächspartnerin Kunstgeschichte studiert und auch als Malerin mit Theorie und Praxis der Abbildung innig vertraut ist.

Vor der gemeinsamen Besprechung einiger Grundideen der PU legten wir ihr ein Dutzend gebräuchlicher Metaphern der deutschen Alltagssprache vor. Obwohl sie die meisten kannte, empfand sie sie fast durchwegs als verschroben und unnatürlich. Das Problem liegt darin, daß man nicht einfach ‚analoge' metaphorische Gebärden-komplexe heranziehen kann; es geht ja gerade darum, was in jedem konkreten Einzelfall als ‚entsprechend' zu werten ist.

C. Papaspyrou, der unser Projekt mit Interesse begleitet, erklärte in Hinsicht auf unsere Schwierigkeiten: „Die – auf Deutsch formulierten – metaphorischen Beziehungen in Wittgensteins Sprachphilosophie müssen deshalb zuerst auf entsprechende Ausdrücke der Gebärdensprache *‚umgedichtet'* werden, bevor man den sprachphilosophischen Inhalt sachlich interpretiert."

4. Innensemantik & Bildersprache

Um die *metaphorische Barriere* zwischen Laut- und Gebärdensprechenden zu umgehen: wäre es nicht besser, zunächst den philosophischen Text zu ‚entmetaphorisieren'? Alles allzu Bildhafte durch ‚Klartext', brute facts zu ersetzen? Wir halten dies bei Sätzen, die vom Wesen sprachlicher Bedeutung handeln, für ausgeschlossen: Solche Sätze sind entweder verkappte syntaktische – oder sie enthalten unreduzierbare Metaphern.

Für eine ausführliche Begründung dieser apodiktischen Absage fehlt hier der Platz. Die Unmöglichkeit, auf bildhafte Umschreibung zu verzichten, wurzelt darin, daß ‚Sprache' in zwei komplementäre Bereiche zerfällt, einen *transparenten* und eine *opaken*, entsprechend nicht-thetischem und thetischem Sprechbewußtsein.

Zur Präzisierung dieser Hypothese fehlt noch der Fachjargon. Wir stehen gleichsam mit einem Fuß auf phänomenologischem und mit dem anderen auf sprachanalytischem Territorium. Doch wird daraus kein Spagat. Denn wir befinden uns in einem Gebiet, wo sich die Wege von Sartre und Wittgenstein überkreuzen: Im Problemfeld von Bewußtsein-Handlung-Leiblichkeit.

Vergleichen wir folgende Ausführung Sartres:

> In Bezug auf meine Hand bin ich nicht in derselben benutzenden Haltung wie im Bezug zum Federhalter. Ich *bin* meine Hand. Das heißt, sie ist der Stillstand der Verweisungen und ihr Abschluß. (Sartre 1943)

mit Wittgensteins Einsicht:

> Das Schreiben ist gewiß eine willkürliche Bewegung, und doch eine automatische. Und von einem Fühlen jeder Schreibbewegung ist natürlich nicht die Rede. Man fühlt etwas, aber könnte das Gefühl unmöglich zergliedern. Die Hand schreibt; sie schreibt nicht, weil man will, sondern man will, was sie schreibt. (Wittgenstein 1984a)

Die Hand kommt im flüssigen Schreiben sowenig vor wie das schweifende Auge im Erfassen der Landschaft. Sie wird im Erleben für-mich gleichsam ‚durchsichtig'. Und pointiert ließe sich sagen: *im* Schreiben habe ich keine Hand; – so, wie Douglas Harding anstelle seines Kopfes die visuelle Welt setzt. Der Kopf – in der 1.Person-Perspektive – verschwindet und macht so Platz für die ganze Welt. (Harding 2002)

Analog dazu verschwindet die artikulierte Sprache im Brennpunkt meiner Rede und macht hier den Platz für *Bedeutung* frei. Semantik, von innen betrachtet, ist sozusagen Syntaktik-für-mich – nicht etwas, das zu wohlgeformten Sätzen hinzukommt, sondern das nicht-thetische Formulieren von Sätzen im Modus *être-pour-soi.*

Daß das Durchsichtigwerden der Sprache gegenüber dem Gemeinten auch das Lesen kennzeichnet, weiß jede(r) mit der Lektüre verschiedensprachiger Texte in raschem Wechsel Befaßte: man kann einfach nicht mehr sagen, ob der zuletzt gelesene Absatz englisch oder deutsch war, obwohl der Inhalt noch als sozusagen gestochen scharfes Nachbild vor dem seelischen Auge steht.

Bedeutung-an-sich, festgestellte Bedeutung, gibt es gemäß der *innensemantischen* Sicht immer nur ex post, in der Reflexion. Das sprechende Menschenwesen befindet sich in der Schieflage von Morgensterns *Blondem Korken*:

> Ein blonder Korke spiegelt sich
> in einem Lacktablett –
> allein er säh' sich dennoch nich'
> selbst wenn er Augen hätt'!
>
> Das macht, dieweil er senkrecht steigt
> zu seinem Spiegelbild!
> Wenn man ihn freilich seitwärts neigt,
> zerfällt, was oben gilt.
>
> O Mensch, gesetzt du spiegelst dich
> im, sagen wir, - im All!
> Und senkrecht! – wärest du dann nich#
> ganz in demselben Fall?
> (Morgenstern 1995)

Wir können unsere Vermutung der ‚zwangsläufig metaphorischen Innensemantik' auch so formulieren, daß die Grammatik von ‚bedeuten' eher der Grammatik psychologischer Verben – als der Beschreibung materieller Zustände oder Relationen ähnelt; deshalb bedarf es bildlicher Vermittlung. Auch diese These muß sich im Zug der Übersetzungsarbeit erst bewähren.

5. ‚Denkspiele' als Erlebensformen

Welchen Stellenwert den Gebärden im ‚Denken' zukommt, darüber herrscht offenbar keine Einigkeit; selbst unter den Gehörlosen mit einer Gebärdensprache als Primärsprache. Viele betonen, daß sie zuweilen *in* Gebärden dächten, weisen jedoch das Bild von ‚inneren Gebärden' – analog dem zu sich selbst Sprechen – oft belustigt zurück.

Während unsere gehörlose Projektpartnerin nach ihrem Selbstverständnis ihre Gedanken mittels Gebärden ausdrückt, verneinte sie kategorisch jede Beteiligung ihrer Gebärdensprache am ‚privaten' – für Lautsprachler: stillem – Denken. Für sie fällt ‚reines Denken' mit einer gesteuerten Abfolge innerer Bilder zusammen.

Das führt zur Frage nach der Ordnung dieser Bilder. C. Papaspyrou meinte im Gedankenaustausch zu diesem Thema, daß jedes Denken – als operativ-rekursive Tätigkeit – auf figurative Ausdrucksmöglichkeiten als Stütze angewiesen sei. Neben Gebärden erwähnte er als Mittel zur Denksteuerung: Schriftsprache, Mathematik, geometrische Formen und Farbsysteme als gebräuchliche ‚Vehikel des Denkens' Gehörloser.

Wir stecken noch in derselben tiefen Verwirrung, die Wittgenstein in den PU angesichts der Memoiren des gehörlosen Mr. Ballard rätseln ließ: „Bist du sicher, daß dies die richtige Übersetzung deiner wortlosen Gedanken in Worte ist?" (Wittgenstein 1984)

Gerade an dieser logischen Bruchstelle - im Niemandsland zwischen unterschiedlichen *‚Sprachsubstanzen'* (Papaspyrou 1990) – wollen wir so wenig wie möglich mit apriorischen Mutmaßungen arbeiten. Nur die Erfahrung des Übersetzungsdialogs kann hier Erhellung bringen; damit wird Licht auf die Familie der ‚inneren' Sprachspiele überhaupt fallen.

Literatur

Harding, Douglas 2002 *On having no head*, Carlsbad: Inner directions

Krausnecker, Verena/Schalber, Katharina 2007 *Sprache Macht Wissen*, Wien: Österreichisches Sprachen-Kompetenz-Zentrum, http:\\www.univie.ac.at/oegsprojekt

Morgenstern, Christian 1995 *Galgenlieder*, Berlin: dtv

Papaspyrou, Chrissostomos 1990 *Gebärdensprache und universelle Sprachtheorie*, Hamburg: Signum

Sartre, Jean Paul 1943 *l'être et le néant*, Paris: Gallimard

Whorf, Benjamin Lee 1970, *Language, Thought&Reality*, Cambridge (Mass.): M.I.T. Press

Wittgenstein, Ludwig 1984 *Werkausgabe Band 1*, Frankfurt am Main: Suhrkamp

Wittgenstein, Ludwig 1984a *Werkausgabe Band 8,* Frankfurt am Main: Suhrkamp

Wundt, Wilhelm 1911 *Völkerpsychologie,* Liepzig: Wilhelm Engelmann

Abbildung und lebendes Bild in Tractatus und Nachlass

Christian Erbacher, Bergen, Norwegen

1. „Der Satz ist ein Bild der Wirklichkeit.“ (4.01)

Zu seiner Verwendung des Begriffs des Bildes im *Tractatus* sagt Wittgenstein im Gespräch mit Waismann vom 9. Dezember 1931, dass sie in zwei verschiedenen Auffassungen wurzelte: zum einen im ‚gewöhnlichen Sinne’ (vgl. 4.011) des Wortes, etwa wenn man von einem gezeichneten Bild spreche; zum anderen im mathematischen Begriff der ‚Abbildung’:

> „Als ich schrieb: ‚Der Satz ist ein logisches Bild der Tatsache’[1], so meinte ich: ich kann in einen Satz ein Bild einfügen, und zwar ein gezeichnetes Bild, und dann im Satz fortfahren. Ich kann also ein Bild wie einen Satz gebrauchen. Wie ist das möglich? Die Antwort lautet: Weil eben beide in einer gewissen Hinsicht übreinstimmen, und dieses Gemeinsame nenne ich *Bild*. Der Ausdruck ‚Bild’ ist dabei schon in einem erweiterten Sinn genommen. Diesen Begriff des Bildes habe ich von zwei Seiten geerbt: erstens von dem gezeichneten Bild, zweitens von dem Bild des Mathematikers, das schon ein allgemeiner Begriff ist. Denn der Mathematiker spricht ja auch dort von Abbildung, wo der Maler diesen Ausdruck nicht mehr verwenden würde.“ (WWK 1989, S.185)

Die vorliegende Untersuchung zeigt, welche Stellen des *Tractatus* mit dem Begriff im mathematischen Sinn in Verbindung stehen und wo die Verwendung wechselt.

2. Mathematische Bestimmungen der Begriffe *Abbildung* und *Repräsentation*

Für die Analyse ist die Erinnerung an einige Begriffsbestimmungen hilfreich. Die Angaben orientieren sich an Orth (1975; vgl. aber auch z.B. Suppes 1988)

Abbildung

Unter einer *Abbildung* von einer Menge *A* in eine andere Menge *B* versteht man eine *Vorschrift* (auch: *Zuordnung*, *Zuordnungregel*), die jedem $a \in A$ genau ein $b \in B$ zuordnet. Da jedem $a \in A$ genau ein $b \in B$ zugeordnet wird, bezeichnet man eine Abbildung auch als *eindeutig*. Wird *A* in *B* abgebildet, so heißt *B* *Bild* von *A* und *A* *Urbild* von *B*.

Wenn es umgekehrt ebenfalls zu jedem $b \in B$ genau ein $a \in A$ gibt, so heißt die Abbildung *umkehrbar eindeutig* (auch: *bijektiv*). Es besteht hierbei also nicht nur eine Abbildung von *A* in *B*, sondern auch umgekehrt von *B* in *A*; man schreibt: $\varphi(a) = b$.

Homomorphe Abbildung (*Repräsentation*) und *Isomorphismus*

Von einer *homomorphen Abbildung* spricht man, wenn nicht nur eine Menge in eine andere abgebildet wird, sondern auch Relationen zwischen den Elementen dieser Menge. Mindestens eine Menge und mindestens eine darauf definierte Relation fasst man als *Relativ* (auch: *Relationenstruktur*) zusammen. Man schreibt hierfür A= $<A, R_1, ..., R_n>$. Eine *Relation* ist eine Teilmenge aller geordneten Paare, die zueinander in einer *bestimmten Beziehung* stehen.

Bei einer homomorphen Abbildung wird also ein Relativ in ein anderes Relativ abgebildet. Man kann dies auch so ausdrücken, dass das Bild einer Relation zwischen zwei Elementen aus der Menge *A* gleich der Relation der Bilder in der Menge *B* ist. Eine entsprechende Definition lautet:

> D1: „Es seien A=$<A, R_1, ..., R_n>$ und B=$<B, S_1, ..., S_n>$ zwei Relative desselben Typs. Eine Abbildung φ von *A* in *B* heißt *homomorphe Abbildung* (oder: *Homomorphismus*) von *A* in *B*, wenn für alle Elemente $a_1, a_2 \in A$ und für alle i = 1, 2, ..., n gilt: $\varphi[R_i(a_1, a_2)] = S_i[\varphi(a_1), \varphi(a_2)]$.“ (Orth, S.16)

Besteht eine homomorphe Abbildung von *A* in *B*, so sagt man auch, dass *A* durch *B* *repräsentiert* wird, und *B* eine *Repräsentation* von *A* ist. *Isomorphe Abbildung* (auch: *Isomorphismus*) wird eine bijektive Repräsentation genannt, also eine umkehrbar eindeutige homomorphe Abbildung.

3. Repräsentation im *Tractatus*

Mit dem so definierten Begriffsinstrumentarium kann man sagen, dass der *Tractatus* die isomorphe Repräsentation der Welt durch Sprache darstellen soll. Dies wird nun an den Stellen des *Tractatus* aufgezeigt, die dem mathematischen Verständnis des Bildbegriffs entsprechen. Die Reformulierung in den oben bestimmten Begriffen ist den Zitaten kursiv vorangestellt:

1. Die Welt wird als Relativ dargestellt, das aus der Menge der Gegenstände und ihren Relationen zueinander besteht:

> „Die Welt ist alles, was der Fall ist.“ (1)
> “Was der Fall ist, die Tatsache, ist das Bestehen von Sachverhalten.“ (2)
> „Der Sachverhalt ist eine Verbindung von Gegenständen (Sachen, Dingen).“ (2.01)
> „Im Sachverhalt verhalten sich die Gegenstände in bestimmter Art und Weise zueinander.“ (2.031)
> „Die Art und Weise, wie die Gegenstände im Sachverhalt zusammenhängen, ist die Struktur des Sachverhalts.“ (2.032)

2. Der Satz wird als Relativ dargestellt, das aus Wörtern und ihren Relationen zueinander besteht:

> „Das logische Bild der Tatsachen ist der Gedanke.“ (3)
> „Im Satz drückt sich der Gedanke sinnlich wahrnehmbar aus.“ (3.1)
> „Das Satzzeichen besteht darin, daß sich seine Elemente, die Wörter, in ihm auf bestimmte Art und Weise zueinander verhalten“ (3.14)

[1] Waismann merkt an, dass dieser Satz nirgendwo genau steht und verweist auf die Stellen 3, 4.01, 4.03.

3. Der Elementarsatz wird als Relativ dargestellt, das die Welt isomorph repräsentiert:

> Diese Reformulierung kann in drei Sätze aufgespalten werden:

a) *Im Elementarsatz bilden Namen Gegenstände ab, d.h. es besteht eine eindeutige Zuordnung von einfachen Zeichen zu Gegenständen*:

> „Im Satze kann der Gedanke so ausgedrückt sein, daß den Gegenständen des Gedankens Elemente des Satzzeichens erntsprechen." (3.2)
> „Diese Elemente nenne ich ‚einfache Zeichen' und den Satz ‚vollständig analysiert'." (3.201)
> „Die im Satze angewandten einfachen Zeichen heißen Namen" (3.202)
> „Der Elementarsatz besteht aus Namen. Er ist ein Zusammenhang, eine Verkettung, von Namen." (4.22)
> „Der Name kommt nur im Zusammenhange des Elementarsatzes vor." (4.23)

b) *Im Elementarsatz werden nicht nur Gegenstände abgebildet, sondern der Elemtarsatz repräsentiert die Sachlage (bildet sie homomorph ab), da auch die Beziehungen zwischen den Gegenständen abgebildet werden:*

> „Die Konfiguration der einfachen Zeichen im Satzzeichen entspricht die Konfiguration der Gegenstände in der Sachlage." (3.21)

c) *Die Beziehung zwischen Sprache und Welt ist nicht nur eine Repräsentation, sondern eine isomorphe Repräsentation, da eindeutige Rückübersetzbarkeit gefordert wird*:

> „Daß es eine allgemeine Regel gibt, durch die der Musiker aus der Partitur die Symphonie entnehmen kann, durch welche man aus der Linie auf der Grammaphonplatte die Symphonie nach der ersten Regel wieder die Partitur ableiten kann, darin besteht eben die Ähnlichkeit dieser scheinbar so ganz verschiedenen Gebilde. Und jene Regel ist das gesetz der Projektion, welches die Symphonie in die Notensprache projiziiert. Sie ist die Regel der Übersetzung der Notensprache in die Sprache der Grammophonplatte." (4.0141)
> „Die Grammophonplatte, der musikalische Gedanke, die Notenschrift, die Schallwellen, stehen alle in jener abbildenden Beziehung zueinander, die zwischen Sprache und Welt besteht." (4.014)

Zusammenfassend kann man sagen, dass der *Tractatus* eine isomorphe Repräsentation von Sachverhalten durch Elementarsätze, von Welt durch Sprache, verlangt. Da der Sachverhalt der Sinn des ihn abbildenden Elementarsatzes ist („Was das Bild darstellt, das ist sein Sinn." (2.221)), kann man auch von einer Konzeption von Sinn als isomorphe Repräsentation sprechen (z.B. Hacker 1981, Glock 2006).

4. Der Wechsel zur gewöhnlichen Begriffsverwendung

Für die Frage, inwiefern im *Tratctatus* durchgängig eine Theorie der isomorphen Repräsentation von Sinn formuliert wird, ist die Betrachtung einer Stelle aufschlussreich, in der von Bild nicht mehr im mathematischen Sinn gesprochen wird. Das ‚gewöhnliche Verständnis' des Begriffs erscheint mit der Beschreibung der Verkettung von einfachen Zeichen zu Elementarsätzen:

> „Der Elementarsatz besteht aus Namen. Er ist ein Zusammenhang, eine Verkettung, von Namen." (4.22)
> „Ein Name steht für ein Ding, ein anderer für ein anderes Ding und untereinander sind sie verbunden, so stellt das Ganze – wie ein lebendes Bild – den Sachverhalt vor." (4.0311)

Es wird hier, wie zuvor, gefordert, dass Namen zu Elementarsätzen verbunden sind; allerdings bleibt die Frage offen, *welche* Relationen zwischen Namen bestehen, *wie* die Verkettung von Namen ein „lebendes Bild" bilden kann. Hierfür scheint der *Tractatus* keine weitere Zuordnungsregel (im mathematischen Sinn) anzugeben. (Die Verknüpfung von Namen durch logische Konstanten kommt hierfür nicht in Frage. Sie sind Operationen, die auf der Menge der *Elementarsätze* definiert sind und von Elementarsätzen zu allen anderen Sätzen führen.)

5. Bild und Abbildung in Manuskripten aus dem Nachlass

Der Blick in Wittgensteins *Nachlass* bestätigt den Eindruck in Bezug auf den Wechsel der Begriffsverwendung wie er oben für den *Tractatus* dargestellt wurde. Betrachtet man die drei Manuskriptbände *Ms101* (09. August 1914 – 30. Oktober 1914), *Ms102* (30. Oktober 1914 – 22. Juni 1915) und *Ms103* (29. März 1916 – 10. Januar 1917), sind Einträge in Verbindung mit den Begriffen Bild und Abbildung vor allem in *Ms101* und *Ms102* zu finden, und hier hauptsächlich in den Monaten September, Oktober und November 1914 (siehe Tabelle 1). Diese frühe Beschäftigung mit dem Thema weist auf seine grundlegende Bedeutung für den *Tractatus* hin.

Tab. 1: Vorkommnisse der Begriffe Bild/bil* und Abbildung/abbil* in *Ms101*, *Ms102* und *Ms103* (abs. Häufigkeiten, nach BEE, diplomatische Version).

	Bild/bil*	Abbildung/abbil*
Ms101	17/22	6/11
Ms102	53/67	4/10
Ms103	2/2	0/0

In *Ms101* scheint eine klare Trennung der zwei Verwendungsweisen mit vornehmlicher Verwendung des Begriffs im mathematischen Sinne vorzuliegen. So spricht Wittgenstein wiederholt von ‚logischem Abbild' (Ms-101,22r, Ms-101,29r, Ms-101,52r)[2] und ‚meiner Theorie der logischen Abbildung' (Ms-101,52r). Wenn von Bild im gewöhnlichen Sinn die Rede ist, dann in Abgrenzung zu dem logischen oder mathematischen Sinn der Abbildung. So heisst es in dem Eintrag vom 29. September 1914 zum Beispiel:

> „Denken wir daran daß auch wirkliche Bilder von Sachverhalten stimmen und nicht stimmen können." (Ms-101,28r, Wittgensteins Unterstreichungen).

Dieser Eintrag entspricht im *Tractatus* einem Satz in 4.011, wo von Bildern „auch im gewöhnlichen Sinne" gesprochen wird. Dass hier von „wirklichen Bildern" und „*auch* im gewöhnlichen Sinne" (meine Hervorhebung) gesprochen wird, legt nahe, dass *ansonsten* von Bild als Abbildung im präzisierten mathematischen Sinne die Rede ist. Es deutet

[2] Zitierweise der Nachlass-Dokumente orientiert sich an den Sigla, die am Wittgenstein-Archiv im Hyperwittgenstein- und DISCOVERY-Projekt (http://wab.aksis.uib.no /wab_discovery.page; http://wab.aksis.uib.no/wab_hw.page/) entwickelt wurden.

sich hier allerdings schon die Verschmelzung der beiden Verwendungsweisen an.

Für die vorliegende Untersuchung liegt der Kulminationspunkt der Manuskripteinträge zwischen Ende Oktober und Anfang November 1914. Der Eintrag vom 30. Oktober macht die entscheidende Rolle der Elementarsätze für die Idee der Repräsentation von Sachverhalten deutlich. Dort heisst es:

> „Vor allem muß die Elementarsatzform abbilden, alle Abbildung geschieht durch diese.“ (Ms-102,3r).

Die erste Hälfte des Eintrages vom 4. November lautet:

> „Wie bestimmt der Satz den logischen Ort?
> Wie repräsentiert das Bild einen Sachverhalt?
> Selbst ist es doch nicht der Sachverhalt, ja dieser braucht gar nicht der Fall zu sein.
> Ein Name repräsentiert ein Ding ein anderer ein anderes Ding und selbst sind sie verbunden; so stellt das Ganze — wie ein lebendes Bild — den Sachverhalt vor.“ (MS-102,17r-18r)

An dieser Stelle geschieht wie in *Tractatus* 4.0311 der Wechsel der Begriffsverwendung. Wittgenstein gibt hier keine Zuordnungsregel für einfache Zeichen zu Elementarsätzen an, sondern spricht von einem „lebenden Bild“.

In dem Notizbuch verwendet Wittgenstein fortan noch häufig den Begriff Bild, und zwar in Zusammenhängen des mathematischen Sinnes; den Begriff Abbildung verwendet er sehr viel weniger. Die beiden zunächst klar unterschiedenen Begriffe scheinen hier verschmolzen. Die deskriptive Statistik der Begriffe in den Manuskripten spiegelt die hier skizzierte Entwicklung sehr gut wider (siehe Tabelle 1).

6. Mathematische Mannigfaltigkeit

Der Rest des Notizbucheintrages vom 4. November beschäftigt sich weiter mit der Verbindung von Dingen und somit auch mit Relationen von Zeichen. Er drückt die Einsicht aus, dass die Verbindungen der Dinge den Relationen der abbildenden Elemente entsprechen müssen:

> „Die logische Verbindung muß natürlich unter den repräsentierten Dingen möglich sein und dies wird immer der Fall sein wenn die Dinge wirklich repräsentiert sind. Wohlgemerkt jene Verbindung ist keine Relation sondern nur das Bestehen einer Relation.“ (MS-102,18r-19r)

Im *Tractatus* wird analog im Anschluss an 4.0311 mit Bezug auf die „mathematische Mannigfaltigkeit“ festgestellt, dass die Relationen der Namen im Elementarsatz den Verbingungen der Gegenstände entprechen können müssen:

> „Am Satz muss gerade soviel zu unterscheiden sein, als an der Sachlage, die er darstellt. Die beiden müssen die gleiche logische (mathematische) Mannigfaltigkeit besitzen. ... „ (4.04)

Liest man Defintion D1 genau, so sieht man dass auch dort diese Forderung genannt ist, denn das abbildende Relativ soll „desselben Typs“ sein wie das abgebildete. Für die Relationen des Relativs der Elementarsätze gibt Wittgenstein im *Tractatus* aber keine Bestimmung. In dem Folgenden Paragraphen heisst es stattdessen:

> „Diese mathematische Mannigfaltigkeit kann man natürlich nicht selbst wieder abbilden. Aus ihr kommt man beim Abbilden nicht heraus.“ (4.0411)

7. Fazit

Der Tractatus gibt Bedingungen einer isomorphen Repräsentation von Sachverhalten an

Einerseits stellt der *Tractatus* die Forderung nach einer isomorphen Repräsentation von Welt dar (vgl. Hacker 1981, Glock 2006). Insofern gibt er Bedingungen für eine Theorie von Sinn als Repräsentation an. Andererseits wechselt Wittgenstein die Verwendung des Begriffs des Bildes von einem mathematischen zu einem gewöhnlichem Sinn, wenn er über die Verkettung von Namen zu Elementarsätzen spricht, also gerade dort, wo die Forderung nach einer Abbildung der Relationen zwischen Dingen erfüllt werden müsste. Abgesehen von der bloßen Forderung der Gleichartigkeit der Verbindungen zwischen Gegenständen und zwischen Namen wird im *Tractatus* keine Zuordnungsregel von Namen zu Elementarsätzen genannt.

Die Frage nach der Konzeption von Wahrheit im Tractatus ist nicht betroffen

Die Frage nach der Konzeption von Wahrheit ist von dieser Analyse nicht betroffen. Die Konzeption von Sinn muss von der Konzeption der Wahrheit im *Tractatus* unterschieden werden (Glock 2006). Insofern ist auch Hintikka (1994, S.223) zuzustimmen, dass es keinen Sinn hat von *der Abbildtheorie* zu sprechen, da verschiedene, voneinander weitgehend unabhängige Ideen unter diesem Titel verhandelt werden. Die hier besprochenen Aspekte betreffen Hintikkas erste von insgesamt sechs Abbild-Ideen („Elementarsätze als Bilder“, S. 224, 227-229). Ihre Formulierung hat zunächst keine Auswirkungen auf Operationen mit Elementarsätzen. Dies drückt auch Wittgenstein aus, wenn er schreibt:

> „Die Schemata No. 4.31 haben auch dann eine Bedeutung, wenn ‚p‘, ‚q‘, ‚r‘, etc. nicht Elementarsätze sind“ (5.31).

‚*Das Leben von Zeichen*‘

Man kann auch die Frage, wie Zeichen ein „lebendes Bild“ bilden können, wie Leben in die Zeichen kommt, ohne Annahme von Elementarsätzen behandeln. Wie wir wissen, beschäftigt sich Wittgenstein mit dieser Frage in späteren Jahren.

Literatur

Glock, Hans Johann 2006 „Truth in the Tractatus“, *Synthese* 148, 345–368.

Hacker, P.M.S. 1981 „The Rise and Fall of the Picture Theory”, in: Irving Block (Hg.), *Perspectives on the Philosophy of Wittgenstein*, Oxford: Blackwell, S. 85–109.

Hintikka, Jaakko 1994 „An Anatomy of Wittgenstein’s Picture Theory“, in Carol C. Gould and Robert S. Cohen (Hrsg.), *Artifacts, Representations and Social Practice*, Dordrecht: Kluwer, S. 223–256.

Orth, Bernhard 1974 *Einführung in die Theorie des Messens*, Stuttgart: Kohlhammer

Suppes, Patrick 1988, „Representation theory and the analysis of structure“, *Philosophia Naturalis* 25, S. 254–268.

Wittgenstein, Ludwig 1963 *Tractatus logico-Philosophicus*, Frankfurt am Main: Suhrkamp

Wittgenstein, Ludwig 1989 Ludwig Wittgenstein und der Wiener Kreis: Gespräche, aufgezeichnet von Friedrich Waismann, Frankfurt am Main: Suhrkamp

Wittgenstein, Ludwig 2001 *Tractatus logico-Philosophicus*, Kritische Edition, Frankfurt am Main: Suhrkamp

Wittgenstein, Ludwig 2000 *Wittgenstein’’s Nachlass: The Bergen Electronic Edition*, Wittgenstein Archives at the University of Bergen (Hrsg.), Oxford: Oxford University Press.

Explaining the Brain: Ruthless Reductionism or Multilevel Mechanisms?

Markus Eronen, Osnabrück, Germany

1. Introduction

In this paper, I will compare and criticize two approaches to reduction and explanation in neuroscience: metascientific reductionism and mechanistic explanation. I will first show that the traditional models of intertheoretic reduction are unsuitable for neuroscience. Then I will compare John Bickle's model of metascientific reductionism and Carl Craver's model of mechanistic explanation, arguing that the latter has a stronger case, especially when supplemented with James Woodward's interventionist account of causal explanation.

2. Intertheoretic reduction

The development of intertheoretic models of reduction started in the middle of the 20th century, in the spirit of logical positivism. The ultimate goal was to show how unity of science could be attained through reductions. In the classic model (most importantly Nagel 1961, 336-397), reduction consists in the deduction of a theory to be reduced (T_2) from a more fundamental theory (T_1). Conditions for a successful reduction are that (1) we can connect the terms of T_2 with the terms T_1, and that (2) with the help of these connecting assumptions we can derive all the laws of T_2 from T_1.

Unfortunately this model fails to account for many cases that are regarded as reductions. The model is too demanding: it is very hard to find a pair of theories that would meet these requirements. Even Nagel's prime example, the reduction of thermodynamics to statistical mechanics, is much more complicated than Nagel thought (see, *e.g.*, Richardson 2007). The classic model also has problems accommodating the fact that the reducing theory often *corrects* the theory to be reduced, which means that the theory to be reduced is strictly speaking false. However, logical deduction is truth-preserving, so it should not be possible to deduce a false theory from a true one.

Problems of this kind lead to the development of more and more sophisticated models of intertheoretic reduction, and finally to the "New Wave reductionism" of P. S. Churchland (1986), P. M. Churchland (1989) and J. Bickle (1998, 2003, 2006). Due to constraints of space, I will not go through these models here. It is sufficient to point out one fundamental assumption that underlies *all* intertheoretic models of reduction, and which leads to serious problems in the case of psychology and neuroscience.

This assumption is that the relata of reductions are exclusively *theories*, and that inter*theoretic* relations are the only epistemically and ontologically significant interscientific relations (see, *e.g.*, McCauley 2007). However, well-structured theories that could be handled with logical tools are rare in and peripheral to psychology and neuroscience. Instead, scientists typically look for mechanisms as explanations for patterns, effects, capacities, phenomena, and so on (see, *e.g.*, Machamer *et al*. 2000 and Cummins 2000). Although there are theories in a loose sense in psychology and neuroscience, like the LTP theory for spatial memory or the global workspace theory, these are not theories that could be formalized, and can hardly be the starting points or results of logical deductions. Therefore looking at the relations between theories is the wrong starting point, at least in the case of psychology and neuroscience.

3. Metascientific reductionism

At least partly for these reasons, John Bickle, the most ardent advocate of New Wave reductionism, has taken some distance from the intertheoretic models of reduction and now emphasizes looking at the "reduction-in-practice" in current neuroscience (Bickle 2003, 2006). He calls this approach "metascientific reductionism" to distinguish it from philosophically motivated models of reduction that are typically used in philosophy of mind.

The idea is that instead of imposing philosophical intuitions on what reduction has to be, we should examine scientific case studies to understand reduction. We should look at experimental practices of an admittedly reductionistic field, characterized as such by its practitioners and other scientists.

According to Bickle, molecular and cellular cognition – the study of the molecular and cellular basis of cognitive function – provides just the right example. The reductionist methodology of molecular and cellular cognition has two parts: (1) intervene causally into cellular or molecular pathways, (2) track statistically significant differences in the behavior of the animals (2006, 425). When this strategy is successful and a mind-to-molecules linkage has been forged, a reduction has been established. The cellular and molecular mechanisms *directly explain* the behavioural data and *set aside* intervening explanatory levels (2006, 426). Higher-level psychology is needed for describing behavior, formulating hypotheses, designing experimental setups, and so on, but according to Bickle, these are just heuristic tasks, and when cellular/molecular explanations are completed, there is nothing left for higher-level investigations to explain (2006, 428).

Metascientific reductionism does not require that the relata of reductions are formal theories, and does not lead to the problem mentioned in the end of last section. However, it is not without its share of problems, as I will show below.

4. Mechanistic explanation

The discrepancies between traditional models of reduction and actual scientific practice in psychology, neuroscience and biology have resulted in the development of alternative models. One alternative that I have just discussed is Bickle's metascientific reductionism. Another approach that has been receiving more and more attention recently is *mechanistic explanation* (*e.g.,* Bechtel & Richardson 1993, Machamer et al. 2000). In this paper I will focus on Carl Craver's (2007) recent and detailed account of mechanistic explanation.

The central claim of advocates of mechanistic explanation is that good explanations describe mechanisms (at least in neuroscience). Mechanisms are ”entities and activites organized such that they are productive of regular changes from start or set-up to finish or termination conditions” (Machamer et al. 2000, 3). A mechanistic explanation describes how the mechanism accounts for the *explanandum phenomenon*, the overall systemic activity (or process or function) to be explained.

For example, the propagation of action potentials is explained by describing the cellular and molecular mechanisms involving voltage-gated sodium channels, myelin sheaths, and so on. The pain withdrawal effect is explained by describing how nerves transmit the signal to the spinal chord, which in turn initiates a signal that causes muscle contraction. The metabolism of lactose in the bacterium *E. coli* is explained by describing the genetic regulatory mechanism of the *lac* operon, and so on.

5. The case of LTP

A paradigmatic example for both Bickle (2003, 43-106) and Craver (2007, 233-243) is the case of LTP (Long Term Potentiation) and memory consolidation. Both authors agree that the explanandum phenomenon is memory consolidation (the transformation of short-term memories into long-term ones), and that this is explained by describing how the relevant parts and their activities result in the overall activity - that is, by describing the cellular and molecular mechanisms of LTP. However, the conclusions the authors draw are completely different.

According to Bickle, the case of LTP and memory consolidation is a paradigm example of an accomplished psychoneural reduction. He describes the current cellular and molecular models of LTP in detail, and argues that they are the mechanisms of memory consolidation. Furthermore, he argues that these mechanisms explain memory consolidation *directly*, setting aside psychological, cognitive-neuroscientific, *etc.*, levels. This is an example of the ”intervene cellular/molecularly, track behaviorally” methodology, and in Bickle’s view a successful reduction.

What makes Bickle’s analysis ”ruthlessly” reductive is the claim that ”psychological explanations *lose their initial status as causally-mechanistically explanatory* vis-á-vis *an accomplished* (and not just anticipated) cellular/molecular explanation” (2003, 110). He argues that scientists stop evoking and developing psychological causal explanations once ”*real neurobiological* explanations are on offer”, and ”accomplished lower-level mechanistic explanations absolve us of the need in science to talk causally or investigate further at higher levels, at least in any robust ’autononomous’ sense” (2003, 111).

Craver’s analysis is quite different. He points out that the discoverers of LTP did not have reductive aspirations – they saw LTP as a component in a multilevel mechanism of memory, and after the discovery of LTP in 1973, there has been research both up and down in the hierarchy. Craver claims that the memory research program has implicitly abandoned reduction as an explanatory goal in favor of the search for multilevel mechanisms. His conclusion is that ”the LTP research program is a clear historical counterexample to those ... who present reduction as a general empirical hypothesis about trends in science” (2007, 243).

What sets Craver’s position in direct opposition to ruthless reductionism is the thesis of *causal and explanatory relevance of nonfundamental things*. That is, he argues that there is no fundamental level of explanation, and that entities of higher levels can have causal and explanatory relevance. This is in sharp contrast to Bickle’s view. Craver’s defense of the causal and explanatory relevance of nonfundamental things relies heavily on Woodward’s (2003) account of causal explanation, which I will briefly present here – the details are available in Woodward’s articles and books.

6. Causal explanation

A key notion for Woodward is *intervention*. An intervention can thought of as an (ideal or hypothetical) experimental manipulation carried out on some variable *X* (the independent variable) for the purpose of ascertaining whether changes in *X* are causally related to changes in some other variable *Y* (the dependent variable). Interventions are not only human activities, there are also ”natural” interventions, and the notion of intervention can be defined with no essential reference to human agency.

Another key concept is *invariance*. Broadly speaking, a generalization or relationship is invariant if it remains intact or unchanged under at least some interventions. Suppose that there is a relationship between two variables that is represented by a functional relationship $Y = f(X)$. If the same functional relationship *f* holds under a range of interventions on *X*, then the relationship is invariant within that range. For example, the ideal gas law “$pV = nRT$” continues to hold under various interventions that change the values of the variables, and is thus invariant within this range of interventions. Invariance is a matter of degree: for example, the van der Waals force law ($[P + a/V^2][V - b] = RT$) is more invariant than the ideal gas law since it continues to hold under a wider range of interventions.

The main point is that according to Woodward, causal explanation requires appeal to *invariant generalizations*. Invariant generalizations are explanatory because they can be used to answer “what-if-things-had-been-different questions” (w-questions). For example, the ideal gas law can be used to show what the pressure of a gas would have been if the temperature would have been different. True but non-invariant generalizations like ”all the coins in the pocket of Konstantin Todorov on January 25, 2008, are euros” cannot be used to answer w-questions. Only if a generalization is invariant under some range of interventions can we appeal to it to answer w-questions. In other words, causal explanatory relevance is just a matter of holding of the right sort of pattern of counterfactual dependence between explanans and explanandum, and invariant generalizations capture these patterns.

If we accept Woodward’s model of causal explanation, we see that Bickle’s claims about higher-level explanations losing their status as causally/mechanically explanatory are unwarranted. In Woodward’s account, things that figure in invariant generalizations have causal explanatory relevance. It is clear that in this sense nonfundamental things can have causal and explanatory relevance even when the ”fundamental” cellular and molecular explanations are complete. For example, the generalizations at the higher levels of the memory consolidation mechanisms will remain invariant even after the cellular and molecular explanations are complete.

In order to counter this argument, Bickle would have to show either that the relevant higher-level generalizations are not actually invariant, or that there is something wrong with Woodward’s account. The latter alternative is the more promising one. Bickle could argue that Woodward’s model is simply wrong, or that there is a stronger notion of causation that applies to the cellular/molecular level. However, a notion of causation like this does not emerge from scientific evidence only (Craver's and Woodward's models are just as much based on scientific evidence as Bickle's), and Bickle seems to be reluctant to provide philosophical arguments for his views.

Furthermore, such a stronger notion of causation would inevitably lead to problems. We can always ask the question: why stop at the cellular/molecular level and not go further down to the chemical/atomic/quantum level? Bickle is conscious of this, and in fact seems to admit that it is possible that in the future causal explanations will be found at the microphysical level (2003, 156-157). This of course means that the cellular/molecular explanations are only temporarily causal explanations. It also suggests that at some point the causal explanations for all human behavior will be microphysical explanations. This kind of a notion of causal explanation strikes me as implausible and unnecessarily restrictive.

On the other hand we have Woodward's notion of causal and explanatory relevance that conforms to scientific practice and is being more and more widely accepted among philosophers of science. The prospects of ruthless reductionism do not look very good.

7. Conclusion

In this paper, I have argued first that intertheoretic models of reduction are inappropriate for neuroscience, mainly because they focus on relations between formal theories. Then I have argued that mechanistic explanation and Woodward's theory of causal explanation taken together present a great challenge to a strongly reductionistic account of explanation in neuroscience.

Literature

Bechtel, William and Richardson, Robert C. 1993 Discovering complexity: decomposition and localization as strategies in scientific research, Princeton: Princeton University Press.

Bickle, John 1998 Psychoneural Reduction: The New Wave, Cambridge, MA: MIT Press.

Bickle, John 2003 Philosophy and Neuroscience: A Ruthlessly Reductive Account, Dordrecht: Kluwer Academic Publishers.

Bickle, John 2006 "Reducing mind to molecular pathways: explicating the reductionism implicit in current cellular and molecular neuroscience", Synthese 151, 411-434.

Churchland, Paul M. 1989 A Neurocomputational Perspective: The Nature of Mind and the Structure of Science, Cambridge: The MIT Press.

Churchland, Patricia S. 1986 Neurophilosophy, Cambridge: The MIT Press.

Craver, Carl 2007 Explaining the Brain: mechanisms and the mosaic unity of neuroscience, Oxford: Clarendon Press.

Cummins, Robert 2000 "'How Does It Work?' vs. 'What Are the Laws?' Two Conceptions of Psychological Explanation", in: Frank Keil and Robert Wilson (eds.), Explanation and Cognition, Cambridge: MIT Press, 117-144.

Machamer, Peter, Darden, Lindley, and Craver, Carl 2000 "Thinking about mechanisms", Philosophy of Science 67, 1-25.

McCauley, Robert N. 2007 "Reduction: Models of cross-scientific relations and their implications for the psychology-neuroscience interface", in: Paul Thagard (ed.), Handbook of the philosophy of psychology and cognitive science, Amsterdam: Elsevier, 105-158.

Nagel, Ernest 1961 The Structure of Science, London: Routledge & Kegan Paul.

Richardson, Robert C. 2007 "Reduction without the structures", in: Maurice Schouten and Huib Looren de Jong (eds.), The Matter of the Mind. Philosophical Essays on Psychology, Neuroscience and Reduction, Oxford: Blackwell Publishing, 123-145.

Woodward, James 2003 Making Things Happen: A Theory of Causal Explanation, Oxford: Oxford University Press.

Occam's Razor in the Theory of Theory Assessment

August Fenk, Klagenfurt, Austria

1. Overview

In this paper I will at first discuss the role of economy, parsimony or simplicity in theory assessment and model selection. This discussion (in Section 2) will amount to a three-dimensional model of theory assessment, including Coombs' (1984) dimensions generality (breadth) and power (depth), and simplicity as the third dimension.

Theory assessment is, most commonly, a matter of the methodology of empirical science. But its principles might also apply to "metaphysical theories", at least in part, as already suggested in Laszlo (1972:389). Thus they might also be applicable, in selfreferential ways, to those meta-theory – the "theory of theory assessment" in terms of Huber (2008:90) – that has invented the above mentioned criteria of model selection and theory assessment. This is exactly what I shall study in Section 3 of this paper, focusing on the key-concepts of *law* and *lawlikeness.* Laws are usually assumed to be a precondition for the reconstruction and explanation of phenomena on the one hand and their anticipation and prediction on the other, but relative frequency will be shown as the proper basis of all our projections to the past and to the future. Evolutionary perspectives are indicated in the last Section 4.

Thus, this paper does not deal with the reduction of theories in the sense of Nagel (1961), or with the problems in the attempts to reduce "emergent" systems to their elements, but rather with the *reduction* of (semantic) complexity and the *elimination* of dispensible components of (meta-)theories. And, in a certain sense, with the "reduction" of *law* to statistical generalizations.

2. Three dimensions of theory assessment

Most theories of theory assessment are two-dimensional, balancing e.g. "empirical adequacy" against "integrative generality" (Laszlo 1972:388) or power against generality (Coombs 1984), and most of the standard methods of model selection provide, according to Forster (2000:205), "an implementation of Occam's razor, in which parsimony or simplicity is balanced against goodness-of-fit".

But there are also some attempts to three-dimensional models: In his above mentioned paper Forster (2000:205) suggests that model selection should, besides simplicity and fit, "include the ability of a model to generalize to predictions in a different domain". In Lewis (1994:480) there is talk about a trade off between the "virtues of simplicity, strength, and fit". And Laszlo's (1972:388) factor "integrative generality" figures as "a measure of the internal consistency, elegance, and 'neatness' of the explanatory framework". Two scientific theories, he says, can be compared with regard to the number of facts taken into account (I), the precision of the accounting (II), and the economy (III) whereby the balance between "integrative generality" and "empirical adequacy" is produced. Economy (III) is, first of all, associated with a small number of "basic existential assumptions and hypotheses" (Laszlo 1972:388). (I) and (II) correspond to Coombs' generality and power, and Coombs' model may be viewed as an appropriate decomposition of Laszlo's factor "empirical adequacy". But it fails to account for Occam's razor.

Considering such arguments I emphasize a three-dimensional model (Fenk 2000) including the dimensions precision, generality (size of domain), and parsimony, as well as a strict distinction between the theory's assertions – the lawlike propositions in the core of any scientific theory – and the theory's "predictive success" (in the sense of Feyerabend 1962:94). Other than in the above mentioned approaches by Forster and by Lewis, goodness-of-fit is not a separate dimension, but the touchstone of the whole theory. According to this model we state an advantage of a theory t2, as compared with a former version or conflicting theory t1, if it achieves at least the same predictive success (number of hits) despite a higher precision of the predictions and/or an extended domain and/or a lower number of assumptions. With regard to Coombs' trade-off between the dimensions „power" and „generality", this idea is illustrated in Fenk & Vanoucek (1992:22f.), though only on the level of single lawlike assumptions.

Popper (1976:98,105) suggests disregarding, at least in epistemological contexts, properties of pure representation as well as the respective conventionalistic, "aesthetic-pragmatical" conceptualizations of "simplicity" or "elegance". But maybe the aesthetic attributes come by the theory's economic functionality, just as in the aesthetic BAUHAUS-principle "form follows function"? And our three-dimensional model actually applies, first of all, to *theory* as a hypothetical representation or construction. It is particularly interesting to see that it none the less fits all of Popper's further arguments regarding the relations between "empirical content", "testability", and "simplicity": The more possibilities ruled out by a sentence ("je mehr er verbietet"; p. 83), the higher its empirical content. "Auf die Forderung nach möglichst großem empirischen Gehalt können noch andere methodologische Forderungen zurückgeführt werden; vor allem die nach möglichst großer *Allgemeinheit* der empirisch-wissenschaftlichen Theorien und die nach größter Präzision oder *Bestimmtheit*." (p. 85) „Einfachere Sätze sind /…/ deshalb höher zu werten als weniger einfache, weil sie *mehr sagen*, weil ihr empirischer Gehalt größer ist, weil sie besser überprüfbar sind." (p. 103) Thus, generality (Allgemeinheit), precision (Bestimmtheit) and simplicity (Einfachheit) turn out to be three different facets of Popper's essential idea of testability and the chance to be falsified.

Are virtues such as "integrative generality" and "economy", as suggested in Laszlo (1972:389), also applicable to "metaphysical" disciplines, i.e. to meta-theories that have to do without the corrective of direct empirical tests? In theoretical semiotics, for instance, a reduced complexity of the terminological framework may allow to solve classificational problems such as the definition of *iconicity* (Fenk 1997), or to solve and communicate them in better understandable ways.[1] Can we apply criteria of scientific progress invented by the philosophy of science even to essential concepts of that philosophy of science?

[1] The latter aspect reminds, in some ways, of the concepts of "userfriendlyness" in Cognitive Ergonomics and of (low) "item-difficulty" in test theory.

3. A reductionistic look on laws and lawlikeness

A general principle „that is applicable to all kinds of reasoning under uncertainty, including inductive inference“ (Grünwald 2000:133) – is such a thing conceivable in view of the problems discussed in the philosophy of science?

I will try that focusing on the key-concepts of *law* and *lawlikeness*. In Goodman (1973:90,108) a hypothesis is lawlike only if it is projectible and projectible when and only when it is supported (some positive cases), unviolated (no negative cases), and unexhausted (some undetermined cases)[2]. But especially the criterion "unviolated" seems to be rather meant for universal laws (Fenk & Vanoucek 1992). What should be considered the negative and the positive cases in view of a weak regularity such as a very severe side-effect of a medicament showing in one of hundred patients in nine of ten studies?

The following outline starts with the universal laws in the Deductive-Nomological (D-N) model by Hempel & Oppenheim (1948). The authors note that their formal analysis of scientific explanation applies to scientific prediction as well. This symmetry between explanation and prediction will outlast. The application of the D-N model, however, is restricted to a world of universal laws – a rather restricted or even non-existent world, if *law* is not understood as a mere proposition but as an empirically valid argument. Thus we see a shift of the focus in the philosophy of science from the universal laws in the D-N model to statistical arguments rendering their extremely high probabilities ("close to 1") to the explanation in Hempel's (1962) Inductive-Statistical (I-S) model. And from here to the reduction of "plausibility" to the relative frequencies observed so far (Mises 1972:114) and to "stable" frequency distributions as a sufficient basis for "objective chances" (Hoefer 2007). Let me carry that to the extremes: If a dice had produced an uneven number in ten of fifteen cases I would, if I had to bet, bet on "uneven" for the sixteenth trial. For if there is a system it seems to prefer uneven numbers, and if there is none, I can't make a mistake anyway (Fenk 1992). But how if the "series" that had produced uneven has the minimal length of only one trial? I would again bet on "uneven". And if I knew that on a certain day in a certain place on the equator the highest temperature was 40° C, I would – if I had to guess in the absence of any additional knowledge – again guess a peak of 40°C for the day after or the day before. The only way I can see to justify such decisions is an application of Occam's razor, or a principle at least inspired by Occam's razor: Do without the assumption of a change as long as you can't make out any indication or reason for such an assumption!

Hardly anybody would talk about laws in the example with the fifteen dices, or in the case of a series of fifteen S1–S2 combinations in a conditioning experiment, and most of us wouldn't even talk about "relative frequency" in our one-trial "series" – despite an ideal "relative frequency" of 1 in the one-trial "series" and in the S1-S2 combinations in the conditioning experiment. But the examples reflect a principle as *simple* as *general*: Use the slightest indication and all your contextual knowledge to optimize your decision but bet on continuity as long as you see no reason to assume that a system might change its output-pattern; generalize the data available to unknown instances! "Laws", "probabilities", and "objective chances" are – beyond a purely mathematical world – nice names for such generalizations and projections, usually based on large numbers of observations. But there is no lower limit regarding the strength of a regularity or the number of data available that ceases the admissibility of this way of reasoning! I can't resist quoting Hempel (1968:117) when he admits that "no specific common lower bound" for the probability of an association between X and Y "can reasonably be imposed on all probabilistic explanation."

4. Evolutionary perspectives

In his commentary on Campbell (1987), Popper (1987) agrees with Campbell's view of the evolution of knowledge systems as a blind selective elimination process. I am not quite sure if this is fully compatible with his remark (p. 120) "that in some way or other all hypotheses *(H)* are psychologically prior to some observation *(O)*". And principles of theory assessment such as Occam's razor might guide a systematic and conscious selection of theories in ways being more efficient and faster than a blind evolutionary process. Any sort of anticipation and of explorative or "hypothesis-testing behavior" imputes regularities and patterns and is successfull only if its heuristics and strategies in turn follow such patterns. The selective pressure was, first of all, on the evolution of mechanisms and strategies for learning risks and chances. In our recent life anticipation plays double a role: still as the cognitive component of any practical decision, and in science as the hypothesis tested systematically in order to improve our knowledge.

Irrespective of whether or not the evolution of knowledge follows a blind selective process: Real progress in nomological science seems to come about relatively slowly (Laszlo 1998), most apparently if predictive success or prognostic performance is taken as the relevant criterion, and in part due to an again "relatively" slow improvement of the respective methods. "Relatively" slow as compared e.g. with "vague but perhaps persuasive forms of explanation in the social and behavioral sciences" and "metaphysical theories of human nature" (Laszlo 1972:389) that cannot claim predictive success. A nice parallel in the evolution of technical equipment: "Using functional and symbolic design features for Polynesian canoes", Rogers and Ehrlich (2008:1) could show "that natural selection apparently slows the evolution of functional structure, whereas symbolic designs differentiate more rapidly."

[2] For cases of two conflicting assumptions both satisfying the above criteria, Goodman (1973:94) suggests deciding for the assumption with the "better entrenched" predicate, e.g. for "all emeralds are green" rather than "... are grue", where "grue" "applies to all things examined before t just in case they are green but to other things just in case they are blue". But this argument is at best relevant if we don't admit any contextual knowledge. Why should we, on the expense of the precision of our predictions, allow all the emeralds having a specified crystal lattice to be either green or blue or to change their "output", i.e. the spectrum of the light reflected?

Literature

Campbell, Donald T. 1987 "Evolutionary Epistemology", in: Gerard Radnitzky and W.W. Bartley (eds.), *Evolutionary Epistemology, Rationality, and the Sociology of Knowledge*, Chicago and La Salle: Open Court, 47 – 89.

Coombs, Clyde H. 1984 "Theory and Experiment in Psychology", in: Kurt Pawlik (ed.), *Fortschritte der Experimentalpsychologie*, Berlin – Heidelberg: Springer, 20 – 30.

Fenk, August 1997 "Representation and Iconicity", *Semiotica* 115, 3/4, 215 – 234.

Fenk, August 2000 "Dimensions of the evolution of knowledge systems", *Abstracts of the VIth Congress of the Austrian Philosophical Society*, June 1 – 4 in Linz.

Fenk, August 1992 "Ratiomorphe Entscheidungen in der Evolutionären Erkenntnistheorie", *Forum für Interdisziplinäre Forschung* 5(1), 33 – 40.

Fenk, August and Vanoucek, Josef 1992 "Zur Messung prognostischer Leistung", *Zeitschrift für experimentelle und angewandte Psychologie* 39(1), 18 – 55.

Feyerabend, Paul K. 1962 "Explanation, Reduction, and Empiricism", in: Herbert Feigl and Grover Maxwell (eds.), *Minnesota Studies in the Philosophy of Science* III, Minneapolis: University of Minnesota Press, 28 – 97.

Forster, Malcolm R. 2000 "Key Concepts in Model Selection: Performance and Generalizability", *Journal of Mathematical Psychology* 44(1), 205 – 231.

Goodman, Nelson ^{3}1973 *Fact, Fiction, and Forecast*, Indianapolis – New York: The Bobbs Merrill Company.

Grünwald, Peter 2000 "Model Selection Based on Minimum Description Length", *Journal of Mathematical Psychology* 44(1), 133 – 152.

Hempel, Carl G. 1962 "Deductive Nomological vs. Statistical Explanation", in: Herbert Feigl and Grover Maxwell (eds.), *Minnesota Studies in the Philosophy of Science* III. Minneapolis: University of Minnesota Press, 98 - 169.

Hempel, Carl G. 1968 "Maximal Specificity and Lawlikeness in Probabilistic Explanation", *Philosophy of Science* 35, 116 – 133.

Hempel, Carl G. and Oppenheim, P. 1948 "Studies in the Logic of Explanation", *Philosophy of Science* 15, 135 – 175.

Hoefer, Carl 2007 "The Third Way on Objective Probability: A Sceptic's Guide to Objective Chance", *Mind* 116 (463), 549 – 596.

Huber, Franz 2008 "Assessing Theories, Bayes Style", *Synthese* 161, 89 – 118.

Laszlo, Erwin 1972 "A General Systems Model of the Evolution of Science", *Scientia* 107, 379 – 395.

Laszlo, Erwin 1998 "Systems and societies: The logic of sociocultural evolution", in: Gabriel Altmann and Walter A. Koch (eds.), *Systems - New Paradigms for the Human Sciences,* Berlin – New York: Walter de Gruyter.

Lewis, David 1994 "Humean Supervenience Debugged", *Mind* 103(412), 473 - 490.

Mises, Richard von ^{4}1972 *Wahrscheinlichkeit, Statistik und Wahrheit*, Wien – New York: Springer.

Nagel, Ernest 1961 *The Structure of Science*, New York: Harcourt, Brace, and Company.

Popper, Karl R. ^{6}1976 *Logik der Forschung*, Tübingen: Mohr.

Popper, Karl R. 1987 "Campbell on the Evolutionary Theory of Knowledge", in: Gerard Radnitzky and W.W. Bartley (eds.) *Evolutionary Epistemology, Rationality, and the Sociology of Knowledge*. Chicago and La Salle: Open Court, 115 – 120.

Rogers, Deborah S. and Ehrlich, Paul R. 2008 "Natural selection and cultural rates of change", PNAS Early Edition, 1 – 5.

Die Nichtreduzierbarkeit der klassischen Physik auf quantentheoretische Grundbegriffe

Helmut Fink, Erlangen, Deutschland

1 Optimistische Meta-Induktion

Die Geschichte der Physik ist eine Geschichte fortschreitender Vereinheitlichung ihrer Grundbegriffe. Dies bedarf sogleich der Erläuterung: "Grundbegriffe" sind hierbei nicht unbedingt solche Begriffe, wie sie in einem Aufbau der Physik nach Prinzipien des methodischen Konstruktivismus am Anfang zu stehen haben, nämlich vorwissenschaftliche Beobachtungen, lebensweltliche Handlungen oder elementare Phänomene. Gemeint sind vielmehr die Grundbegriffe "fertiger" Theorien, wie sie sich in einer nachträglichen rationalen Rekonstruktion zeigen. Idealisierung und Formalisierung haben zur Folge, dass diese Grundbegriffe mathematische Begriffe mit physikalischer Interpretation sind.

Die Vereinheitlichung der Physik ist eine theoretische Vereinheitlichung. Die Phänomene bleiben qualitativ verschieden, ihre Beschreibung offenbart jedoch gemeinsame Strukturen. Je größer der Anwendungsbereich einer physikalischen Theorie, desto größer die Reichweite ihrer Grundbegriffe. Prominente Beispiele solcher Grundbegriffe sind die Potentiale der ("phänomenologischen") Thermodynamik, der Massenpunkt der klassischen Mechanik, die Felder des klassischen Elektromagnetismus.

Umfassendere Theorien können speziellere Theorien etwa als Spezialfall oder Grenzfall enthalten (Scheibe 1997, 1999). Letztere sind dann auf erstere "reduziert", d.h. auf noch fundamentalere Grundbegriffe zurückgeführt. Dabei kann ein "semantischer Rest" bleiben, d.h. ein qualitativer Inhalt der spezielleren Begriffe, der aus den umfassenderen alleine nicht ersichtlich wäre. Dieser Rest darf jedoch zur Rahmentheorie nicht in Widerspruch geraten. Beispiele für solche "schwachen" Theoriereduktionen sind die Rückführung der Wärme auf die Molekularbewegung oder der Lichtausbreitung auf den Elektromagnetismus.

Künftige Fortschritte können aus der gegenwärtigen Physik nicht induktiv erschlossen werden. Die ungeheure Erfolgsgeschichte bisheriger begrifflicher Vereinheitlichungen nährt jedoch die Hoffnung auf einen nächsten Schritt. Die Vereinheitlichung ging bisher immer weiter. Es ist daher vernünftig anzunehmen, dass sie es auch in Zukunft tun wird. Diese Maxime bezeichnen wir als Prinzip (oder Hypothese) der *optimistischen Meta-Induktion*. Sie erscheint zumindest dort gerechtfertigt, wo keine offensichtlichen ontologischen Schwierigkeiten lauern: im Bereich der Theorienreduktion innerhalb der Physik.

Da es nur um Theorien geht, sollte das Verhältnis zwischen den Gegenständen der Mathematik und den Gegenständen der Empirie kein Hindernis für einzelne Reduktionen bieten, denn es betrifft alle Theorien gleichermaßen. Und da es nur um Physik geht, sollte die Erklärungslücke zwischen materieller Konfiguration und subjektivem Erleben kein Hindernis sein, denn die Qualia aus der Philosophie des Geistes kommen in der Physik gar nicht vor.

Die erfolgreichsten Rahmentheorien der modernen Physik sind die klassische Physik (einschließlich spezieller und allgemeiner Relativitätstheorie) und die Quantentheorie (einschließlich Quantenfeldtheorien). Die klassische Physik ist sicher nicht universell, wie ihr Scheitern bei Quantenphänomenen zeigt. Ist die Quantentheorie universell?

2 Hoffen und Bangen des Quanten-Universalismus

Die elementaren Bausteine der Materie werden in quantentheoretischen Begriffen beschrieben. Kernphysik, chemische Bindung, Festkörperphysik, Optik ruhen auf quantentheoretischen Erklärungen. Quantenphänomene können zunehmend auch auf mesoskopischer und makroskopischer Skala herbeigeführt werden. Information erscheint in der Sprache der Quanteninformationstheorie in neuem Licht. Empirisch wird die Quantentheorie überall bestätigt, Grenzen ihres Anwendungsbereichs sind nicht in Sicht.

Der Formalismus der Quantentheorie ist mathematisch, also abstrakt. Der Bezug des Formalismus auf die (bzw. eine mögliche) äußere Realität, d.h. die Interpretation der Quantentheorie, ist nicht so offensichtlich wie die Interpretation der klassischen Theorien. Historisch prägend war die Kopenhagener Interpretation. Sie betont die klassische Beschreibung der experimentellen Anordnung, bestehend aus Präparier- und Registrierapparat. Die Quantentheorie ist in dieser Interpretation konzeptionell nicht selbstständig. Sie scheint nicht auf eigenen Beinen zu stehen, sondern auf klassischen Krücken.

Es war ein naheliegendes Unternehmen, die Grenzen der Quantentheorie auszutesten.

Was in immer neuen Anwendungsfeldern gelang, konnte die eigenen Grundlagen auf Dauer nicht aussparen: eine rein quantentheoretische Beschreibung. Die Kopenhagener Sonderstellung der Apparate erschien zunehmend willkürlich, als historisches Relikt, bestenfalls von pragmatischem Nutzen. Die Sonderrolle von "Messprozessen" erregt Misstrauen, erscheint zunehmend angreifbar, als interpretatorisches Kuriosum, schlimmstenfalls mit anthropozentrischer Botschaft.

Die Theorie erlaubt die formale Einbeziehung der Apparate, ihre Hinzunahme als weiteres Quantensystem, ihre Ankopplung mit Verschränkungseffekt ("Prämessung"), die Definition geeigneter Zeigerobservablen und deren alleinige Betrachtung nach Ende der Messwechselwirkung. Die Grundidee dieser *Quantentheorie der Messung* ist alt: Sie geht auf John von Neumann zurück, der sie als Konsistenztest der Theorie ansah. Zahlreiche formale und begriffliche Verallgemeinerungen wurden seither erarbeitet (Busch et al. ²1996), doch die Gesamtbilanz ist ernüchternd: Messungen haben keine Ergebnisse, wenn sie rein quantentheoretisch beschrieben werden!

Quantenzustände liefern Wahrscheinlichkeiten für die möglichen Messwerte, und das beste, was man erzielen kann, ist dass der Quantenzustand des Messapparats für die jeweiligen Zeigerstellungen genau dieselben Wahrscheinlichkeiten liefert. Das würde den Schluss von der Zeigerstellung auf den gemessenen Wert am ursprünglichen System erlauben — wenn es eine eindeutige Zeigerstellung gäbe. Das Superpositionsprinzip für die Zustandsvektoren reiner Quantensysteme verhindert dies aber. Tatsächlich kann man zeigen, dass die Annahme einer Unkenntnisinterpretation für die Wahrscheinlichkeitsverteilung der Zeigerstellungen (d.h. eine Zeigerstellung liegt objektiv vor und ist nur nicht bekannt) mit der Gesamtbeschreibung unverträglich ist (Mittelstaedt 1998).

Die traditionelle Kopenhagener Reaktion bestand in der Konstruktion der *Neumannschen Kette*, d.h. iteriertes Ankoppeln weiterer Teile der Umgebung ggf. bis zum Gehirn des Beobachters, und im Postulat des *Heisenbergschen Schnitts*, d.h. klassische Beschreibung ab einem (nicht genau festgelegten!) Glied dieser Kette. Das Phänomen der Dekohärenz (Joos et al. ²2003) verspricht ein Verständnis des "Klassischwerdens" durch Berücksichtigung der physikalischen Umgebung. Doch der Widerspruch zwischen der linearen Vektorraumstruktur des quantenmechanischen Zustandsraums und der Eindeutigkeit der klassischen Messergebnisse bleibt bestehen. Das *Messproblem der Quantentheorie* ist ungelöst. Es wurde zum Ausgangspunkt hypothetischer Alternativen für die Zeitentwicklung von Quantenzuständen und bizarrer Interpretationsvorschläge. Wir diskutieren sie hier nicht.

3 Quantentheorie im Phasenraum

Die allermeisten makroskopischen Systeme können im Rahmen der klassischen Physik sehr gut beschrieben werden, auch wenn sie aus Quantensystemen bestehen. Historisch waren Begriffe der klassischen Physik im Bohrschen Korrespondenzprinzip wegweisend beim Aufbau der Quantentheorie. Die grundlegenden theoretischen Strukturen von klassischer und Quantenphysik sind zwar nicht gleich, aber auch nicht völlig verschieden.

Der Zustandsraum der klassischen Physik ist der 2n-dimensionale Phasenraum P, wobei n die Anzahl der Freiheitsgrade des betrachteten Systems bezeichnet. Neben die (verallgemeinerten) Orte treten die (verallgemeinerten) Impulse als kanonisch konjugierte Variablen. Zustände sind Wahrscheinlichkeitsdichten w auf P, reine Zustände entsprechen Phasenraumpunkten. Observablen a sind reelle Phasenraumfunktionen. Erwartungswerte sind Phasenraumintegrale der Observablen, gewichtet mit einem Zustand. Die Zeitableitung eines Zustands ist durch seine Poissonklammer {. , .} mit der Hamiltonfunktion gegeben. Darin stecken die Hamilton-Gleichungen der klassischen Mechanik.

Der Zustandsraum der Quantentheorie ist der (für die meisten Systeme unendlich-dimensionale) Hilbertraum H. Reine Zustände sind Vektoren der Länge 1 in H, allgemeine Zustände W sind positive Operatoren mit Spur 1. Observablen sind selbstadjungierte Operatoren A, deren reelles Spektrum die Menge der möglichen Messergebnisse beschreibt. Erwartungswerte sind von der Form Spur(WA). Die Zeitableitung eines Zustands ist durch seinen Kommutator [. , .] mit dem Hamiltonoperator gegeben. Darin steckt die Schrödinger-Gleichung.

Die Betrachtung von Spezial- oder Grenzfallbeziehungen zwischen zwei physikalischen Theorien setzt die Formulierung beider in gemeinsamen Grundbegriffen voraus. Vergleichbarkeit verhindert Inkommensurabilität. Zur Untersuchung der Beziehung zwischen klassischer und Quantentheorie erscheint es sinnvoll, die mathematischen Grundbegriffe der Quantentheorie auf die historisch vertrauteren Phasenraumobjekte abzubilden. Dabei muss die innere Struktur der Quantentheorie erhalten bleiben. Der Phasenraum wird dann zur gemeinsamen formalen Arena von klassischer und Quantentheorie.

Die bekannteste "Übersetzung" dieser Art (*Phasenraum-Darstellung*) ist die Weyl-Wigner-Abbildung. Generell sind alle Vorschriften interessant, die Hilbertraum-Operatoren W bzw. A linear auf Phasenraumfunktionen w bzw. a abbilden, so dass die Erwartungswerte Spur(WA) zu Phasenraumintegralen über wa werden. Dabei können nicht gleichzeitig folgende drei Bedingungen erfüllt sein (Wigner-Theorem):

(i) Linearität der Darstellung
(ii) Positivität der Darstellung, d.h. aus W positiv folgt w positiv
(iii) Randdichtentreue: Integration von w über Impuls bzw. Ort liefert dieselbe Wahrscheinlichkeitsdichte für Ort bzw. Impuls wie W.

In der Tat verfehlt die Weyl-Wigner-Abbildung Eigenschaft (ii): Wigner-Dichten können negativ werden. Es gibt unendlich viele lineare Phasenraum-Darstellungen der Quantentheorie, von denen manche (ii) und manche (iii) verfehlen. Das aus W gewonnene w kann aber nie als gemeinsame Wahrscheinlichkeitsdichte von Ort und Impuls des Quantensystems interpretiert werden: Erzwingt man die Positivität, so zeigt w dafür Verschmierungen ("Unschärfen") im Phasenraum. Kein Wunder: Die Quantentheorie erlaubt keine gleichzeitige Zuschreibung von Orts- und Impulswerten und keine klassischen Bahnen.

Linearität der Darstellung und Strukturerhaltung der Erwartungswertbildung haben zur Folge, dass Operatorprodukte AB nicht einfach auf Funktionenprodukte ab abgebildet werden können. Auch ist die Phasenraum-Darstellung des Kommutators [A,B] im allgemeinen nicht durch die Poissonklammer {a,b} gegeben, sondern im Fall der Weyl-Wigner-Darstellung durch die Moyalklammer, und in anderen Fällen durch entsprechende Verallgemeinerungen der Moyalklammer. Die Zeitentwicklung quantenmechanischer Systeme im Phasenraum weicht daher von der klassischen Zeitentwicklung ab.

4 Der klassische Limes: Brücke oder Grenze?

Quanteneffekte machen sich (mindestens) überall dort bemerkbar, wo die relevanten Wirkungen in die Größenordnung des Planckschen Wirkungsquantums h-quer kommen. Diese Naturkonstante kennzeichnet den Anwendungsbereich der Quantentheorie. Im Vergleich zu hinreichend großen Wirkungen erscheint sie vernachlässigbar klein. Man erwartet in solchen Fällen die Konvergenz quantentheoretischer Voraussagen, etwa Werteverteilungen geeigneter Messgrößen, gegen die Voraussagen der klassischen (statistischen) Mechanik. Formal wird dabei der Limes h-quer gegen Null gebildet (*klassischer Limes*). Das gelingt für viele physikalisch interessante Situationen (Scheibe 1999).

Theorienreduktion heißt aber mehr: Struktur und Interpretation der gesamten reduzierten Theorie sollen in der reduzierenden aufgehen. Im klassischen Limes sollte die Quantentheorie insgesamt in die klassische Theorie übergehen. Und in der Tat verschwinden Kommutatoren [A,B] inkompatibler Quantenobservablen für h-quer gegen Null, die Struktur der Observablenmenge wird kommutativ, also klassisch. Die optimistische Meta-Induktion, gestützt durch das Parallelbeispiel des nicht-relativistischen Limes, scheint Recht zu behalten.

In den linearen Phasenraum-Darstellungen werden die verallgemeinerten Moyalklammern in diesem formalen klassischen Limes alle zur Poissonklammer und die Zustandsmengen werden alle zur Menge der Wahrscheinlichkeitsdichten auf P, also der klassischen Zustandsmenge. Es scheint, dass sich die zugehörige Interpretation dabei kontinuierlich mitverändern müsste: von einer Welt objektiver Quantenunbestimmtheit über einen Bereich immer kleinerer Unschärfen bis hin zur Welt der klassischen Objekte mit ihren durchgehenden Wertebelegungen aller Observablen (wie z.B. klassischen Bahnen). Der klassische Limes verspricht einen sanften Übergang in die klassische Welt.

Sieht man vom Rahmen der Präparier- und Registrierapparate ab und beginnt die Betrachtung mit der reinen Struktur der Quantentheorie, dann wird der Gegenstandsbereich ihrer Voraussagen im Limes klassisch. Doch für den Messprozess selbst existiert diese Brücke nicht: Hier besitzt die Gesamtbeschreibung der physikalischen Situation eine *semantische Unstetigkeit*, die schon in den Denkvoraussetzungen der Beschreibung steckt und durch Umskalierungen ihres Inhalts nicht beseitigt werden kann. Quanteneigenschaften sind objektiv unbestimmt, Messergebnisse liegen aber als Fakten vor und sind dann objektiv festgelegt. Quantentheoretische Möglichkeiten (etwa Strahlengänge von Photonen) kann man rekombinieren, klassische Daten stehen hingegen fest (und bestehen als Dokumente über die Zeit fort). Das sind qualitative Unterschiede, die nicht eingeebnet werden können.

Die Quantentheorie begegnet der klassischen Theorie also zweimal: einmal als Grenzfall, aber ein andermal als begriffliche Voraussetzung der eigenen Interpretation. Im einen Fall bildet der klassische Limes eine Brücke, im zweiten ist er gar nicht sinnvoll. Die makroskopische Unterscheidbarkeit der Zeigerstellungen macht ja gerade das Spektrum der verschiedenen quantentheoretischen Möglichkeiten sichtbar. Das Faktum des Messergebnisses entsteht dabei unstetig, nicht in einem Limes. Das Faktum ist das abrupte Ende der quantentheoretischen Beschreibung. *Das Faktum bleibt dem Quantum äußerlich.* Der Übergang zur klassischen Beschreibung ist hier eine Grenze der Quantentheorie, nicht ihr Grenzfall.

5 In der Sprache der Quantenlogik

Die Quantenlogik (Mittelstaedt et al. 2005, Kapitel 13) untersucht die Ordnungsstrukturen möglicher Aussagen über Quantensysteme. Alle strukturellen Kennzeichen der Quantentheorie spiegeln sich in ihren Begriffen wider. Der zentrale Strukturbegriff ist dabei der quantentheoretische *Aussagenverband* L(H). Er ist *nicht-Boolesch* (nicht-distributiv) und entspricht dem Verband der Teilräume des Hilbertraums H. Jeder solche Teilraum steht für eine mögliche elementare Aussage (Zuschreibung einer möglichen Eigenschaft). Das Superpositionsprinzip der Zustandsvektoren und die Inkompatibilität von Quantenobservablen werden durch den nicht-Booleschen Charakter von L(H) ermöglicht.

Welches Bild ergibt sich, wenn die klassische Theoriestruktur in diesem begrifflichen Rahmen betrachtet wird? Klassische Theorien sind durch *Boolesche* Aussagenverbände gekennzeichnet. Die darin zusammengefassten Aussagen können immer als objektiv wahr oder falsch, Werte von Observablen daher als objektiv vorliegend oder nicht vorliegend aufgefasst werden. *Die klassische Logik ist die Struktur des Faktischen.*

Der Aussagenverband L(H) eines reinen Quantensystems enthält unendlich viele Boolesche Unterverbände B(H). Die Auswahl eines solchen B(H) kann als abstrakter Ausdruck einer Observablenwahl betrachtet werden. In H entspricht dieser Wahl die Einführung einer Superauswahlregel, d.h. die Auszeichnung eines Systems paarweise orthogonaler Teilräume, zwischen deren Elementen keine Superpositionen erlaubt sind.

Die Struktur der Quantenlogik erscheint somit allgemeiner als die Struktur der klassischen Logik: Letztere kann in erstere eingebettet werden und entsteht aus ihr durch Spezialisierung bzw. zusätzliche Forderungen. Solche Untersuchungen sind auch auf die Struktur der Sprache von klassischer und Quantenphysik, jeweils auch auf relativistischer Raumzeit, ausgedehnt worden (Mittelstaedt 1986). Die Hoffnung des Quanten-Universalismus zeigt sich dabei in der Erwartung einer eigenständigen und fundamentalen Quanten-Ontologie, während die klassische Ontologie als für die physikalische Realität eher untypischer Sonderfall gesehen wird.

Doch die Enttäuschung folgt auf dem Fuß: Auch durch diese strukturelle Einbettung kann die klassische Physik nicht auf die Quantentheorie reduziert werden. Denn das Problem der Faktenentstehung bleibt ungelöst. Der quantenlogische Zugang illustriert im Gegenteil die Notwendigkeit einer klassischen Begriffsbasis auf besonders luzide Weise.

6 Von der Not zur Tugend

Ohne klassischen Beschreibungsrahmen (im Sinne der klassischen Logik für Eigenschaftszuschreibungen für Apparate) hängt die Semantik der Quantentheorie in der Luft. Aussagen über quantentheoretische Möglichkeiten beziehen ihre Bedeutung aus den klassischen Fakten, die zu Beginn schon vorlagen (Präparation) oder am Ende eintreten (Messung). Quantenzustände liefern Wahrscheinlichkeiten, deren Bedeutung ohne Bezug auf die relativen Häufigkeiten der dann tatsächlich gefundenen Messwerte gänzlich unklar bliebe. Quantenobservablen beziehen ihre Bedeutungen und Bezeichnungen aus Transformationseigenschaften, die durch ihre klassischen Entsprechungen definiert sind und sich in Symmetrieeigenschaften der Menge ihrer möglichen Messwerte zeigen.

Der unstetige Übergang zwischen Quanten und Fakten (etwa beim Auftreffen eines Photons auf den Schirm) entspricht dem strukturellen Sprung zwischen L(H) und B(H). Nicht historische Relikte klassischer physikalischer Beschreibungen gilt es daher zu bewahren, sondern nur die methodische Grundlage für die Rede von Fakten. Die Quantentheorie liefert sie nicht, sondern setzt sie voraus. Aus diesem Grund müssen klassische Begriffe in Vortheorien (Ludwig ²1990, 2006) zur Quantentheorie verankert bleiben.

Der Quanten-Universalismus ist selbstzerstörerisch: Er will die Reduktion aller Theorien auf die Quantentheorie und entzieht eben dadurch der Quantentheorie die Grundlage ihrer Interpretation. Denn Interpretation heißt gedanklicher Bezug auf eine mögliche Außenwelt. Und dieser Bezug ist ohne Faktenbegriff nicht zu haben.

Die Reduktion der klassischen Physik auf rein quantentheoretische Grundbegriffe scheitert also. Doch das ist kein Ärgernis, sondern eine methodologische Notwendigkeit. Bohr hat das bereits klar gesehen. Wir müssen es wieder sehen lernen.

Literatur

Busch, Paul, Lahti, Pekka und Mittelstaedt, Peter, ²1996 *The Quantum Theory of Measurement*, Berlin: Springer.

Joos, Erich et al. ²2003 Decoherence and the Appearance of a Classical World in Quantum Theory, Berlin: Springer.

Ludwig, Günther ²1990 Die Grundstrukturen einer physikalischen Theorie, Berlin: Springer.

Ludwig, Günther und Thurler, Gerald 2005 *A New Foundation of Physical Theories*, Berlin: Springer.

Mittelstaedt, Peter 1986 Sprache und Realität in der modernen Physik, Mannheim: BI.

Mittelstaedt, Peter 1998 The Interpretation of Quantum Mechanics and the Measurement Process, Cambridge: Cambridge University Press

Mittelstaedt, Peter und Weingartner, Paul 2005 *Laws of Nature*, Berlin: Springer.

Scheibe, Erhard 1997 Die Reduktion physikalischer Theorien. Teil I: Grundlagen und elementare Theorie, Berlin: Springer.

Scheibe, Erhard 1999 Die Reduktion physikalischer Theorien. Teil II: Inkommensurabilität und Grenzfallreduktion, Berlin: Springer.

Interpretability Relations of Weak Theories of Truth

Martin Fischer, Leuven, Belgium

1. Introduction

Axiomatic theories of truth are understood as extensions of a syntactic base theory which is often taken to be Peano Arithmetic, PA. One way to measure the strength of a theory of truth is to take into account which formulas of arithmetic it proofs. Weak theories in this respect are theories that do not prove more than PA itself, theories that are conservative extensions of PA. The concept of conservativity has gained some interest in formalizing philosophical criteria. This is also the case in the debate on truth, in which conservativity is expected to explain the `no substance' claim of deflationism. For theories of truth conservativity over PA alone seems to be a very crude measure since it does not differentiate between different conservative theories of truth which have quite different properties and prove different formulas containing the truth predicate.

A comparison of the truth-theoretic strength of theories of truth is desirable. A direct comparison of the truth-theoretic strength is the subset relation but it is only a partial order so that not all theories can be compared. Another measure of the strength of a theory of truth would be their interpretability relations to other theories especially their interpretability or noninterpretability in PA. The most famous of these interpretability relations is relative interpretability, introduced in (Tarski et al. 1953), and it is a good measure for PA as base theory. On the one hand the less restricted version of local interpretability collapses in this case into relative interpretability. On the other hand Tarskis theorem of undefinability of truth shows that there is no definitional extension of PA by a one place predicate τ, PA (τ), so that PA (τ) proves $\tau(\varphi) \leftrightarrow \varphi$ for all sentences φ of the language of arithmetic.

2. Axiomatic theories

L_A is the language of arithmetic and $L_\tau := L_A \cup \{\tau\}$. The arithmetical theories are as usual. For the interpretability considerations take the arithmetic theories to be formulated with predicate- instead of functionsymbols.

$$(Ind_P) \quad P(0) \wedge \forall x(P(x) \to P(x+1)) \to \forall x(P(x))$$

Q is Robinson Arithmetic and PA is Peano Arithmetic, that is $Q \cup (Ind_P) \cdot L_A$, where $(Ind_P) \cdot L_A$ is the set of sentences that result from replacing P in (Ind_P) by a formula of L_A with at least one free variable. Accordingly, $I\Sigma_k = Q \cup (Ind_P) \cdot \Sigma_k$.

Assume that L_A contains the relevant syntactical vocabulary: 'Ct' for closed term of L_A, '$Sent$' for sentence of L_A, '$Form_1$' for formula of L_A with one free variable, and so on, such that PA proves the relevant syntactical theorems. Especially if m is the gödelnumber of a formula $\varphi(x)$ with one free variable x and n of a term t, then $m(n)$ is the gödelnumber of the substitution of the free variable x in $\varphi(x)$ by the numeral of t. PA^τ is PA formulated in the language L_τ. A theory of truth T is a L_τ -theory with $PA^\tau \subseteq T$.

$$tot(x) :\Leftrightarrow Form_1(x) \wedge \forall y(\tau(x(y)) \vee \tau(\dot{\neg}x(y)))$$

Disquotational theories of truth are formulated with a scheme of T-biconditionals:

$$(TB_P) \qquad \underline{\tau(\overline{P}}) \leftrightarrow P$$
$$(UTB_P) \qquad \forall x(\tau(P(\dot{x})) \leftrightarrow P(x)).$$

Compositional axioms are the universally quantified versions of the following formulas:

$$(C1) \quad Ct(x) \wedge Ct(y) \to (\tau(x \dot{=} y) \leftrightarrow val(x) = val(y)).$$
$$(C2) \quad Ct(x) \wedge Ct(y) \to (\tau(x \dot{\neq} y) \leftrightarrow val(x) \neq val(y)).$$
$$(C3) \quad Sent(x) \wedge Sent(y) \to (\tau(x \dot{\wedge} y) \leftrightarrow \tau(x) \wedge \tau(y)).$$
$$(C4) \quad Sent(x) \wedge Sent(y) \to (\tau(\dot{\neg}x \dot{\wedge} y) \leftrightarrow \tau(\dot{\neg}x) \vee \tau(\dot{\neg}y)).$$
$$(C5) \quad Sent(\forall yx) \to (\tau(\forall yx) \leftrightarrow \forall z(\tau(x(z)))).$$
$$(C6) \quad Sent(\dot{\neg}\forall yx) \to (\tau(\dot{\neg}\forall yx) \leftrightarrow \exists z(\tau(\dot{\neg}x(z)))).$$
$$(C7) \quad Sent(x) \to (\tau(\dot{\neg}\dot{\neg}x) \leftrightarrow \tau(x)).$$
$$(C8) \quad Sent(x) \to (\tau(\dot{\neg}x) \leftrightarrow \neg\tau(x)).$$

The axiom of internal induction for total formulas is:

$$(I_tI) \quad \forall x(tot(x) \wedge \tau(x(0)) \wedge \forall y(\tau(x(y)) \to \tau(x(y+1))) \to \forall y(\tau(x(y)))).$$

The relevant theories are:

$TB := Q \cup (TB_P) \cdot L_A \cup (Ind_P) \cdot L_\tau$
$UTB := Q \cup (UTB_P) \cdot L_A \cup (Ind_P) \cdot L_\tau$
$PT^r := PA \cup (C1) - (C7)$
$PT^- := PA \cup (C1) - (C7) \cup (I_tI)$
$PT := PA \cup (C1) - (C7) \cup (Ind_P) \cdot L_\tau$
$TC^r := PA \cup (C1) - (C8)$
$TC^- := PA \cup (C1) - (C8) \cup (I_tI)$

TC^r is also known as $PA(S)$ and TC as $T(PA)$.

3. Interpretability

Some basic results:

(i) $TB \subset UTB \subset PT = TC$
(ii) $I\Sigma_1 \cup (C1),(C3),(C5),(C8) \cup (I_tI) = TC^-$
(iii) $I\Sigma_1 \cup (C1) - (C7) \cup (I_tI) = PT^-$
(iv) PT^-, TC^- are finitely axiomatizable.
(v) $TB, UTB, PT^r, TC^r, PT, TC$ are not finitely axiomatizable.
(vi) $TB, UTB, PT^r, TC^r, PT^-$ are conservative extensions of PA.
(vii) TC^-, PT, TC are nonconservative extensions of PA.

Definition
Let S,T be theories formulated in L_S, L_T. Then
T is a *pure extension* of S iff T is an extension of S and $L_T = L_S$.
T is *reflexive* iff T proves Con_Δ for all finite $\Delta \subseteq T$.
T is *essentially reflexive* iff all pure extensions of T are reflexive.
T has *full induction* iff for all formulas φ of L_T : T proves $\varphi(0) \wedge \forall x(\varphi(x) \to \varphi(x+1)) \to \forall x\varphi(x)$.

Full induction and reflexivity are connected in the following way, as shown for example in (Hájek/Pudlák 1993, p.189):

> Lemma 1
> If $PA \subseteq T$ and T has full induction, then T is reflexive.

Let T_I be the minimal theory of truth with full induction: $T_I := Q \cup (Ind_P) \cdot L_\tau$.

Theorem 1
T_I and every pure extension of T_I is essentially reflexive.

Corollary 1
TB, UTB, PT, TC are essentially reflexive.

For other conservative theories with restricted induction it is far more complicated to show that they are reflexive. One example is Tarski's compositional theory with restricted induction. In (Halbach 1999) the conservativity of TC^r over PA is proved by a cut elimination proof. This proof has at least two advantages in comparison to a model theoretic proof along the lines of (Kotlarski et al. 1981). First it can also be used for other base theories especially for all $I\Sigma_k$ with $k \in \omega$. $TC^r(I\Sigma_k) := I\Sigma_k \cup (C1)-(C8)$.

Second it can be formalised in a way that makes it provable in PA. So we get:

Theorem 2
For every $k \in \omega$: PA proves
$\forall x(Sent(x) \wedge \mathrm{Pr}_{TC^r(I\Sigma_k)}(x) \rightarrow \mathrm{Pr}_{I\Sigma_k}(x))$

With this it can be shown that TC^r proves the consistency of all of its finite subtheories.

Theorem 3
PT^r, TC^r are reflexive.

Proof:
Theorem 2 shows that for any $k \in \omega$: PA proves $Con_{I\Sigma_k} \rightarrow Con_{TC^r(I\Sigma_k)}$. Since PA is reflexive, all $I\Sigma_k$ are finitely axiomatizable and $PA \subseteq TC^r$, for every $k \in \omega$: TC^r proves $Con_{TC^r(I\Sigma_k)}$, which is enough to show that TC^r is reflexive. A similar argument shows that PT^r is reflexive. □

Theorem 4
PT^-, TK^- are not reflexive.

Corollary 2
PT^r, TC^r are not essentially reflexive.

For extensions of PA reflexivity and relative interpretability, $\prec$, are connected by Π_1-conservativity in the following way as shown for example in (Lindström 1997):

Theorem 5
Let $PA \subseteq T$. $T \prec PA$ iff T is reflexive and Π_1-conservative over PA.

This shows that:

Theorem 6
TB, UTB, PT^r, TC^r are relatively interpretable in PA.

On the other hand it is easy to see that theories that are not reflexive or Π_1-conservative over PA are also not interpretable in it.

Theorem 7
PT^-, TC^-, PT, TC are not relatively interpretable in PA.

Relative interpretability in PA implies reflexivity and Π_1-conservativity over PA but it does not not imply conservativity over PA.

Theorem 8
$TB + \neg Con_{tb}$ is relatively interpretable in PA but not conservative over PA.

4. Weak Theories of Truth

Considering the set TT of all theories of truth, that is theories formulated in L_τ and containing PA, there are the two subsets with one criterion of weakness:

$CTT := \{ T \in TT \mid T \text{ is a conservative extension of } PA \}$.
$ITT := \{ T \in TT \mid T \prec PA \}$.

The combination of these two criteria allow a more fine grained picture of theories of truth, especially for weak theories. In the preceding sections it was shown that CTT, ITT, their complements and their combinations are nonempty. There are four possibilities of combination. Strong theories of truth not fulfilling either of both criteria will not be investigated here.

The set ITT consists of theories that are deductively weak not only in respect to their arithmetical part but also in respect to their truth theoretic strength. Relative interpretability is sometimes understood as a relation of reduction. The interpretable theories of truth are deductively too weak to be interesting as an explication of a philosophical conception of truth besides a redundancy conception. $CTT \cap ITT$ contains only theories that are also weak in respect to their arithmetical part. Theories of $ITT \cap CTT$, interpretable but nonconservative theories, are not as weak but they are quite artificial. Another reason is that Π_1-conservativity is also a measure of the arithmetical strength and therefore not directly connected to truth-theoretic strength. Interestingly all the theories of ITT are reflexive and not finitely axiomatizable and therefore similar to PA.

Of more philosophical interest are the theories of $CTT \cap ITT$, conservative extensions of PA that are not interpretable in PA. For deflationists conservativity over the base theory is a positive aspect of a theory of truth. It allows truth to be neutral and insubstantial. On the other hand some deflationists claim that the truth predicate fulfills an irreducible expressive function. So it would be an advantage if a theory of truth is deductively strong in respect to its truth-theoretic part. The noninterpretability of a theory in PA would be an indicator that the truth-theoretic part of the theory cannot be ignored. The theories of $CTT \cap ITT$ are also of help in extracting the essentials of truth without influence of their arithmetical part.

The set $CTT \cap \overline{ITT}$ is important for deflationism, but not every theory that is an element of this set is as good as any other. A further investigation which gives more criteria would be of interest. None of the theories in $CTT \cap ITT$ are reflexive and some of them like PT^- are finitely axiomatizable. In this respect PT^- bears a resemblance to ACA_0. There is more than just a resemblance, the two theories are equivalent in the following sense: PT^- is a subtheory of a definitional extension of ACA_0 and the other way around. This can be seen as an argument for the 'naturalness' of PT^-. PT^- is also in other respects promising. Since it contains compositional axioms and a form of induction for formulas with the truth predicate the usual examples to show the deductive weakness of deflationist theories do not obtain.

It is an interesting open question if there are well motivated truth-theoretic sentences not provable in PT^{-}.

Literature

Hájek, Petr and Pudlák, Pavel 1993 *Metamathematics of First-Order Arithmetic*, Berlin: Springer.

Halbach, Volker 1999 “Conservative Theories of Classical Truth”, *Studia Logica* 62, 353-370.

Kotlarski, Henryk, Krajewski, Stanislaw, and Lachlan, Alistair 1981 “Construction of Satisfaction Classes for Non-Standard Models”, *Canadian Mathematical Bulletin* 24, 283-293.

Lindström, Per 1997 *Aspects of Incompleteness*, Berlin: Springer.

Tarski, Alfred, Mostowski, Andrzej, and Robinson, Raphael M. 1953 *Undecidable Theories*, Amsterdam: North-Holland Pub. Co.

Does Bradley's Regress Support Nominalism?

Wolfgang Freitag, Konstanz, Germany

One of the standard arguments against realism about universals is based on Bradley's regress. According to this argument, realism about universals is committed to a vicious regress of instantiation relations. If realism is false and nominalism the only alternative, then, so the argument concludes, nominalism is correct. The strength of this argumentation depends on three things: (1) that commitment to Bradley's regress makes a position untenable; (2) that nominalism as the only alternative to realism is not committed to the regress; and, most importantly, (3) that realism is committed to the regress.

I have three aims in this paper. My proximate aim is to show that if (3) is correct then (2) is incorrect: if the realist is committed to Bradley's regress then so is at least one version of nominalism, namely, trope theory. The demonstration that neither theory is committed to the regress (and hence that (3) is false) is my second aim, attained by the proof that these positions have no commitment to a condition which is generally (and rightly!) held to be necessary for Bradley's regress. As I move along, I shall also claim that there is a widely ignored second condition necessary for the regress, to which – again – neither nominalism nor realism has any commitment. The upshot is this: Bradley's regress problem is independent of the problem of universals. I conclude with an attempt to explain why many philosophers have been misled into thinking otherwise.

1. The regress argument, realism and nominalism

Here, I shall discuss solely nominalism and realism concerning universals, which are understood to be nonrelational or relational properties.[1] For the sake of simplicity, I will focus on nonrelational properties.

Following the tradition, I take realism about universals to be the view that different objects may have the very same, repeatable property. If both the bike and the car are black, then the realist says there is one and the same property, blackness, instantiated by both the bike and the car. Thus, according to realism about universals, a single property may be multiply instantiated in a given world. Nominalism denies this. If the bike and the car are black, then they do not literally speaking have the same property in common. The class nominalist, for example, considers being black as no more than being an element of a certain class of particulars. Instantiation of a property then reduces to membership in a certain class. The trope theorist assumes properties to be much as the realist thinks them to be, except that they are not repeatable: in a given world, no two particulars have literally the same property.

I have encountered the Bradley argument, employed against realism about universals, frequently in personal discussions, and sometimes in print. A very recent formulation of the argument by Gonzalo Rodriguez-Pereyra, a proponent of nominalism, gives me an opportunity to voice my own view on the matter:[2]

> [One argument against universals is this:] Suppose there are universals, both monadic and relational, and that when an entity instantiates a universal, or a group of entities instantiate a relational universal, they are linked by an instantiation relation. Suppose now that *a* instantiates the universal *F*. Since there are many things that instantiate many universals, it is plausible to suppose that instantiation is a relational universal. But if instantiation is a relational universal, when *a* instantiates *F*, *a*, *F* and the instantiation relation are linked by an instantiation relation. Call this instantiation relation *i*2 (and suppose it, as is plausible, to be distinct from the instantiation relation (*i*1) that links *a* and *F*). Then since *i*2 is also a universal, it looks as if *a*, *F*, *i*1 and *i*2 will have to be linked by another instantiation relation *i*3, and so on *ad infinitum*. (Rodriguez-Pereyra 2008)

The argument asserts that instantiation of universals inevitably leads to a regress of ever more instantiation relations, i.e., to what is usually referred to as *Bradley's regress*.[3] The claim that a regress ensues seems to be based on the following two conditions:

(P_u1) Wherever a universal is instantiated, there is an instantiation relation (not identical to one of the *relata*).

(P_u2) The instantiation relation is a universal.

Therefore it seems plausible to attribute to Rodriguez-Pereyra the following line of thought: According to (P_u1), instantiation of a universal demands an instantiation relation. Classifying this instantiation relation as a universal, as done in (P_u2), we are taken back to (P_u1), which then generates another instantiation relation, which together with (P_u2) again takes us back to (P_u1), which generates a further instantiation relation, and so on *ad infinitum*. Rodriguez-Pereyra concludes that realism about universals is in serious trouble. My first aim is to show that if the realist is in trouble, then so is at least one form of nominalism.

One form of nominalism is trope theory. Trope theory distinguishes itself from realism not with respect to the reality of properties, but with respect to the view that properties can be multiply instantiated. Tropes can be instantiated – but only by the sole object having that particular trope. Tropes are "particularised" properties. Now, consider the following pair of conditions:

(P_t1) Wherever a trope is instantiated, there is an instantiation relation (not identical to one of the *relata*).

(P_t2) The instantiation relation is a trope.

[1] Sometimes the dispute is taken to concern not the question of universals in the above sense, but that of the existence of abstract entities. Quine even uses the term 'universal' as synonymous with 'abstract entity'. I shall not enter this different dispute.

[2] For other versions of the argument in print, see Devitt 1980, p. 437, Loux 1998, pp. 38–40, and Moreland 2001, pp. 114–116.

[3] The attribution of such arguments to F. H. Bradley is historically problematic in at least two respects. Firstly, Bradley was concerned with relational properties specifically and not with properties in general. Secondly, he was not the originator of this line of thought. The general type of argument has been known at least since Plato's dialogues. See in particular Parmenides, 127e–130a.

These two conditions differ from (P_u1) and (P_u2) in a single respect only: they contain the term 'trope' where (P_u1) and (P_u2) contain the term 'universal'. If (P_u1) and (P_u2) lead to a regress, then (P_t1) and (P_t2) equally lead to a regress. Instead of speaking of universals or tropes, we can also formulate the matter in general terms, yielding the following pair of conditions:

(P1) Wherever an entity is instantiated, there is an instantiation relation (not identical to one of the *relata*).
(P2) The instantiation relation is an entity.

The regress argument poses a threat only to those who are committed to these two conditions. The trope theorist may deny (P2) as little as the realist. He will understand 'entity' as referring to tropes because he is committed to the view that all relations are particularised relations, hence tropes. A difference between trope theory and realism concerning these conditions can thus at most be given by a difference in commitment to (P1). It will now be shown that there is no such difference.

To see this, we must locate the motivation for (P1), the condition that instantiation demands an instantiation relation. In my view, the motivation lies in the lack of a strict supervenience relation between the existence of the *relata* of instantiation and instantiation itself: given *a* and *F*, it is not determined that *a* instantiates *F*. To illustrate this point, consider the situation in which there are exactly four entities, particulars *a* and *b* and properties *F* and *G*. If we assume that both *a* and *b* individually and contingently instantiate exactly one of the properties *F* and *G*, and nothing else, and if we assume that both *F* and *G* individually are (contingently) instantiated by exactly one of the objects *a* and *b*, and by nothing else, then two situations are possible:

W_1: *a* instantiates *F*; *b* instantiates *G*.
W_2: *a* instantiates *G*; *b* instantiates *F*.

Both situations comprise exactly the same particulars and the same properties. Still, the situations differ; they comprise different facts, different instantiations. This means that the mere existence of particulars and properties does not necessitate a *specific* instantiation. The mere existence of the car and blackness does not necessitate that the car is black. It may still be that the car is green, and what is black is the bike. The existence of particulars and properties may determine *that* facts and instantiations obtain, as some authors (in particular Wittgenstein 1922[4] and Armstrong 1997) maintain. But it does not determine *which* facts, *which* instantiations obtain. As a recent author sums up this point:

> Even if *a* and F-ness cannot exist except in some state of affairs or other, there is nothing in the nature of *a* and nothing in the nature of F-ness to require that they combine with each other to form *a's being F*. (Valicella 2000, p. 238)

Instantiation between two entities does not strictly supervene on the existence of the entities alone, if these entities are considered to be *contingently* related. We need more than the *relata* of instantiation. This need is expressed by condition (P1). (P1) is the reaction to contingent instantiation. The properties *F* and *G* in my example can be understood both as tropes and as universals.[5] It follows that, given contingent instantiation, the trope theorist is as much committed to (P1) as the realist is. David Armstrong has seen this very clearly:

> Suppose that the link between a particular and its tropes is not necessary. Then it is contingent. But if it's contingent, then it seems that we have a clear case of a *relation* between a particular and its trope, and an external relation at that. But then a Bradleian regress ensues […]. (Armstrong 2006, p. 242)

This concludes the argument for my first claim: realism is no more committed to Bradley's regress than at least one form of nominalism, namely trope theory. I now proceed to the argument for my second claim: neither position is committed to the regress.

2. How to avoid Bradley's regress

2.1 Avoiding commitment to (P1)

Contingent instantiation leads to (P1) and starts the regress. In order to avoid (P1), avoid contingent instantiation. Make instantiation necessary. There is a variety of different positions, both nominalist and realist, which conceive of instantiation as being necessary and hence avoid – intentionally or not – commitment to (P1):

(1) One position that makes instantiation necessary is class nominalism. This position, proposed *inter alia* by Anthony Quinton (1957), understands having a property as being a member of a certain class of particulars. The object *a* instantiates *F* iff *a* is a member of the *F*-class. Because classes are identified by their members and class-membership is a necessary relation, instantiation between *a* and *F* strictly supervenes on the existence of the *F*-class alone. In this way, class nominalism can avoid (P1) and thereby the regress. Class nominalism naturally escapes (P1).

(2) Trope theory also has its means of avoiding (P1). In fact, a trope theorist has two options: (2a) Trope theory in combination with a bundle theory of particulars, as defended by, e.g., John Locke and, in more modern times, by D. C. Williams (1953), holds that particulars are sets or bundles of tropes. Consequently, *a* instantiates *F* iff the *F*-trope is in the *a*-bundle. Since the identity of the *a*-bundle is, I take it, defined by the constituting tropes, *a*'s instantiating *F* strictly supervenes on the existence of the *a*-bundle. (2b) The second type of trope theory combines a subject–attribute view with the doctrine of *nontransferable* tropes. A recent proponent of this view is John Heil (2003, chs. 12 and 13), although he prefers the term 'mode' to the term 'trope'. According to this position, a trope is instantiated by the very same object in all possible worlds. Given the nontransferable trope *F* and the particular *a*, the instantiation between *a* and *F* follows by necessity.[6] Again (P1) can be avoided.

[4] Wittgenstein makes this claim with the help of the notion of incompleteness, which he borrows from Frege (1994/1892) but which he applies to all 'objects', properties and particulars alike. Together with the idea that incomplete objects cannot exist on their own, Wittgenstein arrives at his famous view that "[t]he world is the totality of facts, not of objects" (Wittgenstein 1922, 1.1).

[5] That F and G, understood as universals, are, in the case discussed, instantiated only by a single entity, is not of relevance here. To see this, simply change the example accordingly.

[6] This is simplified. There are at least three conceptions of the nontransferability of tropes: (i) F is instantiated in all possible worlds, and it is instantiated in all possible worlds by a. This presumably implies that a must exist in all possible worlds. (ii) F is not instantiated in all possible worlds, but where it is, it is instantiated by a. Option (ii) comes in two varieties: (a) in those worlds in which F is not instantiated, a does not exist; (b) in some worlds in which F is not instantiated, a does exist. The supervenience claim in the main text holds only for (i) and (ii.a).

(3) *Mutatis mutandis,* realism has the same two options as trope theory: (3a) According to a bundle theory based on universals, of which Bertrand Russell (1948, part 4, ch. 8) is a proponent, particulars are understood as bundles of universals. In this view, *a* instantiates *F* iff *F* is a member of the *a*-bundle. Since *F* is a member of the *a*-bundle necessarily,[7] the instantiation relation between *a* and *F* strictly supervenes on the existence of the *a*-bundle. (3b) The second type combines a substance–attribute view with a theory of *nontransferable* universals. According to this position – maintained by, e.g., David Armstrong (2004a, 2004b and 2006)[8] – that *a* instantiates *F* supervenes on the existence of *a* and *F* alone.[9]

Thus, neither nominalism nor realism is committed to the regress. Three of these positions, namely, (1), (2a) and (3a), agree in understanding instantiation to be constituted by class (or bundle) membership. For them necessity of instantiation – and hence the possible denial of (P1) – is built into the ontological conception of instantiation. For the substance–attribute views (2b) and (3b), necessity of instantiation is a feature additional to the basic conception of instantiation and devised, I presume, specifically to avoid (P1).

All of these five options come with heavy ontological burdens. Ignoring their specific difficulties, I shall mention only the problem which they share: necessity of instantiation makes contingency impossible. Whether the substitutes on offer[10] are satisfactory is at least doubtful. So it is worthwhile to investigate whether there might not be another way out of the regress.

2.2 What is necessary for the regress? – A further condition

So far I have acted as if (P1) and (P2) were sufficient for the regress, with the purpose of showing that realism is no more committed to (P1) than trope theory is, and that in fact neither of the two views is committed to (P1). Thus, I hitherto relied on the analysis of Bradley's regress which seems commonly accepted. Now it is time to show that this analysis is flawed. (P1) and (P2) by themselves do not yet yield Bradley's regress. It is quite obvious but frequently ignored: in order for the regress to obtain, it must be given that the instantiation relation is itself instantiated (by the entities it relates). Otherwise, given an instantiation relation, (P1) does not generate a further instantiation relation. To arrive at a regress, we therefore need the further premise

(P3) The instantiation relation is itself instantiated (by the entities it relates).[11]

Conditions (P1), (P2) and (P3) are jointly sufficient for the regress. Are they also individually necessary? I consider (P2) to be superfluous, since any instantiation relation is *ipso facto* also an entity. (P1) and (P3) are hence jointly sufficient for Bradley's regress. I consider them also individually necessary: (P1) states the demand for an instantiation relation given any instantiation, while (P3) makes certain that this instantiation relation demands further instantiation. Thus (P1) and (P3) constitute, I think, the proper analysis of the basis of Bradley's regress.

Given this analysis, there is a second way of avoiding Bradley's regress: accept (P1) and deny (P3); accept instantiation relations and therefore take the first step of the regress, but block the regress by denying that the instantiation relation is itself instantiated. This option should be the natural path to take for substance–attribute views operating with *contingent* instantiation, theories of types (2b) or (3b) albeit without the unnatural condition that instantiation is necessary. There is no space to develop this option here,[12] yet the fact that (P3) is necessary for the regress should eliminate any remaining doubts: Bradley's regress has nothing to do with the problem of universals.

Conclusion

To show that Bradley's regress is neither specific to nor insurmountable for a realist about universals is one thing. To explain why the opposite view has been so compelling to many, is another. So let me end with a suggestion on this point.

The source is the confusion of two different and logically independent senses of the problem of One over Many. There is the *intraworld* version of the problem, which concerns the question whether different particulars *in a single world* can have the very same property *F*. And there is the *transworld* version of the problem, which concerns the question whether different particulars in *different worlds* can have the very same property *F*.

The traditional problem of universals is the intraworld version of the problem of One over Many. Universals can and tropes can't be multiply instantiated within a single world. Bradley's regress, on the other hand, concerns the transworld problem of One over Many. Transferable entities can and nontransferable entities can't be multiply instantiated across different worlds. Keeping these two versions of the problem of One over Many apart, we get a clearer grip on the demands that a satisfying metaphysical theory must fulfil.

Acknowledgments

This paper draws on results from an ongoing research project commissioned by the Landesstiftung Baden-Württemberg (Germany). I am very much indebted for the support provided.

[7] Again, I assume that the identity of a bundle depends on the elements constituting it.

[8] For Armstrong, not only properties but also particulars are nontransferable; particulars have their properties of necessity. Therefore, Armstrong has two independent means to secure the intended supervenience relation.

[9] As in the case of tropes, there are at least three possible conceptions of the nontransferability of universals. The supervenience claim would have to be restricted to the analogues of (i) and (ii.a).

[10] The best, and perhaps only, known way to achieve this is by replacing transworld identity with a counterpart relation for particulars (as David Lewis (1968 and 1986) and Armstrong (2004b) suggest) or for properties, depending on the demands of the theory. Given a suitable semantics, sentences may turn out to be contingent, although instantiations are necessary.

[11] One of the few to recognize the need for this condition is Loux (1998, p. 38).

[12] In (Freitag 2008) I have further explored this possibility.

Literature

Armstrong, David 1997 *A World of States of Affairs*, Cambridge: Cambridge University Press.

Armstrong, David 2004a *Truth and Truthmakers*, Cambridge: Cambridge University Press.

Armstrong, David 2004b “How do Particulars stand to Universals?”, in: D. W. Zimmerman (ed.), *Oxford Studies in Metaphysics*, Vol. 1, Oxford: Clarendon Press, 139–154.

Armstrong, David 2006 “Particulars have their Properties of Necessity”, in: P. F. Strawson and A. Chakrabarti (eds.), *Universals, Concepts and Qualities: New Essays on the Meaning of Predicates*, Aldershot etc.: Ashgate, 239–248.

Bradley, F. H. 1893 *Appearance and Reality*, Oxford: Oxford University Press.

Devitt, Michael 1980 “ ‘Ostrich Nominalism’ or ‘Mirage Realism’?”, *Pacific Philosophical Quarterly* 61, 433–439.

Frege, Gottlob 1994/1892 “Über Begriff und Gegenstand”, in: G. Patzig (ed.), *Gottlob Frege: Funktion, Begriff, Bedeutung*, Göttingen: Vandenhoeck & Ruprecht, 66–80.

Freitag, Wolfgang 2008 “Truthmakers (are Indexed Combinations)”, *Studia Philosophica Estonica*, forthcoming.

Heil, John 2003 *From an Ontological Point of View*, New York etc.: Oxford University Press.

Lewis, David 1968 “Counterpart Theory and Quantified Modal Logic”, in: David Lewis, *Philosophical Papers*, Vol. 1, Oxford: Oxford University Press, 26–39.

Lewis, David 1986 *On the Plurality of Worlds*, Oxford: Basil Blackwell.

Loux, Michael J. 1998 *Metaphysics: A Contemporary Introduction*, London: Routledge.

Moreland, J. P. 2001 *Universals*, Chesham: Acumen.

Quinton, Anthony 1957 “Properties and Classes”, *Proceedings of the Aristotelian Society* 48, 33–58.

Rodriguez-Pereyra, Gonzalo 2008 “Nominalism in Metaphysics”, entry in the *Stanford Encyclopedia of Philosophy*.

Russell, Bertrand 1948 *Human Knowledge: Its Scope and Limits*, London: Allen and Unwin.

Valicella, W. F. 2000 “Three Conceptions of States of Affairs”, *Noûs* 34, 237–259.

Williams, D. C. 1953 “On the Elements of Being I”, *Review of Metaphysics* 7, 3–18.

Wittgenstein, Ludwig 1922 *Tractatus Logico-Philosophicus*, London and New York: Routledge.

Zeitliche Ontologie und zeitliche Reduktion

Georg Friedrich, Graz, Österreich

Ich werde in diesem Beitrag zwei zusammenhängende Fragen behandeln, nämlich (i.) die Frage nach einer *zeitlichen Ontologie.* Darunter möchte ich eine Ontologie verstehen, genauer, ein zeit-räumliches Kategoriensystem, das die zeitlichen Bestimmungen von Dingen ernst nimmt und sie als primär auffasst. Die zweite Frage ist (ii.) die Frage nach den Möglichkeiten einer *zeitlichen Reduktion*. Beide Fragen hängen insofern zusammen, als auch die zeitliche Ontologie als ein Versuch einer zeitlichen Reduktion gesehen werden kann. Sie unterscheiden sich dadurch, dass sie auf unterschiedlichen Ebenen stattfinden.

Die erste Frage ist die Frage nach der ontologischen Sparsamkeit einer zeitlichen Ontologie. Kategoriensysteme dienen der Einteilung der Wirklichkeit. Man kann die Wirklichkeit in vielfacher Weise einteilen, es kann also mehrere Kategoriensysteme geben. Die zeitliche Ontologie reduziert Kategoriensysteme und vereinfacht auch die Kriterien für die Einteilung der Gegenstände in die Kategorien.

Die zweite Frage geht der zeitlichen Reduktion in einem anderen, größeren Umfeld nach. Die Frage im Hintergrund ist, welche Vorteile es haben kann, die Zeit und den Raum bei der Beschäftigung mit ontologischen Fragen zu berücksichtigen, ihnen somit eine größere Bedeutung einzuräumen und gegebenenfalls problematische Begriffe auf zeitliche Begriffe zu reduzieren.

Ein zeit-räumliches Kategoriensystem

Metakategorien, so will ich hier annehmen, kategorisieren ontologische Kategorien, sie dienen der Einteilung von ontologischen Kategorien. Die Metakategorien, des nun folgenden Kategoriensystems, sind Raum und Zeit. Wie man gleich sehen wird, ist die Zeit die bedeutendere der beiden Metakategorien, weshalb man von einer zeitlichen Ontologie sprechen könnte.

Ich gehe davon aus, dass alle Gegenstände im Raum oder in der Zeit sind bzw. nicht sind, d.h. sie stehen immer in irgendeiner Beziehung zu Raum und Zeit. Ein Gegenstand kann im Raum sein oder auch nicht. Ein Gegenstand kann in der Zeit sein oder auch nicht. Führt man, unter besonderer Berücksichtigung der zeit-räumlichen Bestimmungen von Gegenständen, eine kategorische Einteilung durch, so ergeben sich vier mögliche Kombinationen. Es gibt Gegenstände, die …

(1) weder im Raum, noch in der Zeit sind.
(2) im Raum, aber nicht in der Zeit sind.
(3) nicht im Raum, aber in der Zeit sind.
(4) im Raum und in der Zeit sind.

Der erste, bereits erfolgte Schritt, ist die Bestimmung der obersten Kategorien des zeit-räumlichen Kategoriensystems. Interessant, und gesondert zu erwähnen, sind noch die unter Punkt (3a) und (4a) genannten Sonderfälle (andere Sonderfälle können unerwähnt bleiben). Diese Gegenstände sind ewige Gegenstände, wobei "ewig" in zwei verschiedenen Bedeutungen vorkommt. Es sind Gegenstände, die …

(3a) nicht im Raum, aber in der gesamten Zeit sind.
(4a) im Raum und in der gesamten Zeit sind.

Berechtigterweise kann man nun die Frage stellen, warum man genau diese Einteilung vornehmen sollte. Ist sie willkürlich? In diesem Fall könnte ich die Welt genauso gut durch ein Kategoriensystem einteilen, dessen einzige beiden Kategorien rote und nicht-rote Gegenstände sind. Welche Vorzüge hätte die Einteilung der Welt mit Hilfe eines zeit-räumlichen Kategoriensystems? Die Antwort findet sich in (i.) einer Vereinfachung und Reduzierung der ontologischen Kategorien und in (ii.) einer eindeutigen und vereinfachten Zuordnung der Gegenstände. Das muss kurz erklärt werden.

Zur Vereinfachung der Kategorien. In einem Kategoriensystem der einfachsten Art, das in der Philosophie auch verwendet werden könnte, wird man beispielsweise auf folgende Kategorien treffen: physische Gegenstände, psychische Gegenstände und abstrakte Gegenstände. Diese oder eine ähnliche Einteilung hat zumindest zwei Nachteile.

Einerseits ergeben sich Probleme mit der Abgrenzung der Kategorien. Es ist beispielsweise nicht trivial, dass alle Gegenstände entweder physisch, psychisch oder abstrakt sind. Also führt man die Kategorie der physischen Dinge als undefiniert ein. Ich beginne hier mit den psychischen Gegenständen, man könnte aber genauso gut mit den psychischen oder den abstrakten Gegenständen beginnen. Um zu einem vollständigen Kategoriensystem zu kommen, muss man zwischen physischen und nichtphysischen Gegenständen (komplementäre Eigenschaft), unterscheiden wobei man in einem nächsten Schritt die nichtphysischen Gegenstände mit den abstrakten und den psychischen Gegenständen gleichsetzt. Der zweite Nachteil ergibt sich aus dem ersten, denn hat man einmal die Kategorie der physischen Dinge in der eben genannten Weise eingeführt, so ist dadurch festgelegt, dass physische Dinge nicht psychisch und auch nicht abstrakt sind. Ich möchte nicht behaupten, dass es Dinge gibt, die z.B. physisch und psychisch zugleich sind, aber immerhin könnte es sie geben und ihre Existenz wird auch von einigen Philosophen angenommen (Siehe z.B. Searle 1993, 29f.). Auch zeigt schon der Hinweis auf die Möglichkeit solcher Gegenstände, dass die Einteilung in physische, psychische und abstrakte Gegenstände nicht unproblematisch ist. Abgesehen davon macht eine Einteilung, welche physische Gegenstände als nicht psychisch oder abstrakt einführt, eine Voraussetzung. Man könnte indessen meinen, dass im Rahmen der Kategorisierung, in Anlehnung an die Logik, noch keine Voraussetzungen über die Arten der Gegenstände gemacht werden sollten, um auf diese Weise eine Vorauswahl zu vermeiden.

Die Schwierigkeiten gehen weiter, wenn man versucht anzugeben, was z.B. physische Gegenstände sind. Im Alltag kann eine zumindest vage Vorstellung davon haben, was physische Gegenstände sind. Ein Vorschlag zu Bestimmung von physischen Gegenständen könnte sein, diese als in Raum und Zeit lokalisierbar, als sinnlich wahrnehmbar und als ausgedehnt anzunehmen. Zur Erklärung dessen, was physische Gegenstände sind, greift man also von neuem auf Raum und Zeit zurück. Die Bestimmung der räumlichen und zeitlichen Lokalisierbarkeit ist das, was die unter Punkt (4) genannten Gegenstände ausmacht. Diese ist auch die einzig wesentliche Bestimmung, die übrigen könnten weggelassen werden, denn physische Gegenstände als ausgedehnt zu bestimmen ist redundant. Wenn ein Gegenstand im Raum ist, dann ist er notwendigerweise auch ausgedehnt, was in einem Grenzfall bedeuten kann, dass er einen einzigen Raumpunkt einnimmt. Die sinnliche Wahrnehmbarkeit hingegen ist problematisch. Bäume sind sinnlich wahrnehmbar, Elementarteilchen nicht ohne weiteres. Beides scheinen materielle Gegenstände zu sein. Man könnte statt von sinnlicher Wahrnehmbarkeit auch von sinnlicher Wahrnehmbarkeit mit Hilfsmitteln sprechen. Vielleicht ist aber auch nur eine prinzipielle Wahrnehmbarkeit gemeint. Für gewisse Elementarteilchen gilt zudem, dass nur in indirekter Weise auf sie geschlossen wird. Die Frage ist, ob das für eine prinzipielle Wahrnehmbarkeit ausreicht. Es wird jedenfalls zunehmend komplizierter und die Schwierigkeiten beschränken sich nicht auf die physischen Gegenstände; wenn man nämlich beginnt zu fragen, was psychische und abstrakte Gegenstände sind, kommt man über analoge Überlegungen zu ähnlichen Schlussfolgerungen.

Auf welche Gegenstände trifft man in welchen zeiträumlichen Kategorien? Die Gegenstände, die ich in der folgenden Aufzählung nennen werde, sollten vor allem als Illustration verstanden werden. Die Bestimmungen der Kategorien hingegen gelten uneingeschränkt. Es kann sein, dass sich nicht in allen Kategorien Gegenstände finden werden. Die Unterkategorien müssen an dieser Stelle noch offen bleiben.

(zu 1) Die Kategorie der Gegenstände, die weder im Raum noch in der Zeit sind, deckt sich in etwa mit der Kategorie der abstrakten Gegenstände, man beschreibt diese vielleicht besser als atemporale oder zeitlose Gegenstände, da diese Gegenstände, einmal vorausgesetzt sie existieren, unabhängig von bzw. außerhalb der Zeit existieren. Gott, die Idee des Guten, die Zahl 10 und Dodekaeder, sind vermutlich atemporale Gegenstände. A-temporale Gegenstände sind ewige Gegenstände einer ersten Art, sie sind dem zeitlichen Werden und Vergehen nicht ausgesetzt, sie sind auch unveränderlich.

(zu 2) Unter diese Kategorie fallen Gegenstände, die im Raum, aber nicht in der Zeit sind; hier wird man grundsätzlich keine Gegenstände finden können, da alles, was im Raum ist, auch immer schon in der Zeit ist. Dies scheint mir die passende Gelegenheit zu sein, die Frage zu stellen, ob die Zeit gegenüber dem Raum primär ist. Sie wird gleich beantwortet werden.

(zu 3) Gegenstände, die nicht im Raum, aber in der Zeit sind, sind daran zu erkennen, dass die Frage nach ihrem Ort sinnlos ist. Hingegen kann man angeben, wann sie sind. Diese Kategorie umfasst Gegenstände wie Seele, Bewusstsein, Vorstellungen, fiktive Gegenstände. Sie haben irgendwann einen Anfang und ein Ende. Ich meine, es ist ein Vorteil dieser Einteilung und zugleich ein Ergebnis der zeitlichen Reduktion, dieselbe Kategorie für Bewusstsein und fiktive Gegenstände zu haben, denn schließlich werden letztere durch die Tätigkeit des Bewusstseins geschaffen. Sie sind, wie ich meine, auch nicht sonderlich voneinander verschieden. Ihr Unterschied ist vielmehr der Unterschied von öffentlichen bloßzeitlichen Gegenständen und privaten bloßzeitlichen Gegenständen.

(zu 3a) Einige der Gegenstände, die nicht im Raum, aber in der Zeit existieren, könnten die ganze Zeit über existieren. Diese Gegenstände wären ewige Gegenstände im Sinne von zeitlich-ewigen Gegenständen, beispielsweise eine unsterbliche Seele.

(zu 4) Gegenstände, die in Raum und Zeit existieren, sind einerseits das, was man üblicherweise als materielle Gegenstände bezeichnen würde, also Hunde, Menschen, Planeten, Elementarteilchen, aber andererseits auch Gegenstände, die sich nicht so einfach der Kategorie der materiellen Gegenstände zurechen lassen, wie z.B. Ereignisse und Magnetfelder.

(zu 4a) Man könnte sich vorstellen auf Gegenstände zu treffen, die im Raum lokalisierbar sind, und zwar zu allen Zeitpunkten. Man könnte von ewigdauernden bzw. die-ganze-Zeit-über-seienden Gegenständen sprechen. Die Atome Demokrits sind sicherlich Gegenstände dieser Art, denn sie sind unentstanden und unvergänglich.

Die Zuordnung der Gegenstände ist, wie bereits erwähnt, nicht endgültig, sondern Diskussionsgegenstand. Denn die Zuordnung der Gegenstände hängt vor allem von den Bestimmungen der Gegenstände selbst und nicht ausschließlich von der kategorischen Einteilung ab. Als Beispiel: Der christliche Gott ist wahrscheinlich ein Gegenstand der Kategorie (1), Zeus könnte ein Beispiel von (4) sein. Bewusstsein habe ich (3) zugeordnet, Materialisten würden Bewusstsein vermutlich in (4) einordnen.

Noch einige Anmerkungen zu den Besonderheiten dieses Kategoriensystems. Es gibt Gegenstände, die in der Zeit sind und nicht im Raum, aber es gibt keine Gegenstände, die im Raum sind und nicht in der Zeit. Die Frage ist, warum das so ist. Die Antwort könnte diese sein.

> Wenn ich a priori sagen kann: alle äußere Erscheinungen sind im Raume, und nach den Verhältnissen des Raumes bestimmt, so kann ich aus dem Prinzip des inneren Sinnes allgemein sagen: alle Erscheinungen überhaupt, d. i. alle Gegenstände der Sinne, sind in der Zeit, und stehen notwendiger Weise in Verhältnissen der Zeit. (Kant 1974, A34/B51)

Ich meine, das berechtigt zu der Annahme, dass die Zeit die wichtigere Metakategorie ist, weshalb ich von einer zeitlichen Ontologie sprechen möchte, weil die Zeit das jedenfalls enthaltene Element ist.

Teilt man die Gegenstände nach ihrem Verhältnis zu Raum und Zeit ein, so vereinfachen sich sowohl die Kategorien, als auch die Einteilungskriterien, es ist eine ontologisch sparsame Einteilung. Eine solche Einteilung ist sowohl vollständig, als auch eindeutig, d.h. es lassen sich alle möglichen Gegenstände erfassen und sie können zweifelsfrei einer Kategorie zugeordnet werden.

Die zeitliche Ontologie ist auch einfacher in dem Sinn, in dem sie besser verständlich ist. Die Bedeutung von Ausdehnung in Raum und Zeit scheint mir völlig unproblematisch. Hingegen kann man, und das haben die Beispiele gezeigt, darüber uneins sein, was physische, abstrakte oder mentale Gegenstände sind. Zu erwähnen ist noch, dass die zeitliche Ontologie weniger Voraussetzungen macht.

Zeitliche Reduktion

Entgegen der allgemeinen Tendenz, die Zeit selbst auf vielfältige Weise zu reduzieren – zu erwähnen wäre hier zumindest die logische Reduktion, die soziale Reduktion, die physikalische Reduktion und die psychologische Reduktion der Zeit – kann die Zeit ihrerseits dazu verwendet werden ontologische Kategorien zu reduzieren und zu vereinfachen, wie im ersten Teil dargestellt wurde. Des Weiteren eröffnet die Berücksichtigung der Zeit bei der

Betrachtung ontologischer Fragen, Möglichkeiten, problematische Begriffe im Sinne einer ontologischen Sparsamkeit zu reduzieren. Der ontologische Status der Zeit selbst kann zu diesem Zweck noch ungeklärt bleiben.

Ein erstes, hier nur am Rande erwähntes Beispiel für eine zeitliche Reduktion, ist die *zeitliche Reduktion der Modalitäten*. Sie besteht in der Zurückführung der Begriffe “notwendig”, “kontingent” und “möglich” auf zeitliche Begriffe. (Siehe z.B. Rescher und Urquhart 1977, 125ff.). Dieser Ansatz ist für viele Verwendungsweisen der genannten Begriffe nicht unplausibel; so wird “kontingent” interpretiert als “(Es ist der Fall, dass p aber es war nicht immer der Fall, dass p) oder (es ist der Fall, dass p aber es wird nicht immer der Fall sein, dass p).” Zugegebenermaßen sind die Deutungen der Begriffe “möglich” und “notwendig” nicht ganz so unproblematisch. Sie haben aber in jedem Fall den Vorteil der Einfachheit und Klarheit.

Eingehender möchte ich die *Methode der zeitlichen Fragmentierung* betrachten. Alle Gegenstände stehen in irgendeiner Relation zur Zeit. Gegenstände, die in der Zeit sind, können zeitlich fragmentiert gesehen werden. Zeitliche Fragmentierung kann eine Methode der Reduktion sein, wenn es sich um komplexe Gegenstände handelt, die sich in der Zeit mehr oder wenig schell verändern und zudem Gegenstände sind, die man in seiner Ontologie vermeiden möchte. Der Punkt ist, dass man, anstatt von komplexen Gegenständen zu sprechen, nur mehr von Individuen und ihren Eigenschaften zu gewissen Zeitpunkten spricht.

Für die zeitliche Fragmentierung in Frage kommen z.B. komplexe Systeme; auch diese sind Gegenstände, die man irgendwie in einem ontologischen Kategoriensystem unterbringen sollte. Ein Beispiel für ein komplexes System könnte ein soziales System, ein Staat, sein. Die Frage ist, welche Art Gegenstand ein Staat ist. Staaten können als erweiterte Personen gesehen werden, es gibt Versuche sie in Begriffen von Individuen zu definieren oder man könnte sie als logische Konstruktionen verstehen. (Vgl. Prior 1937) Als logische Konstruktionen, und das ist Priors Position, sind Staaten fiktive, unwirkliche Gegenstände. Als solche sind Aussagen über Staaten unter Umständen ersetzbar durch eine Reihe von ähnlichen Aussagen über Individuen, wobei allerdings zu beachten ist, dass dieselben Begriffe in Aussagen über Individuen eine andere Bedeutung haben, als in Aussagen über Staaten.

Über komplexe Gegenstände, wie Staaten, kann man Dinge sagen, die zu ontologischen Verpflichtungen zu führen scheinen. Staaten schließen beispielsweise Verträge ab oder – Priors Beispiel – sie führen Kriege.

> The statement that “England made war on France” […] is not equivalent to “Tom made war on France, Dick made war on France, Harry made war on France, etc.”, but to a set of statements like “Tom made a belligerent speech in the House of Commons”, “Dick dropped a number of bombs on a queue of Parisian women and children”, and “Harry was put in prison for being a conscientious objector”. (Prior 1937, 296)

Ich meine, dass man von Priors Vorschlag ausgehen kann, jedoch sollte man nicht von logischen Konstruktionen sprechen, denn sogleich kann man fragen, was fiktive, unwirkliche Gegenstände sind. Anstatt von England und Frankreich zu sprechen, sollte man, wie Prior meint, von einzelnen Vorfällen sprechen. Hinzuzufügen ist, dass man diese durch eine zeitlich fragmentierte Beschreibung erfassen sollte, und zwar in demjenigen Raum und Zeitintervall, das jeweils interessant ist. Eine zeitliche fragmentierte Beschreibung ist eine erschöpfende Aufzählung aller involvierten Gegenstände und ihrer Eigenschaften zu allen Zeitpunkten. Eine solche Beschreibung würde sehr komplex werden, aber man muss sie nicht durchführen, sondern es reicht aus zu wissen, wie man sie durchführen könnte. (Vgl. Quine, 2002, 282) Und man kann weiterhin über Staaten sprechen.

Es wäre zu überlegen, ob man die zeitliche Fragmentierung auch auf andere Gegenstände anwenden könnte, beispielsweise Ereignisse. Ereignisse sind Gegenstände, die sicherlich in der Zeit lokalisierbar sind, ebenso im Raum, wenn auch nicht genau. Ereignisse sind ontologisch abhängig von in Raum und Zeit lokalisierbaren Gegenständen. Daher könnte es möglich sein über eine zeitlich detaillierte Beschreibung der implizierten Gegenstände zu einer vollständigen und eliminierenden Beschreibung von Ereignissen zu kommen. Analoges kann man sich für andere Gegenstände überlegen.

Schlusswort

Die Ausgangsfrage ist, ob die Berücksichtigung der Zeit bei ontologischen Überlegungen ein Beitrag zur ontologischen Sparsamkeit sein kann. Ich glaube, das ist der Fall. Einerseits hat sich gezeigt, dass ein zeit-räumliches Kategoriensystem ontologisch sparsam ist. Andererseits können durch die Berücksichtigung der zeitlichen Dimension mehrere problematische oder umstrittene Begriffe reduziert werden.

Literatur

Kant, Immanuel 1974 *Kritik der reinen Vernunft* (Werkausgabe Band III), Frankfurt/Main: Suhrkamp.

Prior, Arthur Norman 1937 “The Nation and the Individual”, in: *Australasian Journal of Psychology and Philosophy*, 15, 294-298.

Quine, Willard van Orman 2002 *Wort und Gegenstand*, Stuttgart: Reclam.

Rescher, Nicholas und Urquhart, Alasdair 1977 *Temporal Logic*, Wien: Springer-Verlag.

Searle, John R. 1993 *Die Wiederentdeckung des Geistes*, München: Artemis & Winkler.

Why the Phenomenal Concept Strategy Cannot Save Physicalism

Martina Fürst, Graz, Austria

I start elaborating the main line of the *phenomenal concept strategy* concentrating on the *knowledge argument*. Analyzing the Mary-scenario the crucial particularities of phenomenal concepts are worked out. Next, I argue that only an interpretation of phenomenal concepts which *encapsulate* their referents can capture the decisive uniqueness of these concepts. Finally, the defended account is compared with Papineau's *quotational account* of phenomenal concepts. A careful analysis of this account shows that it has consequences which stand in extreme contrast to the target the physicalist phenomenal conceptualist intends to reach.

1. The phenomenal concept strategy

One of the most famous objections to Jackson's *knowledge argument* (Jackson 1986) is the so-called *two modes of presentation-reply*. The basic idea of this reply – which is the possibility that one single, ontological fact can be known under different modes of presentations – can be easily formulated on the level of concepts. This move leads to the notion of *phenomenal concepts* on the one hand and the notion of *physical concepts* (understood in the widest sense) on the other hand. These two sorts of concepts then are treated in analogy to standard cases of co-reference. Hence, according to the two modes of presentation-reply the brilliant scientist Mary possessed all physical concepts, when being confined to her achromatic room, but gained new phenomenal ones, when enjoying her first colour-experience. Obviously, only type-B-materialist (Chalmers 1997), which grant that phenomenal concepts can not be a priori deduced from physical concepts, can adopt the physicalistic *phenomenal concept strategy* (Stoljar 2005). In other words: physicalists, who intend to save an ontological materialism by granting just a conceptual or epistemic gap, developed this interpretation of the knowledge argument to reach their target.

The physicalist phenomenal concept strategy is based on the idea that the particularities of phenomenal concepts can explain why one can not deduce them a priori from physical concepts, although both sorts of concepts pick out one and the same ontological (ex hypothesi *physical*) referent. Hence, with regard to Mary it can be said that no metaphysical entities such as qualia have to be invoked to explain the scientist's new knowledge – it suffices to point out the uniqueness of phenomenal concepts. For this strategy to work, the decisive features of phenomenal concepts have to be elaborated. These particularities will have to explain why phenomenal concepts are *conceptually isolated* (Carruthers, Veillet 2007) from other concepts, but still pick out physical referents.

In the following I will demonstrate that if we take the uniqueness of phenomenal concepts seriously, we have to conclude that they refer to phenomenal entities and therefore the physicalist phenomenal concept strategy fails. I will start working out the crucial particularities of phenomenal concepts: one particularity concerns the concept-acquisition and the other the very nature of such concepts. Importantly, both particularities of phenomenal concepts are such that they indicate phenomenal referents. In a second step, I will analyze one interpretation of phenomenal concepts which seems to describe the crucial particularities of these concepts adequately: Papineau's *quotational account* of phenomenal concepts (Papineau 2002, 2007). A detailed examination of this account will reveal two possible interpretations: the first interpretation is similar to the herein presented account and therefore leads to a dualistic conclusion. The second interpretation fails to explain the decisive features of phenomenal concepts; such as their semantic stability and the closely linked fact of carrying information about qualitative experiences. Hence, Papineau has to choose between accepting that phenomenal concepts do refer to phenomenal referents or defending a view of phenomenal concepts which leave the crucial particularities of phenomenal concepts and therefore also the Mary-scenario unexplained.

2. The *encapsulation relation* explains the particularities of phenomenal concepts

Let me start my investigation analyzing the particularities of phenomenal concepts. Regarding the *concept-acquisition*, the knowledge argument famously illustrated that we can gain phenomenal concepts only under the condition of attentively experiencing their referents. In other words: one has to stand in the extraordinary intimate relationship of *acquaintance* with the referent a phenomenal concept picks out. Hence, when Mary leaves her achromatic environment, sees for the first time the blue sky and is attentively aware of this colour-experience, she gains a new phenomenal concept. Let me explain this process in more detail: the brilliant scientist, who is aware of her very first blue-experience, *discriminates* this experience from all other current experiences. In my opinion it is this act of attentive discrimination which immediately yields a concept referring to this particular, isolated experience. The close link between an experience and the gained conception of it is a crucial point for my further argumentation.

Regarding the *nature* of phenomenal concepts, a careful analysis reveals an *encapsulation relation* between these concepts and the referents they pick out. The notion of an *encapsulation relation* can be considered as fundamental for the presented account. It is based on the idea that the experience itself is the core of the phenomenal concept referring to it. This fact can be explained by the special way of gaining these concepts: when Mary discriminates a new experience she is acquainted with, this process of isolation implies giving the experience itself a conceptual structure and hence forming a phenomenal concept which encapsulates the very experience itself. Obviously according to this account, both the concept and the referent are occurrences in the subject's mind. The intimate link of encapsulation of the referent in the concept has very particular roots and consequences:

One crucial root of the encapsulation is the self-presenting character of the referent, which enables the direct reference of the concept. It is precisely the fact that an experience is self-presenting, i.e. that it serves as its own presentation, which is responsible for our acquaintance and discriminative awareness of it and hence points towards the close link between experience and phenomenal concept.

The decisive consequences of this account are the following: phenomenal concepts pick out their referents directly and in all possible worlds – facts which are due to the internal constitution of encapsulation. Importantly, since the reference of phenomenal concepts is fixed by their constitution and not by external factors, they carry essential information about their referents. Taking the Mary-scenario into account it becomes evident that the relevant information has to be about the *qualitative character* of experiences because it is precisely this sort of information the scientist lacked in her achromatic room and gained when looking at the sky.

3. Examples of alternative accounts of phenomenal concepts

In my opinion solely the encapsulation relation can explain the particularities of phenomenal concepts. Consider, for example, the fact that released Mary gains a new concept which importantly carries information about the very experience she is undergoing. No *demonstrative account* of phenomenal concepts, such as, for example, the one developed by Levin (Levin 2007), can capture this function of phenomenal concepts. Demonstrative concepts typically refer to the item currently demonstrated at and hence their referents differ from one use to another. Contrary to this, my account of phenomenal concepts makes them pick out their referent necessarily and in all possible worlds. Remember, a phenomenal red-concept should necessarily carry information about phenomenal redness to explain the Mary scenario and demonstrative concepts do not meet this constraint.

If we consider *direct recognitional phenomenal concepts* of the sort invoked by Loar (Loar 1997), we are confronted with another sort of problem: obviously our capacities to discriminate experiences outrun our capacities to recognize experiences. Suppose, Mary has an experience of the shade red21 parallel to shade red23 and can discriminate these two shades introspectively. Nevertheless, she may not be able to recognize these shades when she encounters them. According to the recognitional account of phenomenal concepts Mary has no phenomenal concept of red21 or red23, although she is attentively experiencing these shades and at this moment knows, what it is like to see them. I take this to be a quite implausible conclusion.[1]

These considerations illustrate that no account of phenomenal concepts which neglects the intimate link between these concepts and their referents can successfully explain the particularities of the concepts Mary acquires because of her first colour experience. In addition, accounts which take phenomenal concepts and experiences as separate entities, related to each other only causally, face a further problem: as Balog (Balog, forthcoming) points out, on such accounts it is conceivable that a first-person's application of a phenomenal concept is performed even in the absence of the experience it refers to – and this is quite an absurd way of treating phenomenal concepts. For this reasons, let me return to my thesis of phenomenal concepts encapsulating their referents.

4. The dilemma of Papineaus´s quotational account of phenomenal concepts

In the following I want to focus my attention on a physicalist account, which seems to share the herein elaborated interpretation, but draws physicalistic conclusions from this: the so-called *quotational account* of phenomenal concepts. Papineau developed this account in his book *Thinking about consciousness* (Papineau 2002), but changed some details in a recent article (Papineau 2007).

The quotational account is based on the assumption that phenomenal concepts embed experiences just as quotation marks embed words. If his analogy is worked out in detail, we will see why Papineau faces a dilemma: if his account is understood as a sort of real encapsulation, then he has to conclude that phenomenal concepts pick out phenomenal referents. The reasons for this conclusion are the following: if phenomenal concepts are interpreted as encapsulating their referents, then this unique reference relation has to be explained. According to my analysis, solely an explanation referring to the self-presenting character of phenomenal properties and our special acquaintance relation to them can do this explanatory work. If one wants to avoid this dualistic conclusion, she has to give a physicalistic account of how a concept can encapsulate and directly refer to a physical item and it seems mysterious how this can be done without invoking self-presenting (phenomenal) properties.

The remaining option is to interpret the quotational account as phenomenal concepts just *using* experiences without granting that they are a logical part of the concept itself. In fact, Papineau in his article "*Phenomenal and perceptual concepts*" (2007) doesn't seem to believe anymore that the particularities of phenomenal concepts lie in a unique reference relation, but rather holds that they can be explained by the special (neuronal) vehicle in virtue of which the concept is realized. This suggests that the presence of the experience in the concept should be explained by a physical (neuronal) presence:

> We can helpfully think of perceptual concepts as involving stored *sensory templates*. These templates will be set up on initial encounters with the relevant referents. They will then be reactivated on later perceptual encounters. (Papineau 2007, 114)

Obviously the "stored sensory template" has to be understood as a physical item. At this point some pressing questions arise: firstly, what is meant by "involving" these templates? If this phrase only points at *simultaneous occurrence* of concept and experience, then the concept doesn't carry any information about the qualitative character of the experience. If the citation has to be understood as a constitutional relation, one may wonder a) how a physical item (as a neuronal template) can be encapsulated in the concept and b), how it can carry the relevant information. Ad a) it can be pointed out that on a physicalist account no primitive acquaintance relation can be invoked to explain this constitution and that neural templates are not introspectively accessible. Next, b) has to be explained in more detail: the information a phenomenal concept has to carry surely is not information about a neural state – otherwise Mary would have possessed this concept in her achromatic environment. A phenomenal concept has to carry information about the *qualitative character* of the experience and it is unclear how a physically understood template can do this, without recurring to phenomenal properties. A purely physical description of a (neuronal) template would obviously leave out precisely the sort of information a phenomenal concept has to carry to explain Mary's situa-

[1] My way of arguing shows that I take phenomenal concepts to be singular concepts applying to the very occurring experience. According to my approach, only a generalization-process on the basis of singular concepts yields a general phenomenal concept.

tion. Therefore, if Papineau´s account of phenomenal concepts is interpreted as solely co-occurring with experiences or as involving physical items, then the decisive particularities of the concepts will not be explained adequately anymore.

5. Conclusion

I want to summarize my line of thought: in accordance with most phenomenal conceptualists I showed that the concepts involved in the Mary-case differ in several respects significantly from any other concept the scientist had before her release. But the central point of my analysis – which stands in contrast to target of the physicalist phenomenal conceptualist – was to argue that these differences are such that the new concepts refer (because of their internal structure) necessarily to phenomenal entities.

In a next step, I compared the elaborated account of phenomenal concepts with some physicalistic ones. I demonstrated that the basic assumptions of most physicalist phenomenal conceptualist (as Levin or Loar) can not explain the crucial particularities of phenomenal concepts. Then I focused the attention on the quotational account advocated by Papineau which at first glance seemed to describe these particularities adequately. But a careful analysis illustrated that also Papineau's account has consequences which stand in contrast to the target the physicalist intends to reach: if it is understood as just involving physical items, then it can not meet the constraint of explaining the decisive particularities of phenomenal concepts; such as carrying information about the qualitative character of experiences. But if it is interpreted in accordance with the herein advocated encapsulation relation, then it has exactly the dualistic consequences the physicalist phenomenal conceptualist wants to avoid.

Literature

Balog, Katalin forthcoming "Phenomenal concepts" in: McLaughlin, Brian, Beckermann,

Ansgar (eds.) *The Oxford Handbook in the Philosophy of Mind*, Oxford: Oxford University Press.

Carruthers, Peter and Veillet, Benedicte 2007 "The phenomenal concept strategy", in: *Journal*

of Consciousness Studies, 14 (9-10), 212-236.

Chalmers, David 1997 "Moving forward on the problem of consciousness", in: *Journal of*

Consciousness Studies, 4 (1), 3-46.

Jackson, Frank 1986 "What Mary didn't know", in: *Journal of Philosophy*, 83, 291-25.

Levin, Janet 2007 "What is a phenomenal concept?" in: Alter, Torin, Walter, Sven (ed.) (2007) *Phenomenal concepts and phenomenal knowledge*, Oxford: Oxford University Press,

87-111.

Levine, Joe 1983 "Materialism and qualia: The explanatory gap", in: *Pacific Philosophical*

Quarterly, 64, 354-361.

Loar, Brian 1997 "Phenomenal states: Second version", in: Block, Ned et.alt (eds.) *The nature of consciousness: Philosophical debates,*

Cambridge/MA: MIT Press, 597-616.

Papineau, David 2002 *Thinking about consciousness*, Oxford: Oxford University Press.

Papineau, David 2007 "Phenomenal and perceptual concepts", in: Alter, Torin, Walter, Sven

(Hg.) (2007) *Phenomenal concepts and phenomenal knowledge*, Oxford: Oxford University Press, 111-145.

Stoljar, Daniel 2005 "Physicalism and phenomenal concepts", in: *Mind, Language and*

Reality, 20, 469-494.

Benacerraf and Bad Company (An Attack on Neo-Fregeanism)

Michael Gabbay, London, England, UK

1 Benacerraf on what numbers could not be

In his celebrated paper, *"What numbers could not be"*, Benacerraf presents a challenge to theories identifying numbers with set theoretic constructs. He asks why the numbers should be identified with sequence (1), the Von Neumann ordinals, rather than sequence (2), the Zermelo ordinals.

$$\varnothing,\quad \{\varnothing\},\quad \{\varnothing, \{\varnothing\}\},\quad \{\varnothing, \{\varnothing\}, \{\varnothing, \{\varnothing\}\}\} \qquad (1)$$

$$\varnothing,\quad \{\varnothing\},\quad \{\{\varnothing\}\},\quad \{\{\{\varnothing\}\}\} \qquad (2)$$

Benacerraf concludes that there is no reason why the number 3 should be identified with an element from one construction rather than another. 3 cannot be identified with both as the constructs have incompatible properties. For example in (1) the fourth element has three members, but in (2) the third element has only one member. Since the number 3 cannot be both $\{\varnothing, \{\varnothing\}, \{\varnothing, \{\varnothing\}\}\}$ and $\{\{\{\varnothing\}\}\}$ and there is no fact of the matter whether it is one or the other, it is neither. Thus the attempted identification of numbers with sets has been refuted.

> … if the number 3 is really one set rather than another, it must be possible to give some cogent reason for thinking so. But there seems to be little to choose among the accounts *for the accounts differ at places where there is no connection whatever between features of the accounts and our uses of the words in question*. [Benacerraf 1965]

It is not hard to see that this objection generalises to any theory of numbers that has an ontology containing different sequences of objects that could serve as references of our number language. Realists may escape Benacerraf's argument either by finding a suitably miserly ontology of abstract objects (the ontology of sets is too vast), or simply refusing to get involved in the metaphysics of abstract objects.

I shall argue that Neo-Fregean ontology suffers from Benacerraf's objection in much the same way as the ontology of sets. I conclude, analogously to Benacerraf's original argument, that Neo-Fregean ontology is necessarily too rich and therefore does not provide a satisfactory foundation for arithmetic.

First I shall sketch the Neo-Fregean account of arithmetic, I shall assume that the reader is largely familiar with the formal concepts behind it (in particular, I assume the reader has some knowledge of the workings of *Frege's Theorem* [Wright 2000]).

2 Neo-Fregeanism on what numbers could be

2.1 Hume's principle

The aim of the Neo-Fregean programme is to provide a metaphysics of abstract objects together with an informative account of our epistemic link to them. According to Neo-Fregeanism, reference to the abstract objects that are *the numbers* derives from logic and definitions alone. Logic then entails arithmetic truths and, in this sense, arithmetic is *analytic*.

Neo-Fregeanism promises to provide a realist theory of number that can respond to Benacerraf's argument. According to Neo-Fregeanism, certain abstract objects exist, and we can know and refer to them via *abstraction principles*. The natural numbers are among those abstract objects we can know about via a particular abstraction principle, *Hume's Principle*:

> The number of *F* = the number of *G* iff
> the *F* and the *G* are in one-one correspondence (3)

For each predicate *F,* Hume's principle identifies or allows reference to, an object that is the number of *F*. This formalisation should be familiar to the reader:

$$\forall F \forall G[nx.Fx = nx.Gx \leftrightarrow F1\text{\textasciitilde}1G] \qquad (4)$$

Hume's principle is to be taken as a *definition*, in terms of one-one correspondence, of the binding term-former *nx*.(…). Furthermore, the Neo-Fregeans argue that one-one correspondence is a fundamental application and concept of cardinal numbers. So the abstract entities, reference to which is generated by Hume's principle, really are the cardinal numbers (they are the only abstract entities tied appropriately to the application of counting).

2.2 Frege's theorem

I now sketch how Neo-Fregeans use Hume's principle to provide a realist foundation for arithmetic.

Following Frege, the strategy is to define suitable properties and relations that satisfy the second order Peano axioms of arithmetic. To distinguish the defined terms of this section with the defined terms of Section 3.2, I subscript them with *H* for 'Hume'.

First a successor/predecessor relation is defined:

$$Pre_H(t, t') \text{ means } \exists F \exists z[t' = nx.Fx \wedge Fz \wedge t = nx.(Fx \wedge x \neq z)] \qquad (5)$$

So *t* is the predecessor of *t'* when *t'* is the number of some property *F* and *t* is the number of the *F*s that are not *z*, for some *z*.

Zero is defined to be the number of any inconsistent property, e.g. $0_H = nx.(x \neq x)$, it does not matter which as all empty properties are in one-one correspondence.

A natural number is then defined as being any number in the transitive closure of the predecessor relation from 0_H. More formally, the transitive closure of any binary relation *R* may be defined as:

$$R^*(t, t') \text{ means } \forall F\, [\, (Ft \wedge \forall x \forall y(Fx \wedge R(x, y) \rightarrow Fy)) \rightarrow Ft'\,] \qquad (6)$$

And now we may define the natural numbers as all those objects in the transitive closure of the predecessor relation from 0_H:

$$Nat_H(t) \text{ means } Pre_H^*(0_H, t') \qquad (7)$$

So a natural number is any referent of an abstraction $nx.Fx$ that can be 'reached' from 0_H by following the relation Pre_H. As did Frege, Neo-Fregeans go on to define individual number terms:

$$\begin{aligned} &0_H \text{ means } nx.(x \neq x) \\ &1_H \text{ means } nx.(x = 0_H) \\ &2_H \text{ means } nx.(x = 0_H \vee x = 1_H) \ \ldots \end{aligned} \tag{8}$$

From these definitions we can derive all of Second order Peano Arithmetic, which completely characterises a natural number structure.

3 An alternative abstraction principle

3.1 Benacerraf's principle

Now I turn to the argument that the Neo-Fregean ontology contains too many abstract objects. I do this by presenting an abstraction principle that is similar to Hume's principle. This alternative abstraction principle can do the same work as Hume's principle and in a similar way. But, as with the (1) and (2) above, the two abstraction principles yield two distinct sequences of abstract objects. As was argued in the case of sets, I shall argue that there is nothing to decide which abstraction principle yields the 'true' natural numbers.

The new principle is simpler than Hume's principle, call it Benacerraf's Principle:

the unitariness of F = the unitariness of G iff
neither F nor G are singletons, or
F and G have the same extension. (9)

Say that F is unitary if it has exactly one element in its extension. Then Benacerraf's principle identifies, for each predicate F, an object that is the *unitariness* of F. We can write 'the unitariness of F' as $ux.Fx$, and then Benacerraf's principle is:

$$\forall F \forall G\ [\ ux.Fx = ux.Gx \leftrightarrow ((\neg\exists! xFx \wedge \neg\exists! xGx) \vee \forall x(Fx \leftrightarrow Gx))\ ] \tag{10}$$

Call a property, or concept, *unitary* when only one thing is in its extension. Then Benacerraf's principle is to be taken as a *definition*, in terms of being unitary, of the binding operator $ux.(\ldots)$. The intuition for unitarieness is that one can abstract out of a unitary property its 'unitariness', or the way in which it is unitary. Any non-unitary properties are unitary in the same way: they are not. The way unitary properties are differentiated, in the spirit of Frege's Basic Law V (see (14)) is through their extension.

Unitariness is at least as fundamental to our concept of number as one-one correspondences. After all, a one-one correspondence is a correspondence between unit objects; when we count, we count unit individuals; variables of first order quantifiers range over unit entities; the symbols of the language necessary to express even basic propositions are discrete, discernable units. Without the concept of a unit, a discrete thing, a single entity, we cannot even begin a logical enquiry let alone ground arithmetic in one-one correspondences. Frege himself discusses and rejects the possibility of developing a theory of arithmetic based on units. But his compelling refutations are aimed at theories of numbers as agglomerations or sums of (distinct) units [Frege 1953, §29-§44]. Frege objects that such accounts either make no sense, or fail to generate arithmetic. He did not consider the possibility that the unit, thought of as a property of properties, and derived by a similar abstraction method to Frege's own, could do the same work as his favoured theory of number.

3.2 An analogue of Frege's theorem

I now sketch how Benacerraf's principle can be used to define the numbers along Neo-Fregean lines. To distinguish the defined terms of this section with those of Section 2.2 I subscript them with B for 'Benacerraf'. We begin with zero:

$$\begin{aligned} &0_B \text{ means } ux.(x \neq x) \\ &1_B \text{ means } ux.(x = 0_B) \\ &2_B \text{ means } ux.(x = 1_B) \ \ldots \end{aligned} \tag{11}$$

It is not hard to show that the i_B are derivably distinct. For example suppose that $0_B = 1_B$, then $ux.(x \neq x) = ux.(x = 0_B)$. So by Benacerraf's principle either $\neg\exists! x(x \neq x) \wedge \neg\exists! x(x \neq 1_B)$ or $\forall x(x = 0_B \leftrightarrow x \neq x)$. Each of these is derivably false in even first order logic.

We can go on to define the predecessor relation as follows:

$$Pre_B(t, t') \text{ means } \exists F[t = ux.Fx \wedge t' = ux.(x = t)] \tag{12}$$

A version of Frege's theorem now arises out of adopting Benacerraf's principle rather than Hume's principle. We use (6) to define the natural numbers to be exactly the entities in the transitive closure of Pre_B . This yields the second order Peano axioms.

$$Nat_B(t) \text{ means } Pre_B^{*}(0_B, t') \tag{13}$$

I omit the remaining details here as they are almost identical to those of the proof of Frege's theorem in [Wright 1983].

3.3 The attack on Neo-Fregeanism

Let 0_H, 1_H, 0_H… denote the entities abstracted and defined using Hume's Principle, call them the *Hume-numbers*. Let 0_B, 1_B, 0_B… be the entities abstracted and defined using Benacerraf's principle call them the *Benacerraf-numbers*. It should be clear that an analogue of Benacerraf's original challenge arises. Benacerraf's original argument now applies, both the Hume-numbers and the Benacerraf-numbers serve as characterisations of the natural numbers. Furthermore there is no reason for the natural numbers to be identified with the Hume-numbers rather than the Benacerraf-numbers. Therefore Neo-Fregeanism is to be rejected alongside set theoretic reductionism by a variant of Benacerraf's original argument.

The penultimate claim, that there is no choosing between the Benacerraf and the Hume numbers, is in need of justification. I sketch a justification of it in Section 4 by comparing the systems obtained from the two abstraction principles and showing that there is little that can be done with Hume's principle that cannot also be done with Benacerraf's principle.

4 A comparison of two abstraction principles

4.1 How to avoid bad company

Formally, Hume's and Benacerraf's principles are acceptable abstraction principles. I argue for this here by presenting a condition on good abstraction principles (i.e. I offer a solution to the bad company problem) and show that both abstraction principles satisfy it.

A famous worry for the Neo-Fregean project, called the ‘bad company’ problem, relates to the fact that not all abstraction principles are consistent. A famously inconsistent abstraction principle is Frege’s notorious Basic Law V:

$$\forall F \forall G\ [\ \varepsilon x.Fx = \varepsilon x.Gx \leftrightarrow \forall x(Fx \leftrightarrow Gx)\ ] \tag{14}$$

We can use (14) to derive Russell’s paradox. An argument of Heck [Heck 1992] shows that there are many undesirable abstraction principles. For example, there are many Φ for which the abstraction principle:

$$\forall F \forall G\ [\ \varepsilon x.Fx = \varepsilon x.Gx \leftrightarrow \Phi \vee \forall x(Fx \leftrightarrow Gx)\ ] \tag{15}$$

entails that Φ. It is not hard to find plenty of second order sentences Φ (some of which contain *F* and *G*) that are entailed by an abstraction principle like (15) which we would certainly think ought not to be true. Furthermore, different abstraction principles can be incompatible with each other, although individually consistent; this is strange, as abstraction principles are supposed to be analytic and so ought to be true, and hence compatible, in any context. There is then a question whether some principle can be given to discern the acceptable abstraction principles from the unacceptable ones (see [Weir 2003] for many examples of unacceptable abstraction principles). I now present such a principle.

Let *λ* be any infinite cardinal, then the *consistency constraint* for *λ* is the condition that any abstraction principle should have the form:

$$\forall F \forall G\ [\ \varepsilon x.Fx = \varepsilon x.Gx \leftrightarrow \Psi_\lambda(F, G) \vee \forall x(Fx \leftrightarrow Gx)\ ] \tag{16}$$

Where

i. $\Psi_\lambda(F, G)$ is a second order sentence containing no free variables other than *F* and *G*, and also does contain the ‘new’ abstraction operator *εx*.
ii. Ψ_λ is a transitive and symmetric relation on *unary* predicates. That is:
- $\Psi_\lambda(F, G)$ implies $\Psi_\lambda(G, F)$
- $\Psi_\lambda(F, G)$ and $\Psi_\lambda(G, H)$ implies $\Psi_\lambda(F, H)$
iii. For any model *M* of cardinality *λ*, there are at most *λ* many valuations *σ* such that $\sigma(\Psi_\lambda(F, G)) = \perp$.[1]

Note that the familiar examples of ‘bad’ abstraction principles (e.g. in [Weir 2003]) violate this condition. For example in Frege’s Basic Law V has the form

$$\forall F \forall G\ [\ \varepsilon x.Fx = \varepsilon x.Gx \leftrightarrow \perp \vee \forall x(Fx \leftrightarrow Gx)\ ]$$

which clearly violates this condition for any *λ*. Note also that Benacerraf’s principle and Hume’s principle satisfy the consistency constraint for any infinite *λ*.[2] Now we can show that any abstraction principle satisfying the consistency constraint for *λ* can be interpreted in *any* second order model of cardinality at least *λ*. Let M_λ be a model (that can interpret the language of Ψ) with domain $|M_\lambda|$ of cardinality *λ*. Let *R* be a relation on properties (i.e. a relation on subsets of $|M_\lambda|$) such that

$$R(P, Q) \text{ iff } \sigma(\Psi_\lambda(F, G)) = T$$

for any valuation *σ* such that $\sigma(F) = P$ and $\sigma(G) = Q$.[3] In other words, *R* is the interpretation of Ψ_λ in the model M_λ. Since $\Psi_\lambda(F, G)$ contains no free first or second order variables other than *F* and *G*, *R* does not depend on *σ*.

If $P \subseteq |M_\lambda|$ then let $P_R = \{Q\colon R(P, Q) \text{ or } P = Q\}$. Clearly, P_R is an equivalence class. Now consider the set $A = \{P_R\colon P \subseteq |M_\lambda|\}$ and let μ be its cardinality. If $\mu > \lambda$, then there would be more than *λ* many valuations *σ* that falsify $\Psi_\lambda(F, G)$ (at least one for each of the μ-many pairs of different equivalence classes in *A*). This would violate condition (iii) of the consistency constraint for *λ*. So $\mu \leq \lambda$, i.e. the cardinality of *A* is less than or equal to the cardinality of $|M_\lambda|$. It follows then, that there is an injection *f* from$\{P_R\colon P \subseteq |M_\lambda|\}$ into $|M_\lambda|$.

We may use *f* to identify elements $e_P \in |M_\lambda|$:

$$e_P = f(P_R) \tag{17}$$

It is now a straightforward matter to check that

$$e_P = e_Q \text{ iff } R(P, Q) \text{ or } P = Q$$

It follows that we can extend any second order model M_λ of cardinality *λ*, with an abstraction principle satisfying the consistency constraint for *λ*: we define e_P as in (17) and then extend M_λ to interpret the new language using (18):

$$\sigma(\varepsilon x.Fx) = e_{\{m\colon m \in |M_\lambda| \text{ and } \sigma_{[x/m]}(Fx) = T\}} \tag{18}$$

Let me describe this interpretation in English: Ψ_λ forms equivalence classes of properties; the conditions on Ψ_λ guarantee that there is a one-one function *f* from these equivalence classes into the domain $|M_\lambda|$ of M_λ; we interpret the referent of *εx.Fx* under valuation *σ* as the element *e* which the function *f* assigns to the equivalence class of properties that Ψ_λ forms from the extension of *F*.

We now have in (16) a general criterion for the legitimacy of abstraction principles. This criterion legitimates Benacerraf’s principle *as well as* Hume’s principle. The only difference between them being that extending a model to validate Benacerraf’s principle is slightly more straightforward than Hume’s principle. Say that an abstraction principle is *almost analytic* if it satisfies the consistency constraint for any infinite *λ*. It is now a matter of dispute whether the fact that *λ* has to be infinite detracts from the analyticity of Hume’s principle and Benacerraf’s principle. A point in favour of the Neo-Fregean programme is that we can give an independently motivated formal reasons for treating Hume’s principle as analytic and rule out principles like Basic Law V. However, a point against the Neo-Fregean programme is that Benacerraf’s principle also comes out as analytic, and the argument of Section 3.3 stands.

4.2 The concept of number

Perhaps some argument relating to our concept of number will differentiate between the two principles.

The possibility of any such argument is extremely doubtful, if anything but for the fact that neither Hume’s principle nor Benacerraf’s principle make good analyses of our number concepts. Wright acknowledges this:

[1] This says that the (second order) property represented by $\Psi\lambda$ groups the properties of the domain into at most *λ* many different equivalence classes. In other words, $\Psi\lambda$ is only allowed to distinguish up to extensionality, all properties (i.e. subsets of $|M|$) of cardinality $< \lambda$. $\Psi\lambda$ must be unable to distinguish all but *λ* of the 2λ properties of cardinality *λ*.

[2] We must view Hume’s principle as: $\forall F \forall G\ [\ \varepsilon x.Fx = \varepsilon x.Gx \leftrightarrow F1{\sim}1G \vee \forall x(Fx \leftrightarrow Gx)\ ]$

[3] *σ* assigns elements of $M\lambda$ to first order variables and subsets of $M\lambda$ to second order variables; $\sigma[x/m]$ is a valuation that agrees with *σ* on all variables except that it maps the variable *x* to *m*.

> … no one actually gets their arithmetical knowledge by second-order reasoning from Hume's Principle Rather, the significant consideration is that simple arithmetical knowledge has to have a content in which the potential for application is absolutely *on the surface*, since the knowledge is induced precisely by reflection upon sample, or schematic, applications. [Wright 2000]

The schematic applications Wright has in mind is that of drawing one-one correspondences when counting. The thought might then be that Hume's principle has one-one correspondence, the potential for application of number language, 'on the surface' whereas Benacerraf's principle does not.

But counting is not the only application of numbers. An obvious application that has little to do with counting is when we assign numbers to things to help identify them, perhaps in some ordering. For example, rooms in a hotel may be numbered in such a way as to indicate their location in the building. In such an application, room numbers may serve as no indication of how many rooms there actually are. For example room 1729 on the top floor may be so numbered, in part, because it is on the floor numbered 17, which itself is numbered to indicate it is one up from 16 (and there may not even be 17 floors in the hotel, if there is no 13th floor). What is more important to the hotel-room application of numbers is that they can represent individual units in some successive ordering. It is this potential for application that is absolutely 'on the surface' of Benacerraf's principle.

We can quite easily explain the relation between Benacerraf-numbers and one-one correspondence. One-one correspondence is a learned application of Benacerraf-numbers. The adjectival numerical quantifier can be treated in the obvious way, analysing 'there are *n* apples' as:

> there is a one-one correspondence between the apples and the Benacerraf numbers less than n_B. (19)

(where 'less than' is formalised in terms of Pre_B^*, the transitive closure of the predecessor relation on Benacerraf numbers).

Now, who is to say whether drawing one-one correspondences is part of the 'schema' for applying numbers, or whether it is a further application of a simpler schema relating to units and succession? There are reasons to think of numbers being fundamentally tied to one-one correspondence and equally good reasons to think that they are tied to units and succession. But to which, one-one correspondence or units, are numbers *really* tied? I doubt we could answer one way or the other without making unjustifiable or question-begging assumptions about the psychology of learning a number language. The role of one-one correspondence as an application of numbers will not help us to decide whether Hume numbers or Benacerraf numbers *really are* the numbers.

I conclude this section with the claim that philosophical analysis of the concept of number will not help us decide between Benacerraf's principle and Hume's principle as the *true* abstraction principle for cardinal numbers. At least not without appealing to some disputable intuitions about our psychology of number.

4.3 Distinguishing the abstracts

Perhaps Neo-Fregeans should not try to rule out the Benacerraf-numbers as legitimate references of our number language, but embrace them. There is nothing to stop a Neo-Fregean accepting that in the abstract realm there are at least two number-like sequences of abstract objects. A good line for a Neo-Fregean to take might be that the Hume-numbers are the referents of our numbers-as-cardinals language, whereas the Benacerraf-numbers are the referents of our numbers-as-ordinals language. A Neo-Fregean could then argue that there are two main uses of number language, perhaps even two concepts of number (cardinal and ordinal) and so see no reason to be worried if there are two collections of abstract entities associated with them. Indeed, such a result could be regarded as a success of the Neo-Fregean programme.

The problem is that the Benacerraf-numbers are *not* ordinals: Benacerraf's principle involves no characterisation of ordering or any criterion of position correspondence. To understand Benacerraf's principle we need nothing that is not needed to understand Hume's principle. There is nothing about Hume-numbers that rules them out as being ordinals, and there is nothing about Benacerraf-numbers that rules them out as being cardinals. The concept of a unit is no less important to that of cardinality than the concept of one-one correspondence. Benacerraf's principle and Hume's principle each could be taken as allowing reference to the natural numbers *as cardinals*. But then Neo-Fregeanism must account for why our arithmetic language refers to the Hume-numbers rather than the Benacerraf-numbers (or vice versa). I have been arguing that that we stand in no significant relation to Hume's principle that we do not also stand in to Benacerraf's principle. So if the Neo-Fregean accepts that the two abstraction principles allow reference to different abstract entities, then he has made no progress overcoming the objection of Section 3.3.

5 Conclusion

I have argued that there is no particular abstraction principle that we can associate with the natural numbers. At least two similar, but formally distinct, abstraction principles are capable of lying at the heart of the Neo-Fregean programme. The principles are distinct enough that there is no natural way of equating the abstract objects they give reference to. The principles are however sufficiently similar that there is no principled criterion that identifies one over the other as 'the correct' abstraction principle for elementary arithmetic. I conclude that numbers are not the abstract objects referred to by either abstraction principle, or of any other abstraction principle. The point to emphasise here is that neither the Hume-numbers nor the Benacerraf-numbers are really *the natural numbers*. The whole Neo-Fregean framework of abstraction principles is just another way of generating sequences that *encode* the natural numbers. This conclusion is independent of questions regarding the metaphysics of abstraction and whether abstraction principles really refer to any abstract objects at all.

Literature

Benacerraf, Paul 1965 “What Numbers Could Not Be”, *The Philosophical Review*, 74(1).

Frege, Gottlob 1953 *The Foundations of Arithmetic*, tr. by J. L. Austin, Blackwell, Oxford.

Heck. Richard 1992 “On the Consistency of Second-order Contextual Definitions”, *Nous*, 26.

Weir, Alan 2003 “Neo-Fregeanism: An Embarrassment of Riches”, *Notre Dame Journal of Formal Logic*, 44(1).

Wright, Crispin 1983 *Frege’s Conception of Numbers as Objects*. Aberdeen University Press.

Wright, Crispin 2000 “Neo-Fregean Foundations for Real Analysis: Some Reflections on Frege’s Constraint”, *Notre Dame Journal of Formal Logic*, 41.

Deflationism and Conservativity: Who did Change the Subject?

Henri Galinon, Paris, France

1. The Problem

Deflationists about truth hold that truth is not a substantial property. But what counts as a substantial property? We shall be interested in the thesis that the following is a necessary and sufficient condition for the non-substantiality of truth:

> (Conservativity) The theory of truth of any given theory A is a conservative extension of A.

Suppose that (Conservativity) holds, then the deflationist would have some right to claim that truth is an explanatorily thin property: for it would show that whenever non semantical facts can be explained by a theory being true, they can also be explained by the theory. Suppose (Conservativity) does not hold; then there is a theory A, and sentences in the A-vocabulary that witness non-conservativity; the provability in our theory of truth of those A-unprovable true L_A -sentences would constitute some evidence against the deflationary thesis that truth is explanatorily dispensable.

As a matter of logical fact, some putative theories of truth have the conservativeness property over some given base theories, whereas others don't. But, the conservativity argument against deflationism continues, conservative theories of truth are not acceptable, because they fail to meet an essential requirement. To be sure, the concept of truth features in all these theories, that is to say a predicate satisfying Tarski's convention-T. But having the concept of truth is not enough for a theory to be the theory of the truth of a given theory. For truth ascriptions come with epistemic commitments, and theories of truth must account for them. Ketland (1999), in particular, has argued that holding a theory to be true is not only to hold all its theorems to be true (distributively, so to speak) but also to hold *that* all of its theorems are true (resp. collectively). Consequently the truth theory of A has to prove reflection principles for A: For all x, if Pr(x) then T(x).

Shapiro (1998) has an argument for a related conclusion. He offers a perfectly natural explanation (I'll say of what in a minute) involving the concept of truth, and argues that in any good theory of truth for A we should be able to carry out the reasonning. The example, unsurprinsingly, involves gödelian phenomena: the Gödel sentence conPA, it is well known, is a true and undecidable statement of PA; but why is it that conPA is true? A natural explanation goes like this, according to Shapiro:

> […] all the axioms of PA are true, and inference rules preserve truth. Thus every theorems of PA are true. It follows that 0=1 is not a theorem and so PA is consistent. (Shapiro(1998), p.505. Shapiro uses « A » where I write « PA »).

As we know, there are arithmetical sentences expressing (under codings) the consistency of PA, and these sentences are not theorems of PA. Shapiro's argument is then that they should be provable in the theory of the truth of PA.

To sum up: theories of truth come with some epistemic commitments, and those commitments yield non-conservativity results of truth theories over their base theory; hence truth is not explanatorily thin: knowledge of the truth of an arithmetical theory T yields new arithmetical knowledge beyond T.

We agree that from truth ascriptions consistency ascriptions should follow. We shall argue, however, that the conservativity argument against deflationism is flawed[1].

2. The Fable

Let us go into the fantasy of imagining a concurring civilization where people call themselves earthlungs. Earthlungs resemble us in every respect, except that in mathematics they not only study arithmetic, but also have come to recognize the interest and significance of *arithmutic*. In fact they have come to believe that natural numb*u*rs are the real elementary blocks constituting the universe, and they are for this reason much interested in studying them. A partial axiomatization of arithmutic is obtained by PA + $\neg con_{PA}$, where $\neg con_{PA}$ denotes the negation of a given sentence of the language of PA that is true in N (the standard model of PA) if and only if PA is consistent. Further axioms have been proposed but they are much debated at the moment and so we leave them aside. As it happens, earthlungs mostly use only the PA-part of arithmutic. Moreover, they use the same conventions for formalism as we do when doing logic and, believe it or not, they call PA the partial axiomatisation of arithmutic which is identical to our PA (a nice starting point for a philosophical vaudeville). *We* will sometimes write PA^* to denote their axiomatization and distinguish it from our PA. That is PA^* and PA are formally identical but intentionaly different, the first being intended as speaking about numburs, while the second is to be understood as speaking about numbers. Those people also have two Gudule Theorems that, I have to admit, are just as good as our Gödel Theorems. They usually state them as follows:

> First theorem: If T is a consistent, recursively enumerable and sufficiently rich theory, then T is incomplete.
> Second theorem If PA is consistent then:
> PA does not prove $\forall x\ (\neg Pr_{PA}\ (x), \ulcorner 0=1 \urcorner)$

All this is standard on Urth. The second theorem has especially been welcome since it had long been an open question whether $\neg con_{PA^*}$ was independent of PA*. Now when they hear us say that G-d-l's theorems show that there are true statements undecidable in PA, they agree, but they do not agree that con_{PA} is one of them[2]!

Now Peter, a guy from here who doesn't know much about earthlungs, once decided to engage Puter, one of theirs, in a conversation about the explanatory power of truth (it was raining hard that Sunday). Here's the conversation. (Caveat: I have tried to disambiguate occurences of

[1] Field (1999) has an answer to the conservativity argument and we basically agree with the general lines developped there. We can think of our argument as a variation on his own. We think our version is worth developing, though, since it crucially avoids to take a stand on contentious claims about which axioms are "essential to truth" and which are not (especially in connection with the problem raised by induction axioms involving the truth predicate).

[2] This is just because N is not the salient interpretation of PA in earthlung conversational contexts.

"PA" by using "PA* ", at least at the begining of the conversation. I wrote PA(*) when I was not sure which one one had in mind.)

Peter: Do you believe that the axioms of PA are true?

Puter: Yes, I believe that the axioms of PA* are true.

Peter: And do you believe inference rules to be truth-preserving?

Puter: I do.

Peter: You believe then that all of PA's theorems are true?

Puter: Indeed.

Peter (getting excited): Hence, since PA proves 0≠1, you believe it is true, and thus you believe that PA does not prove 0=1, that is you believe PA to be consistent.

Puter: Yes, I do!

Peter: You agree, then, that your commitment to the truth of PA is a commitment to its consistency, and that a good theory of truth for PA should account for that?

Puter: Yes, it would be nice.

Peter (proud) : Look, Tarski's theory of truth for PA, T(PA), is doing precisely this[3].

Puter (sincerly): I know, that's great indeed! [4]

Peter: But look: PA does not prove the consistency of PA, while T(PA) does. Truth has an explanatory power, it explains new facts that PA can't account for, facts that are expressible in the language of PA. My acceptance of PA does not logically commit me to the acceptance of con$_{PA}$, but once I have acknowledged the truth of PA the acceptance of con$_{PA}$ is forced upon me. There's a new purely arithmetical fact which is explained by my truth-attribution to PA.

Puter (embarrassed): But the consistency of PA$^{(*)}$ is not an arithm*u*tical fact.

Peter: I beg you pardon?

Puter: Well, I agree that your argument is a sound arithmetic reasoning, but it is not an arithmutic reasoning. First, it is false that the sentence con$_{PA}$ expresses the consistency of PA* in arithm*u*tic. And second, you cannot carry your inductive reasoning out in any sound axiomatization of arithm*u*tic, be it PA* or another theory. This is fortunate, since the negation of con$_{PA}$ is a true arithm*u*tical fact!

Peter: I'm not with you here.

Puter: Well, ok. First things first: the sentence con$_{PA}$, it is well-known, expresses the consistency of \$PA* \$ in the sense that it is true in N if and only if PA(*) is consistent. But it is of course *not* the case that con$_{PA}$ is true in arithmutic if and only if PA is consistent! Now the second point. In your argument you apply induction in the following manner: axioms are true, rules are truth preserving, hence all theorems are true. This is a perfectly correct inference of course, but it belongs to arithmetic, not arithmutic, since the induction involves vocabulary beyond the language of PA. To carry this argument out on the theory PA*, we usually use an axiomatic metatheory containing PA-*arithmetic* to talk about strings of PA*, plus a recursive truth theory (Tarski style), and we let the truth predicate appear in the meta-induction scheme. Sometimes we also enrich the logical-mathematical part of our metatheory so as to be able to explicitely define the truth predicate for our base theory. In any case, there is in the metalanguage a sentence con$_{PA*}$ which expresses the consistency of PA* in the sense of being true in the intended model of the metatheory if and only if PA* is consistent, which is provable in the metatheory.

Peter: I'm not sure that I have understood your point correctly. For arithmetic, arithm*u*tic, or what have you, it remains true that the truth-theory of PA non-conservatively extends PA, doesn't it?

Puter: As you can see from my example, T(PA*) is not conservative over PA*. But my point is that in this case it does not mean anything interesting. Although non-conservative over PA*, T(PA*) does not explain anything more than PA* in the sense that arithmutic is left the same before and after we ascend to its theory of truth. It is easy to see that under disambiguation of the vocabulary of PA between arithmutic in the base theory and arithmetic in the metatheory, the non-conservativity phenomenon vanishes.

Peter: Ok. There may have been a misundertanding. I agree with you that the non-conservativity of T(PA*) over PA* is meaningless and may just result from a fallacy of equivocation between PA and PA*. But the point remains that the theory of the truth of PA (and by this I now mean explicitely *arithmetic* PA) should prove the consistency of PA (idem), which PA does not. And in *that* case, since PA is thought of as a theory of arithmetic, and since you admit that consistency of a formal system is an arithmetical fact, you have to admit that the non-conservativity of T(PA) over PA-arithmetic shows truth to have an explanatory power after all!

Puter: I don't think so, for the situation is in fact exactly the same as before. Suppose I'm told that the theory PA is true, but that I'm not sure whether it is PA or PA* which is under discussion. I will *in any case* be able to prove that the theory is consistent in my truth-theoretic metatheory. For not only is T(PA) non-conservative over PA, but so is T(PA*). It is as it should be since the consistency of PA has nothing to do with the the way we think of PA, it has to do only with its formal features. So whether I interpret PA as arithmetic, arithm*u*tic, or whatever in the metatheory, consistency will follow from truth. Now the further claim that we have so derived a truth pertaining to the field of investigation of our base

[3] For reference, T(PA) is the theory obtained by extending PA with the Tarkian recursive axioms for truth, letting the the truth predicate enter the induction scheme.

Equivalently one could get a theory of satisfaction. Moreover, such recursive definitions can be turned into explicit definitions provided that we ascend to a richer theory (when it exists) allowing higher-order variables, or proving existence of sets of higher rank than than any set the existence of which is provable in the base theory . See for instance McGee (1991). It is not very important here which of those strong « theory of truth » for PA one has in mind

[4] Of course he had understood Peter's claim as: T(PA*) proves the consistency of PA* .

theory, that is arithmetic in the case in point, that claim can only be sustained by an argument to the effect that our metatheory is sound as an arithmetical theory: for in general, it is just not true that the restriction of the truth theory of a theory A to the vocabulary of A is sound for the intended model of A! (remember it was not sound as an arithm*u*tical theory). But how do we know that our metatheory is arithmetically sound? There's nothing in our base theory that can guarantee this. Clearly our recognition of T(PA) as arithmetically sound amounts to our acknowledgement of some systems stronger than PA as being arithmetical systems (T(PA) with unduction on unrestricted vocabulary, or second-order arithmetic in the case of an explicitely defined truth-predicate, etc.). In other words, a claim that T(PA) is non-conservative over PA and arithmetically sound amounts to a bold *statement* of new axioms for arithmetic above PA. It is not our commitment to truth which is doing the relevant non-conservative job, but our commitment to PA being arithmetic and to T(PA) being a stronger, arithmetically sound, theory.

3. The Moral

The moral of this story is simple. If someone knows that PA is true, he can conclude that PA is consistent. But he won't be able to convert this into arithmetical knowledge, that is to derive any new arithmetical fact, unless his arithmetical knowledge outreaches PA from the start. And it seems inevitable then to say that it is this hidden knowledge, which is unfolded in the course of building the truth-theory, that does the explanatory work. More generally, knowledge that an interpreted theory A is true yields knowledge that A is consistent. But it will never in itself give any new insights into A-facts, unless one knew from the beginning that A was somehow expressively defective relative to his actual knowledge concerning the intended field of T and had some ways to recognize some extensions of T as being sound relative to his knowledge of T-facts. Were this last condition not met, how could he be sure that one did not change the subject?

It seems fair to conclude that the conservativity argument does not show that truth has any explanatory power by itself. On the contrary, close inspection of the argument tells in favor of the thesis that "true" is a genuine expressive device.

Literature

Field, Hartry. 1999 "Deflating the conservativness argument", Journal of Philosophy 96, 533–540.

Ketland, Jeffrey, 1999 "Deflationism and Tarski's paradise". Mind 108, 69–94.

McGee,Van 1991 *Truth, Vagueness and Paradox*. Indianalpolis: Hackett.

Shapiro, Stewart 1998 "Proof and truth: Through thick and thin", Journal of Philosophy 95, 493–521.

Hard Naturalism and its Puzzles

Renia Gasparatou, Patras, Greece

1. Introduction

Most analytic philosophers today would call themselves naturalists. According to B. Stroud, the minimum commitment necessary is the exclusion of the *supernatural* from their philosophical system. (B. Stroud, 1996) And since today most philosophers seem unwilling to include any supernatural entities such as God or psyche in their accounts of reality or the mind, all could count as naturalists. Yet some forms of naturalism are harder that others. (P.F. Strawson, 1985) The hardest probably being eliminative naturalism suggesting the elimination of all mental language from our everyday vocabulary. This form of naturalism claims that scientific evolution will prove that mental terms are just pseudo-entities. I will argue that even though they strongly depend on science, hard naturalists can hardly account for the evolution of science.

2. Hard naturalism

The term *naturalism* refers to the general view that everything is natural. What gives hard naturalism a more specific touch is how one conceives nature. Hard naturalists take *natural* to mean *physical, material, scientifically explainable*. The claim that *all is natural* then implies that *all is to be studied by the methods of physical science*.

The question is what happens if something stands out against physical explanation. The most worrying example comes from consciousness: mental states resist a purely physical description. To use a crude example, it seems different to say “I am afraid of dogs” than say “seeing dogs produce adrenalin secretion in my brain”. The two sentences have different meanings: They are used in different contexts to draw attention in different aspects of my experience. One important difference being that the former describes the way I feel, providing the phenomenology of the experience from the first person perspective, while the later is a neutral description form the third person perspective.

Now, according to hard naturalists, such as P. M. Churchland, propositions of the former type cannot be translated into propositions of the later type just because the way we approach mental phenomena is already mediated by *folk psychology*. Folk psychology is, according to him, an implicit *theory*; a theory which people use in order to understand, explain and predict their own or other people’s psychological events and behaviour. Following folk psychology, we attribute *desires, fears* or *beliefs* in our attempt to explain our behaviour. Propositional states, such as these, are theoretical constructions and therefore should be evaluated with reference to experience. Like all theoretical entities, *desires* and *beliefs* are open to revision and total elimination, if proven false.

Lots of other folk theories have proved wrong in the past: Folk astronomy claiming that the earth is the centre of the universe, or folk physics talking about phlogiston. Churchland goes on arguing that folk psychology is such a *false* theory, “significantly worse [...] than [...] folk mechanics, folk biology and so forth” (Churchland, 1989, p.231). He compares it with the theory of witches, demonic possession, exorcism and trail by ordeal: *Demons* and *witches* just like *desires* and *beliefs* are theoretical entities. And just as we got rid of the theory of witches, we must now eliminate folk psychology. *Folk psychology is false since it resists physicalistic explanations.* As Churchland writes:

> If we approach *homo sapiens* from the perspective of natural history and the physical sciences, we can tell a coherent story of his constitution, development and behavioral capacities which encompasses particle physics, atomic and molecular theory, organic chemistry, evolutionary theory, biology, physiology, and materialistic neurotheory. That story, though still radically incomplete, is already extremely powerful... And it is deliberately and self consciously coherent with the rest of our developing world picture... But FP [folk psychology] is no part of this growing synthesis. Its intentional categories stand alone, without visible prospect of reduction to that larger corpus. (Churchland, 1981, p.75.)

Churchland clearly aims for a unifying physical theory that can account for all there is. Physical science is the best candidate for such an account. In order to save its *growing synthesis*, then, we should reduce all mental terms about desires, beliefs, fears etc in physical terms about brain activities. If this is not possible, we should eliminate the mental vocabulary from our ordinary language altogether. Neuroscience talk about brain states is supposed to fill in everyday vocabulary about mental states.

It should be clear that when Churchland asks for the elimination of folk psychology, he asks for the abolition of a basic corpus of ordinary dispositions and practices. *Folk psychology* refers to the way we all think and talk about all kinds of issues in our everyday life. It has to do with descriptions and concepts we all use everyday in ordinary language. When we say that the world is round, for example, we express a *belief*, when we take an umbrella before we leave our house, we again reveal our belief that it may rain. So, the implications of Churchland’s views thus go further than his philosophy of mind: Scientific explanations about the physical world are the only kind of *explanation* he is willing to admit.

Physical science is *the only* explanatory principle. Consequently, all kinds of problems people are struggling with (psychological, moral, aesthetic issues etc) should be translated into scientific, materialistic, physical language. If this is not possible, their resistance is strong evidence that they are *pseudo-problems,* which we should abandon *by eliminating* all relevant terms from our vocabulary. Philosophy too is taken in as a branch of theoretical proto science that articulates hypotheses for other sciences to test. (Churchland, 1986) Churchland’ s views then suggest a very strong version of scientism: Physical science is the norm by which the legitimacy of all quests, descriptions and explanations will be measured.

3. Problems with hard naturalism

The question is whether hard naturalism can provide an explanation of scientific evolution. Churchland insists that all questions regarding human consciousness, for example, will be resolved by physical science. His argument is supposedly inductive, for, as it is often said, “induction is the method of science”. So he infers the future of science from its past: Since science has progressed and has managed to illuminate some issues concerning human consciousness, it will evolve more and resolve all relevant questions in the future. Yet, his argument goes beyond induction; it rather appeals to Churchland’ s intuitions about the future of science and of ordinary language. For there is no evidence nowadays that beliefs and desires will be eliminated from our folk vocabulary. We have no clue whether science (perhaps some new branch of science) will embrace them into our common natural history or even whether this whole natural history will prove inaccurate and change. From our current viewpoint all these hypotheses are mere speculation.

Meanwhile, Churchland identifies explanation with the reduction of any phenomenon into physical phenomenon. Yet, he has no full-fledged, specific paradigms of such a reduction to offer. Failing an alternative coherent description of mental phenomena, his insisting on eliminating the ontology of ordinary language seems impracticable. Moreover, the identification of scientific explanation and physical reduction restricts the concept of science, without even defining it conceivably.

The hard naturalist, though, can answer this line of criticism: being a philosopher (and thus a proto-scientist) they don’t need to provide a full-fledged theory to take folk psychology’s place. (Churchland, 1986, p.6). They only need to give an outline of what this theory should be like; and, according to them, this proto-theory is already being built. (Churchland, 1991, p.67)

Yet Churchland views suffer an imminent tension: he takes for granted that many concepts, that are basic for communication and understanding, are pseudo-concepts with no literal meaning. Meanwhile, they are the concepts, which we are brought up with. From day one, we learn to engage those concepts and use them to understand all there is around us, including science. Ordinary language is full of mental vocabulary and the way we approach all human experience is full of folk psychology presumptions and explanations. Official education teaches us to think using such concepts descriptions and explanations. The phenomena we approach are described by them; all our starting hypotheses involve them. These are the concepts Churchland himself uses: when he says that folk psychology is a pseudo-theory he expresses a *belief* of his, there is no other way to say it.

Of course, one would answer that this only goes for now; when folk psychology gets eliminated there will be some other, better way to say it. (Churchland, 1981, p.87) But *for the time being* those are the only concepts we have; it is through them that today’s scientists are trained. If we accuse them of being void, we can no longer sensibly train today’s scientists. Neither can we sensibly articulate today’s hypotheses or theories.

Eliminative naturalists such as Churchland write and teach in a language they consider meaningless. But you cannot teach using a language and simultaneously suggest that most concepts and dispositions embedded in this language are senseless. This only makes what you say senseless as well.

4. Conclusion

Naturalism sees science and scientific method as a valid way people have in their attempt to explain the world. But how do people get engaged into scientific method(s)? Does naturalism manage a theoretical explanation of how scientific education and evolution work?

Hard naturalism identifies scientific explanation with an ideal physicalistic reduction. Yet, hard naturalists such as Churchland offer no strict criteria about what *physical* means: is meteorology a physical science? Is cognitive psychology a purely physical science today? *Science* seems restricted into very few branches and, what’s more, one cannot even know the criterion by which a discipline qualifies as scientific. Churchland offers only some intuitive remarks about how the scientific worldview will be like by proposing the elimination of all terms that today’s science has trouble accounting for.

Moreover, by insisting that all non-reducible terms should be eliminated form our explanatory story, the hard naturalist restricts *the phenomena* in need of explanation into very few. Many questions posed by today’s people (psychological or ethical worries and troubles) are considered pseudo-questions, raised by the pseudo-theory of folk psychology, which our language supports.

Most importantly, Churchland’ s hard naturalism, despite the scientism it implies, does not manage to illuminate the very fact of scientific education and evolution. It makes it incomprehensible that people who teach and think into pseudo-terms produce new good theories and educate new scientists that help science evolve. If our language is full of pseudo-concepts and false ontology, it is a mystery how scientific education was made to work and still continues to do. Consequently, it is a mystery how science progressed and still continues to do so. The conceptual rules used in everyday life are the same rules the scientist uses, even within his technical vocabulary. And despite this very fact, new scientists learn good science, make valid hypotheses and produce compelling theories. Even the most revolutionary among them rely, at least at first, on common world picture. Or, even when they question it, they are articulated in language.

It seems that the primacy ascribed to science comes with a high price: it makes science “stand alone, without visible prospect of reduction to that larger corpus”, to paraphrase Churchland. (1981, p.75) According to him, scientific practice is not *part* of human practices but stands way *above* them. It is the primary explanatory method and the one that will eventually eliminate all other branches. It will eliminate the problems other disciples confront, even the vocabulary that gives rise to those questions. But if one puts science so much higher than any other human practice, they cut its every connection with the community it comes from, the very community that practices it. Hard naturalist’s scientism has to face this paradox: the very primacy of science’s explanatory methods makes it harder to explain how science is communicated and evolved.

Literature

Churchland, P.M. 1981, “Eliminative Materialism and Propositional Attitudes*”, Journal of Philosophy* 78, 67-90.

Churchland, P.M. 1986, “On the Continuity of Science and Philosophy”, *Mind and Language* 1, 5-14.

Churchland, P.M. 1989, “Folk Psychology and the Explanation of Human Behavior”, *Philosophical Perspectives* 3, 225-241.

Strawson, P.F. 1985, *Skepticism and Naturalism: Some Varieties*, London: Methuen & Co. Ltd.

Stroud, B. 1996, **“**The Charm of Naturalism”, *Proceedings and Addresses of the American Philosophical Association*, 70 (2), 43-55.

The Mind-Body-Problem and Score-Keeping in Language Games

Georg Gasser, Innsbruck, Austria

I. The problem

Maybe to no other problem in philosophy so much time and attention has been dedicated in recent years than to the mind-body-problem. Enjoying a personal, subjective, first-person-perspective from which we undergo experiences with a certain phenomenal feel appears like a mystery in a world being fundamentally physical. The objective perspective of physical description lacks all the characteristic features of first-person-perspective.

Purported solutions to the problem tend to assume either a physicalistic-minded or a dualistic-minded form. According to Chalmers, each of these views has its promise, and each view seems to make some ad-hoc assumptions which are hard to spell out in more detail. Take, for instance, type-B materialism and type-D dualism.

Type-B materialism is the view that there is an epistemic gap between the physical and the mental but there is no ontological gap. Saying that there is no ontological gap implies stating identity between the mental and the physical. But how shall identity be stated in the light of the strong intuition that there are the properties of the brain, objects of perception, laid out in space, and, conscious states, defying explanation in such terms? According to Chalmers a type-B materialist is forced to accept the identity between physical states and consciousness as fundamental; it is a sort of primitive principle in hers theory of the world (Chalmers 2003, 254).

Type-D dualism is the view that mental states can cause physical states and vice versa. A very challenging objection to type-D dualism is that the interaction between mind and body is unexplainable. How should anything non-physical be able of interacting with physical things? Dualists have a straightforward answer. They don't know but ignorance should not be taken as decisive argument against their theory: "We should just acknowledge that human beings are not omniscient, and cannot understand everything." (Swinburne 1997, xii; for a similar argumentation, see Foster 1991, 161). In light of the observed connection between physical and phenomenal states it is an inference to the best explanation to assume that there is a psycho-physical nexus, though we are not able to render intelligible how it works.

What does this discussion show? Both, type-B materialism and type-D dualism refer to observed connections between physical and phenomenal states. The conclusions they draw, however, are very different: identity on the one side, psycho-physical interactionism on the other. The reason we cannot go on to investigate such notions in more fundamental terms is that the bottom of the theory in question has been reached. If this characterization is correct, then the various accounts in philosophy of mind seem to result ultimately in an impasse.

In what follows, I would like to ask how we could possibly explain why we permanently seem to end up into such an impasse. In giving a possible explanation I will refer to the concept of 'score-keeping in language game'.

II. Score-keeping in language games

The term 'score-keeping in language games' was introduced by David Lewis. He argues that in a communication process terms and concepts often are partially governed by certain implicit, context relative, parameters. These parameters define the score of a communication, that is, its running well or not. We can compare these parameters with rules in games: The rules define the score of the game. Thanks to the rules it can be told whether a team is doing well or not – whether the score of the game is for or against it. Something similar, according to Lewis, goes on in communication, even though the score is more flexible than the one in games. (Lewis 1979/1983, 240) If Lewis' analysis is correct, then during a communication process we tend to adapt continuously the applied parameters in order to modify the score of the discourse in such a way that its current status is still considered to be successful. A good example to illustrate such continuous adaptations of the conversational score is vague terms such as "bold", "flat" or "big". What is bold at one occasion, is not bold on another, what is flat at one occasion, is not flat on another and what is big at one occasion, may not be considered as big on another: „The standards of precision in force are different from one conversation to another, and may change in the course of a single conversation." (Lewis 1979/1983, 245)

Generally it can be said that our use of standards defining the score is broad and not very restrictive. We could imagine a situation in which subliminally parameters from different contexts are introduced into a single discourse (Horgan 2001 and 2007 argues that the agent exclusion problem is the consequence of such a situation). Thereby an atypical discourse context is created for it is unclear which score which is in use in such a situation. According to which standards should we judge whether a satisfactory score has been reached if the various context parameters in practice do not overlap? Probably we would end up in a kind of discursive cul-de-sac.

III. Application to Philosophy of Mind

Is it plausible to assume that the mind-body-problem is the consequence of such a scenario? Let us focus at possible parameters in the mind-body-problem first. Physical concepts, as we have seen, are developed in a context of objectively accessible phenomena, that is, phenomena generally accessible to science. Normally these phenomena are quantitatively definable, in terms of material and structural composition.

Mental concepts, on the contrary, are qualitatively determined. They are characterized as essentially subjective in the sense that every mental property is principally accessible only from a certain subjective point of view (Nagel 1974, 442).

The mind-body-problem arises out of the tension between concepts apparently as different as the mental and the physical. If someone approaches the mind-body-problem one tends to undergo a series of cognitive steps (I model these steps according to Horgan 2001). A physicalistic-minded person may undergo something like the following:

1. The starting point: The world consists ultimately of nothing but bits of matter distributed over space-time behaving according to physical laws. (Kim 2005, 7)
2. An automatical and subliminal accommodation to the parameter appropriate to this kind of discourse, that is, (micro-)physical explanation takes place.
3. It is not acknowledged that such an accommodation has occurred and that parameters stemming from (micro-)physical explanation are applied to notions such as world, reality or nature.
4. The question: How can there be something such as a conscious experience in a physical world like this?
5. There is, however, no shift in the accommodation of parameters. The discourse continues under the parameters installed at the beginning.
6. It is realized that what is called 'consciousness' or 'the mind' is hard to integrate in the kind of approach under consideration.
7. As a result, the mind appears to be 'special', 'mysterious' or even 'unreal'.
8. Thus, it is intuitively plausible to assume that the mental has a place in our world only if it is identical with something physical. Though the assumption of this identity cannot be illuminated any further, it seems to be the inference to the best explanation.

Crucial components in such a process of reasoning are steps 1, 4 and 5. The question posed at the very beginning introduces parameters which shape decisively the following discourse. Talk about the physical world, bits of matter, space-time and physical laws introduces parameters conforming to scientific discourse where quantitative and structural explanations of reality do not provide any room for subjective and qualitative aspects.

In step 4 a concept with another parameter is introduced. Paying attention to the mind and its characteristic features comes along with parameters pointing towards another score than parameters of a physical context. The parameter-setting under which an entity counts as mental are, for instance, (i) being qualitative and (ii) enjoying a subjective perspective.

In step 5 the way is paved for the puzzlement arising in step 7: It remains unnoticed that talk about the mind goes hand in hand with parameters different from those shaping the overall score of the entire discourse. As long as this conversational score is in use mental phenomena will always fall short of being fully appreciated for there is no way how they can adequately be integrated in a context framed by such parameters.

The same applies to dualistic thinking:

1. The starting point: Physical objects are not conscious; they do not have thoughts and sensations. Men and animals, on the contrary, do enjoy thoughts and sensations. Having a thought or a sensation is not just having some physico-chemical event occur inside one of greater complexity than the physico-chemical events which occur in physical objects. It is not the same sort of thing at all for it is rich in inbuilt colour, smell and meaning. (Swinburne 1997, 1.)
2. An automatical and subliminal accommodation to the parameter appropriate to this kind of discourse, that is, a clear distinction between sentient and non-sentient, conscious and non-conscious takes place.
3. It is not acknowledged that such an accommodation has occurred. The parameters applied to notions such as 'animal', 'man' and 'nature' divide everything up into something mental or physical.
4. The question: How can we explain our experience of mind-body-interaction in the light of the assumption that the mental is so different in nature from the physical?
5. There is, however, no shift in the accommodation of parameters. The discourse continues under the parameters installed at the beginning.
6. It is realized that what might be called mind-body- and body-mind-interaction is hard to integrate in the kind of approach under consideration.
7. As a result, mind-body-interaction appears to be 'special' and 'unexplainable' (dualistic interactionism) or even 'unreal' (epiphenomenalism).
8. Thus, interactionists will argue: It is intuitively plausible to assume that mind-body-interaction takes place. It is just one of the most obvious phenomena of human experience. Not being able to explain how it occurs does not back up the epiphenomenalist conclusion that it does not occur at all or the much stronger claim that the theory is false in principle.

Is it plausible to assume that the mind-body-problem arises out of such scenarios? Let me start with some thoughts from Strawson's *Individuals*. Strawson argues that there exists a categorical framework of our factual everyday thinking which is the realm of meso-scopic entities containing person-like and non-person-like individuals. Person-like individuals enjoy physical and mental properties. If we describe human persons we describe them as a single entity with physical and mental features.

Taken this analysis as a matter of fact we can aim at developing precise theories about mental and physical properties. We can ask how mental and physical properties are to be described more accurately, whether they can consist out of smaller parts, what their differences are. In other words, we can start to reason theoretically about the various features we rather vaguely describe in everyday thinking. Theories in philosophy of mind, according to this story, are theories developed for and framed from a specific theoretical context. In such contexts preciseness, clarity and analyticity are the standards amounting to the score of the discussion. This score, however, is a very different one from the score valid in everyday interaction. As Lewis remarked, in everyday communication we generally tend to be very permissive for we have an interest that communication goes on. In a theoretical setting, on the contrary, we probably are less permissive for communication is judged according to precise definitions and clear argumentation.

If this is correct, then the categorical framework of our factual everyday thinking is open for different theoretical interpretations because the conversational score in everyday thinking is broad and not sharply defined. Saying that human persons have physical and mental properties, for instance, or that human organisms in contrast to other organisms can reason and think leaves it open how these statements are to be spelled out in a more precise way. In such statements the apparently profound differences between mental and physical properties are not thematised any further. From theories in philosophy of mind, however, accurate definition of these aspects is demanded.

I suggest that impasses in philosophy of mind stem from the fact that the variety of our factual categorical framework of everyday thinking is given up in favour of a possible theoretical precision of certain aspects. For instance: What does it mean that biological organisms such as human beings can reason and think? Does this mean that the substance of mental properties is a biological organism? Or does a new entity come into existence, a 'someone' having these experiences? Trying to answer such questions comes along with negligence of other aspects being part of our common categorical framework as well. If a theory is blamed for being counter-intuitive or for not taking into consideration certain aspects of reality adequately enough, then, I guess, the different conversational scores of everyday parlance and theoretical inquiry come into conflict. The widely shared impression that neither physicalistic nor dualistic theories of mind are fully satisfying might have its roots in the fact that the ample categorical framework of our factual everyday thinking cannot be fully integrated into the narrow and specialised frameworks of theories in philosophy of mind. Due to the precision required in philosophical thinking and the lack of precision in everyday communication a theory of mind overlapping in its score with the score of our commonly assumed categorical framework will hardly be available.

IV. Conclusion

This leads to the conclusion that theories of mind will always have an unsatisfying smack. There will always be the feeling that something has not been integrated or that some feature has been turned into something other than what it is.

Physicalistic and dualistic theories are on a pair then – compared with the categorical framework of our factual everyday thinking. Why do philosophers nevertheless have either physicalistic or dualistic tendencies? Following Hardcastle I would argue it is a matter of attitude. (Hardcastle 2004, 801) These divergent reactions turn on antecedent views about what counts as explanatory and what does not. Thus, problems identified in philosophy of mind depend heavily on the perspective out of which we approach the examination of the mind-body-problem. These remarks are not a solution to the mind-body-problem but they explain how the problem arises and why remedy is hard to find.

Literature

Chalmers, David J. 2003 "Consciousness and its Place in Nature", in: David J. Chalmers (ed.), *Philosophy of Mind. Classical and Contemporary Readings*, Oxford: Oxford University Press, 247-272.

Foster, John 1991 *The Immaterial Self*, London: Routledge.

Hardcastle, Valerie G. 1996/2004 "The why of consciousness: a non-issue for materialists", in John Heil (ed.), *Philosophy of Mind. A guide and anthology*. Oxford, 798-806.

Horgan, Terry 2007 „Mental Causation and the Agent-Exclusion Problem", in *Erkenntnis* 67, 183-200.

Horgan, Terry 2001 „Causal Compatibilism and the Exclusion Problem", in *Theoria* 16, 95-116.

Kim, Jaegwon 2005 *Physicalism, or something near enough*, Princeton: Princeton University Press.

Lewis, David 1979 "Scorekeeping in a Language Game", reprinted in David Lewis 1983 (ed.) *Philosophical Papers*, Oxford: Oxford University Press, 233-249.

McGinn, Colin 1989 "Can We Solve the Mind-Body-Problem?", in *Mind* 98, 349-366.

Nagel, Thomas 1974 "What is it like to be a bat?", *Philosophical Review* 83: 435-450.

Wright, Wittgenstein und das Fundament des Wissens

Frederik Gierlinger, Wien, Österreich

In seinem Artikel *Wittgensteinian Certainties* vertritt Crispin Wright eine Position, nach der es eine Klasse von Sätzen gibt, die das Fundament unserer Wissensansprüche ausmachen. Diese fundierenden Sätze (Typ III) werden von Wright unterschieden von Evidenzbeschreibungen (Typ I) einerseits und Behauptungen (Typ II) andererseits. In seiner Schilderung sind es Evidenzbeschreibungen, die herangezogen werden, um Behauptungen zu stützen. Damit aber diese rechtfertigende Verwendung eines Satzes vom Typ I auf einen Satz vom Typ II möglich wird, sind bereits Überzeugungen nötig, die selbst nicht gerechtfertigt werden können. Dies sei anhand des folgenden Beispiels demonstriert:

> Typ I (Evidenz): "Mein derzeitiger Bewusstseinszustand ist von solcher Gestalt, dass hier eine Hand zu sein scheint."
> Typ II (Behauptung): "Hier ist eine Hand."
> Typ III (Hintergrund): "Es gibt eine materielle Welt."

Weil Sätze vom Typ III stets vorauszusetzen sind, befindet Wright, dass die Annahme eines Hintergrunds (i.e. einer Menge von solchen Sätzen des Typs III) notwendigerweise ungerechtfertigt geschieht. Diese Konstruktion und ihr Ergebnis benennt er I-II-III Skeptizismus. Indem Wright des Weiteren behauptet, die Akzeptanz dieser Sätze stehe in keinerlei Zusammenhang mit der Wahrscheinlichkeit ihrer Wahrheit, bestimmt er unsere wissenschaftliche Basis als unsicher. "To be entitled to accept a proposition in this way, of course, has no connection whatever with the likelihood of its truth." (Wright 2004:53) Wir können zur Verteidigung der Akzeptanz dieser Sätze nur vorbringen, dass wir sie aus einer praktischen Notwendigkeit des Lebens heraus akzeptieren. "One's life as a practical reasoner depends upon type III presuppositions. To avoid them is to avoid having a life." (Wright 2004:52f) Diese Berechtigungskonstruktion, die eine wenig spannende Wiederholung der Gedanken David Humes zum skeptischen Dilemma darstellt, nennt er *Entitlement*.

Der Schluss ist somit der, dass wir den skeptischen Zweifel nicht widerlegen können, aber bestimmte Überzeugungen die Welt betreffend haben müssen, auch wenn diese möglicherweise nicht den Tatsachen entsprechen. Ich behaupte, dieser Entwurf ist nicht bloß im Ansatz verkehrt – eine Unterteilung in drei Satzgruppen nimmt keine Rücksicht auf die verschiedenen Umstände, unter denen ein Satz geäußert werden kann – sondern ist eigentlich ganz unverständlich.

Wright behauptet, dass alles ganz anders sein könnte, als wir glauben. Wenn aber jemand sagt, es gibt keine Gewissheit dafür, dass die Dinge sich wirklich so verhalten, wie wir annehmen, dann ist im Grunde nicht klar, *was* hier unter Verdacht steht, anders als angenommen zu sein. Kann denn *alles* angezweifelt werden? Wright vermeint sich zwar im Einklang mit Wittgensteins Bemerkungen in *Über Gewissheit,* wenn er dies ablehnt, aber die Gründe, aus denen er es ablehnt, sind völlig andere als bei Wittgenstein, weshalb von einer Übereinstimmung der beiden keine Rede sein kann. Während Wittgenstein uns darauf hinweisen möchte, dass wir dem Zweifel an Allem keinen Sinn geben können (vgl. ÜG 114), meint Wright, dass der Skeptiker eine ganz und gar berechtigte Frage aufbringt und uns dadurch die Grenzen unserer Rechtfertigungen aufzeigt. "[T]he best sceptical arguments have something to teach us." Nämlich, "that the limits of justification they bring out are genuine and essential" (Wright 2004:50)

Nehmen wir einmal an, jede meiner Überzeugungen ist falsch, d.h. die Dinge verhalten sich tatsächlich anders, als ich glaube – und wir wollen so tun, als verstünden wir für den Moment, was diese Aufforderung von uns verlangt. Nehmen wir zudem an, dass der Dämon, der mich täuscht, eines Tages des Spiels mit mir müde wird und mich erwachen lässt. Warum sollte ich *das* nun Wirklichkeit nennen? Was hindert mich daran, es für einen Traum zu halten? Wie kann ich es überhaupt für irgendetwas halten? Das Problem mit derartigen Überlegungen ist, dass sie dazu verführen, unsere Begriffe "Wirklichkeit", "Wahrheit", "Täuschung", "Skepsis", etc. auf eine Situation anzuwenden, in der diese Begriffe keinen Sinn haben. (vgl. ÜG 36, 37)

Wer des Weiteren behauptet, dass unser Verfahren, Sätze anzunehmen, nichts mit der Wahrscheinlichkeit ihrer Wahrheit zu tun hat, der meint, ohne Kriterium dafür auszukommen, einen Satz als wahr oder falsch zu bestimmen. Für diese Einsicht ist lediglich anzusehen, was es heißen kann, dass wir alle in unseren Überzeugungen falsch liegen. Wenn ich sage: " Du liegst mit deiner Behauptung falsch", so lässt sich mein Einwand prüfen. Wenn jemand aber sagt: "Die Menschheit liegt (möglicherweise) mit allen ihren Behauptungen falsch", so ist zunächst überhaupt nicht klar, wie sich das prüfen ließe. Jede Prüfung bedarf eines geeigneten Maßstabs. Zu sagen, es könne alles ganz anders sein, ist gleichsam der Versuch, eine Länge abzunehmen, ohne ein Längenmaß zu besitzen. Wright bezieht sich auf die Wahrheit als Maßstab, entzieht sie aber zugleich unserem Erkenntnisvermögen. Seiner eigenen Forderung – "Empirical enquiry does par excellence have an overall point, namely [...] the divination of what is true and the avoidance of what is false of the world it concerns." (Wright 2004:43) – ist nicht mehr nachzukommen. Umso mehr, als Aussagen vom Typ 3, an denen alles Weitere ansetzt, weder wahr, noch falsch, weder zu rechtfertigen, noch zu widerlegen sind.

Es lohnt an dieser Stelle, kurz darauf einzugehen, wie Wright das Verhältnis zwischen mathematischem Satz und Wahrheit bestimmt. Zum einen, um besser zu verstehen, weshalb er es überhaupt als nötig empfindet, in dieser Eindringlichkeit auf Wahrheit als Leitidee empirischer Forschung hinzuweisen. Zum anderen, um nachzuvollziehen, wie Wright sich den besonderen Status von Typ 3 Sätzen erklärt. Der mathematische Satz, so Wright, fungiert als Regel, die ein Verfahren definiert. Diese Regel mag zur Erreichung eines bestimmten Ziels *ungeeignet* sein, aber sie ist als Definition nicht mit den Kategorien der Wahr- und Falschheit einzufangen. "The merit of a rule may be discussible: rules can be inept, in various ways. But, since they define a practice, they cannot be wrong." (Wright 2004:43)

Aber wie leitet mich der Satz "2 + 2 = 4" an, ein Verfahren anzuwenden? Ich muss schon verstanden haben, wie mit dem Satz zu verfahren ist, bevor er mir als Regel dienen kann. Indem Wright dies übergeht und den mathematischen Satz der Praxis als Basis zugrunde legt, kommt es zu einem gravierenden Missverständnis. Die notwendige Konsequenz, die Wright richtigerweise selbst daraus zieht, ist, dass Regeln gleichsam sich selbst dienen, sofern wir sie nicht bewusst an ein bestimmtes Ziel anknüpfen. "Rules governing a practice can be excused from any external constraint – so just 'up to us', as it were – only if the practice itself has no overall point which a badly selected rule might frustrate." (Wright 2004:43)

Aus diesen Überlegungen ergibt sich für Wright in der Folge ein wesentliches Problem einer internalistischen Auffassung von Sprache. Wenn es der Fall ist, dass ausschließlich die sprachliche Praxis – und damit die Befolgung bestimmter Regeln – unseren Worten Bedeutung verleiht, dann scheinen wir von dem, wie die Welt wirklich ist, abgekoppelt zu werden. "[I]t is our linguistic practice itself that is viewed as conferring meaning on the statements it involves – there is no meaning-conferrer standing apart from the rules of practice and no associated external goal." (Wright 2004:45). Deshalb meint Wright bei Wittgenstein ein metaphysisches Projekt zu erkennen: Wie wir von der Welt sprechen, könnte davon abweichen, wie es sich *wirklich* verhält. "In taking it for granted [...] that type III propositions 'might just be false' – as a matter of metaphyiscal bad luck, as it were – I-II-III scepticism sets out its stall against the internalism of the *Investigations*." (Wright 2004:45)

Diese Kritik ist aber verfehlt. Regeln und Praxis haben für Wittgenstein Bezug zur Wirklichkeit insofern, als sich etwa ein mathematischer Satz erübrigen würde, wenn Gegenstände sich nicht mehr in gewohnter Weise verhalten. (vgl. BGM I, 37) Wittgensteins Ansatz trennt uns nicht von der Welt und ihren Tatsachen ab, auch wenn die Dinge uns nicht zu einem bestimmten Urteil zwingen können (d.h. es ist keine metaphysische Notwendigkeit darin, welchen Schluss wir aus Erfahrungen ziehen). Der Verweis auf eine externe Wahrheit, nach welcher sich unser Forschen auszurichten hat, ist hingegen ein Versuch, die Grenzen unserer Sprache zu verlassen. Wie soll mit einer externen Wahrheit verglichen werden? Wie soll sie erkannt werden? Die Frage, wonach wir die Wahr- oder Falschheit eines Satzes festlegen sollen, bleibt in einem Modell, wie es Wright vorschlägt, notwendig unbeantwortet. Damit wird allerdings unklar, was "Wahrheit" bedeutet. Die Forderung Wrights, Wissenschaft habe nach Wahrheit zu streben, entpuppt sich in diesem Zusammenhang als sinnlos.

Ich werde nun näher auf Wrights Charakterisierung von Typ III Aussagen als Regeln und damit auf seine Auslegung von *hinge propositions* eingehen. Dabei wird sich zeigen, dass Wrights Beschreibung dieser Sätze nicht stimmig ist. Seine Grundhaltung ist jene, dass *hinge propositions* – zu denen er auch einfache mathematische Sätze zählt – als Regeln gedeutet werden können. (Eine solche Leseart vertritt auch Marie McGinn in ihrem Buch *Sense and Certainty*, wenngleich sie eine von Wright in wesentlichen Punkten abweichende Auffassung des Regelcharakters dieser Sätze verteidigt.)

In welcher Weise aber besitzen *hinge propositions* Regelcharakter? Wright schreibt: "The cases [dies bezieht sich auf seine Unterteilung von *hinge propositions* in drei Gruppen] are [...] unified [...] by their constituting or reflecting our implicit acceptance of various kinds of *rules of evidence*." (Wright 2004:42) und bestimmt *rules of evidence* anhand folgenden Beispiels: "Imagine that you count the pieces of fruit in a bowl containing just satsumas and bananas. You get thirteen satsumas and then seven bananas but when you count all the fruits together, you get twenty-one. Yet you seem to have made no mistake, and no piece of fruit is added or removed – so it seems – during the three counts. [...] According to the mooted account, the necessity of 13 + 7 = 20 is somehow grounded in the fact that such appearances are not *allowed* collectively to stand as veridical. Rather, we inexorably dismiss them out of hand – 'You must have miscounted somewhere', 'Another piece of fruit must somehow have been slipped in', etc." (Wright 2004:34)

Wright behauptet also, dass wir uns weigern, etwas anderes als 20 zum Ergebnis zu nehmen, wenn wir 13 und 7 zusammen zählen. Nur was mit dem mathematischen Satz in Einklang steht, ist für uns als Evidenz zulässig; ein anderes Ergebnis akzeptieren wir nicht. Der mathematische Satz kann nach diesem Verständnis nicht falsch sein und er kann nicht falsch sein, weil wir all jene Fälle, die ihn als falsch zeigen könnten, von vornherein ausschließen. Wright schreibt diese Ansicht Wittgenstein zu.

Untersuchen wir das Beispiel genauer: Wright schildert eine Situation, in der ich einen Fruchtkorb vor mir habe und zuerst einen Stoß mit 13 Mandarinen zähle und dann einen Stoß mit 7 Bananen und dann lege ich beide Stöße zusammen und zähle 21. Möglicherweise wollte ich sicher gehen, wie viele Früchte es sind. Dann bin ich jetzt erst recht unsicher. Ich zähle also nochmals sorgfältig und hierbei wird sich (hoffentlich) herausstellen, dass ich mich entweder hier oder da verzählt habe. Aber, und dies ist entscheidend, ich verwerfe das Urteil, es seien 21, nicht auf Basis meiner Überzeugung, dass gilt "13 + 7 = 20", sondern ich verwerfe das Urteil, es seien 21, weil ich vermute, irgendwo falsch verfahren zu sein. D.h. ich hege keinen Zweifel an der *Korrektheit der empirischen Tatsachen*, sondern an der *korrekten Durchführung* des üblichen Verfahrens, solche empirischen Tatsachen zu untersuchen. (Es könnte beispielsweise sein, dass ich eine Frucht zweimal gezählt habe, etc. Und das kommt ja vor, dass man etwa beim Kartenspiel die Karten zählt und weiß es müssen 52 sein, aber man zählt bloß 51. Man wirft einen Blick in die Schachtel, ob vielleicht noch eine Karte darin liege, und falls nicht, zählt man nochmals mit größerer Sorgfalt. Aber der Evidenz misstraue ich in diesem Fall nicht weil „1 + 1 + 1 + ... = 52“, sondern weil ich weiß, dass im Kartenspiel 52 Karten sein sollen.) D.h.: Wenn wir den mathematischen Satz über Evidenz erheben, so nicht darum, weil wir die Tatsachen an sich in Frage stellen, sondern weil wir einen Fehler in der Anwendung eines Verfahrens, dass wir addieren nennen, vermuten. Es ist zwar denkbar, dass Evidenz in bestimmten Situationen übergangen wird, (vgl. BGM I, 37), aber dies ist nicht, worauf Wright hier anspielt.

Aus dieser defizitären Bestimmung mathematischer Sätze als Regeln ergeben sich in der Folge auch Schwierigkeiten hinsichtlich der Bestimmung von *hinge propositions* als Regeln. Wright behauptet, es verhalte sich mit "Ich habe zwei Hände" auf gleiche Weise, wie mit einfachen arithmetischen Aufgaben. Mit anderen Worten: er versieht *hinge propositions* mit der gleichen Rolle wie mathematische Sätze und schließt, dass wir gewöhnlich an unserer Überzeugung, zwei Hände zu haben, festhalten, auch wenn wir gegenläufige Erlebnisse haben. "My certainty that I have two hands will 'stand fast' above the flow of evidence making [...]. Were I to have a – visual, or tactual – impression that I did *not* have two hands, then I should treat it just on that account as unrealiable." (Wright 2004:36f)

Wright verwechselt hier allerdings zwei Dinge miteinander. Im Fall der Addition steht die korrekte Anwendung eines Verfahrens, das wir zählen nennen, (also etwa die Vorgabe jeden Gegenstand nur einmal zu zählen, etc.) unter Verdacht. Die Überzeugung, dass ich zwei Hände habe, beruht hingegen auf keinem solchen Verfahren und darum ist die Berufung auf eine Regel fehl am Platz. Es kann zwar angemerkt werden, dass die Verwendung des Wortes "Hand" von Regeln geleitet ist, so wie jeder Sprechakt auf Regeln basiert. Damit geht aber nicht einher, was Wright behauptet. In der Mathematik greifen Ergebnis und Verfahren ineinander und es kommt nichts hinzu, dass für die Wahr- oder Falschheit des Ergebnisses Ausschlag gebend ist. Das Wahrheitskriterium ist die korrekte Anwendung eines Verfahrens. "Ich habe zwei Hände" verweist hingegen auf empirische Gegenstände und die Wahr- oder Falschheit des Satzes bemisst sich daran, ob die Dinge sich so verhalten, wie der Satz behauptet, d.h. es wird ein Vergleich mit der Wirklichkeit angestellt. Die korrekte Anwendung eines Verfahrens ist hier kein Wahrheitskriterium des Satzes, sondern entscheidet über dessen Sinnhaftigkeit (i.e. über die Möglichkeit, den intentionalen Gehalt der Aussage zu verstehen).

Damit sollte gezeigt sein, dass die Analogie zwischen den Sätzen "2 + 2 = 4" und "Hier ist eine Hand", die in *Über Gewissheit* an manchen Stellen anzuklingen scheint, nicht aus einem gemeinsamen Regelcharakter dieser Sätze, wie er von Wright charakterisiert wird, herrührt. Und es sollte gezeigt sein, dass die Annahme dieser Sätze uns nicht auf metaphysische Voraussetzungen festlegt, die möglicherweise falsch sein könnten.

Literatur

McGinn, Marie 1989 Sense and Certainty. A Dissolution of Scepticism, Oxford: Blackwell.

Wittgenstein, Ludwig 1984 Über Gewissheit, Frankfurt am Main: Suhrkamp.

Wittgenstein, Ludwig 1984 Bemerkungen über die Grundlagen der Mathematik, Frankfurt am Main: Suhrkamp.

Wright, Crispin 2004 Wittgensteinian Certainties, in: McManus, D. (ed.), Wittgenstein and Scepticism, London: Routledge, 22-55.

Reduction Revisited: The Ontological Level, the Conceptual Level, and the Tenets of Physicalism

Markus Gole, Graz, Austria

1. Reduction and Physicalism in Philosophy of Mind

When the topic of reductionism is addressed, especially within philosophy of mind, one cannot help but summon the topic of physicalism as well. Physicalism, broadly construed, can be defined as the thesis that there is nothing over and above the physical: all there is is physical, in one way or another, and there are no such things as non-physical substances, events, properties and the like which escape the physicalist story. For instance, when I bump into the table in my kitchen and thus hurt my leg, the only story there is to tell is simply the story of the natural sciences. Such a story might go like this: after having bumped into the table, certain physiological mechanisms are activated, *e.g.,* the information of tissue damage is transmitted via nerve fibers from the leg to the brain where certain neurons are caused to fire, which in turn cause other neurons to fire, and finally the statement "Ouch, my leg hurts!" is uttered followed by wincing and groaning. It should be noted that the statement "Ouch, my leg hurts!" is solely used as an abbreviated form of the neuron firing talk and, similarly, wincing and groaning are themselves nothing over and above another neuron firing story. All we need to fully and exhaustively characterize a painful experience is a characterization of the physical events which are solely couched in physical concepts and there does not seem to be a need to use any mental concepts like "pain", "want", "desire", and so on. But why should anyone be a physicalist? The answer to this question leads us to the tenets of physicalism which I take to be ontological parsimony as well as elegance and simplicity in the construction of our theories. The ontological parsimony stems from abandoning non-physical entities, for there is no need to introduce mental entities in order to explain what is going on when someone is in pain. Or, put differently, all mental entities have fallen prey to Ockham's razor. By elegance and simplicity I mean that it gets easier if only one kind of entities, *i.e.,* physical entities, are used to construct theories compared to two kinds of entities, *i.e.,* physical *and* mental entities. Thereby, pain theories become simpler and more elegant once mental entities have been crossed out.

I turn now to the reduction part. In contemporary philosophy of mind, it is widely accepted, both by the physicalist and dualist, that if a mental property can be fully characterized in the language of physics, then the mental property in question is actually a physical property. The translation of mental expressions into physical expressions, or put in another way, the identity between the mental concept and the physical concept, can be thought of as a conceptual reduction and the identity between the mental property and the physical property can be thought of as an ontological reduction. Therefore, reduction represents a relation between two concepts on the one hand and between two things, in this case properties, on the other hand. This relation is the relation of identity, because one concept or property is nothing over and above another concept or property respectively. Moreover, the conceptual reduction is sufficient for the ontological reduction. However, whether the reverse is true is a matter of debate. The present paper is an attempt to tackle that question and it is argued that a conceptual reduction follows from an ontological reduction as well, but only under the background assumption that *a priori* physicalism is true. It is also argued that if the tenets of physicalism are taken seriously, then *a posteriori* physicalism should be dropped in favor of *a priori* physicalism.

2. Conceptual Reduction, Ontological Reduction, and Physicalism

I would like to begin this section by defining *a priori* and *a posteriori* physicalism. Insofar as the physicalism part is concerned, *a priori* as well as *a posteriori* physicalism agree that all there is is physical. Furthermore, proponents of both branches of physicalism are committed to the claim that, necessarily, all the mental phenomena are entailed by the physical phenomena. Thus, if all the physical things are fixed, then all the mental things are fixed, too. Insofar as the *a priori/a posteriori* part is concerned, the discrepancy arises. *A priori* physicalists (*e.g.,* Jackson 1998) hold that all the mental phenomena are entailed *a priori, i.e.,* solely on grounds of the meanings of the words involved. For instance, the mental concept "pain" refers to the mental property "being in pain" and the physical concept "C-fiber stimulation" refers to the physical property "having a C-fiber stimulation". If *a priori* physicalism is true, the concepts "pain" and "C-fiber stimulation" are two words with the same meaning, in fact, they would be synonymous expressions and "pain" could be conceptually reduced to "C-fiber stimulation". Because of their synonymy, both concepts would have the same property as their referent and *a fortiori,* the property of being in pain could be ontologically reduced to the property of having a C-fiber stimulation.

In contrast, the *a posteriori* physicalist (*e.g.,* Loar 1997) argues that the mental phenomena are entailed only *a posteriori* and it is a matter of scientific investigation to find out that the properties in question are actually one and the same property. Because of this *a posteriori* nature of the identity claim, the concepts involved are independent, for it is impossible for the mental phenomena to be entailed by the physical phenomena solely on grounds of the meanings of the words. That is, *a posteriori* physicalists allow and argue for an ontological reduction and in the same vein argue against a conceptual reduction. I think it is safe to say that all *a posteriori* physicalists are sympathetic to Kripke's (1980) framework of necessary *a posteriori* identity claims and his canonical example "water = H_2O". It was an empirical discovery that water is one and the same as H_2O. Nevertheless, conceptual analysis did not get us to say that water is nothing over and above H_2O, and the reason is that the concepts "water" and "H_2O" do not mean the same; they are not synonymous. So, *a posteriori* physicalists see the identity claim "pain = C-fiber stimulation" akin to the identity claim "water = H_2O". Therefore, if *a posteriori* physicalism is true, pain is ontologically, but not conceptually, reducible to a C-fiber stimulation.

The idea of a conceptual reduction is not a new one and the following two examples underline its relevance for a proper understanding of reduction. For instance, logical behaviorists (*e.g.,* Ryle 1949) explicitly state that all mental expressions can be translated into, and thereby reduced to, expressions about behavioral dispositions. Thus, the logical behaviorist is some kind of an *a priori* physicalist in the sense described above, because the mental concept "pain" is translated into a statement about withdrawal behavior resulting from tissue damaging stimuli. Once a conceptual reduction has been accomplished, an ontological reduction will follow. Kim (2005) makes a similar point with his model of functional reduction. According to Kim, the first step in a successful ontological reduction is to define the mental property in question functionally, *i.e.,* in terms of the causal role it occupies. As Kim clearly says, that is a matter of conceptual work and in terms of the present paper it is a conceptual reduction carried out *a priori*. For instance, the mental concept "pain" is translated into a statement about pain and its role in avoiding bumping into more tables in the future. The second step in Kim's model is to find the realizer of the functionalized property, that is, to find the property which fulfills to role of avoiding bumping into more tables. The third and last step is an explanation of how the property which fits the functional specification does its job.

3. An Argument for Conceptual Reduction

Let us assume, in the spirit of physicalism, that pain has been successfully reduced to a C-fiber stimulation on the ontological level. Moreover, every C-fiber stimulation is describable in purely physical concepts. Does it follow that pain is describable in purely physical concepts as well? Let us take a closer look:

(1) Pain = C-fiber stimulation.
(2) Every C-fiber stimulation is describable in purely physical concepts.
(3) Ergo, every pain is describable in purely physical concepts.

This argument claims that as a result of the type identity statement "pain = C-fiber stimulation" and the descriptiveness of every C-fiber stimulation in physical concepts, one is entitled to conclude that also pain is describable in physical concepts. The expression "C-fiber stimulation" in premise (2) can be replaced by "pain" in virtue of their identity established in premise (1). On what grounds could this argument be refuted? One objection comes from the proponents of *a posteriori* physicalism. *A posteriori* physicalists acknowledge that *a priori* physicalists have to accept that argument, and therefore, a conceptual reduction follows from an ontological reduction, but only if *a priori* physicalism is true. However, if *a posteriori* physicalism is true, the argument is a *non sequitur.* (3) does not follow from (1) and (2), because in order to describe pain, one has to use the mental concept "pain". The reason for this is the independence between mental and physical concepts. Once the ontological reduction has been accomplished *a posteriori,* the coreference of the mental and physical concept under discussion has been established as well, but what has not been established is the synonymy. The point is the following: a mental concept can only be conceptually reduced to a physical concept if the physical concept is synonymous with the mental concept. To be synonymous means to have the same *meaning*, and not just to have the same referent (Frege 1892 has famously and convincingly argued for the distinction between *Sinn, i.e.,* meaning, and *Bedeutung, i.e.,* reference). For instance, although water is identical to H_2O on the ontological level, the concepts "water" and "H_2O" do not have the same meaning, but merely the same referent, *i.e.,* H_2O. Due to the lack of synonymy the concept "water" cannot be reduced to the concept "H_2O". *A posteriori* physicalists carry over this analysis to the case of "pain = C-fiber stimulation". Although pain is nothing over and above a C-fiber stimulation on the ontological level, the mental concept "pain" is not reducible to the concept "C-fiber stimulation". The pressing question is whether *a posteriori* physicalism should be the kind of physicalism of choice. This issue cannot be settled easily, yet I want to raise two somewhat related problems for the *a posteriori* physicalist.

First, to ontologically reduce mental properties to physical properties by appealing to the Kripkean necessary *a posteriori* seems to be a red herring, for no account has been given of how mental concepts being independent and distinct from physical concepts fit into the physicalist picture. It seems that the problem has carried over from the ontological level to the conceptual level without losing any of its original force. Instead of asking "How do mental properties fit into the physicalist story?" one must ask "How do mental concepts fit into the physicalist story?". For instance, Loar claims that mental concepts, *e.g.,* "pain", are nothing over and above type demonstratives with the form "*that* kind of experience". Therefore, mental concepts are no concepts *sui generis,* but they are some kind of demonstratives which in turn are not any threat for physicalism. The situation is this: on the one hand, mental concepts are irreducible and therefore independent from physical concepts, but on the other hand, mental concepts *are* demonstratives which are not in conflict with physicalism and thereby can be viewed as some sort of physical concepts. The problem for the *a posteriori* physicalist is that he cannot have both. Either mental concepts are physical concepts or they are not. The *a posteriori* physicalist seems to beat around the bush when trying to answer that question. It is at least a little odd and confusing, if not plainly contradictory, to say that mental concepts are independent from, but at the same time some kind of, physical concepts. Second, one of the aims of every physicalist is parsimony as well as elegance and simplicity in the construction of his theories. Consequently, if these tenets of physicalism are taken seriously, they should also be applied to the conceptual level and the best way of doing so is by means of a conceptual reduction. The reason for this is that spelling out theories gets simpler, more elegant, and more parsimonious with just one kind of concepts, *i.e.,* just physical concepts. Let us make the point clear: scientific psychological theories are not the same as poems. Poems need a lot of fancy words, but scientific theories just do not.

4. Conclusion

To come to an end, *a priori* physicalism is committed to both, an ontological and a conceptual reduction. Thereby, an ideal amount of parsimony, elegance and simplicity has been accomplished. In contrast, *a posteriori* physicalism, *prima facie,* does not require a conceptual reduction. However, as I have argued, denying the need for a conceptual reduction is in tension with the tenets of physicalism, *viz.* parsimony, elegance and simplicity. Moreover, an inherent problem for *a posteriori* physicalists is to give an adequate account of how mental concepts can be independent from physical concepts and at the same time be some kind of special physical concepts. In conclusion, my analysis suggests that *a priori* physicalism is the best option for defending the view that there is nothing over and above the physical. Thus, *a posteriori* physicalism should be rejected in favor of *a priori* physicalism.

Acknowledgments

I would like to thank Johann Marek and Michael Matzer for very helpful comments on earlier versions of this paper.

Literature

Frege, Gottlob W. 1892 "Über Sinn und Bedeutung", *Zeitschrift für Philosophie und Philosophische Kritik* 100, 25-50.

Jackson, Frank 1998 From Metaphysics to Ethics: A Defense of Conceptual Analysis, Oxford: Clarendon.

Kim, Jaegwon 2005 *Physicalism, or Something Near Enough,* Princeton: Princeton University Press.

Kripke, Saul 1980 *Naming and Necessity,* Cambridge: Harvard University Press.

Loar, Brian 1997 "Phenomenal States", in: Ned Block, Owen Flanagan, and Güven Güzeldere (eds.), *The Nature of Consciousness,* Cambridge: MIT Press, 597-616.

Ryle, Gilbert 1949 *The Concept of Mind,* New York: Barnes and Noble.

Reduction and Reductionism in Physics

Rico Gutschmidt, Bonn, Germany

The good old standard definition of reduction is penned by (Nagel 1961) and demands for a theory to be reduced to another that the laws of the first one can be logically deduced by the laws of the latter with the help of bridge laws connecting the different languages of the theories. A theory reduced in this way should then be in principle superfluous - if all their laws are, given the bridge laws, logically contained in the reducing theory, it is in a strict sense not required anymore in our physical description of the world.

But things are not that easy. As (Feyerabend 1962) has shown, this concept is somewhat naïve and there are no interesting examples of reduction in the Nagelian sense. Feyerabend's point is based mainly on two objections. First, the links between physical theories are mostly an approximate derivation of laws rather than their deduction - and there is a great and in the debate of reduction largely overlooked difference between derivation and deduction. And second, the conception of the bridge laws is rather vague: Feyerabend argues that the terms of different theories satisfy not only no identity relation, which could be expressed in bridge laws, but are actually incommensurable and not comparable whatsoever.

Let us take a closer look on these two assertions. First, there is the mathematical problem of approximate derivation: Within physics it seems to be a well established practise to derive laws "only" approximately. But what does this mean in the context of intertheoretic relations? To take an example from the context of gravitation, according to Newton's law of gravitation Galileo's law of falling bodies is strictly speaking false: The acceleration increases instead of being constant. Hence these theories contradict each other, and therefore a deduction is simply impossible and any "derivation" of Galileo's law from Newton's law of gravitation must thus contain some contra-to-fact assumptions. Such assumptions can in this case be and are widely in physical derivations hidden in limiting processes where some parameter, which is not zero or infinite within the law to be derived, is taken to be zero or infinite. In our case the distance to earth of the falling body compared to the earth's radius is taken to be zero - but the law derived under this assumption is strictly speaking only valid for bodies laying down on the earth's surface, while Galileo's law is about falling bodies. Thus the common mode of speaking that this derivation delivers approximate validity only for small distances covers the fact that we haven't deduced Galileo's law but rather established a *comparison* between the two theories under certain circumstances: This is all we can say about "approximate derivations" here and similarly elsewhere.

Nevertheless, Galileo's law is superfluous, not because of being deduced, but rather because Newtonian physics can also describe falling bodies, in a similar manner as Galileo's law as shown in the comparative limiting process. But if "reducing" theories are more complex it is far from certain that they are able to reproduce any statement of a theory to be reduced. Comparisons between theories in the sense of approximate derivation seem to be just comparisons of mathematical structure and not of concrete explanations of phenomena - and the possibility to compare mathematical structure does not include that the "reducing" theory is able at all to deal with the phenomena explained by the theory to be reduced as it is the case in our simple example. If we are able to deduce the laws of a theory we are automatically able to explain their phenomena but we can't expect to be able to do that by virtue of comparisons between mathematical structures as we will see in the closing part of this presentation, which discusses the case of general relativity in this respect.

This point doesn't catch one's eye if one considers simple cases like that of Galileo's law and has therefore widely been overlooked within the debate of reduction. But intertheoretic relations in physics actually *are* in many cases nothing but comparisons between mathematical structures: A look at the details shows, that e.g. in the case of the Newtonian theory of gravitation and general relativity the intertheoretic relation is much more complicated than a limiting process and far from being well established. A mathematical relation in a precise manner between these theories is given e.g. in (Scheibe 1999) in terms of a topological comparison between sets of models of these theories formulated axiomatically (cf. p. 59-108 for the case of general relativity). And while there are only single cases in which explanations of concrete phenomena can be compared (e.g. the planet's orbits, cf. p. 89-101), Scheibe's "reduction" of the whole Newtonian theory of gravitation is not completely worked out and can either way be no deduction but nothing but a very subtle comparison between mathematical structure.

The second of Feyerabend's objections concerns language and the incommensurability of the vocabulary of different theories. In our context, the equation of motion within general relativity is the geodesic equation for neutral test particles whereas Newton's law of gravitation describes a force between two masses. We thus are concerned with two entirely different concepts and the *identification* of the Newtonian gravitational potential with Christoffel Symbols, which can be found in physics textbooks (cf. e.g. (Misner et al. 1973), chapter 12), is a "component manipulation" (l.c., p. 290) rather than a basis for a deduction. Even more concrete, in the example of the planet's orbits their description within the Schwarzschild solution deals with test particles without influence to the overall curvature and thus without gravitational masses whereas their Newtonian description is based on forces between just these masses. Therefore, these concepts cannot be related by any simple identification and we have to concede that we cannot establish reductions via logical deduction with the help of bridge laws.

But nevertheless theories need not to be incommensurable – we can of course compare the concepts of different theories. But this is in general a difficult and not straight forward comparison process far from being able to establish bridge laws suiting for a logical deduction. We can relate the terms of two theories with the help of special case studies and prove that e.g. the Newtonian potential is somehow related to the Christoffel symbols, but such case studies are no self-evident processes and again lead to a comparing relation rather than a deduction, and such a comparison on its own doesn't make any theory superfluous.

All in all, we have seen that reduction via deduction has no interesting examples because intertheoretic relations typically are no deductions but comparisons between both mathematical structure and terms differing completely in its usage. So, if we want to define a relation of reduction we cannot rely on deduction.

Having in mind that in our investigation we are looking for a concept of reduction that is able to support claims of reductionism, one of the most interesting answers to that problem is that of (Schaffner 1967). For him a theory is not reduced to another if their laws are logically deduced, but if it is possible to deduce a new *corrected theory* from the reducing one which is formulated in the latter's vocabulary and *strongly analogous* to the original one to be reduced. In a similar manner, reduction is defined in (Hooker 1981), and recently in (Bickle 1998) as the so called *new wave reduction*, which is a result of merging Schaffner's and Hooker's concept of analogy with the structuralistic approach to physical theories, as e.g. (Endicott 2001) points out.

Let us have a closer look at these concepts. At first, Schaffner's definition rests on a very vague and not clearly defined "strong analogue"-relation between the original theory to be reduced and a "corrected theory" deduced from the reducing one. This relation can surely be made more precise within the structuralistic approach: As stated in (Moulines 1984), a reduction in structuralistic terms yields a "mathematical relationship between two sets of structures" within a "scheme of reduction" which "does not require semantic predicate-by-predicate connections nor deducibility of statements" (l. c., p. 54-55). So this approach delivers indeed a very sophisticated concept of comparing theories that avoids the difficulties of a concept based on deduction, but can on the other hand be nothing but a comparing relation between both concepts and laws. Moreover, this account yields a comparison between two independent theories in such a way, that "we could have a reductive relationship between two theories that are completely alien to each other" (ibid.).

Such a comparing relation now substantiates only reductionistic claims, if it is a comparison between concrete explanations as in our simple case above, but that seems not to be the case if the theories involved are more complex as we will see below considering general relativity. A topological comparison in the sense of (Scheibe 1999), which is possible also between theories "that are completely alien to each other", doesn't make any theory superfluous and hence cannot on its own support reductionistic claims. It is much easier to establish a mathematical-conceptual comparing relation between two theories than to show that the one can explain the phenomena which are typically explained by the other. Now, if it is possible to deduce a "corrected theory" from a reducing one, this would show that the latter is able to cope with the phenomena described by the theory to be reduced. However, comparing relations can be established between theories without deducing a corrected theory or explaining phenomena, and there are indeed theories (as general relativity) not permitting such a deduction or explanation despite being comparable to another one (for instance to Newton's theory of gravitation) – hence the Schaffner-Hooker-Bickle account of reduction seems not to be adequate.

Guided from these observations, I'd like to propose the following two definitions. First it seems to be appropriate to call most of the intertheoretic relations in physics a relation of *compatibility*: One actually can compare two independent theories with each other and mostly such comparisons show that the involved theories are via approximate derivation and related concepts compatible to each other. This term doesn't evoke any reductionistic claim and isn't meant to do so. If we want to find, secondly, a definition of a relation of *reduction* in such a way that a reduced theory is in principle superfluous, it seems that we have to refresh an idea of (Kemeny and Oppenheim 1956). Their definition of reduction is based on the explanation of phenomena ("observable data" in their terms, cf. p.13): Any phenomenon explainable by means of the theory to be reduced must be explainable by the reducing theory. If, furthermore, the explanations of the reducing theory are in a sense better and if the theories involved are compatible and therefore in a way related to each other, it seems legitimate to say that the one is reduced to the other. A theory reduced in this sense is indeed superfluous: "Their" phenomena are explained better by another theory, to which it is compatible (and there is no need to go a long way round via "corrected theories"). This is now surely the case e.g. for Galileo's law of falling bodies or Kepler's laws of planetary motion but not for Newton's theory of gravitation: We will now see that we have "only" compatibility here.

The reason is that in spite of having comparable laws as shown e.g. in (Misner et al. 1973) or (Scheibe 1999), there are many phenomena explained by Newtonian physics but not by general relativity, because no one solved the field equations for them. While the two-body problem is directly solved by Newton's law, it has (and as a matter of fact can have) only numerical solutions within general relativity. And while the orbits of the planets can be described as geodesics within the Schwarzschild solution, their interaction as described by Newtonian physics is not yet explained by general relativity for there are no solutions of the field equations for moving gravitational sources. Similarly, there are no general relativistic explanations for complex formations as star clusters or spiral galaxies either, while they can too be handled with Newtonian physics. It is surely possible to claim that one will find relativistic explanations of these phenomena one day, but because of the difficult and abstract character of the field equations in contrast to the high applicability of Newton's theory we can also put the possibility of such explanations in question. And indeed, as a matter of fact numerical simulations of such phenomena on the basis of the field equations depend due to their complex structure on the heuristic help of Newton's law of gravitation (within the so-called post-Newtonian approximation). Therefore, Newton's law of gravitation can be improved by general relativity, but is not superfluous – it is in our terms *not reduced* to general relativity *despite compatibility*.

Literature

Bickle, John 1998 *Psychoneural Reduction: The New Wave*, Cambridge, MA: MIT Press.

Endicott, Ronald 2001 “Post-Structuralist Angst-Critical Notice: John Bickle, Psychoneural Reduction: The New Wave”, *Philosophy of Science* 68(3), 377-393.

Feyerabend, Paul 1962 “Explanation, Reduction, and Empiricism”, in: H. Feigl and G. Maxwell (eds.), *Scientific Explanation, Space, and Time*, Minneapolis: University of Minnesota Press, 28-97.

Hooker, Clifford A. 1981 “Towards a General Theory of Reduction. Part I-III”, *Dialogue* 20, 38-59, 201-236, 496-529.

Kemeny, John and Oppenheim, Paul 1956 “On Reduction”, *Philosophical Studies* 7, 6-19.

Misner, Charles W., Thorne, Kip S., and Wheeler, John A. 1973 *Gravitation,* New York: W.H. Freeman and Company.

Moulines, C. Ulises 1984 “Ontological Reduction in the Natural Sciences”, in: Wolfgang Balzer, David A. Pearce and Heinz-Jürgen Schmidt (eds.), *Reduction in Science: Structure, Examples, Philosophical Problems*, Dordrecht: Reidel, 51-70.

Nagel, Ernest 1961 The Structure of Science: Problems in the Logic of Explanation, New York: Harcourt.

Schaffner, Kenneth F. 1967 “Approaches to Reduction”, *Philosophy of Science* 34, 137-147.

Scheibe, Erhard 1999 Die Reduktion physikalischer Theorien Ein Beitrag zur Einheit der Physik Teil II: Inkommensurabilität und Grenzfallreduktion, Berlin: Springer.

Physicalism Without the A Priori Passage

Harris Hatziioannou, Athens, Greece

Defenders of a priori entailment hold that physicalism is committed to the following two theses: first, that all macroscopic (high-level) facts, including facts about the mind, are necessitated by the totality of physical (low-level) facts, and, second, that, granting knowledge of the latter set of facts, we can deduce the former without needing any further empirical information. My target in this paper will be the claim regarding the second commitment of physicalism. Specifically, I will argue against two different formulations of the thesis and then conclude with some suggestions regarding the way in which we may understand a posteriori physicalism and the determining relation between low and high-level facts that it posits.

A number of attempts have been made to analyze the thesis that physicalism is committed to the idea that physical facts a priori determine all other facts. One prominent example is David Lewis's account (Lewis 1972), which, relying on the Ramsey - Carnap method of defining theoretical terms, appeals to the functional definability of high-level terms in order to *deduce* them from the terms of the reducing theory. By this procedure, the theoretical terms in question are understood in terms of the relations that they hold among each other, as these are expressed in a vocabulary of which we had prior understanding: they are explicitly defined as the unique entities, whatever these may be, that occupy the causal roles specified by the theory. In the mind/body case, the theory under reduction, 'folk psychology', is supposed to include all commonly known platitudes about the mind, platitudes that are built in our a priori understanding of these terms. In Lewis's view, these platitudes do nothing more than specify the position of each mental state in the causal nexus in which it partakes; thus, each mental state can be in principle explicitly defined in terms of its characteristic causes and effects. With the expected advancement of science, and when these same causes and effects are given a physical characterization, we will be able to identify the states picked by the two different sets of terms, thus effecting the reduction of our folk psychological theory of mind to the more comprehensive physical theory.

Now, Lewis's method provides a clear picture of the way in which the a priori entailment of facts about the mind by physical facts could be understood. Given the functional definitions, the identification will clearly be the result of a deductive inference. However, Lewis's contention that mental terms can be explicitly defined in terms of their causal role is hardly convincing, not only to non-physicalist, but also to many physicalist philosophers. The problem with such explicit analyses is threefold: First, they seem to misconstrue the conventional meaning of such terms; our concepts picking conscious states have strong non-causal connotations, so that a functional definition a la Lewis is bound to miss some part of the meaning we conventionally associate with them. Second, they ignore the possibility of multiple realization. There is widespread agreement, at least since Wittgenstein, that we can find explicit analyses in terms of necessary and sufficient conditions only for few of our mental or other everyday concepts: *automobile*, *life*, or *belief that X*, are all concepts that seem to be multiply realizable, in the sense that it is impossible to specify in a finite non-trivial way the conditions of application that will capture all and only their referents. The third objection points to the fact that conceptual analyses such as Lewis's seem to hold future empirical research into the nature of mind hostage to a priori meaning considerations. Raising the widespread platitudes about the mind to the status of a priori definitions of mental terms pays no heed to the fact that our concepts evolve continuously in the light of novel empirical as well as conceptual developments. The moral to be drawn from the foregoing considerations, is that the explicit analysis of mental and other terms that we commonly use in order to describe the world, and their corresponding concepts, is something that we cannot aspire to, since any proposed analysis is likely to miss an essential part of their content.

That same moral has been drawn by Chalmers and Jackson (Chalmers & Jackson 2001, Chalmers 1996, Jackson 1998), who have proposed an alternative scheme for reduction. Their scheme eschews such finite explicit analyses, being based instead on a priori intensions, which are understood as functions from possible worlds to extensions, and that cannot be put into any explicit linguistic description. These functions are supposed to capture our implicit knowledge of the application conditions of our concepts, the kind of knowledge that allows us to judge, on a case by case basis, whether they apply to a certain situation or not. So, given a non-trivial neutral description of a possible world, considered to be the actual world, we can (ideally) determine the extension of our concepts. For example, given a description of a world where the salient transparent, odourless, drinkable etc. liquid ('the watery stuff') in the environment is H_2O, the a priori intension of the concept 'water' refers to H_2O; given a description of a world where the watery stuff is XYZ, it refers to XYZ. In other words, the component of meaning that is a priori associated with any given term or concept has no explicit description; it is encompassed in the term's a priori intension, the function that determines the term's extension in every world considered as actual. Accordingly, no appeal to explicit analyses needs to be made in accounting for the relation of a priori entailment that supposedly holds between low-level and high-level facts: this relation can simply be analysed as one between functions from possible worlds to extensions.

Chalmers's and Jackson's account thus avoids Lewis's questionable commitment to the explicit definability of theoretical terms. However, the problem is that it seems to have lost the transparency that characterized Lewis's way of analysing the relation of a priori entailment. With the latter method, the question whether a certain instance of inter-level reduction succeeds has a clear answer: if the proposed functional analyses of the respective sets of terms are in place, then the deductive inference and, consequently, the inter-level reduction, can be carried through. Chalmers's and Jackson's method, by repudiating explicit conceptual analyses, has lost this virtue: for any proposed instance of inter-level reduction, the answer whether it succeeds or fails lies in the viability of our intuitions regarding the referents of our concepts in various counteractual scenarios. But this method is always going to be open to scrutiny, since a sceptic may appeal to indeterminacy, different intuitions, or even knowledge deficit, in order to question a proposed reduction of one theory to another, or a suggested failure thereof. A priori entailment, understood as in Chalmers's and Jackson's way, gives us no principled way to demonstrate the success or failure of reductions.

What is more, it seems to me that Chalmers's and Jackson's account fails to ground the metaphysical necessity that, according to physicalism, connects physical to mental states, on relations between concepts, propositions, or whatever else can become objects of knowledge, whether that be a priori or a posteriori. This is because, by construing the relation of a priori entailment as one between functions from possible worlds to extensions, their account seems to reverse the required order of explanation, analysing a supposedly logical - conceptual relation in metaphysical terms, whereas what was being advertised originally was exactly the opposite. To say that knowledge of the complete physical description of the world would allow us to have a priori knowledge of all macroscopic facts, which is what the original thesis about a priori entailment included, is to say that our understanding of the terms that we use in order to describe the world in physical terms, allows us to deduce, without looking at the world any further, its complete description in macroscopic terms. But, an essential component of this thesis seems to be that it is our *a priori understanding of the concepts involved* that grounds the derivation, not the objects that are referred to by these concepts. In Chalmers's and Jackson's account, the burden of the derivation is transferred to the level of extensions, not the way that these extensions are represented by us. This, in my view, renders the account unsuitable to be used in explicating the way metaphysically necessary relations are grounded on logical - conceptual relations such as that of entailment.

In later work, Jackson tries to accommodate similar criticisms, by attempting to reconcile the metaphysical nature of his proposed relation of entailment with the apriority that supposedly characterizes it. Thus, he calls the relation of a priori entailment that he endorses *de re,* claiming that it is a type of metaphysical necessitation between *properties*, not necessitation between *sentences* or *concepts*. Here is the relevant definition that he appeals to, quoted directly from his paper:

> P_1 a priori necessitates P_2 iff one's grasp of what it is to be a P_1 and what it is to be a P_2 allows one to see that if P_1 is instantiated then so is P_2.
> (Jackson 2005, pp. 252-3).

However, I do not think that this move can serve Jackson's argument. It is obvious that, contrary to his pronouncements, his characterisation of de re a priori necessitation, by appealing to our 'grasp of what it is to be' a given property, proceeds via concepts; given this, Jackson clearly has to suggest a way in which these concepts are related. Since he repudiates explicit conceptual analyses, he has to represent the relation between them as a relation between functions, i.e. a priori intensions. But, as has been argued above, lacking explicit analyses, the relation between these intensions can never be cognitively obvious, since it operates at the limit, i.e. as a relation between the concepts' extensions across possible worlds. This cannot be reasonably thought of as a relation between mental or linguistic representations; rather, it seems to be closer to what he and Chalmers would call a metaphysically 'brute' one.

But, if the commitment of physicalism to a priori entailment is to be rejected, so that our true high-level descriptions of the world cannot be derived from the complete physical story, then how can the thesis ever be vindicated? If physical properties fail to account completely, in a fully reductive account, for the successful explanations that we give in macroscopic terms, in what sense can they be considered superior in the explanatory scheme of things? A full answer to these questions certainly needs a systematic investigation into the relation between properties that pertain to different levels of explanation. However, I think that the key for accepting physicalism without an a priori passage is to realize that, even within the domain of physics, the descriptions of complex systems can very rarely be derived in an a priori way from the descriptions of their constituents. The derivation must necessarily involve idealizations, simplifying assumptions, approximations, and brute numerical methods, all techniques that rupture the smooth mathematical derivation of properties that pertain to more complex systems, thus rendering impossible the a priori passage from one level of description to another. The failure is already apparent in the classical three-body problem, which has exact solutions only in some restricted forms. And, of course, it is patently obvious in more complex systems: even knowing the momentary positions and momenta of all the fundamental particles that comprise an iron atom, say, we have no hope at all of a priori deducing, on the basis of our best laws of quantum theory, their dynamical evolution in time. To arrive at the physics of such complex systems from the physics of more simple ones, we simply have to resort to methods which are, at least in part, justified by appeal to empirical data. Physics simply does not have an analytically solvable equation that describes the behaviour of every system that falls within its scope. In fact it has such equations only for unrealistic, highly idealized systems, which are encountered in tightly controlled experimental situations (see Cartwright 1999).

I think that the important point to keep here is the following: If the passage from simple to complex systems is not a priori guaranteed even within physics, then we should not expect that the passage from each neurophysiological to its corresponding mental state will be thus guaranteed either. I am aware that this point raises doubts concerning the way in which an antireductionist account such as this could support physicalism: if no smooth reductions are forthcoming, not even within the domain of physical theory, how can we ever be confident that it is the *physical* properties that account for the causal and other characteristics of mental states? I believe that, just like in the cases of the classical three-body problem and the iron atom we have reason to believe that the features of the world that are responsible for the dynamical evolution of these systems are *physical* (albeit ones that cannot be precisely quantified), so in the case of a psychological state we have reason to expect that the features that are responsible for its causal outcomes are also physical, even if there is no way to move from the physical description to the psychological one, except by appealing to 'brute' a posteriori knowable necessitation. Thus, we may view the psychological description as capturing, vaguely and imprecisely, the salient features of the physical system, while at the same time expecting that it will be the description of the underlying complex physical state that will ultimately fully account for mental phenomena, in the sense of providing the exact sufficient causes for them, the effective mechanisms that are present in the world. It is true that we need independent arguments to warrant this expectation, but, of course, plenty of these have already been given in the literature, and there is no space to discuss them here. I hope at least that these sketchy suggestions point towards a viable understanding of a posteriori physicalism.

Literature

Cartwright, Nancy 1999: *The Dappled World*. Cambridge.

Chalmers, David 1996: *The Conscious Mind.* Oxford.

Chalmers, David & Jackson, Frank 2001: ‘Conceptual Analysis and Reductive Explanation’, in: *The Philosophical Review* 110, 315-61.

Jackson, Frank 1998: From Metaphysics to Ethics. A Defense of Conceptual Analysis. Oxford.

Jackson, Frank 2005: ‘The Case for A Priori Physicalism’, in: N. Christian and A. Beckermann (eds.), *Philosophy–Science–Scientific Philosophy*. Main Lectures and Colloquia of GAP.5, Fifth International Congress of the Society for Analytical Philosophy, 2003 Paderborn: Mentis (pp. 251–265).

Lewis, David 1972: ‘Psychophysical and Theoretical Identifications’, in: *Australasian Journal of Philosophy* 50, 249-258.

Wittgensteins Projektionsmethode als Argument für die transzendentale Deutung des *Tractatus*

Włodzimierz Heflik, Krakau, Polen

Einleitung

Die Projektionsmethode, die ich in diesem Beitrag untersuche, wird von Wittgenstein in den Thesen 3.11 - 3.14 des *Tractatus* eingeführt, und daraufhin von einer anderen Seite her in der These 4.0141 besprochen. Der Hauptgedanke dieser Methode ist im folgenden Abschnitt enthalten:

> „Wir benutzen das sinnlich wahrnehmbare Zeichen (...) des Satzes als Projektion der möglichen Sachlage. Die Projektionsmethode ist das Denken des Satzsinnes." (3.11)

Die ontologische Basis für die Projektionsmethode bestimmen die Thesen über die Abbildungsform als das, „was das Bild mit der Wirklichkeit gemein haben muss" (vgl. 2.151; 2.16 u. 2.17). Die Form der Abbildung wiederum hat ihre endgültige Begründung in einfachen Gegenständen als „die Substanz der Welt" (vgl. 2.021).

Das Ziel dieses Beitrags ist, die Projektionsmethode nicht nur im Bereich des Systems des *Tractatus* zu zeigen, sondern auch einen Versuch zu unternehmen, diese Methode im Bezug auf Kants Philosophie darzustellen. Auf diese Weise möchte ich festlegen, ob die transzendentale Interpretation der sogenannten ersten Philosophie Wittgensteins berechtigt ist.

Jetzt stelle ich drei Bemerkungen als Hypothesen dar, in denen drei Analogien formuliert werden, die sich zwischen der Problematik der Deduktion der Kategorien bei Kant und der Frage nach Wittgensteins Methode der Projektion des Sinnes beobachten lassen:

(1) Die von Wittgenstein angenommene Hauptvoraussetzung über das Vorkommen der Abbildungsform, die etwas Gemeinsames für das Bild/Satz und Tatsache ist, ist ein Analog der Hauptthese der metaphysischen Deduktion der Kategorien, und diese lautet: „Dieselbe Funktion, welche den verschiedenen Vorstellungen in einem Urteile Einheit gibt, die gibt auch der bloßen Synthesis verschiedene Vorstellungen in einer Anschauung Einheit, welche (...) der reine Verstandesbegriff heißt." (A79/B105).

(2) Aussagen darüber, dass (i) „das Bild die Wirklichkeit erreicht", und (ii) über „Zuordnungen" der Elemente des Bildes den Gegenständen" (vgl. 2.15 - 2.1515), sind analog zum Hauptgedanken der transzendentalen Deduktion – wie sich die Kategorien auf die Gegenstände der Erfahrung beziehen.

(3) Die Thesen des Tractatus drucken die Projektionsmethode aus (vgl. 3.11 -3.13 u. 3.1431), entsprechen dem, was Kant unter dem Leitwort des „Schematismus" versteht. Er führt die sogenannte Schematisierung der Kategorien durch und hebt die Rolle der Einbildungskraft im Prozess des Bezugs der Kategorien auf Erscheinungen, bzw. Gegenstande der Erfahrung hervor.

Jede der obigen Bemerkungen verlangt die Entwicklung und Rechtfertigung, die ich nun darbiete.

1

Jenen wesentlichen Abschnitt der metaphysischen Deduktion (A79/B105) kann man - in Bezug auf einige im *Tractatus* auftretende Ideen - folgendermaßen verstehen. Diese von Kant genannte Funktion kann mit der logischen Form, d.h. auch mit der Form der Abbildung gleichgesetzt werden. Daran wird deutlich, dass hier ein weitgehender Einklang mit der von Wittgenstein gegebenen Bestimmung vorliegt: „Die Form ist die Möglichkeit der Struktur" (2.033). Diese Form bzw. Funktion, die eine Struktur (1) den Tatsachen und (2) Sätzen als deren Bilder gibt, verleiht zugleich den Tatsachen und auch Sätzen Einheit. Die Struktur besteht in einer Verbindung der Elemente - deren Konfiguration. Wir haben also mit solch einer Verbindung zu tun: sowohl auf der Seite der Tatsachen, d.h. Erscheinungen als auch auf der Seite ihrer Bilder, d.h. es kommt zu einer Verbindung der Zeichen in Form des Satzzeichens. Das Vorkommen dieser Verbindung der Elemente - ob im Urteil oder in der Anschauung/Tatsache - ist Kant zufolge gleich, mit dem Angeben der Einheit bzw. der Beziehung auf die Einheit.

Diese Einheit nennt Kant „Kategorie". Die Elemente des Urteils, die durch die Kategorien verbunden werden, sind empirische Begriffe, d.h. unanschauliche Vorstellungen, d.h. Zeichen im Sinne Wittgensteins. Diese Elemente als Vielheit und Mannigfaltigkeit werden zuerst zusammengesetzt. Das heißt: Sie werden synthetisiert, aber ohne dass ihren eine Struktur gegeben wird. Erst der Bezug auf Kategorien ermöglicht ihnen eine Gestallt des Urteils zu erreichen. Analog dazu verläuft der Prozess bei den Anschauungen, d.h. Erscheinungen oder Tatsachen. Diesen beschreibe ich nun in Wittgensteins Terminologie. Tatsachen sind also bestehende Sachverhalte, genauer gesagt – Mitvorkommen zugleich der vielen elementaren Sachverhalte (vgl. Brief an Russell, Cassino 19.08.1919). Der Sachverhalt ist eine Bindung der Gegenstände (2.03) oder deren Konfiguration (2.02071) - eine Synthese ersten Grades. Die Tatsache wiederum als Mitvorkommen vieler Sachverhalte, d.h. ihr Produkt, ist die Synthese zweiten Grade. Dabei aber liegt auch das Angeben der Einheit vor. Das bedeutet: Diese Vielheit der Sachverhalte wird als Einheit erfasst, die Tatsache ist.

Die Wittgensteinschen Tatsachen entsprechen den Gegenständen der Erfahrung bei Kant; beide sind Erscheinungen. Jeder Gegenstand der Erfahrung, d.h. *phenomenon*, ist ein Ergebnis einer zweigradigen Synthese; und zwar jede einzelne Vorstellung, im Augenblick vorkommende, ist eine bestimmte Mannigfaltigkeit. Die Erscheinung wiederum ist eine Vereinigung vieler Mannigfaltigkeiten im Einen, durch die Beziehung auf das Eine, das eine Kategorie ist, die endgültig auch die transzendentale Einheit der Apperzeption ist.

Zusammenfassend weist die betrachtete Analogie auf die Funktion hin, die zwei Ebenen der Vorstellungen auf eine transzendentale Grundlage bezieht. Diese Ebenen sind: (1) Urteile und (2) Erscheinungen bei Kant, (1') Sätze und (2') Tatsachen bei Wittgenstein. Diese Grundlage wird durch Kategorien und transzendentale Apperzeption festgelegt bei Kant, hingegen bei

Wittgenstein durch einfache Gegenstände und logische Form.

2

Die zweite Analogie spricht von der Abbildung: ‘Bild→Tatsache’, die in der Wittgensteinschen Terminologie erfasst ist; diese Abbildung wird dann im Licht der Kantschen transzendentalen Deduktion dargestellt. Anders gesagt: Dies ist ein Versuch eine Antwort auf die Frage im Still Kants zu geben: Wie bezieht sich das Bild auf Tatsache/Wirklichkeit? Die Schlussformulierung Wittgensteins sagt, dass das Bild bis zur Wirklichkeit reicht. Es sei allerdings dabei erinnert, dass Wittgenstein zuvor folgende Thesen aufgestellt hat:

> „Wir machen uns Bilder der Tatsachen.” (2.1)
> „Das Bild ist eine Tatsache.” (2.141)

Daran kann man deutlich erkennen, dass das Bild der Tatsachen auch eine Tatsache ist, demzufolge ist das Beziehen des Bildes auf die Tatsache eine Relation, die zwischen zwei Tatsachen besteht. Die Grundfrage lautet: Auf welche Weise erreicht eine Tatsache (Bild) die andere Tatsache, d.h. die Wirklichkeit? Wittgenstein erklärt:

> „Die abbildende Beziehung besteht aus den Zuordnungen der Elemente des Bildes und der Sachen.” (2.1514)
> „Diese Zuordnungen sind gleichsam die Fühler der Bildelemente, mit denen das Bild die Wirklichkeit berührt.” (2.1515)

Diese Zuordnungen können als Vektoren verstanden werden. Diese Vektoren sind an Elemente des Bildes befestigt und in die Richtung einzelner Dinge in der Wirklichkeit gerichtet; so dass das Vektorende den Ort berührt, an dem sich das bezeichnete Ding befindet. Demzufolge stellt These 2.1514 fest, dass das Verhältnis zwischen dem Bild und der Tatsache wesentlich nur eine Summe der Zuordnungen ist.

Allerdings genügt diese Summe der Zuordnungen, d.h. die abbildende Beziehung, allein noch nicht, um ein Bild auszumachen. Das Verhältnis der Abbildung ist eine notwendige Bedingung des Bildes, aber keine ausreichende Bedingung. Umgekehrt, erst wenn wir ein Bild haben, können wir in diesem diese Zuordnungen erkennen. Kurz gesagt: Die abbildende Beziehung bzw. die Summe der Zuordnungen gibt noch keinen Sinn. Der Sinn also ist weder diese Summe, noch lässt er sich auf diese Summe reduzieren; der Sinn ist etwas mehr.

3

Erst die Projektionsmethode führt die Antwort auf die Frage aus, wie sich das Bild auf die Wirklichkeit bezieht. Wir befassen uns hier hauptsächlich mit den Sätzen als einer besonderen Art der Bilder.

Bereits in den Tagebüchern erscheint ein Vermerk, der auf Wittgensteins Interesse an der Frage nach dem Suchen des im Satz verbogenen Mechanismus hinwies. Dieser Mechanismus bewirkt, dass der Satz über eine Kraft der Abbildung verfügt:

> „Jener Schatten, welchen das Bild gleichsam auf die Welt wirft: Wie soll ich ihn exakt fassen? Hier ist ein tiefes Geheimnis. (...)
> Der Satz ist eben nur die Beschreibung eines Sachverhalts. Aber das ist alles noch an der Oberfläche.” (15.11.1914)

Der Leitfaden der Projektionsmethode ist in den darauffolgenden Thesen des Tractatus beschrieben:

> „Wir benutzen das sinnlich wahrnehmbare Zeichen (...) des Satzes als Projektion der möglichen Sachlage.
> Die Projektionsmethode ist das Denken des Satz-Sinnes.” (3.11)
> „Das Zeichen, durch welches wir den Gedanken ausdrücken, nenne ich das Satzzeichen. Und der Satz ist das Satzzeichen in seiner projektiven Beziehung zur Welt.” (3.12)
> „Zum Satz gehört alles, was zur Projektion gehört; aber nicht das Projizierte.
> Also die Möglichkeit des Projizierten, aber nicht dieses selbst.
> Im Satz ist also sein Sinn noch nicht enthalten, wohl aber die Möglichkeit ihn auszudrücken.
> (Der Inhalt des Satzes heißt der Inhalt des sinnvollen Satzes.)
> Im Satz ist die Form seines Sinnes enthalten, aber nicht dessen Inhalt.” (3.13)

Das Wesen der Projektionsmethode lässt sich auf die Konstruktion des Sinnes des Satzes zurückführen. Das Grundproblem besteht darin, festzustellen, wie Wittgenstein den Sinn versteht und wie viele Gemeinsamkeiten seine Konzeption des Sinnes mit Freges Theorie hat und wie weit von dieser die Stellung Wittgensteins entfernt ist.

Die Bestimmung des Sinnes führt Wittgenstein eher ein, d.h. vor dem Angeben der Beschreibung der Projektionsmethode:

> „Was das Bild darstellt, ist sein Sinn.” (2.221)

Um festzustellen, was sich hinter dem Terminus „Sinn” verbirgt, soll zuerst die Bestimmung „was das Bild darstellt” aus einer anderen Perspektive betrachtet werden. These 2.201 erscheint dabei hilfreich:

> „Das Bild bildet die Wirklichkeit ab, indem es eine Möglichkeit des Bestehens und Nichtbestehens von Sachverhalten darstellt.”

Aus beiden obigen Thesen ergibt sich: Sinn ≡ Möglichkeit des Bestehens und Nichtbestehens von Sachverhalten. Der Sinn gehört zur Sphäre des Möglichen im Gegensatz zum Satzzeichen und der Sachlage in der Welt; diese sind Tatsachen und gehören zur Wirklichkeit. Der Sinn ist gerade der „Schatten”, der vom Satz auf die Wirklichkeit geworfen wird.

Lassen wir vorübergehend das Problem des weiteren Präzisierens der Bestimmung des Sinnes und konzentrieren wir uns auf einem wichtigen Unterschied und zwar den zwischen dem Satz und dem Satzzeichen. Im Lichte von These 3.12: Der Satz = {Satzzeichen + Projektive Beziehung zur Welt}. Vor diesem Hintergrund mag überraschen, was Wittgenstein in folgender These sagt:

> „Zum Satz gehört alles, was zur Projektion gehört; aber nicht das Projizierte. Also die Möglichkeit des Projizierten, aber nicht dieses selbst.” (3.13)

Was ist das „Projizierte”? Kann man dieses mit dem Sinn gleichsetzen? Auf diese Gleichsetzung scheint das nächste Fragment derselben These hinzuweisen:

> „Im Satz ist also sein Sinn noch nicht enthalten, wohl aber die Möglichkeit ihn auszudrücken."

Ob die obige Gleichsetzung berechtigt ist, wird weiter erörtert.

Die Projektionsmethode samt dem Satzzeichen konstruiert den Sinn, der, obwohl er konstruiert worden ist, über eine eigenartige Autonomie hinsichtlich des Satzes verfügt. Außerdem scheint es auch berechtigt, die projektive Beziehung vom Sinn zu unterscheiden. Die projektive Beziehung kann man auch als Intention interpretieren, die dieses Satzzeichen wieder lebendig macht (vgl. Ammereller 2001, 132). Das Wesen des Sinnes wiederum wird in folgenden Thesen des *Tractatus* erleuchtet:

> „Sehr klar wird das Wesen des Satzzeichens, wenn wir es uns, statt aus Schriftzeichen, aus räumlichen Gegenständen (etwa Tischen, Stühlen, Büchern) zusammengesetzt denken. Die gegenseitige räumliche Lage dieser Dinge drückt dann den Sinn des Satzes aus." (3.1431)
> „Der Konfiguration der einfachen Zeichen im Satzzeichen entspricht die Konfiguration der Gegenstände in der Sachlage."(3.21)

Diese Thesen bringen uns zu folgender Verstehensweise des Sinnes. Der Sinn ist eine reine und bloße Konfiguration, abgetrennt von seinem Träger, der aus einfachen Zeichen besteht. Es lässt sich eine Ähnlichkeit feststellen, die zwischen der Konzeption des Sinnes bei Wittgenstein und der Auffassung von Husserl vorkommt. Husserl zufolge ist der Sinn *noemat*, die projektive Beziehung hingegen entspricht der Richtung des Intentionsstrahles. Im Gegensatz zu Frege schlägt Wittgenstein vor, Sinn vielmehr für das, was durch den Denkakt konstruiert wird, zu halten, als was nur in diesem Akt als ewiges, fertiges Objekt erfasst wird. Wittgenstein würde vielmehr sagen, dass ewig die Möglichkeit des Sinnes sei, aber nicht der Sinn selbst (vgl. 3.13). Dieser Unterschied der Ansichten über die Problematik des Sinnes zwischen Wittgenstein und Frege hat auch seinen Ursprung in verschiedenen Annährungen zur Frage nach dem Sinn der Namen. Frege setzt voraus, dass Namen ähnlich wie Sätze auch Sinn haben; Wittgenstein wiederum ist der Meinung, dass nicht Namen Sinn haben, sondern nur Sätze (vgl. Ishiguro 1989). Dieser Vergleich mit der Auffassung Freges hebt hervor, dass der Sinn - in Wittgensteins Auffassung - sehr stark durch die Konfiguration der Gegenstände bedingt wird. Wenn wir dagegen mit einem Namen zu tun haben, der einem Gegenstand als einfaches Objekt bezeichnet, können wir von keiner Konfiguration reden, also – auch von keinem Sinn.

Es bleibt jedoch eine gewisse Doppeldeutigkeit des Terminus „Projektionsmethode" zu klären. Man kann nämlich diese Methode und das Satzzeichen in zweierlei Erfassung betrachten: (1) als nur gemeinsam zusammengesetzt aber nicht verbunden, oder als (2) durch den Gedankenakt miteinander verbunden. Im ersten Fall bleibt die Projektionsmethode nur eine abstrakte Regel, die erst anzuwenden wäre. Im zweiten Fall wird der Gedankenakt mit der Projektionsmethode gleichgesetzt. Im zweiten Fall schafft also der Gedankenakt eine Konfiguration. In These 4.0141 finden wir die Bestätigung, dass Wittgenstein diese Doppeldeutigkeit des Terminus „Projektionsmethode" zulässt:

> „Das es eine allgemeine Regel gibt, durch die der Musiker aus der Partitur die Symphonie entnehmen kann (...), darin besteht eben die innere Ähnlichkeit dieser scheinbar so ganz verschiedenen Gebilde. Und jene Regel ist das Gesetz der Projektion, welches die Symphonie in die Notensprache projiziert (...)"

Diese These zeigt auch, dass sich das „Projizierte" ganz außerhalb des Satzes befindet. Daher ist der Sinn wahrscheinlich etwas Anderes als das Projizierte. Also ging der letzte Vorschlag, den Sinn mit dem Projizierten gleichzusetzen, ging eindeutig zu weit.

Um klarzumachen, was Wittgenstein unter dem „Projizierten" versteht, muss man auf den Zusammenhang zwischen dem Sinn des Satzes und den einfachen Gegenständen achten. In den *Tagebüchern* können wir lesen:

> „Die Forderung der einfachen Dinge ist die Forderung der Bestimmtheit des Sinnes. (...) Wenn es einen endlichen Sinn gibt, und einen Satz, der diesen vollständig ausdrückt, dann gibt es auch Namen für einfache Gegenstände." (18.06.1915)

Das ist offensichtlich, weil der Sinn eine mögliche Konfiguration dieser Gegenstände ist (vgl. 2.0272). Wittgenstein vertritt eine ähnliche Ansicht wie Leibniz in der *Monadologie* (vgl. §§1,2 dieses Werks). Wenn wir keine einfachen Elemente zeigen würden, könnten wir die Konfiguration dieser Elemente nicht bilden. Also könnten wir den Sinn nicht zeigen!

Es kann das „Projizierte", ebenso wie der Sinn, einfach nicht mit dem Sachverhalt gleichgesetzt werden. Ein Sachverhalt ist nämlich eine wirkliche Verbindung der Gegenstände. Daher scheint, dass die Bestimmung „die bestehenden Sachverhalte" (vgl. 2.04 u. 2.05) redundant ist! Demzufolge wird deutlich, dass das ‘Projizierte' kein Sachverhalt zu sein braucht. Falls das ‘Projizierte' Sachverhalt sein müsste, dann könnten wir mit Hilfe der Projektionsmethode nur (bestehende) Sachverhalte rekonstruieren. Es könnte dagegen unmöglich sein, solche Konfigurationen zu konstruieren, die keine wirklichen Verbindungen ausdrücken, d.h. eine Gruppe von Gegenständen, die miteinander nicht verbunden sind. Das ist ebenfalls das Problem der Falschheit und Negation. Daher schreibt Wittgenstein in den *Tagebüchern*:

> „Die Realität, die dem Sinne des Satzes entspricht, kann doch nichts Anderes sein, als seine Bestandteile, da wir doch alles Andere nicht wissen." (20.11.1914)

Wir wissen also nicht, ob das ‘Projizierte' ein Sachverhalt oder nur eine Gruppe von einfachen Gegenständen ist, von denen wir eine falsche Hypothese formulieren, dass diese Gegenstände einen Sachverhalt bilden.

Fassen wir zusammen: Es erscheinen Vieldeutigkeiten, indem wir festzustellen versuchen, wie der Sinn, das ‘Projizerte' und die Projektionsmethode verstanden sein sollen. Diese Schwierigkeit, die den Terminus „Sinn" begleitet, besteht darin, dass der Sinn gleichzeitig: (1) universell und (2) konkret sein muss. Daher kann man der ersten Bedingung zufolge anerkennen, dass Sinn eine reine Struktur/Konfiguration ist. Die zweite Bedingung hingegen ordnet den Sinn als Konfiguration samt der intentionellen projektiven Beziehung an. Das heisst, als einen auf ein bestimmtes Fragment der Wirklichkeit geworfenen „Schatten". In dieser zweiten Erfassung kann wohl der Sinn mit dem ‘Projizierten' gleichgesetzt werden.

Schlussbemerkungen

Die Projektionsmethode, die in diesem Beitrag analysiert wurde, ähnelt in vielen Punkten dem transzendentalen Schema bei Kant. Dieses Schema, als ein Erzeugnis der transzendentalen Einbildungskraft (vgl. A140/B179) bestimmt die Art, wie Kategorien auf Erscheinungen angewendet werden sollen. Die Projektionsmethode bestimmt dagegen die Art der Konstruktion des Sinnes dank der Regeln, denen zufolge zuerst eine Konfiguration der Zeichen, d.h. Satzzeichen, gebildet werden muss. Dieses Zeichen wird dann entsprechend interpretiert, damit der konstruierte Sinn als „Schatten" auf das beabsichtigte Fragment der Wirklichkeit/der Welt geworfen wird.

Ähnlich wie Wittgenstein den Sinn von der Projektionsmethode unterscheidet, grenzt Kant Bilder der sinnlichen Gegenstände und Schemata ab (vgl. A140/B180). Das Schema bedeutet für Kant „eine Regel der Synthesis der Einbildungskraft" und dieses existiert nur „in Gedanken" (vgl. ebenda). Die Projektionsmethode, ähnlich wie das transzendentale Schema, erfordert die Handlung des Gemüts, [um sie anzuwenden.] Wittgenstein erwähnt dabei „das Denken des Sinnes des Satzes", Kant die Handlung der Einbildungskraft. In beiden Fällen ist die Grundlage dieser Handlung - als psychischer Akt- das, was apriorisch und transzendental ist, d.h. ein endgültiger Beziehungspunkt. Bei Kant ist dieser Punkt die transzendentale Einheit der Apperzeption. In Wittgensteins System scheint hingegen die logische Form eine analoge Rolle zu spielen.

In der Projektionsmethode nimmt Wittgenstein an, dass wir zu den einfachen Gegenständen einen direkten Zugang haben. Eine Schwierigkeit, die mit dieser Voraussetzung verbunden ist, besteht darin, dass der Philosoph nicht deutlich genug darauf hingewiesen hat, wie diese Gegenstände verstanden werden sollen. Aufgrund einiger weiteren Thesen des *Tractatus* und Notizen aus den *Tagebüchern*, kann man jedoch voraussetzen, dass die transzendentale Deutung der einfachen Gegenstände, als den Kantschen Kategorien analoger Objekte überzeugend genug ist. Eine Entwicklung und Begründung der Frage nach dem Status der einfachen Gegenstände wurde den Rahmen dieses Beitrags überschreiten.

Literatur

AMMERELLER, Erich 2001 „Die Abbildende Beziehung", [in:] *Tractatus logico-philosophicus. Klassiker Auslegen*, hrsg. von Vossenkuhl W., Berlin, s. 111-139

ISHIGURO, Hide 1989 „Die Beziehung zwischen Welt und Sprache: Bemerkungen im Ausgang von Wittgensteins Tractatus", [in:] *Grazer Philosophische Studien* 33/34, s. 49-66

KANT, Immanuel 1990 *Kritik der reinen Vernunft*, Frankfurt am Main

WITTGENSTEIN, Ludwig 1984 Tractatus logico-philosophicus. Werksausgabe Band 1, Frankfurt am Main

Rule-Following and the Irreducibility of Intentional States

Antti Heikinheimo, Jyväskylä, Finland

1. Reduction through Functional Definition

It is not always clear what exactly is meant when it is said that something mental is reducible to something physical. Thus, when debating about reductionism, it is important to keep in mind just which kind of reduction one is talking about. One clearly defined and plausible notion of reduction comes from Jaegwon Kim. Reducibility is often taken to be a relation between two "levels", such as the mental and the physical level. Kim argues, plausibly in my opinion, that so called bridge-laws that connect the two levels with empirical regularities, do not amount to reduction (Kim 2005, 103-5). This is because both the higher- and the lower-level phenomena need to be mentioned in a statement of a regular connection between phenomena at two different levels, whereas reduction requires an account of the higher-level phenomenon solely in terms of the lower level. I take this much to be common ground between most reductionists and non-reductionists – that it is not enough for the reductionist to establish empirical connections between the mental and the physical. He/She needs something stronger. In Kim's view this stronger requirement is:

> Conceptual connections, e.g., definitions, providing conceptual/semantic relations between the phenomena at the two levels. (Kim 2005, 108)

These conceptual connections serve as the first step of a reductive explanation, in terms of the "base" level, of the phenomenon to be reduced. The reductive explanation consists of three steps:

> Step 1 (functionalization of the target property) Property M to be reduced is given a *functional definition* of the following form: Having M $=_{def.}$ having some property or other P (in the reduction base domain) such that P performs causal task C.
>
> Step 2 (Identification of the realizers of M) Find the properties (or mechanisms) in the reduction base that perform the causal task C.
>
> Step 3 (Developing an explanatory theory) Construct a theory that explains how the realizers of M perform task C. (Kim 2005, 101-2)

On this model, then, the reduction of a higher-level property, such as being a gene, consists of (1) a functional definition, such as "being a gene = $_{def.}$ being a mechanism that encodes and transmits genetic information"; (2) finding the realizers for the causal-functional role – in this case, DNA molecules; and (3) a theory – in our case molecular biology – that explains how the realizers – the DNA molecules – fulfil this role (Kim 2005, 101). In the mind-body case, the higher-level properties in question are such as "being in a mental state S".

Although Kim's notion of reduction through functional definition is not, by any means, the only intelligible concept of reduction, I will make it the target of my following discussion on reductionism. In the end of this paper I will include a very brief comment on theory reduction and reduction through mind-body identity. There are a few things to notice about this reduction schema. First, the functional definition should, of course, be adequate to the established meaning of the higher-level concept. It is sometimes said that, because of some indefiniteness of everyday-language concepts, they can not, strictly speaking, be defined. Since this is obviously not the real issue between reductionists and non-reductionists, 'definition' here should be understood in a relaxed sense, meaning something like "rough characterization". Second, it is the attainability of the functional definition in step 1 that is essential to the philosophical issue of reductionism vs. non-reductionism. If step 1 can be completed, i.e. adequate definitions of the higher-level properties can be given through causal roles, but the reduction nevertheless fails in steps 2 and 3, the resulting position will not be non-reductionism (at least not in the usual sense of that word), but eliminativism (if there are no realizers for the roles specified)[1]. Third, the philosophical debate over reductionism (or at least the one I have in mind) concerns the *in principle* or *theoretical* attainability of the functional definitions, not their attainability in practice.

We are now in a position to see what would constitute a conclusive argument for either side in the reductionism debate. The mind-body reductionist needs to show that

> MBR[2] It is in principle possible to define mental properties, adequately to the established meaning of the concepts in question, with recourse to causal-functional roles, not using mental property concepts in the *definiens*.

The non-reductionist, respectively, needs to show that MBR is not true, i.e. that it is not possible, even in principle, to give such definitions.

According to Kim, functional definitions are not attainable for concepts of phenomenal properties, but are attainable for concepts of intentional/cognitive properties, such as believing that p or desiring that q (Kim 2005). I will argue that functional definitions are not attainable in the case of intentional properties either, that is, that MBR does not hold for intentional properties.

2. The Normativity Argument

My argument is based on the discussion on rule-following in Saul Kripke's *Wittgenstein on Rules and Private Language* (Kripke 1982). Kripke's question was, approximately, "what makes it the case that, in saying 'plus' and using the + symbol, I mean addition and not some other function?" His answer was, roughly, that there is nothing, no fact, short of the whole practices of attributing meanings and doing addition in the community of language-users that makes the difference between my meaning the one thing or the other. Kripke specially considers one sort of facts that might be thought to make the difference. Namely, facts about my dispositions to use the word 'plus' and the + symbol. Now these dispositions are exactly the kind of causal-functional roles that appear in Kim-style reductive explanations. Furthermore, functionally defining

[1] That is, if we have conclusive grounds for claiming that there are no realizers for the causal roles. If we have just not yet managed to find the right realizers, then, of course, we do not have to give in to eliminativism.

[2] For mind-body reductionism.

intentional states requires functionally defining meaning something instead of something else. For surely we need to be able to differentiate the contents of intentional states in order to differentiate the states themselves. And if a definition does not enable us to tell the difference between, say, believing that there is a cow in front of me and believing that there is a horse in front of me, then it is clearly not adequate to the meaning of the concept of belief. Those who think that mental content does not depend on public language might object that considerations of word meaning do not apply to intentional states. I believe that mental content does depend on public language. But even if it does not, in order to have reductive explanation, we need to be able to publicly refer to specific mental contents. So the distinction between different mental contents needs to be done in public language. Thus similar considerations apply. So let us take a look at Kripke's argument against dispositional analyses of meaning.

Kripke's main argument against dispositionalism is the normativity argument, which I will now lay out. In order to make it the case that I mean anything by a word, the meaning-determining fact needs to make the difference between right and wrong uses of the word. It needs to justify my using the word the way I use it (if I actually am using it correctly). But dispositions can not do this. If what I mean by a word was determined by the way I am disposed to use it, then whatever I say would be correct (Kripke 1982, 24). I could not mistake a cow for a horse, for if I called a cow 'horse', then that particular cow would, *for that very reason,* be included among the things I mean by 'horse'. So there would be no distinction between using a word correctly, in accordance with its meaning, and using it incorrectly. From this it follows that there would be no such thing as meaning anything by a word.

There are, of course, other candidate solutions for the rule-following problem, besides the Kripkean community view and the simple dispositional view. The most promising such solutions will not, however, help the case of reductionism, since they do not offer causal-functional analyses of meaning. I have in mind here primarily the accounts of Crispin Wright and Philip Pettit, which are, in essence, versions of the community view (see Kusch 2006, ch. 7). The reductionist needs a solution close enough to the simple dispositional view to yield functional definitions.

The lesson to be learned from the normativity argument is this: Meaning is normative. In order for a word to mean something, there must be correct and incorrect ways to use the word. Any functional definition of meaning must maintain this distinction between correctness and incorrectness. Similarly, any functional definition of intentional states must maintain the distinction between fit and misfit with actual states of affairs (in case of belief this amounts to the distinction between true and false beliefs, in case of desires, satisfied and not satisfied desires, and so on). Next I will take a brief look at some causal-functional analyses of intentional states, and how the normativity argument shows them to be defective.

3. Functional Analyses of Intentional States

The first functional analysis I will consider is W.V.O. Quine's behavioural semantics (Quine 1960). Quine, of course, intended his analysis to be an analysis of the meaning of sentences, for he did not believe in intentional states (see Quine 1960, 221). It is, however, quite straightforward to extend the behavioural account also to mental content. Quine's basic idea was that the (stimulus) meaning of a sentence is the set of stimuli, presented with which a language user would, if queried, affirm the sentence in question (Quine 1960, 32). So it is natural to say that the same set of stimuli constitutes the content of a belief of the language user. In other words, that he/she believes the sentence to be true. Functional definitions of other intentional states along these lines may be more complicated, but it does not matter to my argument. If the behavioural account fails in the case of belief, which is the simplest case, then there is not much hope for it in other cases either. Now it is easily seen that the normativity argument refutes the behavioural account. For the behavioural account is really nothing more than the simple dispositional account already discussed. If whatever stimulus that prompts me to affirm a sentence is counted as partly determining the meaning of the sentence, then it is not possible for me to make a mistake by affirming the sentence. So in the case of belief, all my beliefs will be true, for their contents are determined by whatever the facts happen to be when I express the beliefs. Quine, of course, tried to make room for mistakes, but even he had to acknowledge that from the behavioural account follow all kinds of indeterminacy in meaning, so that it would often have to be more or less arbitrarily decided whether someone is mistaken or uses a word in an unusual way.

Another possible source for functional definitions is a sentences-in-the-head view. According to such a view, intentional states are brain states that somehow resemble public language sentences. The most important example of such a view is Jerry Fodor's language of thought -hypothesis (Fodor 1976). There are at least two possible ways to conceive of sentences in the head. They could have content in virtue of their non-causal properties, such as some kind of isomorphism with public language sentences. Or they could have content in virtue of their role in controlling behaviour. If content of brain states is due to non-causal properties, this will not help the reductionist, for the reductionist needs causal-functional definitions. If, on the other hand, content is due to causal role in controlling behaviour, the reductionist still faces the problem of defining intentional states in terms of behaviour. And as we just saw, because of the normativity condition, that problem seems hard to solve. So it seems that sentences in the head will not be of much help to the reductionist. This, of course, is not a problem for Fodor, since he is not a reductionist.

Still another reductionist theory of mental content is teleosemantics, which purports to account for content in terms of evolutionary selection history (see e.g. Millikan 1984). But teleosemantics is a historical, not a causal-functional theory. This means that, in the teleosemantic view, content does not supervene on the totality of causally relevant facts about the present (see Dretske 2006, 75). And this rules out the possibility of causal-functional definitions of intentional states. So teleosemantics is not an option for a Kim-style reductionist. Accordingly, teleosemantics does not aim at reduction through functional definition, but reduction through identity.

4. Conclusion

I hope my discussion this far to have shown that there are some *a priori,* philosophical grounds to doubt the possibility of mind-body reduction through functional definition. I believe, though limitations of space prevent me from elaborating the point, that similar considerations apply against theory reduction – the view that a correct theory of the mental could in principle be derived from an all-encompassing theory of the physical – since I see no other route to theory reduction besides functional definitions of the higher-level properties. Still it might be thought that the sentences-in-the-head view, as well as teleosemantics, might facilitate reduction through mind-body identity. But I think there are difficulties for this project, too. Reduction through identity is supposed to be based on an empirical discovery to the effect that some higher-level phenomenon is in fact identical with some lower-level phenomenon, as in the case of water = H_2O. But the water = H_2O identity rests precisely on the fact that the characteristics of water can be explained in terms of water being H_2O. And the normativity argument shows that similar explanation of the characteristics of intentional states in terms of brain states is not to be expected. The purpose of these remarks on theory reduction and reduction through identity has been merely to hint at the direction where I think the problems are, and they are not intended to be at all conclusive.

Literature

Dretske, Fred 2006 “Representation, Teleosemantics, and the Problem of Self-Knowledge” in: Graham MacDonald and David Papineau (eds.), *Teleosemantics,* Oxford: Clarendon Press, 69-84.

Fodor, Jerry 1976 *Language of Thought,* Hassocks: Harvester Press.

Kim, Jaegwon 2005 *Physicalism, or Something near Enough,* Princeton: Princeton University Press.

Kripke, Saul A. 1982 *Wittgenstein on Rules and Private Language,* Cambridge: Harvard University Press.

Kusch, Martin 2006 A Sceptical Guide to Meaning and Rules, Chesham: Acumen.

Millikan, Ruth Garrett 1984 *Language, Thought, and Other Biological Categories,* Cambridge: M.I.T. Press

Quine, Willard Van Orman 1960 *Word and Object,* Cambridge: Technology Press of the M.I.T.

Relating Theories. Models and Structural Properties in Intertheoretic Reduction

Rafaela Hillerbrand, Oxford, England, UK

1. Introduction

The Russian doll model of scientific progress is very appealing: When a new and more profound theory is able to reproduce and refine the results of one or several well-established theories or even exceeds the scope of the old theories, this is seen as a clear instance of scientific progress. The older theories t_i, $i=1,2, \dots$, nest in the new and more profound theory *T* just like a Russian doll nests inside the next bigger one; the old theory or theories are said to be *reduced* to *T*. For simplicity, this paper considers the case $n=1$; $t_i=t$. Not only within the philosophical literature, but also among many scientists and non-scientists alike such a reduction from *t* to *T* is perceived as a central part of progress in science.

In many instances, the reducing theory *T* is a more fine-grained, `microscopic' description of the system under consideration: For instance, in the wake of Lucas critique (Lucas 1976), microeconomics aims at founding large parts of macroeconomics; molecular biology strives to explain classical genetics; … While the `microscopic' theories are seen as fundamental, the coarse-grained ones – macroeconomics just as classical genetics – are often disdainfully referred to as `mere phenomenological'. The alleged success of fine-grained theories in reduction explains at least partly the great hopes and fears associated with advances on micro-sciences like molecular biology or nanoscience (cp. Schmidt 2004).

However, reducing one theory to another is not a piece of cake and closer inspection reveals a plethora of unsettled questions. Likewise, all examples mentioned above have been subjected to heavy doubts as to whether they indeed fulfill the criteria of reduction. These criteria are commonly equated with the ones given by E. Nagel (1974). I follow this notion and identify reduction roughly with Nagelian reduction.

Despite various criticisms, the paradigm of successful reduction of an alleged phenomenological to a microscopic theory remains the merging of thermodynamics in statistical mechanics. My arguments will be developed along these two theories. By choosing a highly mathematized science like physics, I hope to provide arguments that can be carried over to other, less formal sciences in a straightforward way. In particular, I want to point to two omissions of the classical account on intertheoretic reduction: Firstly, it is often not theories that are reduced; rather, *models* deriving from adequate theories are related in a way that may be called `reductionist'. Secondly, the common view on reduction focuses on different descriptive entities appearing in the mathematical formulation of the theories *t* and *T*. These entities – the theories' furniture of the world – are correlated via so-called correspondence (or bridge) principles. The structural properties of the theories are commonly overlooked, whereas I will contend that a successful reduction must at least correlate some of the *structural properties* of the theories *t* and *T*.

2. A Tale of Two Models: Models as Mediators in Intertheoretic Reduction

The core idea of Nagel-type reductions is that some theory *T* reduces another *t* only if the laws of *t* are (logically) derivable from those of *T*. In the case of thermodynamics and statistical mechanics just like in many other instances of intertheoretic reduction, the descriptive vocabulary of *T* and *t* differ. Terms like entropy or temperature, for example, are defined in very dissimilar ways in both theories – one speaks of heterogeneous reduction.

For heterogeneous reductions, the requirement of connectability involves the provision of correspondence (or bridge) rules connecting the vocabulary of *T* to the one of *t*. Within the philosophy of physics, the debate on nature and status of the bridge rules results in a heated debate on what it actually means to reduce thermodynamics to statistical mechanics. The original approach of Nagel and others has been dismissed as too simplistic and Nagel's requirements for a successful reduction turned out too stringent a criterion. The only aspiration we can reasonably hope for is that statistical mechanics gives us an approximation of the laws of thermodynamics (e.g. Callender 2001, Frigg 2008, cp. Schaffner 1976): *T* does not actually reduce *t*, but reduces a modified version *t'*. For instance, from a statistical theory no strict universal laws given by thermodynamics can be deduced. Consider a system characterized by intensive state variables. Statistical physics tells us that the corresponding extensive variables can be only specified as mean values. No matter how sharp this mean value is for a macroscopic system, in the statistical approach the extensive variable never becomes a state variable as this is a non-stochastic variable.

How exactly the approximated theory *t'* that connects to *T* via correspondence laws actually relates to the original theory *t*, raises serious questions. In this paper, I want to contend that it is not *t* that is reduced to *T*: not theories reduce or become reduced – rather a concrete *model* of *T* can be related to a *model* of *t* in such a way that the connection between these models qualifies as a reduction. Only for concrete models does the notion of reduction make sense.

Take as an example the bridging of the concepts of temperature in statistical mechanics and in thermodynamics. To determine the correspondence principles, concrete models of the considered physical system are set up – a model deriving from statistical mechanics, another from thermodynamics. Let us begin with the former and focus on an ideal gas. The model considers gas particles confined to a container. This allows deriving an explicit formula for the pressure of the gas via the force the particles exert on the idealized and rectangular walls of the container. By averaging, we obtain a formula relating the (microscopic) pressure of the gas to the volume of the container and the mean kinetic energy of the gas particles.

Conversely from thermodynamics, deriving a concrete model that allows to specify the temperature of a concrete system amounts, amongst other things, to choosing a certain temperature scale. Then the formal analogy between the thermodynamic ideal gas law – a combination of Boyle's (Mariotte's) law and Charle's (Gay-Lussac's) law – and the statistical law relating pressure, volume, and mean kinetic energy allows correlating thermodynamic with statistical pressure as well as thermodynamic temperature with the mean kinetic energy and the number of degrees of freedom of the individual gas particles. Only by settling for a concrete temperature scale, are we able to identify Rydberg constant and thus Boltzmann constant; only by choosing a concrete realization of statistical mechanics were we able to relate (statistical) pressure to volume and mean kinetic energy. Note that the need to invoke the latter model was also noted by S.W. Yi (2003).

The (not purely) formal analogy between the model equations allows *identification* of the corresponding quantities, yielding the well-known bridge concept that relates thermodynamic temperature to the mean kinetic energy of microscopic particles per degree of freedom. Note that we follow here a distinction introduced by L. Sklar (1993): Bridge rules may merely *correlate* the involved quantities, or they can *identify* a concept in *T* with a corresponding one in *t*. Following Yi (2003), any identification of terms between various theories is to be rejected as a metaphysically heavily loaded concept with many nontrivial and far from obvious assumptions on how theoretical terms can make sense outside the theory they are embedded in. Nonetheless, Sklar is right when he points out that without further specification the term *correlation* is so vague that it begs the question as to how the terms are actually related. The preceding analysis showed a way out of this dilemma: It is not terms within theories that are mapped in one way or another, but in the narrow setting of concrete models, various descriptive terms can be *identified* on (not merely) formal analogies without the metaphysical ballast bothering us if the identification were on the level of the involved theories.

3. Reducing structural properties

Even assuming that the commonly suggested correspondence rules successfully reduce models of *t* to models of *T*, this is not yet the end of the story of reduction to be told here. Not only the observational vocabulary stated in theoretical terms like temperature or pressure that take on specific numerical values needs to be correlated; also (parts of) the structural properties of *T* have to be mapped to those of *t*.

Structural properties refer to those properties of theories that do not turn on arbitrary choice of units, like the choice of a certain temperature scale, but concern intrinsic features of a system. Consider the thermodynamic concept of quasistatic changes, meaning that the system goes through a sequence of states that are infinitesimally close to (thermodynamic) equilibrium. It is not straightforward what the equivalent of a quasistatic transformation in statistical mechanics might be (cp. Frigg 2008). However, implicit correspondence rules map this structural property of thermodynamics to that of statistical mechanics. The claim for quasistatic transformation within the thermodynamic framework translates to the requirement that on the microscopic level the relaxation time t_{pa} of the particles is much smaller than the typical time scale t_{av} at which changes occur in the coarse-grained, averaged quantities. Hence the microscopic condition corresponding to the thermodynamic requirement of quasistatic changes is: $t_{pa} >> t_{av}$, implying $t_{pa}/t_{av} \rightarrow 0$.

Following R. Batterman (2002), this limiting procedure is a special kind of explanation common within physics, a so-called asymptotic explanation. As C. Pincock (2007) noted these belong to the broader class of *abstract explanation* appealing primarily to the ``formal relational features of a physical system'' and thus account to what I have referred to as structural properties.

It is indispensable that reduction accounts also for (some of) the abstract explanations of the different theories involved. One obvious objection against this claim contends that the content of a theory is identified with its empirical content, embedded in the observables that take on numerical values. However even then some of the structural properties need to be bridged. Any model or theory makes predictions only within some range of applicability. Beyond theory-external conditions, there are always specifications of the range of applicability internal to the theory or the model under consideration. In specifying this range of applicability, we fall back on the structural properties of the theory. Thermodynamics, for example, makes predictions about the state of a system if the undergone changes are quasistatic. A successful reduction requires that at least those structural properties of the reduced theory required to specify the range of applicability are connected to the reducing theory, and vice versa.

Concluding this section, it is worth noting that there is a genuine difference in how the connection of the descriptive vocabulary and the structural properties of two theories describing the same physical system are treated within the sciences. While for the former explicit correspondence or bridge rules are stated – as for example in relating the microscopic and the thermodynamic temperature discussed above – the relation between the formal relational features of two mathematical descriptions is mostly given implicitly and often remains among the `tacit knowledge' of the scientists, shared by the practice of doing a specific scientific research.

4. Conclusion

This paper argued that even when well established theories exist, the reduction might not be an inter*theoretic* one. Rather a concrete model of a theory *T* is correlated with a model of theory *t* in a way that qualifies as reductionist. Our discussion of the alleged reduction of thermodynamic to statistical mechanics thus explicitly showed how some of Kitcher's classical criticism of Nagelian reduction within biology translate into the more formal sciences. With a view to the debate within philosophy of physics, the central role of models as regards intertheoretic reduction can be taken as a hint to not only ``take thermodynamics less seriously'' (Callender 2001), but also take statistical mechanics as a theory less serious.

Turning to more general debates within philosophy of science, both points made in this paper – the central role of models in the process of reduction and the necessity to also connect (some of the) structural properties of *T* and *t* – reveal a more complex picture of scientific progress as commonly recognized within philosophy of science. Although the raised points do not refute the hopes and fears advanced in the microscopic, reducing theories like molecular biology or nanotechnology, they do raise serious doubts as regards the common view that `microscopic' theories are generally more embracing.

Literature

Batterman, Robert (2002) *The Devil in the Detail*, Oxford: Oxford University Press.

Callender, Craig (2001) ``Taking Thermodynamics too Seriously'', in: *Studies in the History and Philosophy of Modern Physics* 32, 539-553.

Frigg, Roman 2008 ``A Field Guide to Recent Work in the Foundations of Statistical Mechanics'', to appear in: Dean Rickles (ed.), *The Ashgate Companion to Contemporary Philosophy of Physics*, London: Ashgate.

Kitcher, Philip (1984) ``1953 and All That: A Tale of Two Sciences'', in: *Philosophical Review* 93, 335-373.

Lucas, Robert (1976) "Econometric Policy Evaluation: A Critique", in: *Carnegie-Rochester Conference Series on* Public Policy 1: 19-46.

Nagel, Ernest (1974) ``Issues in the Logic of Reductive Explanations'', in: *Teleology Revisited*, New York: Columbia University Press, 95-113.

Pincock, Christopher (2007) ``A Role for Mathematics in the Physical Sciences'', *Noùs* 41:2, 253-275.

Schaffner, Keneth (1976) ``Reduction in Biology: Prospects and Problems'', in: *Proceedings of the Biennial Philosophy of Science Association Meeting 1974*}, 613-632.

Schmidt, Jan C. (2004) ``Unbounded Technologies: Working Through the Technological Reductionism of Nanotechnology'', in: D. Baird, A. Nordmann and J. Schummer (eds.), *Discovering the Nanoscale*, Amsterdam: IOS Press.

Sklar (1993) Physics and Chance: Philosophical Issues in the Foundation of Statistical Mechanics, Cambridge: Cambridge University Press.

Yi, Sang Wook (2003) ``Reduction of Thermodynamics: A Few Problems'', *Philosophy of Science* 70, 1028-1038.

The Constitution of Institutions

Frank Hindriks, Groningen, The Netherlands

Institutions depend on human beings for their existence. They are human constructs that would not be there if it were not for us. The challenge is to unpack this. Are they mere social constructs, or do they have a reality beyond our social categorizations? Institutions involve human beings and their (inter)actions. Can they simply be reduced to these? Or do they have a reality that goes beyond them? I shall suggest that institutions present us with a number of puzzles that justify a serious investigation into these issues.

A US president has the power to veto laws not due to any superior physical or mental abilities, but because he is granted this power by the American people. Apparently, the powers of presidents do not depend on their intrinsic features only. Examples such as this one pose problems for a straightforward reduction of institutions to human beings and their (inter)actions. Taking constitution to be a (non-reductive) relation of unity without identity, I argue that such puzzles dissolve once institutions are taken to be constituted by human beings, their mental states, their interactions, and their surroundings.

1. Against Identity and Mereology

Presumably institutional properties supervene on physical ones. Supervenience, however, is a fairly innocent relation. Both reductive and non-reductive materialists about the mental accept that mental properties in some sense supervene on physical ones. In order to provide an adequate ontology of institutions, then, relations other than supervenience have to be considered. The first one that I consider is identity, which is a relation between entities rather than between properties. I shall argue that institutions are not identical to the entities they consist of or are composed of (where these latter notions are used in a metaphysically innocent sense). For the purposes of this paper I focus on the case of organizations leaving other kinds of institutions for another occasion.

Consider the United Nations (UN). The UN consists of countries, which are its members that are united by the Charter of the UN. Is the UN identical to the set of its members? Presumably not. The UN can enlarge its membership, while sets cannot. A set that has more members than another set is numerically distinct from that other set. The UN remains the same entity when it acquires a new member.

What about organizations that have only one member? Are they identical to their members? A limited liability company (LLC) can consist of only one individual. Even if it does, however, an LLC is not identical to that person. An individual can create and later dissolve an LLC that has only one member, herself. Someone who does so existed before the LLC did, and she outlives it. So the persistence conditions of organizations differ from the entities they are made of. In other words, there is a difference regarding what accounts for their identity over time.[1]

Are organizations mereological sums of their members? At least on some conceptions of them, mereological sums they do not have any (causal) properties their parts do not have (Lewis 1986). However, the Security Council of the UN has the power to adopt resolutions, but none of its members does. The UN can have a code of conduct and ensure compliance to it without any of its members doing so (recall that its members are countries). Similar claims hold for other organizations. An LLC can sue another company without any of its members doing so. And a choir can sing a cantata without any of its members doing so. The upshot is that organizations can have causal properties none of their members have.

One can, of course, conceive of mereological relations in a more substantial way. The part-whole relation might be an ontological relation between parts and wholes each of which exist. Suppose it is also granted that wholes can have causal properties none of its parts have in isolation. Even then it would be inappropriate to conceive of the relation between organizations and their members as a mereological relation. This is because the part-whole relation is transitive while the membership relation as it applies to organizations is not. Countries consist of people, the UN consists of countries, but the UN does not consist of (exactly those) people: as a Dutch citizen, I am a member of the Netherlands; I am not, however, a member of the UN.

To sum up, the relation between organizations and (collections of) human beings is not identity, nor is it a mereological relation. They differ in persistence conditions and causal properties. Furthermore, the relation between them is non-transitive. In the next section, I shall argue that, in order to accommodate these features, the relation between organizations and (the collections of) their members should be conceived of as constitution.

2. Constitution

Constitution is a relation of unity without identity. It obtains, for instance, between a statue and the lump of clay of which it is made. These are united in that they consist of the same material. They are not identical to one another: the lump of clay can even when the statue does not; the statue can survive gradual replacement of the clay of which it consists resulting in a situation in which the statue still exists even though it contains none of the material of which it consisted originally. These two features can be captured in terms of a condition of material coincidence (1) and a modal condition (2), a condition that captures possibilities such as the non-existence of a statue in the presence of a lump of clay.

In order to account for the fact that a particular lump of clay does constitute a statue further conditions have to be added. A notion of favourable conditions can serve a useful purpose here. Statues owe their status of art object to their surroundings. At a general level, they bear some relation to an art-world (Baker 1997). They might, for instance, have been commissioned as art objects. This is one of the conditions that are favourable for an object to constitute a statue. Such conditions explain why particular objects are statues. They account for this in that they can

[1] Cf. Ruben (1985) and Uzquiano (2004).

be invoked in response to the question: why does this lump constitute a statue? This can be done in virtue of the fact that, necessarily, if a lump of clay is in statue-favourable conditions it is a statue. These conditions are such that they confer the status of statue on lumps of clay. Thus, favourable conditions play an explanatory role in relation to the instantiation of the constitution relation.

One condition that has to be added, then, is another modal condition – one that states that in the relevant favourable conditions a constituter necessarily constitutes the constituted object (3). In order for it to do any work, this has to be combined with the condition that the relevant conditions actually obtain (4). (Whether or not this condition is satisfied is a contingent matter. In spite of the second modal condition, then, constitution as such is a contingent relation.)

The constitution relation is usually taken to be irreflexive, asymmetric, and transitive. The first modal condition accommodates the irreflexivity of the constitution relation. This 'possibility condition' amounts to the claim that, if conditions were not favourable the constituter would not constitute the constituted object (the lump of clay can exist without there being a statue). No object can have such a relation to itself. Asymmetry can be captured by adding 'an impossibility condition', a condition concerning the impossibility of the existence of the constituted object without a constituter (5): a (clay) statue cannot exist without a lump of clay existing at the same time. Note, however, that statues can also be made of other material than clay including marble. Such multiple realizability can be accommodated by specifying the property (properties) that is (are) characteristic of the constituter in a sufficiently general way. This could be a characterization in terms of a disjunction, or in terms of properties that can be satisfied by several kinds of objects.

Before commenting on transitivity, let me present the account of constitution implicit in the preceding discussion:[2]

> *a* constitutes *b* at *t* if and only if *a* is *F* and *b* is *G* and (1) – (5) hold:
> 1. *a* and *b* coincide materially at *t*.
> 2. It is possible for *a* to exist in the absence of an *x* that is *G* and materially coincides with *a*.
> 3. Necessarily, if an *x* that is *F* is in *G*-favourable circumstances, there is a *y* that is *G* that coincides materially with *x*.
> 4. *a* is in *G*-favourable circumstances at *t*.
> 5. It is impossible for *b* to exist in the absence of an *x* that is *F* and materially coincides with *b*.

Conditions 1 and 5 account for the unifying character of the constitution relation. Condition 2 reveals that the relation is distinct from identity. Finally, conditions 3 and 4 explain why the one object constitutes the other.

How does this account apply to organizations? The first two conditions imply that organizations coincide materially with their members and that it is possible for these individual members to exist without an organization of type *G* existing that coincides materially with them.[3] What about the other conditions? How should we conceive, for instance, of favourable conditions of organizations? What does it take, for instance, for one or more persons to form a limited liability company (LLC)? The answer to this question can be found in the Revised Uniform Limited Liability Company Act 2006.

The central conditions are the formulation of an operating agreement that regulates the relations between the members (which they are 'deemed to assent to' as soon as the company exists), and the drafting of a certificate of state (which needs to be submitted to the Secretary of State). Together these constitute what I call 'the statute' of a particular LLC. The act in which all this is specified in very precise terms provides the means for a non-circular specification of the relevant favourable conditions (to which I henceforth refer as LLC-favourable conditions). It is not possible for an LLC to exist without these conditions being satisfied for a particular (collection of) individual(s).

The satisfaction of favourable conditions accounts for the persistence conditions of constituted objects, and for their causal properties. A collection of individuals can exist prior to the existence of the LLC they end up constituting because at the time they did not yet constitute it they were not in LLC-favourable conditions: they had not yet formulated an operating agreement, or they had not yet submitted a certificate of state to the Secretary of State. Furthermore, it is only because of the LLC-favourable conditions that their liability is limited. This has real consequences in any lawsuits in which they might be held accountable. The upshot is that the LLC-favourable conditions account for the differences between an LLC and its members.

Earlier it was noted that the constitution relation is usually taken to be transitive. In section 1, I dismissed the part-whole relation as inadequate for capturing the membership relation because of its transitivity. This seems to make it problematic for me to invoke another relation that is transitive in order to characterize the relation between organizations and their members. This appearance is deceiving.

It is the aggregate of members that constitutes the UN, not any of the members themselves, at least not directly. As a consequence, it is somewhat misleading to say that the relation between organizations and its members is one of constitution. Instead, the relation between an organization and the aggregate of its members is one of constitution. This in turn implies that the membership relation should not be cashed out in terms of constitution, at least not directly.

Consider, for purposes of comparison, a case in which an organization does consist of single-member organizations. Suppose a chess player has to incorporate him or herself in order to participate in a chess tournament, which only admits single-member foundations. The chess tournament is organized by a society created for this very purpose that only has foundations as its members. In this case, the people who constitute the foundations that

[2] This account owes a lot to Baker (1997, 2007). Let me comment on some of the differences. First, I do not require *F* and *G* to be what Baker calls 'primary kind' properties, which are properties that the relevant objects have essentially. Second, I include the impossibility condition in order to account for the asymmetry of the constitution relation. Baker (2007, 163-65) believes the necessity condition ensures asymmetry. The idea is that there simply are no favourable circumstances that account for constitution as a top-down relation. Rather than appealing to a (supposed) metaphysical fact that is external to the account, I build asymmetry explicitly into the definition of constitution. Third, Baker has expanded on the coincidence condition so as to rule out that a constituter might constitute two objects of the same kind. I do not include such a condition, because one and the same collection of individuals can constitute two different organizations. See note 3 for another difference.

[3] In fact, I believe that the condition of material coincidence is problematic for organizations and their members. In my 2008 I argue it should be replaced by an enactment condition.

constitute the chess society are members both of those foundations (each of his or her own, that is) and of the society.

This discussion reveals that it can be confusing to say that the constitution relation is transitive. Even in case of the chess society, individual human beings constitute the foundations, but the chess society is constituted by an aggregate of foundations. So it is not the case that one thing constitutes another one that in turn constitutes some further object. To be sure, there are such cases. Perhaps a human body can constitute a person, which in turn can constitute a limited liability company. However, in many, if not most, cases of constitution, the constituter is an aggregate rather than a constituted object.

Literature

Baker, L.R. 1997 'Why Constitution is Not Identity', *Journal of Philosophy* 94, 599-621.

Baker, L.R. 2007 *The Metaphysics of Everyday Life*. Cambridge, Cambridge University Press.

Hindriks, F. 2008, forthcoming 'The Status Account of Corporate Agents', in: K. Schulte-Ostermann, N. Psarros, and B. Schmid (eds.), *Concepts of Sharedness – New Essays on Collective Intentionality*, Frankfurt: Ontos Verlag.

Lewis, D. 1986 *On the Plurality of Worlds*. Oxford, Basil Blackwell.

Ruben, D.H. 1985 *The Metaphysics of the Social World*, London: Routledge and Kegan Paul.

Uzquiano, G. 2004 'The Supreme Court and the Supreme Court Justices: A Metaphysical Puzzle', *Nous* 38, 135-53.

Do Brains Think?

Christopher Humphries, London, England, UK

1 Introduction

The motivating idea of B&H's 2003 *Philosophical Foundations of Neuroscience* ('PFN') is that a clear view of the relationship between neuroscience and human psychology is not possible without a correct analysis of the psychological concepts and categories involved in the descriptive understanding of mental life. The authors find that these concepts are often misconstrued or misapplied by neuroscientists and their philosophical allies. Defective understanding and misguided questions may, at worst, render research futile by misdirection of experimentation and misunderstanding of its results. It is the authors' constructive intention that their conceptual analysis should 'assist neuroscientists in their reflections antecedent to the design of experiments.'

A leitmotif of PFN is the identification of a persistent mistake of construing the brain, or components of the brain, as subject or locus of mental predicates. For B&H, the ascriptions properly belong to the person or animal. The mistake institutes a sort of Cartesian revanchism, with the old error of ascribing psychological attributes to a mental substance replaced, in the new materialist version, with the error of ascribing them to a physical substance. Brain/body dualism is incoherent, like talk of the East Pole. Thus (PFN: 71): 'Only of a human being and what resembles (behaves like) a living human being can one say: it has sensations; it sees, is blind, is deaf; is conscious or unconscious.' (Wittgenstein 2001: I §281); and 'Perhaps indeed it would be better not to say that the soul pities or learns or thinks but that the man does *in virtue of the soul.'* (Aristotle 1986: 408b).

The neuroscientist's reply might be that talk of brains and their neural circuits seeing shells flying and deciding to take cover is an innocent *façon de parler*; a harmless and amusing shorthand that leads to no practical error. Its value is metaphorical: for example, when describing neural mechanisms, it can harness the insights that have accrued to neuroscience from the field of information technology. For B&H, this last is further confusion: brains are not computers, and computers do not enact rule-governed manipulation of symbols. Computers are artefacts that 'produce results that will coincide with rule-governed, *correct* manipulation of symbols.' (Bennett and Hacker 2007: 151). The projection of the designer's perspective into the operation of the computer is a version of the very error of thought currently in view.

B&H assert a sharp line between investigation of the logical relations between concepts – the philosopher's trade, having to do with the distinction of sense and nonsense – and the scientist's investigations, which have to do with empirical truth and falsehood. But the orthogonality of truth and sense is assailable: e.g. are not answers to conceptual questions true or false? (Dennett 2007: 79-82) Again, B&H's claim that conceptual truths delineate the logical space in which the facts are located, and are prior to them, (129) could be met by the simple objection that the concept of colour is not prior to colour facts (cf. PFN: 129-130). At the opposite pole to B&H is the Quinean view. Abandonment of the 'two dogmas of empiricism' results in a 'blurring of the supposed distinction between speculative metaphysics and natural science.' Thus it is nonsense 'to speak of a linguistic component and a factual component in the truth of any individual statement.' Conceptual scheme and the deliverances of sense interpenetrate within our 'total science'. (Quine 1961)

2. An Inner Process

According to B&H, it only makes sense to ascribe mental predicates to what is or resembles a living human being. Following Wittgenstein, behavior is taken to provide logical criteria for the application of mental concepts. Only the person (the rational, responsible being), and not the brain, satisfies these criteria (PFN: 83). Searle takes this Wittgensteinian move to be at the heart of the argument that leads to the impossibility, for B&H, of consciousness, qualia, feelings etc. existing in and being predicable of brains (Searle 2007: 102). Further, Searle takes B&H to identify pain (let's say) with the criterial basis for pain, i.e. its external manifestation. Then, because the pain is seen to be identified with its criterial manifestation, Searle takes B&H to think that it cannot be the subject of neurological investigation. On this understanding, the PFN programme amounts to criterial behaviourism:

> Just as the old-time behaviourists confused the behavioral evidence for mental states with the existence of the mental states themselves, so the Wittgensteinians make a more subtle, but still fundamentally similar, mistake when they confuse the criterial basis for the application of the mental concepts with the mental states themselves. That is, they confuse the behavioral criteria for the *ascription* of psychological predicates with the *facts ascribed by these* psychological predicates, and that is a very deep mistake (103)… The fallacy, in short, is one of confusing the rules for using the words with the ontology (104)…. I think that once this basic fallacy is removed, then the central argument of the book collapses. (105)

I don't think this charge sheet will hold up in court. In the first place, B&H nowhere explicitly make the equation between behavior and the subject ontology of mental predicates. The former is criterial for the latter, not identical with it. Pain behavior is a manifestation of pain, and a criterion of it, but is not the pain itself. Moreover, the charge of behaviourism is refuted if the behavioral criterion is 'defeasible', i.e. only partly constitutive of its object. Thus, if I'm reciting the alphabet 'in my head', there *is* no behavior. B&H display the defeasibility of behavior when they say: 'an animal may be in pain and not show it or exhibit pain behavior without being in pain. (We are no behaviourists.)' (Bennett and Hacker 2007: Note 18 p211).[1] Secondly,

[1] Wittgenstein takes behavior as criterial for the mental, but not to be equated with it ontologically or causally: the relation is logical and normative. Thus behavior, expressed by the body, is the window of the soul. (Wittgenstein 2001: II §178) Only to a being that has capacities can mental concepts be ascribed. But a being that has capacities can exercise them or not: the matter is not causally determined. Behaviourism is therefore no apt theory of such a being. See Glock: 55-58 and Hacker 1990: 224-254. Thus Wittgenstein is not a metaphysical behaviourist. Logical behaviourism (asserting semantic equiva-

B&H do not deny that it is possible to mount neurological investigations of pain etc. ‘Research on the neurobiology of vision is research into the neural structures that are causally necessary for an animal to be able to see and into the specific processes involved in its seeing.’ (Bennett and Hacker 2007: 161).

This last point raises the question of the causal relationship between pain etc. and neurophysiology. Searle acknowledges that B&H look for causally necessary conditions for consciousness, but insists that a causally sufficient account is what is required, and uses this assumption in the construction of his case for B&H’s Wittgensteinian behaviourism. But that requirement seems to beg the question about the dichotomy of mental and neurophysiological predicates, with a tacit assumption that a correct theory of mind must be physically reductive. For reduction requires explanatory connection between explanandum and explanans, together with bridge laws connecting the relevant properties. A causally sufficient explanation of consciousness in terms of physical law would deliver both of these requirements for reduction straight away. Reduction, thus established, would dissolve the possibility of a division of categories between the inner and the outer. So the hidden reductive assumption covertly imports the conclusion into the premises.

Searle’s reply to this might follow his chapter essay ‘Reduction and the Irreducibility of Consciousness’ (Searle 1992: Chapter 5). Consciousness is there described as a ‘causally emergent property of systems’ on a notion of emergence that denies that an emergent has any causal powers that cannot be explained by the causal powers of the physical base. Searle says that this type of emergence usually delivers causal-explanatory reduction, from which ontological reduction follows. However, Searle continues, in the case of consciousness, the ontological reduction does not work, because the subjectivity of experience cannot be explained in third-party causal terms. But this failure of ontological reduction he claims to be unimportant, because it is a trivial consequence of our definitional practices. We cannot, following the usual reductive procedure, redefine consciousness in causal terms which, being causal, discount the appearances that are characteristic of the reduced domain, because in this case the appearances are what are of interest. On this argument, mental predicates related to consciousness are not ontologically reduced, and so the question is not begged.

However, even if this argument is accepted, it still does not go far enough. This is because, although it says why reduction is harmless in case of consciousness, it does not show why reduction would be harmless in case of normativity. B&H, as already noticed in the discussion of information technology, take normativity to be external to the causal realm as exemplified by the computing artefact. So Searle’s critique of B&H’s psychological ontology still begs the question for mental predicates related to normativity.

In ascribing mental predicates to the animal rather than the brain, B&H are proclaiming that the predicates, together with their ontological subjects, belong to a separate and distinct logical category. Searle criticises B&H’s expression ‘mereological fallacy’, pointing out that brains are not proper parts of persons: what B&H are attacking is a would-be Rylean category mistake. Precisely so. Most of the category mistakes on the table in PFN are simple logical mispredications, not requiring a specifically Wittgensteinian unmasking.

For Searle, it is ‘more or less educated scientific common sense’ that conscious states ‘exist in the brain’, being produced causally as ‘higher-level or system features’ (2007: 99). For B&H, neither does a mental predicate attach to a brain as subject or agent, nor is the mental fact referred to located *in* the brain. Thoughts do not occur in the brain, they occur in the study (PFN: 179-180). The claim that to deny the brain-location of thinking is like denying the stomach-location of digesting (Searle 2007: 109) exemplifies the tacit reductivism already noticed.

3. Persons

For B&H, the proper subject of mental predicates is the person, though no extended analysis is offered of what a person is. For that desideratum we may borrow a page from the patriarchs (Strawson 1957). Thus material objects are found to be the basic particulars – identifiable and re-identifiable without reference to other particulars and partly constitutive of the ‘uniquely pervasive and comprehensive’ system of individuation provided by time and space. Persons are a separate and distinct class of particulars, ascribing to themselves actions, intentions, thoughts, feelings, perceptions etc. These are predicated of a single entity, which is grammatically and, by argument, ontologically the same entity as that to which are ascribed the physical characteristics of the person. (‘I am happy; I am thin.’) This entity, the person, is logically prior to the individual consciousness, for if the priority is taken the other way round, no experience can ever be attributed other than to oneself. The ontological priority of the person must be accepted, not to avoid scepticism, but ‘in order to explain the existence of the conceptual scheme in terms of which the sceptical problem is stated.’ (Strawson 1957: 106) (Hacker rejects this ‘dichotomous division of predicates’ as ‘overly Cartesian’ and prefers a more vague definition of the person as a subcategory of the animal, having capacities of reason, will and morality; see Bennett and Hacker 2007: 312-3.)

A person, then, is subject of both physical and mental predicates. The Strawsonian analysis upholds the division of category between physical and mental predicates, while uniting them in the person. B&H’s central point about mispredication is not a uniquely Wittgensteinian insight, but flows from a distinction of categories that is fundamental in the descriptive metaphysics of mind and body. The point does not therefore stand or fall with the various peculiarities of PFN, such as the claim that two people can share the same pain (in the way that two pillar boxes can share the same colour, see PFN: §3.8) and that subjective qualities of consciousness (qualia) do not exist (qualia not being properties of consciousness but of objects: ‘quale’ equivocates between the subjective quality of an experience and the experience itself, see PFN: §10.3).

lence of mental predicates and behavioral dispositions) is a stronger tendency, though less so in Wittgenstein’s later thought.

4. Conclusion

So do brains think or don’t they? B&H think not, and I have argued that their conclusion does not depend on their specifically Wittgensteinian account of contemporary neuroscience. The proposition can be denied, as a category mistake, from an alternative descriptive-metaphysical approach.

I think B&H’s arguments are stronger than Searle’s critique of them. But the Quinean point made above disrupts the neat conceptual taxonomy. The way in which scientific knowledge influences the *a priori* conceptual scheme is a large question, that cannot be analysed here. But this work is needed, because if the conceptual and the empirical are orthogonal in the way that B&H claim, then there is nothing further to be said about the ontology of mind: enquiry is brought to a close by their strictures.

Literature

ARISTOTLE 1986 *De Anima*, tr. Lawson-Tancred, H., London: Penguin Books.

BENNETT, MAXWELL, DENNETT, DANIEL, HACKER, PETER AND SEARLE, JOHN 2007 *Neuroscience and Philosophy: Body, Mind and Language*, New York: Columbia University Press.

BENNETT, M.R. AND HACKER, P.M.S. 2003 *Philosophical Foundations of Neuroscience*, London: Blackwell.

DENNETT, DANIEL 2007 “Philosophy as Naïve Anthropology: Comment on Bennett and Hacker” in: Bennett et al. 2007.

GLOCK, HANS-JOHANN 1996, *A Wittgenstein Dictionary*, London: Blackwell.

HACKER, P.M.S. 2007 Human Nature: The Categorical Perspective, London: Blackwell.

HACKER, P.M.S. 1990 *Wittgenstein: Meaning and Mind*, London: Blackwell.

QUINE, WILLARD VAN ORMAN 1961, “Two dogmas of empiricism”, in *From a Logical Point of View: Nine Logico-Philosophical Essays*, 2nd ed., Cambridge Mass: Harvard University Press.

SEARLE, JOHN 2007 “Putting Consciousness Back in the Brain: Reply to Bennett and Hacker” in: Bennett et al. 2007.

SEARLE, JOHN 1992 *The Rediscovery of the Mind*, Cambridge Mass: The MIT Press.

STRAWSON, PETER 1957 Individuals: An Essay in Descriptive Metaphysics, London: Routledge.

WITTGENSTEIN, LUDWIG 2001 *Philosophical Investigations*, tr. Anscombe, G.E.M., London: Blackwell.

How Metaphors Alter the World-Picture – One Theme in Wittgenstein's *On Certainty*

Joose Järvenkylä, Tampere, Finland

Metaphor is a topic that is not usually connected with Wittgenstein's thought. He used many metaphors throughout his works but he never presented any theory considering them. Of course it is said that his philosophical method does not consist of theorizing at all, and he explicitly said that "[in Philosophy] we may not advance any kind of theory" (PI, § 109). Yet it seems that later Wittgenstein had a positive account of metaphor which is connected to the idea that Wittgenstein used metaphors exactly to avoid theorizing. This view of metaphors is connected to what Wittgenstein writes on the world-picture[1] in *On Certainty*.

In this paper I have tried to reconstruct what might be called Wittgenstein's positive view of metaphor on *On Certainty*. In order to do so I have compared Wittgenstein's remarks on the picture of the world with some elements of Donald Davidson's theory of metaphors. I believe that both philosophers would have agreed that metaphor has only literal meaning. They would have also accepted that the impact metaphor has on recipient does not belong in the analysis of metaphor, But unlike Davidson, Wittgenstein is not interested in analysis. Instead he shows the influence metaphors have by using metaphors himself. At first I will examine Davidson's theory of metaphor and later on I will conjoin it to views Wittgenstein had.

1. Metaphor to Davidson and Wittgenstein

In its most austere sense, "metaphor is a figure of speech, in which a word or phrase that literally denotes one thing is used to denote another, thereby implicitly comparing the two things" (Woltersdorff 1999, 562). From this it follows that metaphor has the propositional form because it states that something is something, but it lacks meaning because it is impossible to apply conceptions of true or false to it. Hereby it is not an ordinary bipolar proposition. It says that world is organized in particular way, but you cannot compare world and it in any eligible way.

It is sometimes said that metaphor has its own peculiar meaning and that only through this meaning an acceptable interpretation can be provided. Probably the most prominent candidate to challenge this view is Donald Davidson, whose polemic claim is that metaphor has only literal meaning. Still metaphors might have an effect on us; they can make us notice some aspects of things we have not seen before. (cf. Davidson 1979, 43) Thus metaphors can affect on our understanding of the world. This is an insight Wittgenstein would have approved.

While saying that metaphor has only literal meaning, Davidson's intention is to show that even if metaphors can make us grasp new insights, there is no such insight connected to the content of metaphor. He singles out the way metaphors are used from what they mean, and claims that only latter is of the interest of philosophy. (See Davidson 1979, 29-30)

Davidson's views are near to Wittgenstein's, but one clear difference remains. For Wittgenstein, asking the meaning of metaphor is not as interesting as how they are actually used. While Davidson seems to take Wittgenstein's famous remark "[For a large class of cases] the meaning of a word is its use in the language" (PI, § 43) as a basis of his theory of meaning, I rather believe that Wittgenstein just wants us to grasp that instead of asking what is the meaning of a word we just have to look how they are actually used in the language. (See PI, § 130) For Wittgenstein investigating a metaphor is precisely throwing light into those language-games where metaphors are used.

For Wittgenstein proposition is a meaningful sentence, something which can be legitimately called to be true or false. Therefore metaphors, for one thing, cannot be called propositions. Similarly there are sentences that function only as a norm of description and whose use is not regulated by other sentences. These sentences are not bipolar so they are not propositions in Wittgensteinian sense. Wittgenstein's famous metaphor states that these sentences are like hinges on which questions we raise and our doubts depend on (OC 341).

Now there is a certain analogy between these hinges and metaphor. They both look like propositions but lack a meaning. Danièle Moyal-Sharrock says, stressing the nonsensicality of hinges, that they are ineffable but have a propositional doppelgänger which can be meaningfully mediated inside language game. (Moyal-Sharrock 2005, 94-97)

I would like to say that there are similarities between hinges and metaphors and sometimes metaphor can work as a hinge or at least as its doppelgänger. The uttered metaphor is like doppelgänger of hinge which has a propositional form but which does not refer on any fact on the world. But to fully understand the content of metaphor it must be interpreted and within this process of interpretation those ineffable hinges taken to be certain may change.

Avrum Stroll has said Wittgenstein's late realization to be that by creating mental pictures metaphorical language can break with the logical model thus opening important new dimensions of communication (Stroll 2004, 23). Metaphors can replace one logical model on another in a sense that the certainties are like axioms which regulate our use of language. By changing model new axioms arises. The metaphor creates mental picture which challenges earlier pictures we lean on and this is just the sense how metaphors can enlarge our understanding of the world.

2. World-picture

World and picture are constant theme in Wittgenstein's philosophy. Early Wittgenstein said that language is the picture of the world and later Wittgenstein rejected this idea as misguiding. In his latest philosophy he realised that *world as a picture* is also illuminative metaphor which can be used to illustrate how insights are mediated in philoso-

[1] Anscombe and Paul translate Wittgenstein's "Weltbild" into "picture of the world", but to avoid confusions with other philosophical usage of the concept I rather use world-picture.

phy. This is exactly why Wittgenstein introduces concept "world-picture".

If I read him correctly, in his book *Moore and Wittgenstein on Certainty* Stroll seems to identify World-picture with community. (Stroll 1994, 170) This is strange and somewhat inaccurate because world-picture itself is a metaphor which cannot be reduced into such concepts. Stroll admits that this is merely a hint of positive account of what world-picture might be. (Ibid.) In his later writings he seems to stay on negative accounts saying: "it is a deep Wittgensteinian point that a philosophical model does not give rise to new facts, but may change one's 'picture' of the world" (Stroll 2004, 20). Wittgenstein would have accepted that world-picture can be altered with metaphors. In this sense metaphor seems to have its original Greek meaning "to carry over" or "to transfer". Metaphor carries us over the limits of our World-picture.

Wittgenstein's first remark of world-picture in *On Certainty* goes as follows:

> Everything that I have seen or heard gives me the conviction that no man has ever been very far from the earth. Nothing in my world-picture speaks in favour of the opposite. (OC 93, Paul and Anscombe translates here *picture of the world*)

From this quote it seems that the world-picture is a context where we decide whether some belief is true or false. Because nothing in my world-picture speaks in favour of the opposite, I believe that no man has ever been very far from the earth. However, world-picture is not any system of beliefs, as Wittgenstein is quick to point out, but rather "the inherited background against which I distinguish between true or false" (OC 94).

It does not follow from any conscious decision-making that we end up supporting some world-picture, as is the case with the scientific picture of the world. Sentences that describe my world-picture are not propositions in which we can say if they are true or false, but they are like a rules of a game which "can be learned purely practically, without learning any *explicit* rules" (OC 95, emphasize mine). In this sense when we choose between two different pictures of the world, we already must have some world-picture to lean on and this is what Wittgenstein means by calling world-picture "the inherited background".

Nevertheless, though world-picture seems to be adherent, it is not totally solid. Wittgenstein compares it to the mythology that "may change back to the state of flux, the river-bed of thoughts may shift" (OC 97). Later he describes how this change can take place:

> It is clear that our empirical propositions do not all have the same status, since one can lay down such a proposition and turn it from an empirical proposition into a norm of description.
> Think of chemical investigations. Lavoisier makes experiments with substances in his laboratory and now he concludes that this and that takes place when there is burning. He does not say that it might happen otherwise another time. He has got hold of a definite world-picture – not of course one that he invented: he learned it as a child. I say world-picture and not hypothesis, because it is the matter-of-course foundation for his research and as such also goes unmentioned. (OC 167)

If it would happen that Lavoisier's experiment does not support his hypothesis, it would also mean that Lavoisier is forced to correct his scientific picture of the world. Yet he has a hold of the definite world-picture, which goes unmentioned and thus is not disposed to such changes. In contrast, world-picture can alter through a change in the status of some propositions. By turning some empirical proposition as a norm of description we are also changing that background against which we distinguish between true or false, and for Wittgenstein this background is the world-picture. Therefore we can identify world-picture also with the hinges we take for certain.

This is where metaphor enters the picture. As Stroll points out, metaphors do not just conjoin two seemingly diverse objects, but they also create a kind of mental image. (Stroll 2004, 20) They trigger our imagination and lure us to imagine in what respect two conjoined objects are similar. The textbook example of a metaphor is "girl is a rose", and if we ask how this can be, the answer might be that she is beautiful or her nature is spiky or she smells good etc. In each of these cases we create a sort of mental image, not actual picture of girl being a rose (what this could mean?), but she being a rose in certain way, in certain context. The metaphor grasps something essential of girl's nature and therefore it also enlarges our way of understanding the world, understanding the girl as a rose.

It seems that it is quite arbitrary when some sentence is used as a metaphor and when it is used as a literal statement. For Wittgenstein it is the matter of context whether the sentence "earth came into being 50 years ago" is or is not a metaphor. Literal interpretation of this sentence would say that it is true or untrue empirical statement, whereas metaphorical interpretation would ask in what sense world can be seen existing for just 50 years. While talking of 50 years old fantasy book it is legitimate to say that the fantasy-world it describes came into being 50 years ago. Also it could be that inside the fantasy universe it has existed only 50 years.

But understanding the meaning of a metaphorical sentence in a context also means sentence's end as a metaphor, because the mental stance towards the world changes so that in the new context metaphor has only literal meaning. So understanding the metaphor presumes that we change context in which we interpret it, and within this context metaphor has eligible meaning. But the change of this context is somewhat peculiar process:

> I can imagine a man who had grown in quite special circumstances and been taught that the earth came into being 50 years ago, and therefore believed this. We might instruct him: the earth has long… etc. – We should be trying to give him our world-picture. This would happen through a kind of *persuasion*. (OC 262)

Wittgenstein uses expression "a kind of" because the persuasion in question is not typical argumentation. If our world-picture alters it must also mean that those hinges we take for certain must change and if our certainties change this cannot be due to rational process because there cannot be any criterion they lean on. But metaphor creates a picture which may replace the earlier picture. The alteration of the world-picture does not interfere with the facts on the world, but nevertheless our mental stance towards it changes. In this sense philosophy can have some positive content.

To summarize, Wittgenstein and Davidson would have agreed that metaphor has only literal meaning, but while Davidson is more interested in the content of metaphor, Wittgenstein is interested in the context. Wittgenstein uses world-picture as a metaphor of the pictorial form of the widest imaginable context. Also metaphors can be seen to create a picture which sometimes replaces our world-picture. Therefore Wittgenstein himself gives us a picture trying to persuade us to realize that philosophy works with pictures and by changing them we also change our understanding of the world.

Literature

Davidson, Donald 1979 "What Metaphors Mean", in: Sheldon Sacks (ed.) *On Metaphor*, Chicago: The University of Chicago Press.

Moyal-Sharrock, Danièle 2005 *Understanding Wittgenstein's On Certainty*, Basingstoke, UK: Palgrave MacMillan.

Stroll, Avrum, 1994 *Moore and Wittgenstein on Certainty*, New York and Oxford: Oxford University Press.

Stroll, Avrum 2004 "Wittgenstein's Foundational Metaphors" in: Moyal-Sharrock, D. (ed.) *The Third Wittgenstein*. Aldershot, UK: Ashgate, 2004, 13-24.

Wittgenstein, Ludwig 2001 *Philosophical Investigations*, translated by G.E.M. Anscombe, 3rd edn Oxford: Blackwell. [PI]

Wittgenstein, Ludwig 1975 *On Certainty*, edited by G.E.M. Anscombe and G.H. von Wright, translated by D. Paul and G.E.M. Anscombe, Oxford: Blackwell. [OC]

Woltersdorff, N. 1999 "Metaphor", in: Robert Audi (ed.) *The Cambridge Dictionary of Philosophy*, 2nd edition, Cambridge: Cambridge University Press, 562.

The Modal Supervenience of the Concept of Time

Kasia M. Jaszczolt, Cambridge, England, UK

Partial arguments in support of the supervenience of the concept of time on the concept of degrees of probability are ample. Moens and Steedman (1988) and Steedman (1997) contend that temporality is supervenient on the concepts of perspective and contingency and that tense and aspect systems are founded on the same conceptual primitives as evidentiality which, by our definition, is a concept overlapping with that of epistemic modality. Slightly more remote from this thesis is that of van Lambalgen and Hamm (2005) who argue that the past, the present and the future are linked by means of the imposition of goals, planning, and causation. Temporality supervenes on what is intended, desired as *present*, as well as on the cause-and-effect relation between events and states that are arranged on the line with relations such as earlier-than, later-than, or overlap. Finally, Nuyts (2006: 19) proposes that modality occupies a higher place than time in the hierarchy of semantic categories, which means that it is of a higher level of abstraction.

It is by no means a new idea that time and modality are interconnected. But it is much less often claimed, and much more controversial, that the concept of time supervenes on the concept modality or that time *is* modality. Nevertheless, we can find plenty of arguments in support of this view if we are prepared to look through different domains, including the behaviour of languages from diversified language families, and collect all extant information. In this paper I assess some arguments and evidence according to which time and modality are related by supervenience relation and end with speculating on the possibility that they are one conceptual category.

Peter Ludlow (1999) argues that the future is predictability or potentiality, 'disposition of the world', and hence is to be regarded as a modal concept. He analyses the future-tense morphemes in Spanish as consisting of an *irrealis* marker *ar* and a 'future' ending. For example, *hablaré*, 'I will speak', is analysed not as *habl* + *aré*, but instead *habl* + *ar* + *é*. Moreover, as he points out, in Italian, to express futurity, one standardly uses a present tense form (e.g. *vado*, 'I go') reserving the future tense form (*andrò*, 'I will go') for situations of lesser probability or uncertainty. Similarly, in English, futurity can be expressed with any of the forms listed as (1)-(4), where the present-tense forms in (1) and (2) express higher certainty (see Jaszczolt 2005, ch. 6).

(1) Peter goes to London tomorrow morning.
(2) Peter is going to London tomorrow morning.
(3) Peter is going to go to London tomorrow morning.
(4) Peter will go to London tomorrow morning.

Such scales pertaining to degrees of speaker's commitment to the proposition and the degrees of certainty with which the speaker issues a judgement testify to a very intimate connection between time and modality. And since these scales are scales of modality, modality is the basis for temporal supervenience in the case of expressions of the future.

In spite of the rather unquestionable unreal character of the future, not all languages express it as equally 'unreal'. As de Haan (2006: 41-42) points out, a Native American language Caddo treats the future as a *realis* category. The future morpheme *-ʔaʔ* is combined with the *realis* prefix *ci-* as in (5):

(5) cííbáw-ʔaʔ
ci yi bahw ʔaʔ
Realis *1Sg* see *Fut*
'I will look at it.'

In a Californian language Central Pomo, on the other hand, the future can be accompanied either by *realis* or by *irrealis*, depending on the speaker's judgement concerning the degree of probability of the described event (see *ibid.*: 42). This freedom of combination with *realis* or *irrealis* constitutes a strong argument in favour of the underlying modal character of the future: states of affairs are described as more, or less, certain. This explanation is further supported by the fact that there are languages in which there is a choice between different future morphemes to express different degrees of certainty (see *ibid.*: 50 for examples). The pairing with the *realis* category in Caddo, on the other hand, is more difficult to explain without a more detailed analysis of the devices available in that language. It may, for example, signal that in different languages there is a different degree of reliance on the epistemology of time. When the degree is high, the internal, psychological time and the *irrealis* prevails; when it is low, the ontology of time and the B series surface out as *realis*. The fact that generally in languages of the world the future pairs with modality (van der Auwera and Plungian 1998) appears to testify to the strong cognitive reasons for the predominance of the internal time.

The past is governed by the same principle of supervenience on modality. Although it is a little more difficult to see because, unlike the uncertain future, the past may seem to consist of what 'actually happened' and is subject to judgements of truth or falsity, the supervenience is there nevertheless. Ludlow (1999: 160) points out that 'in most non-Indo-European languages the so-called past is generally just some form of aspectual marker'. Similarly, in English the past-tense morpheme *-ed* is the leftover from a perfect aspectual marker. Next, past tense is used in counterfactuals to express an alternative present state of the world (or a certain *now* of an alternative possible world) as in (6).

(6) If I *had* more time, I would meet my friends more often.

Ludlow provides pertinent references to the accounts on which the past is taken to mean 'remoteness', 'remoteness from reality', and 'exclusion'. But here is where Ludlow's analysis differs from mine. For Ludlow, states of affairs can be 'remote in time' or 'remote in possibility'. Hence, he speculates that there is 'some deeper third element [that] underlies both tense and counterfactual modality' (p. 161). He proposes evidentiality as this underlying parent category: all past-tense morphology is morphology of evidential markers. This recourse to evidentiality is, however, superfluous when we redefine epistemic modality as inferential evidentiality. Evidence that we have *now* about what happened *in the past* allows us to use indicators of the past tense but by the same token we are detaching ourselves from the *now* in the sense of diminished probability as compared with that of a statement in the present tense. Hence, the situation with the past is analogous to that with the future described above: the truth of *now* is given *in* and *by* the *now*: the truth about the future and about the past is given *in* the *now* and by what we *now* remember about the past, or anticipate about the future (see Dummett 2004). This is how the modal detachment is created and cannot be escaped. De Haan (2006: 51) reports on the reconstruction of the tense-aspect-modality system in Proto-Uto-Aztecan where the *irrealis* morpheme is the same as the past tense morpheme and both are founded on an abstract conceptual feature called *dissociative*: past tense marks a dissociation from *now*, just as *irrealis* marks a dissociation from reality. This construal derives from the Aristotelian view according to which statements can be true or false even though we may not be in a position to know the truth value (cf. Kaufmann *et al.* 2006). Varieties of this view include evaluations of anticipations and memories discussed below or versions of temporal logic where representing the past as modality has also been successfully attempted. Thomason (2002) proposes to view pastness as *historical necessity* founded on the model of forward-branching time: with the passage of time, *historical possibilities* diminish monotonically.

Last but not least, it is necessary to mention languages in which formal indicators of time are optional. In such languages we should investigate not only expressions of time but also the semantic category of temporality which is often realised through pragmatic inference. In Thai, both tense and aspect can be left out of the sentence and the specification of these can be left to the addressee's pragmatic inference. For example, *f3on t1ok*[1] ('rain fall'), can express a wide range of temporal and modal commitments from 'it is raining', through 'it was raining' and 'it will rain', to 'it might rain'. When a modal marker is present, its meaning can also vary and the contextual accommodation normally allows the addressee to recover the speaker's intentions without giving rise to ambiguity. A lexical item *d1ay1II*, with the lexical meaning 'to receive', can perform the function of a modality marker expressing ability. Sentence (7) can be translated as a statement of Gremlin's (the cat's) ability but the temporal location is not specified, as (7a) and (7b) indicate.

(7)	k1r3eml3in	c1ap	ng3u:	d1ay1II
	Gremlin	catch	snake	*d1ay1II*

(7a) Gremlin *was able to catch* a snake (and he caught it).
(7b) Gremlin *can catch* a snake (if he wants to).

(from Srioutai 2006: 109; see also Jaszczolt and Srioutai forthcoming.) Contextual information allows the addressee to opt for (7a) or (7b). In addition, as Srioutai (2006) demonstrates, *d1ay1II* comes with a salient, preferred meaning of past tense. In other words, when context does not suggest otherwise, (7) is taken to mean (7a). Pastness is the default, but cancellable, interpretation. It is not encoded, it is merely recovered as the preferred and more common interpretation. Similarly, a Thai word *c1a*, normally translated as the English *will*, is not necessarily a marker of futurity. Just as the English *will*, *c1a* can assume the meaning of epistemic necessity (as in 8) or the habitual meaning, also called dispositional necessity (as in 9).

(8)	m3ae:r3i:I	kh3ong	c1a	d1u:	'1op1e:r3a:I	y3u:I	t1o'nn3i:II
	Mary	may	*c1a*	see	opera	*Prog*	now

'Mary will be in the opera now.'

(9)	b1a:ngkh3r3angII	m3ae:r3i:I	c1a	p1ay1	d1u:	'1op1e:r3a:I
	Sometimes	Mary	*c1a*	go	see	opera
	n3ay2	ch3udw3o'm				
	in	tracksuit				

'Mary will sometimes go to the opera in her tracksuit.'

(from Srioutai 2006: 125). Unlike the English *will*, *c1a* incorporates readily into the Thai grammatical system and expresses modality with predominant future reference, just as *d1ay1II* expresses modality with predominant past reference. This behaviour of modals, combined with the situation in which the language itself does not have an obligatory marking of tense, provides a strong argument for the supervenience of temporality on modality in the sense of conceptual and semantic inheritance: modal detachment is grammaticalized, and temporal detachment follows as defaults or context-driven non-default interpretations.

The past, present and future, the A-theory terms, are terms pertaining to human experience. While in reality time exists but does not flow, for human agents it is the *now* that has the privileged status; I am experiencing the symptoms of flu *now*, I perceive the clock on my mantelpiece *now*, I hear its ticking *as I am writing these words*. It is the privileged status of the *now* that forces us to conceptualize the *not now* not as experience, but as an anticipation or a memory of an experience. To turn to McTaggart (1908: 127) again:

> 'Why do we believe that events are to be distinguished as past, present, and future? I conceive that the belief arises from distinctions in our own experience.
> At any moment I have certain perceptions, I have also the memory of certain other perceptions, and the anticipation of others again.'

Unless they are illusory, perceptions are real and certain. Memories of perceptions and anticipations of perceptions are removed from this certainty to some degree, just as the past and the future are removed from the very central experience of the *now*. In this paper I considered a selection of arguments in support of treating the semantic category of time as derived from modality. There is only a small step from there to the thesis that internal time itself, i.e. the psychological future, present and past, are modalities. For this step we will have to utilize the premise that semantic categories are a window on conceptual categories – in agreement with the rich tradition in various strands of semantic theory, from broadly defined cognitive (e.g. Jackendoff 2002) to dynamic truth-conditional (Hamm *et al.* 2006). This semantic analysis of temporal expressions, albeit pertinent, is a topic for another occasion (see Jaszczolt forthcoming, ch. 4).

[1] 1,2,3 and I, II stand for tone markers.

Literature

van der Auwera, Johan and Vladimir A. Plungian 1998 “Modality’s semantic map”, *Linguistic Typology* 2, 79-124.

Dummett, Michael 2004 *Truth and the Past*. New York: Columbia University Press.

de Haan, Ferdinand 2006 “Typological approaches to modality”, in: William Frawley (ed.). *The Expression of Modality*. Berlin: Mouton de Gruyter, 27-69.

Hamm, Fritz, Hans Kamp and Michiel van Lambalgen 2006 “There is no opposition between Formal and Cognitive Semantics”, *Theoretical Linguistics* 32, 1-40.

Jackendoff, Ray 2002 Foundations of Language: Brain, Meaning, Grammar,

Evolution. Oxford: Oxford University Press.

Jaszczolt, K. M. 2005. Default Semantics: Foundations of a Compositional Theory

of Acts of Communication. Oxford: Oxford University Press.

Jaszczolt, K. M. forthcoming Representing Time: An Essay on Temporality as

Modality. Oxford: Oxford University Press.

Jaszczolt, K. M. and Jiranthara Srioutai forthcoming “Communicating about the past through modality in English and Thai”, in: Frank Brisard and Tanja Mortelmans (eds). *Cognitive Approaches to Tense, Aspect and Modality*. Amsterdam: J. Benjamins.

Kaufmann, Stefan, Cleo Condoravdi and Valentina Harizanov 2006 “Formal approaches to modality”, in: William Frawley (ed.). *The Expression of Modality*. Berlin: Mouton de Gruyter, 71-106.

van Lambalgen, Michiel and Fritz Hamm 2005 *The Proper Treatment of Events*. Oxford: Blackwell.

Ludlow, Peter 1999 Semantics, Tense, and Time: An Essay in the Metaphysics of

Natural Language. Cambridge, MA: MIT Press.

McTaggart, J. Ellis 1908 “The unreality of time”. *Mind* 17. Reprinted in: J. Ellis

McTaggart 1934 *Philosophical Studies*. London: E. Arnold, 110-31.

Moens, Marc and Mark Steedman 1988. “Temporal ontology and temporal reference”. *Computational Linguistics* 14, 15-28.

Nuyts, Jan 2006 “Modality: Overview and linguistic issues”, in: William Frawley (ed.). *The Expression of Modality*. Berlin: Mouton de Gruyter, 1-26.

Srioutai, Jiranthara 2006 Time Conceptualization in Thai with Special Reference to D1ay1II, Kh3oe:y, K1aml3ang, Y3u:I and C1a. PhD Thesis. University of Cambridge.

Steedman, Mark 1997 “Temporality”, in: Johan van Benthem and Alice ter Meulen (eds). *Handbook of Logic and Language*. Amsterdam: Elsevier Science, 895-938.

Thomason, Richmond H. 2002 “Combinations of tense and modality”, in: Dov Gabbay and Franz Guenthner (eds). *Handbook of Philosophical Logic*. Vol. 7. Dordrecht: Kluwer. 205-34.

The Determination of Form by Syntactic Employment: a Model and a Difficulty

Colin Johnston, London, England, UK

1.

An entity's logical character, for Russell and Wittgenstein, is a matter of the ways in which it may combine with other entities to form atomic facts. Where Russell gives a theory of the logical constitution of atomic facts, however, Wittgenstein asserts that the ways in which entities combine in facts can be known only *a posteriori* through the process of analysis.[1] Russell was thus mistaken in Wittgenstein's eyes in laying out as he did his logical variety of particulars and the various kinds of universal. Pressing the Tractarian position, Ramsey claims that we know "nothing whatever about the forms of atomic propositions". We do not know, for example, "that there are not atomic facts consisting of two terms of the same type" (Ramsey 1990 p29).

I shall suggest that this Tractarian agnosticism is in tension with the Tractarian doctrine that the logico-syntactic use of a sign determines a logical form. Imagine a 'world' in which there are only two forms (that is, logical types) of object and only one mode of combination, a mode in which a single object of each form is combined. The symmetry of this world is such that the two object forms are internally indistinguishable. The internal character of each form is exhausted by its being the form of an object whose only possibility for combination is in a certain mode with an object of the other form, and the internal character of the mode of combination is exhausted by its being a mode of combination of one object of each form. Wittgenstein's agnosticism regarding the forms of reality means that he cannot say in advance that reality does not, like our imagined 'world', include distinct but internally indistinguishable forms. A logico-syntactic use, however, is to determine a logical form by virtue of determining the internal nature of that form only. If reality turns out to include internally indistinguishable forms it follows that the determination as envisaged of logical form by logico-syntactic use will not everywhere be possible.

To bring this concern into focus I want to develop a simple, semi-formal account of syntactic use, of form, and of the place of syntactic use in the determination of form. The account will be appropriately general to accommodate Wittgenstein's ignorance of the nature of the forms of reality. I do not claim that the semi-formal work is at every point implicit in the Tractatus. Rather, the work is intended as an elucidatory model of certain Tractarian ideas.

2.

The notion to be developed is of an *atomic syntactic system*. An atomic syntactic system S has:

> a vocabulary V of signs, and
> a set T = $\{M_j: j\in J\}$ of sign types

where each $M_j\subseteq V$ and J is an indexing set. Signs here are typographically identified marks. Further, the system S has:

> a set C of manners of sign combination.

A manner of sign combination $c\in C$ will be a manner of combination of a determinate, finite number of ordered signs. The combination in mode c of the signs $s_1, s_2, \ldots, s_n$ so ordered is denoted by $c(s_1, s_2, \ldots, s_n)$. Finally for S there is, for each manner of combination $c\in C$, a rule of the form:

$$x_1\in M_{f(c,1)},\ x_2\in M_{f(c,2)},\ \ldots,\ x_n\in M_{f(c,n)} \Leftrightarrow c(x_1, x_2, \ldots, x_n)\in F$$

where f is some (appropriately partial) function from $C\times\mathbb{N}$ to J. Set F is the set of formulae of S; it contains no members besides those provided by the system's rules of combination. Note that the rules for membership of F have the form of equivalences. What is not allowed in a syntactic system is, say, $c(s,t)\in F$ and $c(u,v)\in F$, but $c(s,v)\notin F$. Each position in each manner of combination determines a set of signs which figure in that position in a formula, and whether or not a combination in a mode of C of signs from V is a formula of the system depends on the signs' positions in the combination and their membership of such sets only.

Next we want to reach an idea of the structure of a syntactic system, abstracting away from the signs and manners of combination deployed in any particular system instantiating that structure. The thought here is that what is of structural interest is simply the number of positions belonging to each sign type in each manner of combination. Thus let's say:

> $X\in T$ occurs n (≥ 0) times in combination c if, and only if, $X=M_j$ and exactly n of f(c,i) are equal to j

And with this we make the definition:

> Two atomic syntactic systems S1 and S2 with manners of combination C1 and C2 and sets of syntactic mark-types T1 and T2, are *isomorphic* if, and only if, there exists a bijection $\alpha: C1\rightarrow C2$ and a bijection $\beta: T1\rightarrow T2$ such that, for all $c\in C1$ and $X\in T1$, (X occurs n times in c) $\Leftrightarrow$ ($\beta(X)$ occurs n times in $\alpha(c)$).

Such a bijection $(\alpha,\beta): C1\times T1\rightarrow C2\times T2$ is an isomorphism from S1 to S2.

The notion of an atomic syntactic system and its structure is now given. Let's take a look at what its interest might be.

[1] See Wittgenstein 1961 5.55 – 5.5571. See also Wittgenstein 1993 pp. 29-30 and Wittgenstein 1979 p. 42.)

3.

An atomic syntactic system is a system of combinations of marks. Whether and how certain marks in a system's vocabulary may combine with each other to make formulae of the system depends on what types of marks they are. With this in view we could say: a *syntactic use* of a sign is a role that sign has in some syntactic system S as a possible element of members of F, the set of formulae of S, by virtue of its membership of a particular sign type of S. Of course, such a use is bound to the modes of combination and sign types of S, but this tie is something we can abstract away from. If S1 and S2 are isomorphic systems with isomorphism (α,β), then the role a mark has in S1 by virtue of its membership of a sign type X of S1 is structurally equivalent to the role in S2 had by a member of $\beta(X)$ by virtue of its membership of that set. And at first glance one might think to say here: the two syntactic uses determine the same *form*. With its syntactic use in a certain system, a sign determines a place in the abstract combinatorial structure instantiated by that system – it determines a form.

On closer inspection, these last two sentences will be seen to be slightly hasty. But let's not worry about that right away. Rather, let's run with them and look instead at a few concrete examples of atomic syntactic systems, beginning with the case of a Russellian system. A Russellian atomic system has:

$$F_R = \cup_{n\geq 2} \{c_n(x_1, x_2, \ldots, x_n): x_1 \in U_{n-1}, x_2, \ldots, x_n \in P\}$$

U_n here is the set of universal signs of degree n, and P is the set of particular signs. In line with traditional scripts, one might use P = {'a_i'}, U_n = {'$R^n{}_i$'} and set $c_n(x_1, x_2, \ldots, x_n)$ to be the combination that the x_i are written in order. (Thus F_R would contain such formulae as '$R^1{}_1a_1$', '$R^2{}_4a_1a_2$', '$R^3{}_2a_5a_1a_6$'.) Of course, many other sign types and combinatorial modes could be used; the resulting systems would, however, all bear the same structure.

Systems with structures quite different from the Russellian structure can of course be readily concocted. Ramsey envisages the possibility of atomic facts consisting of two entities of the same type. Forms answering to this description would arise within such (non-isomorphic) systems as S1, S2 and S3 defined by:

$$F1 = \{c1(x, y): x, y \in A\},$$
$$F2 = \{c2(x, y): x, y \in B\} \cup \{c3(x, y): x \in C, y \in D\}, \text{ and}$$
$$F3 = \{c4(x, y): x, y \in E\} \cup \{c5(x, y, z): x \in E, y \in F, z \in F\}$$

Ramsey's claim against Russell is that we have no more reason to believe that *logical* forms – the forms of reality – are those generated in F_R any more than they are those generated by such entirely different systems as F1, F2 and F3.

4.

Pausing on the system S2, an interesting possibility may come into view. An atomic syntactic system, one will notice, can be non-trivially self-isomorphic. A mapping (α,β):{c2, c3}×{B, C, D}→{c2, c3}×{B, C, D} set to identity other than $\beta(C) = D$ and $\beta(D) = C$ is an isomorphism from S2 to itself. Similarly we might consider a system S4 defined by:

$$F4 = \{c6(x, y): x, y \in G\} \cup \{c7(x, y): x, y \in G\}$$

This system is again non-trivially self-isomorphic with a non-trivial isomorphism taking G to G, c6 to c7, and c7 to c6.

With such possibilities in mind, let's make a few further definitions. Consider a system S with manners of combination C and set of sign types T. Then for each $t \in T$ and $c \in C$ let

$$\Lambda_t = \{x \in T: \text{there is an isomorphism } (\alpha,\beta): C \times T \to C \times T \text{ such that } \beta(t)=x\}$$
$$\Gamma_c = \{x \in C: \text{there is an isomorphism } (\alpha,\beta): C \times T \to C \times T \text{ such that } \alpha(c)=x\}$$

From this we may say that a system S with manners of combination set C and set of syntactic mark-types T is *symmetrical* with respect to $K \subseteq T$ if, and only if, there exists $t \in T$ such that $K=\Lambda_t \neq \{t\}$. Similarly S is symmetrical with respect to $L \subseteq C$ if, and only if, there exists $c \in C$ such that $L=\Gamma_c \neq \{c\}$. If S is not symmetrical with respect to any set then S is asymmetrical.[2]

How are we to place the possibility of symmetry within atomic syntactic systems? Well, Wittgenstein envisages the possibility of distinct objects which are internally indistinguishable.[3] In a similar vein we imagined above a 'world' (call it W1) in which there are only two forms of object and only one mode of combination, a mode in which one object of either form is combined. The two object forms of this world are distinct but the symmetry of the combinatorial situation is such that they are internally indistinguishable. Alternatively we could imagine a world W2 in which there is a single form of objects and two modes of combination, each mode being a mode of combination of two objects. Here the two modes are distinct but internally indistinguishable. And what is in general being imagined with such indistinguishabilities, we can see, are precisely worlds whose structures are instantiated by symmetric syntactic systems. S4 above, for instance, instantiates the structure of W2 and is symmetrical with respect to {c6, c7}. The structure of W1 is instantiated by a system S5 defined by

$$F5 = \{c8(x, y): x \in H, y \in I\}$$

which is symmetrical with respect to {H, I}.

5.

It would appear that we should revise the general thought above that a sign in use in an atomic syntactic system determines a place in the abstract combinatorial structure instantiated by that system, that is that it determines a form. Take the system S2. This system has a structure with three forms; two of these three forms are, however, internally indistinguishable. A sign of S2 which is a member of B determines as such the distinguishable of these three forms; in use as a member of B the sign has that form. Members of C and D, however, determine as such only the class of the two indistinguishable forms: their syntactic use gives the shared nature of the two forms but

[2] Note that the Λt and Γc partition T and C respectively. They cover T and C, for $t \in \Lambda t$ and $c \in \Gamma c$ (put (α,β) to identity). Next, if sign type $q \in \Lambda r \cap \Lambda s$ then there exist isomorphisms $(\alpha 1,\beta 1)$ and $(\alpha 2,\beta 2)$ on S such that $\beta 1(r)=\beta 2(s)=q$. Then $(\alpha 2\text{-}1.\alpha 1, \beta 2\text{-}1.\beta 1)$ is an isomorphism on S such that $\beta 2\text{-}1.\beta 1$ (r)=s. (The inverse of an isomorphism is an isomorphism (as defined), and the composition of isomorphisms is an isomorphism.) Now take some $u \in \Lambda s$. There exists an isomorphism $(\alpha 3,\beta 3)$ on S such that $\beta 3(s)=u$. But then $(\alpha 3.\alpha 2\text{-}1.\alpha 1, \beta 3.\beta 2\text{-}1.\beta 1)$ is an isomorphism on S such that $\beta 3.\beta 2\text{-}1.\beta 1(r)=u$. Thus $u \in \Lambda r$ and so $\Lambda s \subseteq \Lambda r$. Similarly $\Lambda r \subseteq \Lambda s$, and so $\Lambda r=\Lambda s$. In the same way, if there is a mode of combination $d \in \Gamma e \cap \Gamma f$ then $\Gamma e=\Gamma f$.

[3] See Wittgenstein 1961 §2.0233. Indeed, he envisages the possibility of two entities which are externally as well as internally indistinguishable (Wittgenstein 1961 §§2.02331, 5.5302).

does not select between them. Noting the possibility of such a situation one might move to say that a syntactic use determines not a form but a form type. Taking up this description of the matter one needs, however, to bear in mind that the number of 'tokens' had by a particular 'form type' is internal to the type. Where the type has only one token, then, the determination is of nothing less than the token.

In whatever terms one chooses to weaken the general claim that syntactic uses determine forms, the Tractarian position that a sign in *logico*-syntactic use determines a *logical* form comes under threat. Wittgenstein does not know what the logical forms are; he does not know the logical structure of reality. Therefore he does not know that the structure of reality is not symmetrical with regard to certain object forms. But if reality is so symmetrical, a logico-syntactic employment of a sign – that is, a syntactic employment of a sign in a system instantiating the structure of reality – will not always determine a unique logical form.

The point might be thought to be somewhat nitpicking. A logico-syntactic use is guaranteed to determine, as said, a 'form type', even if it is not certain that all logico-syntactic uses will determine a single form. Is this not good enough for Wittgenstein? Well I cannot here follow through what all the repercussions might be for his system if the thesis of the determination of logical form by logico-syntactic use is relaxed as mooted. We can quickly note, however, that on pain of the possibility of nonsense Wittgenstein will have to allow that what one symbol – that is a sign in logico-syntactic use – can mean might depend on what other symbols of the language actually mean. To see this note first that two signs in the same use may not refer to entities of distinct types: two signs in the same use will be intersubstitutable in propositions, and so their reference to entities of distinct types would entail the possibility of nonsense propositions. Now suppose that reality has two internally indistinguishable forms. In a language instantiating the structure of reality there will, under this supposition, be a logico-syntactic use u which determines the type of these indistinguishable forms but does not select between them (in fact there will be two such uses). A sign in use u will not, however, be free to refer to an object of either of these two forms: it will, on pain of the possibility of nonsense, be constrained to refer only to objects of the same form as those referred to by other signs in the same use.

Literature

Ramsey, Frank P. 1990 *Philosophical Papers*, Mellor (ed.), CUP, Cambridge.

Wittgenstein, Ludwig 1961 *Tractatus Logico-Philosophicus*, Pears and McGuinness (tr.), Routledge: London.

Wittgenstein, Ludwig 1979, *Wittgenstein and the Vienna Circle*, B. McGuinness ed, Blackwell: Oxford,

Wittgenstein, Ludwig 1993, *Philosophical Occasions*, J.C. Klagge and A. Nordmann eds, Hackett: Indianapolis

Zwischen Humes Gesetz und „Sollen impliziert Können“ – Möglichkeiten und Grenzen empirisch-normativer Zusammenarbeit in der Bioethik (Teil I)*

Michael Jungert, Bamberg & Tübingen, Deutschland

Vorbemerkung

Die Beiträge „Zwischen Humes Gesetz und „Sollen impliziert Können“ – Möglichkeiten und Grenzen empirisch-normativer Zusammenarbeit in der Bioethik, Teil I und II“ stellen eine Sinneinheit dar, sind jedoch aus technischen Gründen in zwei Abschnitte unterteilt.

1. Empirie und (Bio-) Ethik – Alte Probleme und neue Dringlichkeiten

Bioethik ist in den letzten Jahren zu einem „Exportschlager“ der Philosophie avanciert. Diese Entwicklung hat unter anderem zur Folge, dass sich neben Philosophie und Theologie als ethischen Stammdisziplinen auch zahlreiche andere Fächer aus dem Bereich der Natur- und Sozialwissenschaften zunehmend mit bioethischen Fragestellungen auseinandersetzen.

Die dadurch entstandene Multidisziplinarität hat zum einen zwar wesentlich zur gegenwärtig großen Akzeptanz und institutionellen Verankerung der Bioethik beigetragen, verleiht zum anderen aber auch der Frage nach dem Verhältnis zwischen Empirie und normativer Theorie neue Brisanz. Besonders intensiv werden dabei Debatten um das Verhältnis von *sozial*wissenschaftlicher Empirie und normativer Theorie geführt. Dies hängt in erster Linie damit zusammen, dass insbesondere die empirischen Sozialwissenschaften, worunter neben empirischer Soziologie auch Disziplinen wie Psychologie, Ethnologie, empirische Anthropologie etc. fallen, traditionell einen schweren Stand in normativ-bioethischen Debatten hatten: Viele philosophische Ethiker befürchteten, der Einfluss soziologischer Kontextualisierung auf normative Theoriebildung führe zwangsläufig zu ethischem Relativismus (vgl. Borry et al. 2005: 60).

Zwar sind solche Vorbehalte mittlerweile *faktisch* in den Hintergrund gerückt, systematische Analysen von Möglichkeiten und Grenzen empirisch-normativer Zusammenarbeit sind jedoch weiterhin dringend notwendig, vor allem, weil in den bisherigen Debatten zentrale philosophische Argumente und Theorien nicht angemessen berücksichtigt werden, was zahlreiche begriffliche und argumentative Unklarheiten zur Folge hat.

Vor diesem Hintergrund werden wir im ersten Teil dieses Beitrags zunächst drei idealtypische Modelle empirisch-normativer Zusammenarbeit vorstellen. Daran anschließend entwickeln wir wissenschaftstheoretische und formallogische Kriterien einer adäquaten Zusammenarbeit. Die formallogischen Analysen werden im zweiten Teil fortgesetzt, um darauf aufbauend eine Bewertung der drei Modelle vornehmen zu können. Abschließend skizzieren wir drei konkrete Modi adäquater normativ-empirischer Zusammenarbeit.

2. Formen und Kriterien empirisch-normativer Zusammenarbeit

G. Weaver und L. Trevino folgend lassen sich drei Ansätze empirisch-normativer Zusammenarbeit unterscheiden, welche die Autoren als *symbiotisch, parallel* und *integrativ* bezeichnen (vgl. Weaver & Trevino 1994).

Symbiotische Ansätze postulieren bestimmte *Kontaktstellen* zwischen Empirie und (Angewandter) Ethik, ohne jedoch an der grundsätzlichen theoretischen Eigenständigkeit beider Gebiete zu rühren. Dementsprechend zeichnet sich eine zulässige Zusammenarbeit dadurch aus, dass theoretische und methodologische Kerne beider Disziplinen strikt getrennt bleiben, Empirie und Ethik aber dennoch bzw. gerade deshalb auf eine Kooperation angewiesen sind.

Vertreter des *parallelen* Ansatzes sprechen sich explizit gegen eine Zusammenarbeit von Empirie und normativer Theorie aus und fordern eine *strikte Trennung* beider Bereiche. Neben pragmatischen Faktoren, wie etwa mangelnder Kenntnis der Theorien und Methoden des jeweils anderen Faches, werden vor allem fundamentale theoretische Aspekte als Begründung angeführt. Dazu gehören die notwendige Unterscheidung zwischen Fakten und Normen und das Wertfreiheitspostulat empirischer Wissenschaft sowie der auf David Hume zurückgehende Sein-Sollens-Fehlschluss.

Integrative Ansätze postulieren schließlich das Gegenteil. Durch die *Verschmelzung* der theoretischen Kerne beider Wissenschaftsbereiche sollen die Grenzen zwischen Empirie und normativer Ethik aufgelöst werden. Ein Beispiel für diese Position ist der sogenannte „Integrated Empirical Ethics“-Ansatz von B. Molewijk et al. (vgl. Molewijk et al. 2004). Eine grundlegende Annahme dieses Ansatzes besteht darin, dass Fakten und Normen nicht (klar) voneinander getrennt werden können, weil Fakten in der sozialen Praxis immer normativ aufgeladen sind (vgl. Molewijk et al. 2004: 58). Dies führt zur Forderung einer engen Kooperation zwischen Sozialwissenschaft und normativer Ethik mit dem Ziel, Moraltheorie und empirische Daten miteinander zu verflechten, um letztlich normative Konklusionen unter Rückgriff auf die jeweils relevante Sozialwissenschaft zu ziehen (vgl. Molewijk et al. 2004: 57). Diese Forderungen gehen einher mit der Behauptung, Humes Gesetz stelle kein grundsätzliches Hindernis für eine derartige Zusammenarbeit bzw. Integration dar.

3. Wissenschaftstheoretische Grundlagen empirisch-normativer Zusammenarbeit

Eine Bewertung der genannten Ansätze muss also letztlich auf der Beantwortung der basalen Frage beruhen, ob empirisch-normative Zusammenarbeit in (bio-)ethischen Fragestellungen möglich ist und, wenn ja, wie sie aussehen kann.

* Der zweite Teil dieses Aufsatzes findet sich weiter unten; siehe den Beitrag von Sebastian Schleidgen.

Eine fundamentale Rolle kommt dabei der Unterscheidung zwischen Fakten und Normen zu, die aktuell Gegenstand zahlreicher Debatten ist. Offenkundig hängt es wesentlich von der Position bezüglich dieser Kategorien ab, ob überhaupt sinnvoll über eine Zusammenarbeit gesprochen werden kann. Die Rede davon macht offenbar nur dann Sinn, wenn von zwei nicht-identischen Entitäten gesprochen werden kann, d.h wenn der Bereich des Normativen in einem gewissen Sinn selbstständig und unabhängig vom Bereich der Fakten ist. Dies gibt uns Gelegenheit zur Klärung einiger wichtiger Punkte: Unseren nachfolgenden Überlegungen liegt die Ablehnung jedweder objektivistischer Normativitätskonzeptionen zugrunde. Die Frage nach der Geltung von Normen kann nicht, wie im Fall von Fakten, durch die Untersuchung der Beschaffenheit der Welt geklärt werden. Normen werden nicht entdeckt oder aufgespürt, sondern im Kontext verschiedenster Maßstäbe *definiert*. Die Anerkennung dieser „menschlich-kulturellen Leistung“ (Birnbacher 2004: 6) führt weder zu inakzeptablen ontologischen Dualismen, noch steht sie in einem Zusammenhang mit Letztbegründungsansprüchen. Wenn etwa Gerhard Engel als Vertreter naturalistischer Ethik schreibt, es gelte zunächst „die Realität als Normquelle anzuerkennen“, um daran anschließend die Frage zu stellen „Warum soll der Philosoph moralische Werte erst begründen müssen, wo sie doch in überreicher Anzahl vorzufinden sind?“ (Engel 2004: 52), so ist die Antwort darauf: Natürlich sind moralische Werte in der (sozialen) Wirklichkeit vorzufinden. Jedoch ist damit noch nichts über die *Richtigkeit* moralischer Normen gesagt. Hier zeigt sich ein kategorialer Unterschied, der bei der Frage nach Fakten und Normen häufig übersehen zu werden scheint: der Unterschied zwischen Feststellungen und Begründungen. Während Faktenwissenschaften erstere untersuchen können, bedürfen letztere immer eines nicht-faktischen Kontextes, aus dem Maßstäbe und Kriterien zuallererst generiert werden.

Werfen wir zur weiteren Erhellung dieses Unterschiedes zunächst einen Blick auf die wissenschaftstheoretischen Grundlagen normativer Theoriebildung: Grundsätzliches Ziel von Moraltheorien ist es, im Hinblick auf einen normativ konstruierten moralischen Idealzustand handlungsleitende Normen bzw. Prinzipien zu entwerfen. Die Rede vom moralischen Idealzustand meint dabei *keine* Letztbegründbarkeit moralischer Normen, sondern lediglich jenen Zustand, in dem die spezifische Zielvorgabe einer Moraltheorie vollständig erreicht ist: Der utilitaristische Idealzustand bestünde beispielsweise darin, stets die Nutzensumme aller von einer Handlung Betroffenen zu maximieren. Das bedeutet jedoch nicht, dass ausschließlich konsequentialistische Theorien einen moralischen Idealzustand anstreben. So definiert z.B. Kant den Idealzustand als einen Zustand, in dem ausschließlich im Sinne der reinen praktischen Vernunft gehandelt wird.

Handlungsleitende Normen im Hinblick auf einen moralischen Idealzustand zu entwerfen, bedeutet nun zunächst, „Idealnormen“ (Birnbacher 1988: 16) zu entwickeln, die auf idealtypische Akteure ausgerichtet werden. Die Umsetzung der Idealnormen würde folglich in den avisierten moralischen Idealzustand münden. Jedoch handelt es sich bei den Akteuren der Alltagspraxis nicht um ideale Akteure, weil sie „in ihrem Denken und Handeln kognitiven und motivationalen Beschränkungen unterworfen sind“ (Birnbacher 1988: 16). Deshalb müssen Moraltheorien in einem zweiten Schritt die zuvor entwickelten Idealnormen in Praxisnormen übersetzen, um den Grenzen menschlichen Denkens und Handelns gerecht zu werden. In diesem Übersetzungsprozess sind sie auf empirische Methoden angewiesen, welche die jeweils relevanten Grenzen menschlichen Denkens und Handelns erfassen können. Zentral ist hier, dass Praxisnormen immer auf Idealnormen basieren müssen. Denn ein moralisches Sollen kann ausschließlich durch Idealnormen etabliert werden, Praxisnormen hingegen dienen dazu, die menschliche Praxis so weit wie möglich an dieses Sollen anzupassen und können selbst kein moralisches Sollen etablieren. Dies liegt in erster Linie daran, dass eine Entwicklung von Praxisnormen ohne zugrunde liegende Idealnormen beliebige Beschränkungen menschlichen Denkens und Handelns anführen, mithin beliebige Normen für die Alltagspraxis „begründen“ könnte. Liegen hingegen Idealnormen zugrunde, lassen sich im Zuge der Übersetzung in Praxisnormen nur solche Beschränkungen anführen, die eine Umsetzung dieser spezifischen Idealnormen faktisch verhindern würden.

Man könnte hier einwenden, das Moment der Beliebigkeit werde dadurch lediglich in den Bereich der Idealnormen verlagert. Dies ist insofern richtig, als wir Idealnormen keinerlei objektivistische Annahmen zugrunde legen, die den Beliebigkeitsverdacht ausräumen würden. Dennoch ermöglicht bzw. erleichtert die Verschiebung von Begründungen in den Bereich der Idealnormen einige zentrale moraltheoretische Leistungen: Zum einen eröffnet sie einen *Reflexionsraum*, der im Gegensatz zur Diskussion einzelner Praxisnormen fundamentale systematische Analysen unter Berücksichtigung von Kriterien wie Kohärenz, Reichweite etc. ermöglicht. Der letztlich unvermeidbare Begründungsabbruch findet dadurch im Idealfall auf einem vergleichsweise hohen Reflexionsniveau statt. Zum anderen erleichtert der Rekurs auf den empiriefreien Bereich der Idealnormen eine *Fokussierung* auf die jeweiligen Kernprobleme, die durch die empirische Komplexität im Bereich der Praxisnormen nahezu unmöglich ist.

Bei der Betrachtung von Idealnormen ist allerdings zu berücksichtigen, dass diese oftmals um sogenannte *Brückenprinzipien* erweitert werden. Brückenprinzipien sind Sätze nach dem Schema „Eine Handlung H ist moralisch geboten gemäß der Norm N genau dann wenn das empirisch zu überprüfende Kriterium K gegeben ist“. Sie binden demnach die Geltung einer Norm an ein situationsspezifisch empirisch zu überprüfendes Kriterium K (vgl. Ruß 2002: 119). Dabei ist es gleichgültig, welche moralischen Vorschriften N formuliert. Wesentlich ist, dass das mit ihr verknüpfte Brückenprinzip eine nur empirisch zu leistende Überprüfung von K verlangt, um festzustellen, ob N in der vorliegenden Situation Geltung hat.

Mit Hinblick auf unsere Fragestellung ergibt sich, dass ausschließlich normative Ethik in der Lage ist, Idealnormen zu entwickeln, die Grundlage jeder angemessenen Moraltheorie sind. Schließlich sind es normative Theoretiker, die qua ihres Methodenrepertoires „mit moralischen Begriffen, Argumenten, Normen und Wertsystemen umzugehen“ verstehen (Birnbacher 2003: 61). Auf der anderen Seite ist aber eine Umsetzung moralischer Normen in der Praxis nur auf Basis empirischer Daten möglich: Einerseits im Zuge der *Anpassung* von Idealnormen an die einschlägigen Beschränkungen menschlichen Denkens und Handelns. Und andererseits im Zuge der Klärung von *Anwendungsbedingungen* einer Norm, sofern ihre Geltung an empirisch zu überprüfende Kriterien gebunden ist. Empirische Analysen sind jedoch nicht im Methodenrepertoire normativer Forschung enthalten, weshalb sie genau an diesen Stellen notwendig auf eine Zusammenarbeit mit empirischen Wissenschaften angewiesen ist.

Es stellt sich dann die Frage, welche Rolle empirische *Sozial*wissenschaften in diesem Zusammenhang spielen können. Erstens lassen sich mit ihren Mitteln *interne* kognitive und motivationale Potentiale und Grenzen menschlicher Akteure bestimmen. Darüber hinaus ist eine Erfassung *extern* bedingter Handlungsmöglichkeiten und -beschränkungen möglich, d.h. Rahmenbedingungen einer spezifischen Handlungssituation, die den Handlungsspielraum zwar mit strukturieren, auf die der Akteur aber keinen Einfluss nehmen kann.

Zweitens sind empirische Sozialwissenschaften in der Lage, *kollektive* Prozesse und Veränderungen zu erfassen, beschreiben und erklären, d.h. soziale Prozesse gesamt-, teilgesellschaftlicher oder handlungsraumspezifischer Natur. Das bedeutet beispielsweise, die Auswirkungen bestimmter Normen und Regeln auf das tatsächliche Verhalten der Akteure zu messen oder Fragen nachzugehen, wie in bestimmten Situationen faktisch gehandelt wird, welche moralischen Vorstellungen der Akteure dabei zum Tragen kommen oder welche „neuen“ moralischen Problemstellungen sich – z.B. aufgrund neuer technologischer Entwicklungen – innerhalb einer Gesellschaft ergeben.

Aus den Zielsetzungen und methodologischen Möglichkeiten ergeben sich jedoch auch die immanenten Erkenntnisgrenzen empirischer Sozialwissenschaften in der ethischen Auseinandersetzung: An der normativen Genese moralischer Normen können empirische Sozialwissenschaften *per definitionem* nicht beteiligt sein. Aufgrund ihres Erkenntnisinteresses und Methodenspektrums sind empirische Sozialwissenschaften in der Auseinandersetzung mit ethischen Problemen auf die Erfassung, Beschreibung und Erklärung der sozialen Wirklichkeit festgelegt. Aus diesem Grund ist der normativ interessierte Sozialwissenschaftler (so er denn nicht in Personalunion beides vereint) immer auf eine Zusammenarbeit mit normativen Theoretikern und ihren methodologischen Möglichkeiten angewiesen.

4. Formallogische Grundlagen empirisch-normativer Zusammenarbeit I: Humes Gesetz

Die bislang auf wissenschaftstheoretischem Wege dargelegten Grenzen und Möglichkeiten empirisch-normativer Zusammenarbeit können durch logische Überlegungen weiter untermauert werden. Eine zentrale Grenze empirisch-normativer Zusammenarbeit ergibt sich aus *Humes Gesetz*. Hume schreibt in *A Treatise of Human Nature*: „In every system of morality, which I have hitherto met with, I have always remark'd, that the author proceeds for some time in the ordinary ways of reasoning, and establishes the being of a God, or makes observations concerning human affairs; when of a sudden I am surpriz'd to find, that instead of the usual copulations of propositions, *is*, and *is not*, I meet with no proposition that is not connected with an *ought*, or an *ought not*. This change is imperceptible; but is however, of the last consequence. For as this *ought*, or *ought not*, expresses some new relation or affirmation, 'tis necessary that it shou'd be observ'd and explain'd; and at the same time that a reason should be given; for what seems altogether inconceivable, how this new relation can be a deduction from others, which are entirely different from it.” (Hume 1992: 469)

Ohne Humes Argument an dieser Stelle einer detaillierten metaethischen Analyse unterziehen zu können, bleibt festzuhalten, dass er eine logische Unmöglichkeit konstatiert, direkt von Fakten auf normative Aussagen zu schließen. Seit G.E. Moores Erläuterungen zum naturalistischen Fehlschluss (Moore 1993) versteht man diese These im Wesentlichen auf Basis der Unterschiede im logischen Status von Seins- und Sollens-Aussagen: Während sich deskriptiven Prädikaten Wahrheitswerte zuordnen lassen, ist dies für normative Prädikate nicht möglich (Engels 2008: 134). Daher sind *direkte* logische Umformungen von Seins- auf Sollens-Sätze unmöglich. Wer also aus empirischen Daten – ohne weitere Prämissen –normative Konklusionen zieht, begeht einen logischen Fehlschluss. Allerdings sind *explizit genannte* und *begründete* Sein-Sollens-Schlüsse prinzipiell zulässig: Hume ist der Meinung, dass ein Sein-Sollens-Schluss plausibilisiert werden kann, wenn er hinreichend begründet und erklärt wird (vgl. Hume 1992: 469). Demzufolge müssten zwischen einer deskriptiven Prämisse und einer normativen Konklusion weitere begründete und logisch gültige Schlüsse stehen. In der von Moore geprägten Lesart des Humeschen Gesetzes bleibt aber festzuhalten, dass Schlüsse von rein-deskriptiven auf rein-normative Aussagen prinzipiell unzulässig sind (vgl. Engels 2008: 134). Darauf ist auch und insbesondere im Rahmen normativ-bioethischer Explikationen zu achten, die aufgrund ihres immanenten Anwendungsbezugs in besonderem Maße der Gefahr ausgesetzt sind, gegen Humes Gesetz zu verstoßen (vgl. Engels 2008: 125).

Daraus folgt allerdings nicht, dass eine Zusammenarbeit zwischen (sozial-)wissenschaftlicher Empirie und normativer Theorie prinzipiell unmöglich ist. Es folgt lediglich, dass von (sozial-)wissenschaftlich gewonnenen Fakten nicht direkt auf normative Aussagen geschlossen werden kann. Zu diesem Ergebnis waren wir bereits während unserer Auseinandersetzung mit Erkenntnismöglichkeiten und -grenzen normativer Theorien und sozialwissenschaftlicher Empirie gekommen. Nun können wir dieses Ergebnis um eine logische Komponente erweitern: Normative Theorien und sozialwissenschaftliche Empirie kommen aufgrund der ihnen jeweils immanenten Methoden zu wissenschaftlichen Aussagen, deren logischer Status sich wesentlich unterscheidet.

An dieser Stelle könnte man einwenden, dass solche logischen Fehlschlüsse zwar in abstracto zu befürchten sind, in der Praxis jedoch nur selten vorkommen und daher häufig gegen imaginäre Gegner gekämpft wird. Darauf lässt sich Folgendes entgegnen: Tatsächlich werden naturalistische Fehlschlüsse deutlich seltener begangen als behauptet. Diese Feststellung sollte eine genauere Analyse der jeweils verwendeten Prämissen und Konklusionen zur Folge haben, um zu verhindern, dass der Vorwurf eines naturalistischen Fehlschlusses inhaltliche Diskussionen prinzipiell unmöglich macht. Das ist deshalb wichtig, weil „verdächtige“ Argumente häufig Brückenprinzipien beinhalten, die – einmal expliziert – den Vorwurf des naturalistischen Fehlschlusses entkräften und das jeweilige Argument einer inhaltlichen Diskussion zugänglich machen können. Allerdings spielt diese Feststellung nicht, wie man zunächst meinen könnte, denjenigen in die Karten, die für den Einfluss empirischer Fakten auf normative Theoriebildung plädieren. Vielmehr zeigen solche Fälle einmal mehr, dass Deskription und Normation in entscheidender Hinsicht getrennt sind: Hat man beispielsweise den Verdacht eines naturalistischen Fehlschluss der Form „x ist moralisch gut, weil x eine natürliche Eigenschaft darstellt“ dadurch ausgeräumt, dass man die normative Prämisse „natürliche Eigenschaften sind im moralischen Sinne gut“ in den Syllogismus einfügt, hat man genau diese Trennung vorbildlich aufgezeigt.

Sobald nämlich eine solche normative Prämisse eingeführt wird, muss sich die Diskussion zwangsläufig mit deren Richtigkeit bzw. Falschheit befassen. Zu dieser dezidiert normativen Frage kann Empirie, wie wir bereits gezeigt haben, nichts beitragen, da sie sich auf einer kategorial davon verschiedenen Ebene bewegt, auf die empirische Wissenschaft aufgrund ihres Gegenstandsbereichs und Methodenrepertoires keinen Zugriff hat.

Im nachfolgenden zweiten Teil wird zunächst eine Analyse der „Sollen impliziert Können“-Annahme durchgeführt, eine Bewertung der o.g. drei Modelle empirisch-normativer Zusammenarbeit vorgenommen sowie schließlich drei konkrete Modi dieser Zusammenarbeit vorgestellt.

Literatur

Birnbacher, Dieter 1988 *Verantwortung für zukünftige Generationen*, Stuttgart: Philip Reclam jun.

Birnbacher, Dieter 2003 *Analytische Einführung in die Ethik*, Berlin, New York: Walter de Gruyter.

Birnbacher, Dieter 2004 “Prognosen statt Normen? Das Zusammenspiel von Normen und Fakten in der Angewandten Ethik” in: Christoph Lütge and Gerhard Vollmer (eds.), *Fakten statt Normen? Zur Rolle einzelwissenschaftlicher Argumente in einer naturalistischen Ethik*, Baden-Baden: Nomos: 3-13.

Borry, Pascal, Schotsmans, Paul, and Dierickx, Kris 2005 “The Birth of the Empirical Turn ion Bioethics”, *Bioethics* 19: 49-71.

Engel, Gerhard 2004 “Von Fakten zu Normen: Zur Ableitbarkeit des Sollens aus dem Sein” in: Christoph Lütge and Gerhard Vollmer (eds.), *Fakten statt Normen? Zur Rolle einzelwissenschaftlicher Argumente in einer naturalistischen Ethik*, Baden-Baden: Nomos: 43-59.

Engels, Eve-Marie 2008 “Was und wo ist ein ‚naturalistischer Fehlschluss’? Zur Definition und Identifikation eines Schreckgespenstes der Ethik” in: Cordula Brand, Eve-Marie Engels, Arianna Ferrari, and László Kovács (eds.), *Wie funktioniert Bioethik?*, Paderborn: Mentis: 125-141.

Hume, David 1992 *A Treatise of Human Nature*, Hong Kong: Oxford University Press.

Molewijk, Bert, Stiggelbout, Anne M., Otten, Wilma, Dupuis, Heleen M., and Kievit, Job 2004 “Empirical Data and Moral Theory. A Plea for Integrated Empirical Ethics”, *Medicine, Health Care, and Philosophy* 7: 55-69.

Moore, George Edward 1993 *Principia Ethica*, Cambridge: Cambridge University Press.

Ruß, Hans Günther 2002 Empirisches Wissen und Moralkonstruktion. Eine Untersuchung zur Möglichkeit und Reichweite von Brückenprinzipien in der Natur- und Bioethik, Frankfurt a. M., München, New York: Hänsel-Hohenhausen.

Weaver, Gary, and Trevino, Linda 1994 “Normative and Empirical Business Ethics: Separation, Marriage of Convenience, or Marriage of Necessity?”, *Business Ethics Quarterly* 4: 129-143.

Assessing Humean Supervenience

Amir Karbasizadeh, Tehran, Iran

1. Humean Supervenience:

Humean Supervenience is a central article of faith for David Lewis, who defines it thus:

> "Humean supervenience is named in honor of the greater [sic] denier of necessary connections. It is the thesis that all there is to the world is a vast mosaic of local matters of fact, just one little thing and then another…We have geometry: a system of external relations of spatio-temporal distance between points. Maybe points of spacetime itself, maybe point-sized bits of matter or aether fields, maybe both. And at those points we have local qualities: perfectly natural intrinsic properties which need nothing bigger than a point at which to be instantiated. For short: we have an arrangement of qualities. And that is all. All else supervenes on that." (Lewis 1986 p. *x)*

The "all else" includes nomic facts (laws, physical necessity, causation, etc.). The gist of Lewis' suggestion is that every contingent property instantiation supervenes on the arrangement of perfectly natural properties. One may ask what Lewis means by a "perfectly natural property." Recall that Lewis has a rather hybrid conception of properties, being an amalgam of two very different property conceptions. (1) On the one hand, Lewis has a conception of properties according to which a property is just the set of all of its instances, this-worldly and other-worldy. So the property of being a donkey is the set of all donkeys, both donkeys from our world and other-worldly donkeys. To have this property is to be a member of the class of donkeys. This conception of properties is abundant because on this view, "any class of things, be it every so gerrymandered and miscellaneous and indescribable in thought and language, and be it ever so superfluous in characterizing the world, is nevertheless a property (Ibid, p.192) Concerns from many fronts (e.g., Lewis' desire to formulate viable theories of laws, causation and events) require that there be some way to distinguish the properties that ground objective resemblances and which are causally efficacious from those which are not. (2) In light of this, Lewis has supplemented his abundant conception of properties with a sparse conception of properties. Although his hope is that a viable nominalistic sparse theory of properties is formulable, Lewis would settle for, (roughly) Armstrongian universals as well. With this contrast in mind, we can finally grasp the conception of "perfectly natural properties" operative in Lewis' conception of Humean Supervenience. Lewis gives the following sufficient condition for a property being perfectly natural:

> A property, F, is perfectly natural if its members are all and only those things that share some one universal.

Properties like mass, charge and spin, at least at the present point of scientific development, seem to be apt candidates for being perfectly natural properties.

2. Humean supervenience: Two Independent Theses

Although he does not mention it, Lewis's Humean supervenience has two logically independent theses. The first, which we may call *Separability*, claims that spatio-temporal relations are *the only fundamental external physical relations*. To be precise:

> *Thesis 1 (Separability):* The complete physical state of a non-alien world is determined by (supervenes on) the intrinsic physical state of each spacetime point (or each pointlike object) and the spatio-temporal relations between those points.

Separability posits, in essence, that we can chop up space-time into arbitrarily small bits, each of which has its own physical state, much as we can chop up a newspaper photograph into individual pixels, each of which has a particular hue and intensity. As the whole picture is determined by nothing more than the values of the individual pixels plus their spatial disposition relative to one another, so the world as a whole is supposed to be decomposible into small bits laid out in space and time.

The thesis of Separability concerns only how the total physical state of the universe depends on the physical state of localized bits of the universe. The second component of Lewis's Physical determination takes care of everything else:

> *Thesis 2 (Physical Determination)*: All facts about a non-alien world, including modal and nomological facts, are determined by its total physical state.

I have employed the new terminology "Physical determination" to distinguish Thesis 2 from Physicalism. Physicalism holds that two worlds which agree in all physical respects (i.e. with respect to all items which would be mentioned in a perfected physics) agree in all respects. Thesis 2 essentially adds to Physicalism the further requirement that all physical facts about the world are determined by its total physical state, by the disposition of physical properties. If one holds[1], for example, that the laws of nature do not supervene on the total physical state of the world (at least so far as that state can be specified independently of the laws), then one can be a Physicalist while denying Physical determination. One can hold that worlds which agree on both their physical state and their physical laws agree on all else, while denying that the laws are determined by the state. Lewis's Humean Supervenience importantly maintains the stronger claim.

3. Physicalism and Physical determination

In order to clearly distinguish Thesis 2 from Physicalism, we must remark that the following condition on acceptable analyses is accepted by the Physical determinationist, but not by the Physicalist:

> *Non-circularity condition*: The intrinsic physical state of a non-alien world can be specified without men-

[1] Cf. Carroll 1994

> tioning the laws (or chances, or possibilities) which obtain at the world.

When conjoined with the thesis of Separability, the non-circularity condition implies that the physical state of every spacetime point is metaphysically independent of the laws that govern the world. This in turn implies that the fundamental physical quantities, such as electric charge, mass etc., are metaphysically independent of the laws of electromagnetism, gravitation, and so on. This is a controversial thesis, but one that Lewis accepts. It will not come in for further notice here.

The interest in dissecting Humean supervenience into Separability and Physical determination arises, in the first place, from the remarkable fact that contemporary physics strongly suggests that the world is not separable. This discovery casts the question of *motivating* a desire to defend Thesis 1 into a peculiar light, for one knows beforehand that the motivations, whatever they may be, turn out to lead away from the truth. So before asking why one might want to be Humean, we shall review the evidence that the world is not Humean. Only then will we seek the motivations for defending Separability, and then lastly turn to the possible motivations for Physical determination.

4. Non-Separability in Quantum Theory

The central challenge which quantum theory poses for Separability is the following. Suppose there are a pair of electrons, well separated in space (perhaps at opposite ends of a laboratory) which are in the Singlet State. If the principle of Separability held, then each electron, occupying a region disjoint from the other, would have its own intrinsic spin state, and the spin state of the composite system would be determined by the states of the particles taken individually together with the spatio-temporal relations between them. But, it can be shown, no pure state for a single particle yields the same predictions as the Singlet State, and if one were to ascribe a pure state to each of the electrons, their joint state would be a product state rather than an entangled state. The joint state of the pair simply cannot be analyzed into pure states for each of the components.

5. Lewis’s Reaction and the Motivation for Separability

Lewis is aware that the quantum theory poses a threat to Separability, and says he is prepared to take the consequences:

> But I am not ready to take lessons in ontology from quantum physics as it now is. First I must see how it looks when it is purified of instrumentalist frivolity, and dares to say something not just about pointer readings but about the constitution of the world; and when it is purified of doublethinking deviant logic; and—most of all— when it is purified of supernatural tales about the power of observant minds to make things jump. If, after all that, it still teaches nonlocality, I shall submit willingly to the best of authority.” (Lewis 1986 p. *xi*)

If we take Lewis at his word, then he should abandon Separability (and hence his version of Humean supervenience) forthwith. For one can see how quantum physics looks when purified of instrumentalism, and quantum logic, and consciousness-induced wave collapse. This has been done in several quite different ways: in David Bohm's so-called ontological interpretation (see, e.g. Bohm and Hiley 1993), in the (mind-independent) spontaneous collapse theories of Ghirardi, Rimini and Weber (1986), even in the Many Minds theory of David Albert and Barry Loewer (see Albert 1992). These theories all have fundamentally different ontologies and dynamics, but all agree that the physical state of the world is not Separable, for they all take the wavefunction seriously as a representation of the physical state. This is not to say that Non-Separability is absolutely forced on us by empirical considerations: it would not be impossible to construct a Separable physics with the same empirical import as the present quantum theory. But no one is trying to do it, and there seems to be no reason to start: the quantum theory (in a coherent formulation) is elegant, simple and empirically impeccable. Lewis would not elevate his preference for Separable theories into some *a priori* constraint which could dictate to physics, as the quote shows. Given the definition of materialism cited above, contemporary materialism (i.e. metaphysics built to endorse the approximate truth and descriptive completeness of contemporary physics) must deny Separability.

This leaves us with two questions. First, what drew Lewis to Separability in the first place? Since the thesis appears to be false, we ought to consider carefully the grounds upon which it was thought to be established, or at least rendered plausible. Second, and more importantly, what of Physical Determinism? This second component of Humean supervenience remains as yet untouched by any criticism, and one could continue to insist upon it even while abandoning Separability. Perhaps the physical state of the universe does not supervene on the local intrinsic states of its point-like parts together with spatio-temporal relations, but yet the "modal properties, laws, causal connections, chances" (ibid., p. 111) all are determined by the non-Separable total physical state of the universe. Perhaps. The considerations in favour of Humean supervenience already led us astray with respect to Separability, so why think they are likely to be any more reliable with respect to Physical determination? Before we can even begin to take up this question, we must answer the first: what considerations seemed to support Separability in the first place?

Fortunately, the answer to this question is clear, simple and intelligible. It has, indeed, already been stated. Lewis wants a metaphysics built to endorse the ontology of physics. And, as the quotation from Einstein above forcefully illustrates, *classical* physics *is* Separable. Classical mechanics and field theory do postulate that the physical state of the whole universe is determined entirely by the dispositions of bodies, their intrinsic physical properties (such as charge and mass) and the values of fields at all points in space through time. Taking one's ontology from classical physics does entail Separability. But the advent of the quantum theory, as we have seen, has superseded that argument; it is irreparably damaged, and Lewis has nothing more to say.

6. Counter-examples to Physical Determination

Our survey of Humean supervenience would not be complete unless we consider the second thesis, namely physical determination. In the following sections, we consider a putative knock-down argument against Physical Determination due to John Carroll (1994), which he calls mirror argument. I will argue that this argument does not succeed in bringing out a surprising consequence of the physical determination thesis; it fails as a refutation of that thesis. Physical determination is a very strong negative thesis: it claims that there do not exist any two possible worlds that match with respect to the non-nomic details but have different laws of nature. So one way to argue against it is simply to try to describe a pair of possible worlds that constitute a counter example. This strategy is employed by Michael Tooley (1977, pp.669-672) and also is used in Carroll (1990). I will consider the so called mirror argument against Physical Determination here.

The argument begins with a possible world, U1, that consists of five X-particles and five Y-fields. When each particle enters its Y-field it acquires spin up. All of the particles move in a straight line for all of eternity. But close to the route of one particle there is a mirror on a swivel. (Call this particle "particle b"). The mirror is in such a position (call it "position c") that it does not get in the way of the trajectory of particle b. It seems plausible that the following is a law in U1:

(L) All X-particles subject to Y-fields have spin up.

Now consider U2, a world that is just like U1 except that particle b does not acquire spin up upon entering the Y-field. Hence, L is not true at U2. Now, U1 and U2 do not pose a problem for the MRL view because the worlds differ in their particular matters of fact. The problem stems from considering what would have occurred in each of the worlds had the mirror been in position d, stopping particle b from entering the field. Consider the nearest possible world to U1, U1*; here, it seems reasonable to say that L is a law because the worlds only differ in that the mirror blocks the particle from entering the field. Now consider the closest world to U2, U2*, where the mirror blocks the particle. It seems that although L is true in U2* it is an accident because had the mirror not been in the way L would be false. U1* and U2* are identical in their particular matters of fact; yet it seems that L is a law at U1* but not at U2*. Hence, laws do not supervene on particular matters of fact. Therefore, Physical determination is false.

I guess there is one reasonable response to the Mirror Argument given by Humeans. The Humean just retorts that any counterintuitiveness is not a strike against Physical determination, for such intuitions presuppose an anti-determinationist vantage point. Helen Beebee offers this sort of response to the Mirror Argument. She writes:

> As a friend of supervenience, I have no desire to find a way of grounding the 'fact' that L is a law in U1*, but not in U2*, since I think L is a law in U2* and not an accident. This commits me to the apparently unacceptable claim that the position of the mirror in U2 affects what the laws of nature are, since I am committed to the truth of the counterfactual 'if the mirror had been in position d, L would have been a law.' But I truly see no harm in that … As I said earlier, part of the Humean creed is that laws of nature depend on particular matters of fact and not the other way around; it is no surprise to the Humean, then, that by counterfactually supposing the particular matters of fact to be different one might easily change what the laws of nature are too. The intuition that's really doing the work in this counterexample, then, is the intuition that laws are not purely descriptive … But to describe the example in those terms is not to describe it in neutral terms but to describe it in terms which explicitly presuppose an anti-Humean starting point …

At first blush at least, it is clear why the Humean would feel compelled to assert this. After all, there is a sense in which they are being told that their view is false simply because it doesn't say the laws govern. The Humean thinks the anti-Humean position is in the grip of an intuition which is ultimately incorrect.

7. Conclusion

I have considered Humean Supervenience and its two components and their plausibility. I conclude that the first component of Humean Supervenience namely Separability is untenable. However, I see no reason not to believe in the second component of it.

Literature

Albert, D. Z., 1992, *Quantum Mechanics and Experience*. Harvard University Press.

Beebee, Helen., 2000, "The Non-Governing Conception of Laws of Nature", *Philosophy and Phenomenological Research*, LXI, No.3.

Carroll, John., 1994,.*Laws of Nature* Cambridge: Cambridge University Press.

Lewis, D., 1973, *Counterfactuals*, Cambridge: Harvard University Press.

-----------, 1983, "New Work for a Theory of Universals, *Australasian Journal of Philosophy*, 61: 343-377.

-----------, 1986, *Philosophical Papers*, Volume II, New York: Oxford University Press.

Roberts, John., 1998. "Lewis, Carroll and Seeing Through the Looking Glass" *Australasian Journal of Philosophy* 76(3).

Tooley, M.,1977, "The Nature of Law", *Canadian Journal of Philosophy,* 7: 667-98

Zu Carnaps Definition von ‘Zurückführbarkeit’

Roland Kastler, München, Deutschland

In den *Principia Mathematica* versuchen Whitehead und Russell (Whitehead, Russell 1957[2]) die Begriffe der Mathematik in jene der Logik (Typenlogik bzw. Logik plus Klassentheorie)einzubetten. Rudolf Carnap erweiterte das Anwendungsgebiet der in den *Principia Mathematica* eingeführten Methode der logischen Konstruktionen, indem er in seinem Werk *Der logische Aufbau der Welt* (Carnap 1966[3]) die Grundzüge eines Projektes darstellt, welches die Begriffe der Welt auf unmittelbar Gegebenes zurückzuführen intendiert. Die von Carnap entwickelte Konstitutionstheorie nimmt dabei nicht nur im allgemeinen Bezug auf Russell und Whitehead, und zwar in dem Sinne, in dem Carnap sich beispielsweise dem Phänomenalismus Ernst Machs verpflichtet fühlt (Vgl. Carnap 1993, 29.), sondern er formuliert bezüglich der „*Principia*“, daß jene ‘ein „Konstitutionssystem“ der mathematischen Begriffe’ darstellen (Vgl. Carnap 1966, 47f.).

Gegen Carnaps Konstitutionstheorie wurden nun innerhalb der Literatur eine Vielzahl von Argumenten vorgebracht, wobei insbesondere auch seine Definition von ‚Reduktion’ bzw. ‚Zurückführbarkeit’ im Mittelpunkt der Kritik stand. Bevor aber auf das prominenteste dieser Argumente gegen sein Reduktionskonzept eingegangen werden soll, seien vorerst Beispiele für konstitutionale Definitionen aus der „*Principia*“ und dem „*Aufbau*“ angeführt:

(K1) $0 =_{\mathrm{Def}} \{\varnothing\}$, $1 =_{\mathrm{Def}} \{\alpha \mid \exists x\, (\alpha=\{x\})\}$, $2 =_{\mathrm{Def}} \{\alpha \mid \exists x \exists y\, (x \neq y \wedge \alpha=\{x,y\})\}$
$\mu+\nu=\{\zeta \mid \exists\alpha\exists\beta\, (\mu\in nc \wedge \nu\in nc \wedge \alpha\in\mu \wedge \beta\in\nu \wedge \alpha\cap\beta=\varnothing \wedge \zeta=\alpha\cup\beta)\}$

(K2) $\mathrm{gesicht} =_{\mathrm{Def}} \{\alpha \mid \exists\lambda(\lambda\in \mathrm{sinn} \wedge \mathrm{Dzp}(\, 5, \lambda, \alpha, \mathrm{Umgr'Aq}))\}$

(K1) gibt also die Definition von Zahlausdrücken sowie des additiven Funktionszeichens an, wobei ‘nc’ die Klasse aller Kardinalzahlen bezeichnet. Es ist hier die vereinfachte Carnapsche Version der „*Principia*“ dargestellt (Vgl. Carnap 1929, 52.), welche aber die zugrundeliegende Idee präzise widerspiegelt (Vgl. zur „*Principia*-Version“: Russell, Whitehead 1957, Vol. II, 72.). In (K2) wird der Gesichtssinn als diejenige Sinnesklasse von Qualitäten bestimmt, deren Ordnung der Qualitäten in Bezug auf die durch *Aq* bestimmte Umgebungsrelation die Dimensionszahl 5 hat (Vgl. Carnap 1966, 155.).

Russell formuliert zur oben angegebenen Definition an anderer Stelle, dass aufgrund dieser Definition die Zahl 2 die Klasse aller Paare *ist* (Vgl. Russell 1975, 29.). Eine der entscheidenden Fragen, welche sich in Bezug auf Carnaps Begriff der Zurückführbarkeit stellen wird, ist jene, ob wir die konstitutionalen Definitionen so interpretieren müssen, wie sie jene Äußerung von Russell - welche hier natürlich völlig aus dem Zusammenhang gerissen präsentiert wird - nahezulegen scheint.

Carnap definiert zunächst folgendermaßen:

> Unter einer „konstitutionalen Definition“ des Begriffes a auf Grund der Begriffe b, c verstehen wir eine Übersetzungsregel, die allgemein angibt, wie jede Aussagefunktion, in der a vorkommt, verwandelt werden kann in eine umfangsgleiche Aussagefunktion, in der nicht mehr a, sondern nur noch b, c vorkommen. (Carnap 1966, 47.)

Und in engem Zusammenhang zum Begriff der konstitutionalen Definition steht jener der Zurückführbarkeit:

> Gibt es zu jeder Aussagefunktion ausschließlich über die Gegenstände a, b, c,... (wobei b, c ... auch fehlen dürfen) eine umfangsgleiche Aussagefunktion ausschließlich über b, c ..., so heißt a „zurückführbar“ auf b, c, ...
> [...] Unter einer Aussage oder Aussagefunktion „ausschließlich über die Gegenstände a, b ...“ verstehen wir eine solche, in deren schriftlichem Ausdruck als nichtlogische Zeichen nur „a“, „b“, ... vorkommen (Carnap 1966, 47.).

Gegen einen derartigen Reduktionsbegriff könnte man einwenden, dass extensionale Identität nicht das Kriterium sein kann, da ja beispielsweise Zahlprädikate innerhalb einer Zahlentheorie auf Zahlen zutreffen und nicht auf Mengen genauso wie Mengenprädikate innerhalb einer Mengentheorie auf Mengen zutreffen und nicht auf Zahlen, weshalb es zu keinem Zahlprädikat ein umfangsgleiches Mengenprädikat geben kann.

Es gilt hier also zunächst zu prüfen, ob die obige Reduktionsdefinition überhaupt derart interpretiert werden kann.

Carnap arbeitet bereits 1928 am zweiten Teil zu den *Untersuchungen zur allgemeinen Axiomatik* (Vgl. Bonk, Mosterin 2000, 47.), in welchen es u. a. um eine Formulierung von Extremalaxiomen in der Objektsprache geht. Ein Beispiel für ein Extremalaxiom wäre innerhalb des Hilbertschen Axiomensystems der euklidischen Geometrie das sogenannte Vollständigkeitsaxiom. Dieses behauptet, dass die Grundgegenstände des Axiomensystems bei Aufrechterhaltung sämtlicher anderer Axiome nicht erweitert werden können (Vgl. Carnap 1936, 166f.). Wir können demnach festhalten, dass für Carnap Theorien zunächst Satzmengen darstellen, welche sich auf einen Gegenstandsbereich beziehen. Im *Abriß der Logistik* formuliert Carnap u. a., wie er das Konstitutionssystem der „*Principia*“ mittels des Explizitbegriffes eines (mathematischen) Axiomensystems zu erweitern versucht. Er stellt dies u. a. am Beispiel des Hausdorffschen Axiomensystems dar, welches als Klasse der Hausdorffschen Umgebungssysteme einen rein logischen Begriff darstellt (Vgl. Carnap 1929, 76ff.). Eine derartige Formulierung kann aber nun in der Sprache des „*Aufbaus*“ als eine „Zurückführung von mathematischen auf logische Gegenstände“ betrachtet werden. Nehmen wir nun also an, die Relata der Zurückführbarkeitsrelation beziehen sich auf die Gegenstandsbereiche zweier verschiedener Theorien. Da Carnap zu jener Zeit, in welcher der „*Aufbau*“ ausgearbeitet wurde, Theorien immer in der typenlogischen Sprache formuliert (Vgl. Carnap 1929, 70 – 90.), setzen wir dementsprechend zusätzlich voraus, dass unsere fraglichen Theorien in einer Typenlogik gegeben sind. Und aufgrund der Tatsache, dass insbesondere dem Problem der Reduzierbarkeit von Grundgegenständen der einen Theorie auf Gegenstände der anderen Theorie besondere Relevanz zukommt, da eine derartige Reduktion eines der wesentlichen Ziele eines konstruktionalen Systems darstellt, soll die Frage an diesem Spezialfall verdeutlicht werden.

Aus der obigen Definition folgt, dass, wenn ein Gegenstand a auf einen Gegenstand b reduzierbar ist, es zu jeder Aussagefunktion über a eine umfangsgleiche Aussagefunktion über b gibt. Also gibt es beispielsweise zur Aussagefunktion ‘... $\in\{y \mid y=a\}$’ eine solche ausschließlich über b. Weiters gilt, dass, wenn a zurückführbar auf b ist, es eine konstitutionale Definition des Gegenstandsnamens von a mittels einer Formel über b gibt (Vgl. Carnap 1966, 47.). Konstitutionale Definitionen sind aber bei Carnap Identitätsaussagen, welche in der Objektsprache des Systems formuliert sind. Da nun Identität symmetrisch ist, müssen sowohl die Extension des Definiendums als auch jene des Definiens Elemente des Vorbereichs der Identitätsrelation sein, somit gemäß der Russellschen Typentheorie bzw. der Carnapschen Fassung derselben vom selben Typ sein. Damit aber ist klar, dass die obige Definition nicht von der Relation der Zurückführbarkeit zwischen Gegenständen verschiedener Theorien handelt, da eine derartige Auffassung einem wichtigen Grundprinzip der Konstitutionstheorie widerspricht. Ist nämlich a zurückführbar auf b, so soll a einen logischen Komplex von b darstellen, welcher beispielsweise mit der Klasse von b gegeben wäre (Vgl. Carnap 1966, 48.). Es muss also die Extension des Definiendums von höherem Typ als b sein, was aber nicht möglich ist, da der Gegenstand a ja einen Grundgegenstand darstellt, somit vom Typ 0 ist, wie eben auch der identische Gegenstand, welcher Extension des Definiens ist.

Ebenso wenig kann ‘Zurückführbarkeit’ bedeuten, dass auf verschiedene Gegenstandsbereiche der Konstitutionstheorie Bezug genommen wird, wenn man darunter zwei getrennte Grundgegenstandsbereiche und ihre daraus gebildeten Gegenstände versteht. Denn auch hier kann mit Hilfe des obigen Argumentes und mit einem ähnlichen Argument für höherstufige Gegenstände gezeigt werden, dass unter einer derartigen Interpretation der Begriff der Zurückführbarkeit einen leeren Begriff darstellen würde.

Die Relata der Zurückführbarkeitsrelation sind also zunächst lediglich Objekte der Konstitutionstheorie, und zwar jene, deren sie bezeichnende Ausdrücke im Konstitutionssystem definiert werden, sowie die Extensionen der im Definiens vorkommenden Ausdrücke, welche sich letztlich auf die Grundgegenstände beziehen. In Carnaps erläuterndem Beispiel für die These, dass jeder wissenschaftliche Begriff eine Klasse oder Relation ist, welche sich alleine mit Hilfe der Grundausdrücke formulieren lässt, wird dementsprechend auch genau dieser Sachverhalt veranschaulicht (Vgl. Carnap 1966, 118ff.). Eine derartige Auffassung ist des Weiteren durchaus verträglich mit seinen allgemeinen Charakterisierungen eines Konstitutionssystems. Denn ein Konstitutionssystem ist zunächst ein System, in welchem die Begriffe bzw. Gegenstände schrittweise aus den Grundbegriffen abgeleitet werden (Vgl. Carnap 1966, 2.), was wiederum bedeutet, dass jedes Axiomensystem, welches diesen Sachverhalt der schrittweisen Ableitung erfüllt, ein Konstitutionssystem darstellt (Vgl. Carnap 1929, 70f.). Axiomensysteme können aber für sich interpretiert werden und weisen nicht notwendigerweise einen Bezug zu einem anderen Axiomensystem auf, womit in einem solchen Fall eine Zurückführbarkeitsrelation sich zur Gänze auf die im System definierten Ausdrücke beziehen würde.

Wenn nun aber die Gegenstände mit Ausnahme der Typen keine weitere Sortierung erfahren, dann bedeutet ‘Zurückführbarkeit’ lediglich, dass bestimmte Klassen oder Relationen bzw. Klassen von Klassen usw. von Gegenständen durch (klassen-)logische Operationen aus Gegenständen, Klassen oder Relationen bzw. Klassen von Klassen usw. von Gegenständen gewonnen werden können. Carnap aber will, wie er im Aufbau an vielen Stellen erklärt, eigentlich Zahlen auf Klassen, Physisches auf Psychisches und umgekehrt sowie rationale Zahlen auf natürliche Zahlen usw. zurückführen. Der fehlende Zwischenschritt wird nun klar, wenn man zusätzlich beachtet, dass die konstitutionalen Definitionen im Aufbau jeweils eine definitorische Erweiterung der Konstitutionstheorie bedeuten, also eine Erweiterung ihres Vokabulars. Demnach kann man nun ‘Zurückführbarkeit’ folgendermaßen charakterisieren:

(Z) Der Gegenstandsbereich der Theorie T ist genau dann auf jenen der Konstitutionstheorie K zurückführbar, wenn K durch konstitutionale Definitionen derart definitorisch erweitert werden kann, sodass gilt: T ist Teilmenge der durch die Definitionen erweiterten Konstitutionstheorie.

Mit dieser Bestimmung (Z), wobei natürlich noch eine Anpassung der Typenindices erfolgen muss, welche dann je nach gewünschter Reichweite des Kriteriums verschieden formuliert werden kann, ist es zunächst im Prinzip durchaus verträglich, dass die Gegenstände von T und jene von K verschieden sind. Kriterium ist zunächst nur die Strukturerhaltung. Dem aber scheinen die vielen „identifizierenden“ Aussagen Carnaps zu widersprechen, wenn er etwa meint, dass in Bezug auf den Leib-Seele-Dualismus im Konstitutionssystem des *„Aufbaus“* ein Monismus von Physischem und Psychischem gilt (Vgl. Carnap 1966, 223f.), oder er an anderer Stelle formuliert, dass bei der Zurückführung von Physischem auf Psychisches dem physischen Gegenstand *seine* wahrnehmbaren Kennzeichen zuzuordnen sind (Vgl. Carnap 1966, 78.), oder er intentionale Gegenstände als komplexe Ordnungen von Erlebnissen darstellt (Vgl. Carnap 1966, 227.). Man scheint also durchaus gerechtfertigt zu sein, ihn dahingehend zu interpretieren, dass er eine Identifizierung der Gegenstände der in (Z) formulierten Zurückführbarkeitsrelation vornimmt. Demgemäß ist ‘a ist zurückführbar auf b’ so zu verstehen, dass a im Grunde nichts anderes ist als b.

Dagegen aber richtet sich nun der Haupteinwand Goodmans, ein Einwand, den übrigens auch Quine formuliert (Vgl. Quine 1997^7, 212f.), bezüglich des Konstruktionssystems des *„Aufbaus“* und ähnlicher Theorien. Ist nämlich ein Gegenstand derart beschaffen, dass er mittels verschiedener Konstruktionswege gebildet werden kann, so bedeutet dies zunächst, dass seine entsprechenden konstitutionalen Definitionen aufgrund der Identität in verschiedenen Konstruktionssystemen formuliert werden müssen. Aber da er weiters im Grunde nichts anderes ist als der eine und auch im Grunde nichts anderes ist als der andere Gegenstand, müssen auch diese beiden Gegenstände identisch sein, was aber aufgrund der eindeutigen Identitätskriterien von Klassen nicht sein kann. Ein anschauliches Beispiel innerhalb elementarer Sprachen stellt diesbezüglich die Reduzierbarkeit der Peano-Arithmetik auf ZF mittels der von Neumannschen Version und der Zermeloschen Version dar (Vgl. Goodman 1966^2, 9 oder auch 22.).

Nun ist es zwar richtig, dass im Konstitutionssystem (im engeren Sinn) aufgrund des Sachverhaltes, dass konstitutionale Definitionen Identitätsaussagen sind, keine voneinander verschiedenen Konstitutionswege desselben Gegenstandes formulierbar sind. Und es ist auch richtig, dass Carnap, nach unserer Interpretation aufgrund von (Z), „ontologische“ bzw. „identifizierende“ Aussagen trifft. Dies stellt jedoch nur dann ein Problem dar, wenn innerhalb der ontologischen Interpretation die Identität von

Klassen und abstrahierten Entitäten vorausgesetzt wird. Betrachten wir dazu die folgende Bemerkung Carnaps:

‘*Die Klassen sind als Extensionen Quasigegenstände*. Die Klassenzeichen haben keine selbständige Bedeutung, sie sind nur ein zweckmäßiges Hilfsmittel, um allgemein über die Gegenstände, die eine bestimmte Aussagefunktion befriedigen, sprechen zu können, ohne sie einzeln aufzählen zu müssen. *Das Zeichen einer Klasse repräsentiert also gewissermaßen das diesen Gegenständen, ihren Elementen, Gemeinsame*.’ (Carnap 1966, 44.)

Die Frage, welche sich nun also stellt, lautet, ob ‘Repräsentation’ Identität bedeutet, ob also der Klassenterm die abstrakte Entität bezeichnet. Es ist aber gerade das Programm des *„Aufbaus“*, welches an dieser Stelle klarmachen sollte, dass eine derartige Gleichsetzung nicht adäquat ist, da ja Carnap nicht müde wird zu betonen, dass dieselben Objekte auf verschiedene Art und Weise konstituiert werden können, nämlich beispielsweise auf physischen oder psychischen Basen von Konstitutionssystemen. Die Bestimmung, dass die Identitätskriterien von abstrakten Entitäten gleich jenen von Klassen sind, ist daher, zumindest was den *„Aufbau“* betrifft, unangemessen. Wenn wir aber eine derartige Identifizierung nicht vornehmen, dann könnten die in Frage stehenden ontologischen Interpretationen folgendermaßen paraphrasiert werden:

Auf Basis eines Konstitutionssystems K und des Prinzips (Z) gilt, dass der Gegenstand der zu reduzierenden Theorie S, nennen wir ihn ‘a’, identisch ist mit jener abstrakten Entität c, welche durch einen bestimmten Quasigegenstand (Klasse) der Konstitutionstheorie K repräsentiert wird.

Mit dieser Bestimmung aber, wobei wir hier im Unterschied zum oben angeführten Zitat Repräsentation als eine Relation zwischen Objekten verstehen, steht durchaus nicht im Widerspruch, dass a identisch ist mit einer abstrakten Entität, welche durch einen Quasigegenstand einer anderen Konstitutionstheorie repräsentiert wird. Daraus würde dann eben nur folgen, dass diese beiden Entitäten identisch sind.

Das Problem also, welches innerhalb der Literatur oft mit Carnaps Definition von Zurückführbarkeit verbunden wird bzw. als Haupteinwand formuliert wird, scheint also bei einer Unterscheidung von drei Ebenen, nämlich des Konstitutionssystems im engeren Sinn, des Prinzips (Z) und der ontologischen Interpretation, wobei Carnaps Definition sinnvoll auf der ersten und dritten Ebene angewendet werden kann, nicht einen logischen Widerspruch aufzuweisen, sondern vielmehr darauf hinzuweisen, dass der *„Aufbau“* fragmentarisch ist. Und zwar nicht nur in dem Sinne, dass Teile, wie etwa die Konstitution des physischen Raumes, nicht ausgeführt sind, sondern auch dahingehend, was natürlich die Leistung des *„Aufbaus“* in keinerlei Weise schmälern soll, dass für die konstituierten abstrakten Entitäten noch Identitäts- und Einzigkeitsbedingungen zu formulieren sind. Dies weist meiner Meinung nach wieder einmal darauf hin, dass der *„Aufbau“* in erster Linie erkenntnistheoretisch, das heißt als eine Theorie des Erkennens bzw. der Erkenntnisprozesse, zu interpretieren ist.

Literatur

Bonk, Thomas and Mosterin, Jesus 2000 “Einleitung”, in: Rudolf Carnap, *Untersuchungen zur allgemeinen Axiomatik*, Darmstadt: Wissenschaftliche Buchgesellschaft, 1-52.

Carnap, Rudolf 1929 *Abriß der Logistik*, Wien: Springer.

Carnap, Rudolf 1966^3 *Der logische Aufbau der Welt*, Hamburg: Meiner.

Carnap, Rudolf 1993 *Mein Weg in die Philosophie*, Stuttgart: Reclam.

Carnap, Rudolf and Bachmann, Friedrich 1936 “Über Extremalaxiome”, *Erkenntnis* 6, 166-188.

Goodman, Nelson 1966^2 *The Structure of Appearance*, Indianapolis: Bobbs-Merrill.

Quine, Willard v. O. 1997^7 “Ontological Reduction and the World of Numbers”, in: Willard v. O. Quine, *The Ways of Paradox and Other Essays*, Cambridge (Mass.): Harvard University Press, 212-220.

Russell, Bertrand 1975 *Einführung in die mathematische Philosophie*, Wiesbaden: Vollmer.

Whitehead, Alfred N. and Russell, Bertrand 1957^2 *Principia Mathematica*, Vol. I-III, Cambridge: Cambridge University Press.

Ding-Ontology of Aristotle vs. *Sachverhalt*-Ontology of Wittgenstein

Serguei L. Katrechko, Moscow, Russia

In the history of philosophy, we can distinguish three possible types of ontology[1]. The first claims that the *world* 'is made up' of *things* that are considered its initial elements. In the Antiquity, the *ontology of things* tendering to nominalism was presented by Aristotle and Democritus, and their concepts, despite the differences, belong to the same type of ontology. They differ only on a scale of 'thing-ism', the former postulating that the world consists of things, the latter considering that the world is built up of atomic 'bricks', or particles, regarded as *micro–things*, that, in their term, make up ordinary things that we are used to. And it is *the ontology of things* that has been an overwhelming ontology of contemporary natural science.

The second and the third types of ontology are based on predicate interpretation of *being*, and postulate a non-object character of the world. If we take the classical approach to structure a sentence (resp. the world) as '*S is P*' the ontology of things is accented on '*S is*— ', with '*S is'* representing an inseparably linked complex, and *S* — the essence of the thing which acts as a *substance* for predicates of the thing (resp. grammatically *S* acts as the *subject* of the sentence). Predicate ontology is the one of the '— is P_x' type where 'things' (resp. *subjects* of the sentence) become secondary formations, and are determined not by the essence but by their predicates. Accordingly, *esse* is being related here not with the *subject* but with the prior–predicate 'is' (resp. link–verb of the sentence), and produces as *transcendental condition* for the rest 'real' predicates of the thing (P_1, P_2, $P_{3\ldots}$).

We can also mark out two subtypes in the predicate ontology. The first one presents *being* as *property* (resp. unary $P^1{}_n$ predicate), with *property* interpreted as '— is $P^1{}_n$' inseparably linked predicative complex. It is the property of things that is considered prior in ontology while things, being secondary, act as 'intersections' of properties (bundle theory of substance). For instance, *the table* is something *that is made of wood, right–angled, yellow, and used for writing*. Here, *something* is predetermined by its properties (as unary predicates). Plato's ontology has been the first and forming theory of the type so, this type of ontology can be named Platonic ontology. It's possible to demonstrate that such interpretation of Plato's idealism does not sound like idealism at all. Priority of Plato's 'world of ideas' might be understood as a mere acceptance of priority of properties (predicative complex 'is $P^1{}_n$') in respect of thing (*S*–subject). Furthermore, similar concepts are more realistic as compared to both, Aristotelian ontology and natural ontology of contemporary natural science presuming virtual existence of matter (as universal), for we experience not 'disguised' (latent) Aristotelian *essence* and not hypothetic *matter* postulated by up-to-date physics (e.g. dark matter in astronomy) but *real properties* of objects that we can discover objectively, through perception or by means of instruments.

It's worth noting, that Aristotelian and Platonic ontology do not reject but, rather, enrich each other. These are two different approaches to the world defining its diverse sections, and each of two has the right to be (similar to corpuscular wave dualism in physics). They produce two necessary (transcendental) conditions of being of objects. The first one postulates the presence of single self-identical essence as an indispensable 'sublayer' (sub–stance) for property of thing, while the second one dictates the necessity 'to partake' things into the world of ideas which terminates the possibility to take into possession this or that set of properties.

Explication of the second predicate ontology was made much later on. It interprets *being* as relation (n–ary predicate $P^k{}_{n}$.), with *Tractatus* being one of its variations pronouncing that the world is the totality of facts, not of things (prop. 1.1). Here, *facts* act as something different from things, as a sort of relation between things or 'combinations of things' (prop. 2.01). Hence, under *Tractatus*, it is relation that is believed to be prior while object is defined through a set of relations it could become a constituent part of, and, according to Wittgenstein, the possibility of that must be already written into the object itself (prop. 2.011–2.0121).

This proposition differs from Aristotelian ontology considering *essence* prior to *relation* that may be added to it as random (accidental) characteristic: 'that which is per se, i.e. substance, is prior in nature to the relative for the latter is like an off-shoot and accident of being' (Aristotle, *Nicomachean Ethics,* 1096a20; Boethius, *On the Trinity,* § 5). Challenging such an underestimation of 'relation' category Plato, in anticipation to *Tractatus* speculations, could have argued that 'anything which possesses any sort of power to affect another, or to be affected by another… had real existence' (Plato, *Sophist*, 247e), i.e. really exists only that which is able to interact. Then, one more transcendental condition is being revealed, that is relations (interactions) deprived of which not a thing could exist (resp. *esse* acts here as grounds for any 'real' relations)[2]. We can register *things* into ontology status of being provided that all three transcendental condition are observed.

Thus, we can separate three types of ontology: the ontology of things, attributes and relations. Each of them is correlated with a particular type of language. In the ontology of things the key structure belongs to the noun. *Esse* interpreted as property, the key position goes to the adjective. With *esse* conceived as relations, separate words (nours, verbs or adjective) give way to the whole sentence structure, expressing the facts of relationship between objects. For instance, the sentence '*The stone is falling'* which purportedly postulates existence in the world of Aristotelian 'initial entities' (here: stones) able to act, will

[1] The term 'ontology' will be used here in two related but a bit different meanings. The initial meaning of the word is concerned with the doctrine (teaching) of "being qua being" (Aristotle, *metaphysica generalis*). The second one studies how the universe is made up, i.e. what ontological commitments we accept (*metaphysica specialis*). Further on, we'll speak of concrete ontology (the second meaning) which depends on the general understanding of existence.

[2] To put it more precisely, had an object ever existed, since it never becomes a constituent part of any relation, we couldn't have learnt about it since learning is also a relation between an object and a subject.

be substituted, in predicate interpretation of *being* as relations, by a verbal modified sentence like *'It's stoning'* [3].

Let's pass now to a more detailed analysis of the ontology of the *Tractatus* which could be conceived in different ways. So, interpretation of the ontology of the *Tractatus Logico-Philosophical* (TLP) proposed by the author of this article will be presented above. Basically, it will be proceeded from the point that the TLP ontology gives a logical description of the world, i.e. is a logic ontology; it does not postulate that there are some unchangeable 'entities' like Aristotelian 'things', i.e. it is of non-substantial character[4].

Peculiarity of the TLP ontology is that it strives to describe the world as a *system of combinations* (interacting bodies), choosing isomorphism of the world and the language as an important heuristic principle. To solve this problem a definite balance of statics (synchronism) and dynamics (diachronism) is needed. Logic analysis allows to fix a snapshot of the present 'state of affairs' (Sachlage) or facts (Tatsache). Further, this 'picture of the world' may be specified and enriched by discovering new facts and observing results (consequences) of the old ones. The proposed interpretation of the *Tractatus* emerged from reflections on the translation of the term 'Sarhverhalt' which was translated into Russian not with a standard equivalent '*sobytie*' but with an original term '*so-bytie*' (where *'bytie'=esse* (Lat.), or ='*Being*'; Wittgenstein, 1994; Kozlova, 1995) thus, marking co–existence (*common existence/being*) of objects, i.e. inevitability of existence of an object 'excluded from the possibility of combining with others' (prop. 2.0121). While classical ontology thinks of an object as something self-sufficient (closed), the TLP ontology regards it open, inviting other 'things', demonstrating 'so–bytie–nost'' (in the Russian language that is similar to 'co–existence–ness'); each thing is predetermined by its own system of correlations for it 'is, as it were, in a space of possible states of affairs' (prop. 2.013).

'Simple facts' (Sachverhalte), ultimately accumulated into totality, determining the logical world of the TLP ontology (since complex facts consist of simple ones), are isomorphic to simple sentences describing a state of affair (*A book is on the table*), i.e. has a «*A–x–B*» structure, with «*–x–*» denoting a particular combination (relation) between *A* and *B* (prop. 2.01). Propositions 2.01 — 2.02 grasping main differences between the TLP ontology and the ontology of things seem to be essential in conceiving the specificity of the TLP ontology. While Aristotle believes that initial elements of the world are unchangeable entities «A» and «B» predetermining «*–x–*», Wittgenstein states the priority of the «–x–» *functional relation*. However, we cannot imagine objects in isolation, and their extracting ('exclusion') from the combined system leads to a gross idealization inadmissible in a common case, which Wittgenstein tries to fight.

What makes the basis for the ontology *turn* carried out in the *Tractatus*? If we take the book ('lying') on the table as the object of our analysis, we can use conventional body and visual language, since all objects of our world belong to the 'continuous substance' (Descartes). Substantiality, in this case, means that the changes taking place with the book might be neglected; the book will stay a book, with its identity predetermined by its essence. But if we need to describe a *process*, say, an electric current impulse and its magnetic field, the current (or the magnetic field) stops to be a *constant thing* with its own unchangeable *essence* existing like a *book*. The example with the current encourages us to be more critical to postulating Aristotelian entities, although the essence of the Wittgenstein's turn is connected not with the dynamic nature of the *current* but, rather, with the fact that it demonstrates the example of an imperceptible 'thing' (contrary to the example with the book) which indicates lack of means for body and visual description typical for classical ontology; the world consists not only of spacial objects but of non-visual 'state of affairs' as well, and while describing the world, should the description has a claim on adequacy and universality, those points must be observed. According to Wittgenstein, there is a universal language of description, and that is *logics*, in the broad sense, conceived as the teaching of *functions* (G. Frege). E.g., the 'state of affairs' of the *current* can be described by a formula (compare with *Sachverhalt*) showing the dependence of the *strength of the current* on the *tension* and *resistance*, acting here as basic constituencies. In a general case, any *state of affairs* is given by a *logic and functional space*, with strength lines 'combining' the 'things' within and, by that, predetermining their characteristics while their 'intersections' correspondingly constitute a particular 'thing'.

In a sense, correlation of the ontology of things and the ontology of facts is similar to correlation of atomic particles and field structure in physics. The ontology of facts postulating the priority of field structures in respect of particles is *holistic* contrary to the elementary ontology of things. According to up-to-date mathematics, difference between the ontology of things and the ontology of facts can be explained in the following way. The present mathematics based on the theory of sets grounds is well concorded with classical ontology: the set is seen as a specific *objectification* of properties, i.e. is being interpreted as a *meta-object*, and the sign giving the set acts as its intrinsicality. For the *ontology of 'state of affairs'*, the language of the mathematic theory of categories seems to be more appropriate since today it is considered as a serious alternative to the theory of sets approach. Within this theory, the objects are defined not by an *internal* but by an *external* mode, through the system of arrows, corresponding to combinations (relations) of the given object.

It's clear that in the ontology of facts (the predicate ontology, generally), the status of things become different. At the initial stage of cognition, no individual things, in their usual sense, occur but there are indefinite objects – *quasi-things* – interacting with each other that can be presented in the mode of fussy sets. While the facts are accumulated in the process of experience the borders of the sets will be specified due to class division and 'intersections' of the one-type facts, and, at a stage, they will be detailed so that *quasi-things* will turn into ordinary, habitual for us – individual – things.

Let's explain the aforesaid on an example with a hammer. Under prop. 1.1 of the *Tractatus*, initially we have no *'hammer' thing* (resp. *notion* of a hammer), we have only a fact of 'hammering in *something with the help of something*' which can be described by the «*A–x–B*» sentence. Here, the *hammer* is correlated with the active component of the fact '*that–with–which–is–hammered–in'*, this function will correspond to, for instance, the following

[3] Here, we don't distinguish the ontology connected with the language of verbs and the ontology of facts (resp. sentences) that can be identifies. Each sentence states this or that fact of the action done. Though, this do not exclude distinguishing the ontology of actions as a separate type of ontology.

[4] For the first time, this kind of interpretation was proposed on the forum '*The World of Tractatus'*, as well as in the works [Katrechko, 1999; 2002].

set — {a hammer, a stone, a roll of paper, a vase…} consisting of objects that potentially can be involved into the act (fact) of hammering something in[5]. But after we've tried to hammer a 'nail' into a wall with our hammer[6], i.e. check whether the fact is true, it will occur that the roll of paper is torn out and the crystal vase is broken. That's why, at the second stage, the roll of paper and the vase are excluded, and the *hammer* corresponds to a narrower set — {a hammer, a stone}. And if we try to hammer a *nail* into a concrete wall the stone as one of possible candidates to become a *hammer* will not be able to execute the *hammer*–function — as a result, it must be excluded from the initial set. Hence, in the process of accumulating facts quasi–*hamme*r will be gradually redefined which means its 'turning' into an ordinary tool – a *thing* – an *individual hammer.*

At a language level the described procedure corresponds to accumulation of facts, as is «A_1–*x*–*B*», «A_2–*y*–*C*», «A_3–*z*–*D*»…, and developing on the quasi-thing *A* will, at first, correspond to A_1, then to the 'intersection' $A_1 \cap A_2$, then to $(A_1 \cap A_2) \cap A_3$, etc. So, the indefinite character of the quasi-thing in the *functional* ontology of the *Tractatus* indicates the possibility for further specification while Aristotelian things predetermined by its essences logically stay *all the same*.

Specificity of the ontology of the *Tractatus* can be expressed clearly enough by the metaphor which belongs to John Wheeler, a prominent physicist and theorist of the XX century. He suggests two variants of a game in '20 questions'. The first variant corresponding to standard ontology, gives the *thing* in advance and we, by answering 20 questions in the mode of constructing an appropriate classification tree, have to guess what the given thing was. In the second variant corresponding to the *world of the Tractatus*, no thing is given but, since the *answers* (resp. physical experiments) to consequently asked *questions* must coordinate with each other, the 'totality' of answers (resp. *Sachverhalt*) *gives* the required thing so, the inquirer can also 'guess' and, to be more exact, constitute the initially indefinite thing (though, if the sequence of questions changes the required thing might also change). In this sense, the ontology of the *Tractatus* corresponds not only to the *logic* but also to the quantum mechanic picture of the world with its postulate on the importance of the *observer* in *cognition*.

Literature

Katrechko, Serguei 1999 *Wave Ontology as the Forth Type of Ontology* /Proceedings of the 2nd Russian Philosoph. Congres, Ekaterinbourg.

Katrechko, Serguei 2002 *Functional Ontology of the Tractatus Logico-Philosophical* /Proceedings of the 3rd Russian Philosoph. Congress, Rostov/Don.

Kozlova, Maria 1995 On the Translation of Philosophical Works of Wittgenstein /Journal Publ., «Put'» No.8, 1995.

Wittgenstein, Ludwig 1994 *Tractatus Logico-Philosophical* (trans. by M. Kozlova) //Ibid. Philosophical Works P. 1, Moscow, Gnozis.

[5] Within the ontology of things we can, in a general case, randomly give the name of *hammer* to any of those objects; in the ontology of facts, a *thing* is given primarily through its *function*.

[6] It's clear that the *nail* is also a quasi-thing (to accent that we use quotation marks) since it's also defined through its function as *something–that–is–being–hammered–in*, but we just omit it here to make the story easier.

How do Moral Principles Figure in Moral Judgement? A Wittgensteinian Contribution to the Particularism Debate

Matthias Kiesselbach, Potsdam, Germany

1. Introduction: What is moral deliberation?

One of the key debates in current moral philosophy focuses on the role of moral principles in moral deliberation. Among the many opinions on the table, we find the theses of Universal Weak Particularism (UWP) and Universal Weak Generalism (UWG), which can be formulated as follows:

> (UWP) Generally, the application of moral principles is *not sufficient* for correct moral judgement.
> (UWG) Generally, the application of moral principles is *necessary* for correct moral judgement.

Obviously, these theses are mutually consistent. Moreover, both are conclusions of strong arguments: (UWP) is inductively supported by the fact that, so far, for every candidate of a suitably general and non-trivial moral principle, it has been possible to devise a scenario in which the principle's strict application would strike us as simply wrong. This is true for both *all out* and *pro tanto* principles (see Dancy 2004). (UWG) is supported by the fact that our aim of consistency in ethical learning, debate and judgement is not just a piece of ideology, but an actually attainable goal. Consistency between particular moral judgements, however, is nothing but the existence of principled relations among them. If these arguments are successful, we have good reason to accept both (UWP) and (UWG).

However, the combination of (UWP) and (UWG) does not seem to appeal to many commentators. Their reservation is, I think, due to the thought that we lack a theory of moral deliberation which implies both theses at once. What *are* moral principles, they ask, if moral judgement cannot be *reduced* to their application, and yet *depends* on the latter? In this paper, I want to argue that the work of the later (and latest) Ludwig Wittgenstein gives rise to an interesting and plausible answer to this question. It revolves around the ideas that moral principles can be interpreted as grammatical propositions, and that moral problems can be interpreted as instances of grammatical inconsistency and, hence, as occasions for grammatical revision. Moral judgement, on this view, is a matter of following grammar, but it is also a matter of adequately revising it in the face of grammatical tension.

2. Grammatical statements, grammatical tension and grammatical evolution in Wittgenstein's work

We are surely warranted to take seriously Wittgenstein's insistence that his project is one of philosophical *therapy,* aiming to free us from our urge to philosophise by unmasking our seemingly deep metaphysical ideas as mere grammatical confusions. However, in order to be able to read Wittgenstein in this way, we cannot help but ascribe to him a certain number of theoretical commitments regarding the workings of language. In this section, I want to review, as quickly as possible, key elements of Wittgenstein's mature conception of language, and to show that they comprise ideas of *grammatical tension* and *grammatical evolution*.

Wittgenstein's return to philosophy in 1929 marks a radicalisation of the view that natural language is best analysed as a practical calculus embedded in and continuous with non-linguistic practice. While the *Tractatus* still entertained the idea that some (namely the “atomic”) propositions stand in isomorphic relations with aspects of the world, the later Wittgenstein thinks of the calculus of language as fully autonomous. All utterances are now conceived as practical manoeuvres, connected via rules with other such manoeuvres as well as with non-linguistic phenomena and doings in their vicinity. On this view, all talk of “meaning” or “content” is just a way of discussing the role which an expression plays within the practical calculus of language.

This idea poses an obvious threat to the distinction between analytic and empirical content. Traditionally, the meaning of a proposition was thought to be a two-component object. There was the empirical component on the one hand, and the analytical (logical, conceptual) component on the other. With the idea that the meaning of an expression is exhausted by the logical or, as Wittgenstein has it: *internal* (TLP 4.125ff., 5.131, 5.2ff.) role within the calculus, it becomes an open question how empirical content is at all possible, or what it would amount to.

Moreover, in attacking the traditional analytic-empirical distinction, Wittgenstein's move threatens our everyday practice of distinguishing between *misunderstanding* and *disagreement*. If communication, as Wittgenstein writes, depends on “agreement not only in definitions but also (queer as this may sound) in judgements” (PI 242), then it seems that every time speakers diverge in their propositional judgements, they turn out to play different games and thus *talk past each other.* Can *this* be true?

While many thinkers have taken this threat as a pure and simple *reductio ad absurdum,* Wittgenstein holds fast to his interpretation of language as a practical calculus and looks, in his later writings, for a pragmatic way to re-erect the traditional distinctions in question. Wittgenstein's eventual solution centres around the idea that if philosophers took into account ordinary speaker's actual *employment* of the calculus of language, they would soon notice that speakers do not just *draw* on its rules, but constantly *develop* them further. They are always, he thinks, in the business of coining *new* linguistic manoeuvres, such as new propositions. Of course, it has long been known that our language is compositional. i.e. that it comprises sub-propositional components (such as *concepts*) which can be regrouped to form new, yet immediately understandable, sentences (and other utterances). But since concept rules are, on a calculus account of language, bound up with proposition rules, this does not show how empirical content or the possibility of proper disagreement comes into the picture. Wittgenstein's idea, now, is that we can use propositions to *alter* the rules governing concepts – i.e. their meanings – and *thus* convey empirical content.

To give a very simple example: although the meaning of a concept like “dog” is fully determined by true propositions like “Every dog is a mammal”, “Dogs don't lay eggs” (and so on), we *can,* in a novel proposition, bring a new predicate to bear on “dog”. If we do this, we propose to (slightly) alter the concept rules (meanings) of both “dog” and the new predicate. *This way, we pass on new information about both.* Since in a way, only *“new”* propositions are interesting, it makes sense to say that every interesting proposition brings with it new rules for the use of concepts. However, Wittgenstein makes clear that there is also a use for “old” propositions. “Old” propositions clarify how already-established concepts are used and thus serve as interpretation guidelines for new propositions drawing on these concepts. “Every dog is a mammal” is a good example of this. To put the distinction in Wittgenstein's own words: a generally accepted proposition “is removed from the traffic. It is so to speak shunted onto an unused siding. / Now it gives our way of looking at things, and our researches, their form.” (OC 210f., see also 96ff.)

It is *here* that the mature Wittgenstein re-introduces the distinction between substantive (empirical) content and logical (conceptual) rules, the latter now being called *grammar.* Grammar is comprised of rules of concept use as established in “old” (or “hardened”, OC 96) sentences, while substantive content resides in the proposed rules of concept use displayed in “new” (or “fluid”, OC 96) sentences. I have attached quotes to the terms “new” and “old”, because these terms are relative to particular conversations. A proposition can be long accepted (“hard”) in some contexts, but strikingly novel (“fluid”) in others. Wittgenstein, keenly aware of this fact, confirms that

> Sentences are often used on the borderline between logic and the empirical, so that their meaning changes back and forth and they count now as expressions of norms, now as expressions of experience. (For it is certainly not an accompanying mental phenomenon ... but the use, which distinguishes the logical proposition from the empirical one.) (RC I:32, see also III:19, OC 309)

Here, then, we see Wittgenstein's way of re-erecting the analytical-empirical distinction within a practical calculus account of language. From here, of course, it is not difficult also to re-erect the distinction between misunderstanding and disagreement: if two speakers diverge with respect to a proposition which we (the interpreters) take to be a piece of grammar, we call their divergence a misunderstanding. If the proposition in question is interpreted (by us) as a proposal of new concept rules, we call their divergence a disagreement.

Importantly, the sketched conception of language is dynamic. It holds that once a proposition is accepted (within a particular conversation), every re-iteration will be a grammatical utterance. The account thus includes a commitment to the *evolution* of language. I now want to stress that according to Wittgenstein, language does not always evolve smoothly. There are situations in which a new empirical proposition involves a *violation* of a piece of grammar – without thereby being rendered senseless. To see what I have in mind, consider again the idea that communication rests on “agreement in judgements” (PI 242). Wittgenstein's clearest example of this is colour discourse (see RC I:66, III:42, III:86ff., III:94, III:127). Clearly, when someone claims to have seen a patch of “bluish orange” (RC III:94), we would conclude that either the speaker is crazy, or that her way of speaking is in need of translation into *our* vocabulary (perhaps colour vocabulary, but perhaps she does not speak about colour at all). “There is, after all,” says Wittgenstein, “no commonly accepted criterion for what is a colour, unless it is one of our colours.” (RC III:42) And yet, Wittgenstein insists (in many passages), there *are* possible cases in which we *would* allow for *different* colours.

> It is quite possible that, under certain circumstances, we would say that people know colours that we don't know, but we are not forced to say this... (RC III:127)

Wittgenstein goes on to supply two analogies.

> This is like the case in which we speak of infra-red 'light'; there is a good reason for doing it, but we can also call it a misuse. And something similar is true with my concept 'having a pain in someone else's body'. (RC III:127)

From this last passage, we can glean an implicit theory of grammatical evolution through grammatical tension. To see this, take the infra-red case. We can easily imagine two opposing factions who argue as follows: “Light makes objects visible. Infra-red does not make objects visible. Therefore it is not light.” versus “Light is the kind of radiation which helps us navigate and is processed by the eyes. This is the case with infra-red (if we use night-sight devices). Therefore, this radiation comes under the concept of light.” The important point to notice is that we have, here, two premises which are clearly taken to reflect the grammar of shared language, yet which, along with uncontroversial minor premises, come into conflict with one another in the face of the invention of infra-red radiation. If this description is correct, then we have an example of a situation in which two sets of grammatical norms turn out to be such that following them beyond a certain point leads to conflict. This conflict demands a grammatical revision *in the form of a new empirical judgement*. Since this amounts to a change of the game of language, every proposition uttered or written before the revision must be carefully tested and, if necessary, translated into the new language.

3. Moral principles as grammatical norms, moral problems as grammatical tension

I concede that this short reconstruction of Wittgenstein's mature conception of grammar deserves both a stronger exegetical appraisal and much more discussion of its details. In this paper, however, rather than paying these debts, I want to show that its core idea is capable of providing just the account of moral deliberation we need to counter the reservation discussed above. If we interpret moral principles as grammatical rules and moral problems as grammatical tensions along the lines of Wittgenstein's infra-red example, we see how it can be the case that moral judgement relies both on moral principles *and* on the capacity to make sensible revisions in the face of practical conflict, making both (UWP) and (UWG) true. The idea is that moral judgements follow norms of grammar just as closely as in colour discourse, only that in moral discourse, they are less settled and less harmonic.

To see that this interpretation is not just a wild stipulation, consider that grammatical norms do not usually come as traditionally analytic propositions, like “A bachelor is unmarried” or (to take a moral example) “Justice is to give to each person her due”. On the contrary, *every* proposition can serve, once established and accepted as true, as a reminder of a piece of grammar. If this is true for all propositions, it is clearly true for the following remarks:

"A promise must be kept", "A promise must be kept, unless this would involve the breach of a right", "If a proposition constitutes a promise, that counts in favour of doing the act to which it refers." These, of course, are paradigm examples of moral principles.

In the face of the particularist insistence that we can always devise scenarios in which following a principle like these turns out to be morally objectionable, we can now lean back. If, for example, the mentioned promise turns out to have been given under torture, we can *agree* that on this condition, the fact that some utterance constitutes a promise counts *against* committing the act in question, making even the weakest of the three principles (the *pro tanto* principle) false. The important point to notice, however, is that we can take this situation as one of *grammatical tension* analogous to the case in which the old grammar surrounding the concept "light" was confronted with the new realities of infra-red. In other words, we can see in the case an occasion for a controlled revision of the grammar surrounding the concept "promise", i.e. a partial revision of its very meaning and thus of our language game. On this view, both (UWG) and (UWP) come out true. Every judgement involves following norms; but some judgements necessarily involve re-developing them pragmatically. I want to submit to your consideration the thesis that this is a suitable interpretation for all moral problems.

The main attraction of this view, besides yielding a plausible account of tragic choices, is that it tells a story about moral discourse which very closely parallels an emerging consensus about science, according to which the body of scientific knowledge and the meanings of scientific terms evolve together.

Literature

Baker, G.P. and Hacker, P.M.S. 1985 *Wittgenstein. Rules, Grammar and Necessity,* Oxford: Blackwell

Dancy, Jonathan 2004 *Ethics without Principles,* Oxford: Oxford University Press

Wittgenstein, Ludwig 2002 *Tractatus Logico-Philosophicus*, London: Routledge [TLP]

Wittgenstein, Ludwig 2001 *Philosophical Investigations,* Oxford: Blackwell [PI]

Wittgenstein, Ludwig 1972 *On Certainty,* New York: Harper & Row [OC]

Wittgenstein, Ludwig 1978 *Remarks on Colour,* Oxford: Blackwell [RC]

“Downward Causation”: Emergent, Reducible or Non-Existent?

Peter P. Kirschenmann, Amsterdam, The Netherlands

1. Introduction

It is common to view reality as a hierarchy of levels, such as the physical, biological, psychological level. Entities (plus their properties) at higher levels (of “organization” or “complexity”) consist of entities of lower levels, but are supposed to be “emergent” – form “wholes that are more than the sum of their parts” – and, possibly, have “downward causal” influences on their parts.

All these views, often not well-articulated, are contested. For hard-headed (eliminative) materialists or physicalists, there is no emergence, let alone “downward causation”. For reductionists, they are “nothing but” material processes. Emergentists often rest content with arguing for their *possibility*. “Non-reductive physicalists” recognize emergent phenomena, but insist on their being physically based.

I myself think that the occurrence of a hawk catching a mouse *is* a macroscopic emergent phenomenon, though I doubt that it involves “downward *causation*”. I used to be a fan of layered ontologies (e.g. Nicolai Hartmann’s). Yet lately, I think that we should avoid as much as possible imposing a level structure on reality. Surely, a distinction between macrolevel and microlevel often is very sensible, but it can misfire.

I shall discuss two original views as well as the muddled conceptualization and terminology, analyze two examples, comment on the “causal exclusion argument”, and conclude with a computer analogy.

2. The Muddle of ‘Downward Causation’

Donald T. Campbell, in 1974, first used the expression ‘downward causation’ (cf. Hulswit 2006, 266), for the view that “all processes at the lower level of a hierarchy are restrained by and act in conformity to the laws of the higher levels”. His (biological) hierarchy ran from molecules and cells up to populations and evolution. He noted himself that the expression was at odds with our usual concept of efficient causation: his higher-level “laws” selectively restrain lower-level processes, unlike events causing other events.

Around that time, Roger W. Sperry, after his split-brain investigations, started defending his “emergent interactionism” (cf. Emmeche et al. 2000, Ripley 1984, Hulswit 2006: 269). To clarify the “form of control” that conscious phenomena exert over neural events he used the examples of a wheel running downhill and an eddy in a stream. The movement and fate of the constituent molecules are “determined very largely by the holistic properties of the whole wheel” or the whole eddy, though without any change in the lower-level molecular laws. Similarly, the component parts of “an excitatory neural process are carried along and thus controlled by dynamic properties of the whole system”, namely by unitary mental experiences. Sperry was not quite happy with the term ‘interactionism’, as it smacked too much of a Cartesian dualism.

What ‘downward causation’ or the downward part of ‘interaction’ is for these authors remains unclear. Meanwhile, the terminological diversity has been exploding. Thus, the higher level has been said to cause inasmuch as it restrains, constrains, controls, organizes, structures, determines, governs, delimits, bounds, entrains, enslaves, harnesses or selects the lower-level phenomena (cf. Hulswit 2006: 279f.), while said to emerge from, cannot be explained by, reduced to, or predicted from, them and their laws. Since several of those “causal influences” hardly are examples of efficient causation, some regard them as Aristotelian formal or final (functional) causes. The most general and neutral term, though in need of appropriate qualifications, is ‘determines’. Another disconcerting variation concerns the relata of the alleged downward causality: entities, substances, events, processes, states, types or instantiations of properties, patterns, structures, laws or regularities.

No doubt, novel things have *diachronically* “emerged” during the evolution of the universe: stars, heavy elements, planets, life, consciousness. We focus here on *synchronic* emergence: “higher levels of reality” of some permanence. Crucial in identifying emergent entities or systems, I think, are their significant, possibly lawful “horizontal” interactions with entities at the “same level”, rather than “downward” influences.

A useful distinction is between three possible meanings of ‘downward causation’ according to their *strength* (cf. Hulswit 2006: 280, also Emmeche et al. 2000), all concerning an active system constituted by a set of active elements:

> *weak* – “downward explanation”: the behavior of the elements cannot adequately be *explained* without reference to the system;
> *medium* – “downward determination”: it is partly *determined* by the system;
> *strong* – “downward causation”: it is partly *brought about* by the system.

3. Examples – Analyzed

The most intriguing questions of emergence concern life and mind. Yet, phenomena of the non-living world have raised similar and more easily analyzable questions. They often are supposed to attest to a continuity of emergence and downward causation through all layers of reality (e.g. Rockwell 1998). I have my doubts.

Sperry’s wheel is an example in point. The macroscopic properties of the wheel, together with gravity, slope etc., determine the movement of the wheel and thus *a fortiori* that of its constituent molecules. Its *stable circular shape*, due to the cohesive arrangement of its constituents, is an emergent property – for not being a property of the constituents – in an innocent sense, a “*structural* emergence. It can thus be “explained”, without getting *reduced* to properties and arrangement of molecules. It is a factor in macroscopic causal relations, as when the wheel should bump into a wall. Yet, *pace* Sperry, there is no “interaction” between the wheel as a whole and its constituents, let alone its “downwardly causing” their movement, except in the weak sense of downward explanation. Note, *pace* Campbell, that the laws of mechanics, holding for the wheel, need not be considered as emergent, as they equally hold for the free movement of molecules or stable collections thereof.

There are many more macroscopic, *structurally* emergent properties, like fluidity, viscosity, solidity which can be analyzed in a corresponding manner.

A quite different, well-worn example is the Bénard instability. It concerns an *open* system, a fluid heated from below, far from thermodynamical equilibrium. At a critical temperature difference between bottom and top, heat conduction changes into convection, in the form of cylindrical rolls, with the top closed, or in that of Bénard cells, with the top open. These patterns of movement are emergent in a sense that goes beyond structural emergence. Some say (cf. Rockwell 1998) that such a pattern is created by the coordination of the motions of molecules, which in turn, “downwardly”, influences their behavior, and call this a case of ‘circular causality’.

Yet, there is at best some downward determination: groups of molecules, tending to move (first) upwards, are coaxing, and are being coaxed by, neighboring molecules into the emerging motion pattern and kept in it. As possibility and actuality, this pattern is a determining circumstance, though not an extra force. One should not say that the “system as a whole” does much in determining the behavior of its “parts”, since the decisive, but “outside”, driving forces are the heat flow and gravity, acting under the geometrical conditions of the pan of fluid and the possible motions of the molecules. A hard-headed reductionist might want to take the system plus total environment as the real “whole” to be analyzed in terms of processes of their micro-constituents.

The cylindrical rolls can either turn clockwise or counterclockwise. There is a “bifurcation point”: it is up to chance which way they turn. This spontaneous, unpredictable “choice”, at the point of instability, can be taken to enhance the emergent character of the resulting pattern.

What about the pattern being a significant factor at its “own level” of macroscopic causality? It is responsible for the rapid heat flow; it can be photographed and printed as illustration.

The Bénard instability represents a nonlinear dynamic system. The emergence of the patterns constitutes a symmetry break. Such systems are dissipative structures, since energy is not conserved within them. They also count as examples of “self-organization”, mostly a misnomer, inasmuch as the driving forces are heteronomous. The behavior of nonlinear dynamic systems can lie in chaotic regimes, but also in regimes representing various kinds of “attractors”, comparable to the patterns in Bénard cells.

Organisms also are open systems, but differ from such non-living examples by their enormous degree of internal regulation, coordination and integration of all their constituent processes. They possess emergent features, such as multiplication, growth, self-repair etc., which are determining factors in all kinds of causal relationships, as when we catch a cold or when moles disfigure our lawns. Yet, I will not discuss claims that their lives might manifest a special kind of “downward causation” and turn to more abstract considerations.

4. Supervenience and Critique

A decisive argument for emergent properties is the multiple realizability argument. The standard example is the mental state of being in pain, which can be realized by different physical states in the same or different persons. This provides an argument *against* (type) identity theories of mind and brain, reductive or eliminative accounts of the mind and *for* the now widespread view of “non-reductive physicalism”. Similarly, biological functions can be physically realized in different ways in different organisms.

The discussion of the relation between properties at different “levels” has been dominated by the rather formal notion of supervenience, which accommodates the general idea of multiple realizability, but is chiefly used in the philosophy of mind.

One definition is: “A set of properties *A* supervenes upon another set *B* just in case no two things can differ with respect to *A*-properties without also differing with respect to their *B*-properties.” (McLaughlin, Bennett 2005). It states a dependency, not further specified. Important versions of the notion differ in the kind of necessity attributed to this dependency.

Jaegwon Kim, a prime elaborator of the notion, used it in his “causal exclusion argument” against “downward causation”, or, to show that “non-reductive physicalism” is self-contradictory (cf. Rockwell 1998). Consider the following schema:

M causes M*
P causes P*

Here, a mental event (instantiation of, or change in, a mental property) M is assumed as causing another mental event M*. Yet, M supervenes on (is realized by) a physical (neurophysiologic) event P, which causes P*, the physical realization of M*. Kim argued, put simply, that the mental layer would do no causal work: P causes P* all by itself; M in no way “downwardly causes” P*; the mental layer is at best epiphenomenal, if not non-existent.

In its strongest form, the argument presupposes the “causal closure” of the physical world. Yet, if this was just the world of the most fundamental physical entities, then everything else, not just the mental realm, would be epiphenomenal. The argument also assumes causation to be the only kind of determination.

What the supervenience approach totally ignores, as Mark Bickhard with Donald Campbell (2000) rightly pointed out, are the external relations of the systems concerned, which, as we saw, are especially important in the case of open systems. More generally, they criticize it for still assuming a basic level of fundamental particles, thus a metaphysics of substances, whereas modern physics forces us to adopt a metaphysics of fields. Fields are continuously in process, which is “inherently and necessarily organized” or patterned. “So, delegitimating process organization as a potential locus of emergence renders *all* reality epiphenomenal”. This absurdity amounts to an argument for the reality of all patterns of processes, also for the plausibility of the emergent reality of mind.

Non-trivial emergence, for the authors, is “emergent (novel) causality”, which (in contrast to my view) will, thus as a criterion, “necessarily involve downward causality”. Unfortunately, they fail to articulate some alternative notion of pattern causality. The kinds of “downward causation” they survey are all cases of *constraints*, thus at best cases of “downward” *determination*.

I also think that especially living beings, in which physical constituents are continuously replaced, are patterns of coordination and integration of processes rather than substances. Yet, for understanding them, it is inessential whether one considers molecules as entities or, in their turn again, as patterned field processes.

There are a number of alternative proposals concerning emergence, supervenience and “downward causation”, which I cannot take up here.

5. Yet Another Computer Analogy

Analogies with (information processes in) computers have been very conspicuous in the philosophy of mind. They can be instructive for our present issues. Software – programs, essentially algorithms – can be multiply realized in diverse computers. No one would call the algorithms ‘emergent’; they get artificially, thus contingently, imposed on the hardware, whereas in natural systems one supposes some lawful connections.

A computer runs through a series of physical states which is isomorphous to the logical steps in the algorithm (cf. Jongeling 1997). Clearly, the algorithmic program *determines* the physical operation of the computer, but we need not call it a ‘*downward* determination’. *How* a computer, say, can do a calculation, must be reductively explained in terms of electronics. The logics of the program, however, is apriori establishable, not even in need of being reduced to electronics.

Without pretending to solve the riddle of the mind, I want to draw some analogies. Our experiencing and thinking depends on our “hardware”: when we are tired, we “cannot think straight”. Furthermore, thoughts or experiences do not causally follow upon each other, but rather in dependence of their content. For instance, when making plans, our thinking proceeds by association or goal-directed deliberation, albeit not algorithmically. Still, I should say that the “logics” of such thought processes can be understood in themselves and need not be reduced to brain processes. *How* precisely they are realized in the brain, is the riddle. Yet, given that the relation between thoughts is not causal, the inference to a “downward” causal influence cannot even get started.

Thoughts, no doubt, play a role in determining (changes of) states of our brain and body, as well as, when plans get realized, meaningful macroscopic alterations in the physical world. As in the case of the vaguely comparable Bénard cells, I conclude that we should speak here at most of ‘downward *determination*’ and certainly not of ‘downward *causation*’.

Literature

Anderson, P.B., Emmeche, C., Finnemannn, N.O. and Christiansen, P.V. (eds.) 2000 *Downward Causation. Minds, Bodies, Matter*, Aarhus: Aarhus University Press.

Bickhard, M.H. with Campbell, R.L. 2000 “Emergence”, in: Andersen et al. (eds.) 2000, 322-348.

Emmeche, C., Køppe, S. and Stjernfelt, F. 2000 “Levels, Emergence, and Three Versions of Downward Causation”, in: Andersen et al. 2000, 13-34.

Hulswit, M. 2006 “How Causal is Downward Causation?”, *Journal for the General Philosophy of Science* 36, 261-287.

Jongeling, B. 1997 “Wat is reductionisme?”, in: W.B. Drees (red.), *De mens: meer dan materie? Religie en reductionisme*, Kampen: Kok, 38-54.

McLaughlin, B., Bennett, K. 2005 “Supervenience”,

http://plato.stanford.edu/entries/supervenience/

Ripley, C. 1984 “Sperry’s Concept of Consciousness”, *Inquiry* 27, 399-423.

Rockwell, T. 1998 “A Defense of Emergent Downward Causation”

http://users.california.com/~mcmf/causeweb.html

On Game-theoretic Conceptualizations in Logic

Maciej Tadeusz Kłeczek, Nottingham, England, UK

Game-theory is a rich mathematical framework formalizing real-life and intuitive concepts. It comes with a set of slogans such as: winning, losing, dynamics, interaction, process, choice. The basic ontology is that of players acting according to certain definitiary rules of the relevant game. How one reacts to the merging of game-theory and logical concepts depends on one's philosophical assumptions. Some philosophers of logic view with suspicion the general anthropomorphic flavor and procedural elements involved.

At least two different levels of analysis are present in the literature on logic games. Certainly, games are processes and can be described by process theories, such as modal logic with some form of bisimulation as the invariance relation[1]. However my concern in this paper is more classical and focuses, after preliminary exposition of paradigmatic logic games, on the interaction of properties of semantic games with '~'.

On the standard Tarskian account, truth in a structure is understood as a certain abstract relation holding between some particular structure and a formula (relative to some assignment if the relevant formula is an open formula). Truth and/or satisfaction conditions (1) $M \vDash \Phi[\alpha]$ are provided in a compositional manner.

The game-theoretic account of truth in a structure is given as follows:

> (1') $M \vDash^{+} \Phi[\alpha]$ if and only if there is a winning strategy for the initial Verifier (called II) in a semantic game $G(M, \Phi, \alpha)$. Falsity is defined dually:
> (2) $M \vDash^{-} \Phi[\alpha]$ iff there is a winning strategy for an initial Falsifier (called I) in a semantic game $G(M, \Phi, \alpha)$.

The definitional rules of the game of semantic evaluation are given as follows:

> (1) If Φ is an atomic formula no action is taken. II wins iff $M \vDash \Phi$; otherwise I wins.
> (2) $G(\sim\Phi, M, \alpha)$ — the game is played as on $G(\Phi, M)$ except that the roles of the players are transposed.
> (3) $G(\varphi_1 \wedge \varphi_2, M, \alpha)$ — I makes the first choice of a conjunct from $\Omega \in \{1, 2\}$. The game continues with the conjunct chosen.
> (4) $G(\varphi_1 \vee \varphi_2, M, \alpha)$ — II makes the first choice of a disjunct $\Omega \in \{1, 2\}$. The game continues with the disjunct chosen.
> (5) $G(\forall x\Phi, M, \alpha)$ — I chooses the witness individual a from $|M|$. The game continues $G(\Phi, M, \alpha \cup \{x, a\})$.
> (6) $G(\exists x\Phi, M, \alpha)$ — I makes the first choice of individual a from $|M|$. The game continues $G(\Phi, M, \alpha \cup \{x, a\})$.

Assuming the axiom of choice[2] (1) and (1') are equivalent. Proof proceeds by induction on the complexity of a formula. It is evident that game theoretic content is relevant at the level of non-atomic formulas.

It is important to note that each first-order game of semantic evaluation: (1) terminates in a finite number of moves; (2) is a two person zero-sum (the payoffs assigned to players at the terminal nodes of a game tree equal zero); (3) exemplifies perfect information, which amounts to lack of ignorance about previous moves of the opponent; (4) is non-cooperative. Taking (1), (2) and (3), it follows from general game-theory that each FO semantic game is determined: one of the players is in possession of a winning strategy. Determinancy guarantees the validity of the Law of Excluded Middle: (1) $M = \varphi \vee \sim\varphi$ if and only if I has a winning strategy in $G(M, \varphi \vee \sim\varphi)$ (2) if and only if I has a winning strategy in $G(M, \varphi)$ or $G(M, \sim\varphi)$ (3) if and only if I has a winning strategy in $G(M, \varphi)$ or II has a winning strategy in $G(M, \varphi)$. The definitional rule for negation involves switching the roles of players and winning/loosing conventions. Under perfect information this rule respects the classical semantics for negation according to which: $\sim\varphi$ is true in M if and only if φ is false.

A structure comparison can be captured by means of a certain game of perfect information. For a given vocabulary L and L-models M, N a partial injective function $f: M \rightarrow N$ is a partial isomorphism if there is an isomorphism $f^*: [\mathrm{dom}(f) \rightarrow \mathrm{rng}(f)]$, which extends f. Of course, the isomorphism f* is unique. An alternative formulation is in terms of a back and forth set. If L is a vocabulary, M and N are L models then M and N are partially isomorphic if and and only if there is a back and forth set for M and N. A back and forth set for M and N is a set $P \subseteq \mathrm{Part}(M, N)$ satisfying the two following conditions: (1) $\forall f \in P \forall a \in M \exists g \in P(f \subseteq g \text{ and } a \in \mathrm{dom}(g))$ (2) $\forall f \in P \forall b \in N \exists g \in P(f \subseteq g \text{ and } b \in \mathrm{rng}(g))$.

Conditions (1) and (2), imposed on the back and forth set, can be characterized by means of a Ehrenfucht-Fraisse game. There are two players who move alternately, choosing elements from a domain of assigned models. Analogously to semantic games, the task of players is conflicting. I tries to show structural difference, whereas II seeks structural similarities. As one can expect:

> (1) $M \cong_p N$ if and only if player II has a winning strategy in $EF_\omega(N, M)$.

A more fine-grained concept is that of a back and forth sequence of some finite length P_n: $(i \leq n)$ such that: (1) $\varnothing \neq P_0 \subseteq _{\dots} P_n \subseteq \mathrm{Part}(M, N)$ (2) $\forall f \in P_i \forall a \in M \exists b \in N \exists g \in P_{i+1}(f \cup \{a, b\}) \subseteq g$ for $i \leq n$ (3) $\forall f \in P_i \forall b \in N \exists a \in M \exists g \in P_{i+1}(f \cup \{a, b\}) \subseteq g$ for $i \leq n$. A back and forth set implies a back and forth sequence but not vice versa. All of this summarized in the following chain of equivalences:

> (1) $M \cong^{p}_{n} N$ are partially isomorphic if and only if there is back and forth sequence of length n if and only if there exists a winning strategy for II in $EF_n(M, N)$.

So far the notion of a game has been used in an informal way. Restricting attention to finite cases and assuming a set of possible actions A, a game of perfect information is a triple $G = \langle N, H, P \rangle$, where: (a) N is a set of players; (b) H is a set of finite sequences (histories) over a set

[1] Game trees can be seen as relational models. Lets M = <W, R, V> and M' = <W', R', V'> be two relational models. Bisimulation is a non-empty relation $E \subseteq WxW$ such that if wEw': (1) if w and w' satisfy the same proposition letters (2) if wRw* then there exists w*' such that w'Rw*' and w*Ew'* (3) and vice versa.

[2] The axiom of choice is needed in connection with $\forall\exists$ formulas in the direction from Tarski to GTS.

A with distinguished subset Z of terminal histories; (c) P is a function assigning a player to the elements of H\Z. A strategy is a function from the preimage of P to a set A. A winning strategy is a strategy, which leads to a win for a player irrespective of the opponent's strategy.

As was mentioned above, game-theoretic characteristics are equivalent to standard ones. This fact certainly makes them adequate. However, it induces doubt about whether game-theoretic conceptualizations introduce any genuine content into logical theory. The choice of technique seems to be arbitrary.

This explains the somewhat pessimistic but instructive remark found in Johan van Benthem's paper: "... all these games are useful didactic and heuristic tools in this area, but no significant logical results or insights so far appear to rest on them exclusively. But again this may reflect the poverty of the notion of game involved so far in current literature". [van Benthem, 1990]

It all depends on what kinds of results are expected. The games outlined above vindicate what can be called a strategic viewpoint in logical theory. Consider two graphs G = <{1, 2, 3}, V> and G' = <{1, 2, 3, 4}, V>, where V is an irreflexive and antisymetric relation. Obviously, those graph structures are non-isomorphic. There is no bijection between them. Thus I has a winning strategy in game EF_n(G, G'). Actually, I has two winning strategies. He can force a winning outcome in two or three rounds of EF_n(G, G'). We find an analogous case in semantic games. It often happens that II or I are in possession (obviously not simultaneously) of more than one winning strategy. Hence, winning strategies are more fine-grained semantic values than truth-values. Although those facts can be hardly regarded as striking results.

The situation changes when the assumption of perfect information is abandoned and lack of information about the opponent's previous moves is allowed. A winning strategy is no longer built up on all the possible moves of the opponent.

A game of imperfect information is the following tuple: G = <N, H, P, $I_{i,\ i\in N}$>. The new element is a set $I_{i,\ i\in N}$ whose elements are called information sets. Information sets induce partitions (equivalence relation) on a set of histories such that for all h, h'∈I∈I_1, A(h) = A(h'), where A(h) denotes the set of possible actions after history h. Histories belonging to particular information sets occupy the same depth on a game tree.

The difference between the semantic game of perfect as opposed to imperfect information can be seen by two styles of skolemizing branching-quantifier formulas: (1) ∃f∃gφ(x, f(x), z, g(x, z)) (2) ∃f∃gφ(x, f(x), z, g(z)). It is assumed that Skolem functions correspond to winning strategies for II. All of this was made explicit by Jaakko Hintikka who introduced IF FO logic (being a proper extension of standard FO logic) where the failure of perfect information is syntactically expressed by the new element '/' i.e. (∀x)(∃y/∀x). Allowing imperfect information in a process of semantic evaluation provides a formalisation of the idea of dependence/independence relations holding between quantifiers and attached variables.

Formulas involving slashes introduce equivalence a relation E(Φ, M) on a particular structure M. Those equivalence relations correspond to information sets. Whenever some elements or sequences[3] of elements belong to such such an equivalence class it is required that a strategy function f satisfies a uniformity condition: i.e. E(a, b) → f(a) = f(b).

Determinacy does not survive in this setting. It is possible that neither player is in possession of a uniform winning strategy when he is deprived of knowledge about all previous moves of his opponent. Thus, the inference from the lack of a winning strategy for one player to the existence of a winning strategy for the other is no longer valid. Despite the fact that the game rule for '~' is the same as it is in the classical case.

The logic of imperfect information is not closed under contradictory negation. Consequently the Law of Excluded Middle does not hold in IF logic and truth-valueless sentences are allowed. This enables a distinction between false/ non-true and non-false / true. Game-theoretic negation under imperfect information satisfies the Strong Kleene evaluation schema: (a)~(t) = f (b)~(f) = t (c)~(u) = u. More exactly LEM fails in non-FO fragment of IF logic. A canonical example of a truth-valueless sentence is: ∀x(∃y/∀x) x ≠ y, where |M| ≥ 2.

The expressive strength of IF logic, which is equal to the existential fragment of second order logic, opens the possibility of truth definitions for IF language in the signature of Peano's arithmetic in the language itself. Game-theoretic non-determinancy guarantees that the Liar Paradox will not arise.[4] Applying the standard procedure of Godel numbering and the Fixed Point Theorem to the Liar sentence λ we obtain the paradoxical looking sentence ~Tr[λ] where neither I nor II is in possession of a winning strategy.

Moreover, IF logic is strictly more expressive than FO logic[5]. However it validates (a) Downward Skolem-Lowenheim (Let M be an L-model of infinite cardinality k and let k' be a cardinal such that k > k' ≥ |L|. Then M has an elementary submodel of cardinality k') (b) Compactness (Let W be a set of IF formulas. If every finite subset of W has a model, then the entire set W has a model). Thus according to the Lindstrom Theorem its expressive strength should not exceed those of FO logic. But the crucial assumption of the Lindstrom theorem is that the logics he considers are closed under contradictory negation. Here game-theoric conceptualization shows once again its force.

In this paper the employment of game-theoretic apparatus in the core areas of logical theoretizing was presented. Certainly, the fact that the existence of a winning strategy corresponds to an assertion that such and such logical property holds is surprising. At first blush those concepts might be regarded as categorically different. All of this introduces an entirely new methodological perspective, which can, should and is pursued in a systematic way.

As emphasized earlier, varying purely game-theoretic properties of a relevant game (affecting the conditions a strategy to be a winning strategy) entails change of meaning of the most important logical constant: negation. As it is self-evident this result is far from trivial and hardly can be seen as some form of heuristics and/or didactic enterprise.

[3] Finite ordered sequence is a function f from the finite set of natural numbers to some finite set A.

[4] As it is rightly emphasized by Hintikka the problem of Liar Antinomy does not concern truth predicate in isolation but the truth predicate in interaction with negation. Consider Truth-Teller sentence: "(1) This sentence is true". It does not give rise to any kind of paradoxical conclusion.

[5] Two logics L1 and L2 are equal in the expressive strength if every class of models definable in L1 is definable in L2, and vice versa.

Literature

van Benthem, Johan 1990 “Computation versus Play as a Paradigm for Cognition”, *Acta Philosophica Fennica* 49, 236 - 251

van Benthem, Johan (eds) 1997 “Handbook of Logic and Language”, Cambridge Massachusetts: The MIT Press

Hintikka, Jaakko 1998 “The Principles of Mathematics Revisited”, Cambridge: Cambridge University Press

Vaananen, Jouko 2007 “Dependence Logic”, Cambridge: Cambridge University Press

A Metaphysically Moderate Version of Humean Supervenience

Szilárd Koczka, Miskolc, Hungary

Humean Supervenience (HS) is the doctrine that the nomic features of our world supervenes on the arrangement of basic, non-nomic properties. David Lewis's motivation in elaborating HS was to present a view which does not refer to properties and relations alien to physics, so it is extremely important for the defender of HS to make it compatible with current physics. Contrary to Lewis, philosophers known as *necessitarians* or *governists* claim that laws express necessary relations. These claims involve an implicit ontological commitment that Lewis summarizes as follows *„...there are more things in heaven and earth than physics has dreamt of."* (Lewis 1994,474).

Within a non-humean ontological framework there can be states of affairs which are metaphysically related. A non-humean approach can be plausible only if the defenders of this view can put forward a convincing argument to the point that the very notion of a natural law involves a non-humean ontology. Whether one can acquire such an argument is dubious, and as humean analyses do not require such extra entities to explain the role of laws in science, it seems that the ontologically less comitted humean view is the promising one. For the rest of the paper I will assume that there is no such argument around.

It is clear, that philosophical accounts of scientific laws appeal to metaphysical intuitions, but there is a significant difference between invoking intuitions or metaphysical theories. If one tries to analyse the notion of scientific laws in terms of his favorite metaphysical theory, the proposed analysis assumes that the background metaphysical theory is already accepted. As Lange (2000) emphasises: without a previously established metaphysical framework, this kind of approach is unilluminating. But if we keep in mind that a philosophical account of the natural laws is to be fitted with scientific practice, then it seems the weaker metaphysical commitment the philosophical account has the better. Scientists do not employ fully developed metaphysical theories, but they surely know how to work with these laws.

Formulations of Humean Supervenience

Lewis's original formulation of HS can be summarized as follows: In the actual world, everything supervenes on local qualities instantiating at certain points of space-time. There is no difference between two possible worlds without some difference in these local qualities. As Loewer (1996) points out, Lewis's motivation was to defend physicalism against philosophical challenges. Thus HS is motivated by a tendency to avoid appeal to non-physical entities. However, Lewis's (1986) own formulation postulates *local* properties which seems to contradict current theories in quantum mechanics. It seems that if our current quantum mechanics is correct, then we have to abandon the view that there are local properties at all. Lewis in his (1986,xi) admitted that his formulation – as a physicalist doctrine – can be falsified by empirical research, if it excludes the possibility of his local properties; in this sense HS is at best a contingent truth.

Loewer argues that for an acceptable doctrine we need to refine the Humean Supervenience thesis in order to make it compatible with recent physics. Loewer(1996) claims that the problem of defining the Humean Base (HB) on which everything else supervenes can be solved by appealing to the mathematical apparatus of quantum physics. According to his solution, the HB can be characterized as consisting of points of the fundamental mathematical space which is used by current quantum mechanics. Earman and Roberts (2005) have some misgivings about Loewer's version of HS. They argue that different physical theories use different mathematical spaces, and there is no uncontroversial way to determine which mathematical treatment is the fundamental one. If we try to specify the HB as the fundamental magnitudes of the best future physical theory HS – without an account of what makes a theory 'physical' – will be vacuous. It is possible that future physics postulates several kinds of entities which are alien to our current theories.

There are two possible answer to the question of HS's compatibility with recent physical theories: (i)To make HS compatible with recent physics a neo-humean philosopher have to define the HB on which everything else supervenes. As Earman and Roberts emphasize: at this point our original metaphysical problem turns to be an epistemological one. (ii)There is a possible alternative to the epistemological account that Schaffer (2008) has in his mind while he defends reductionism as a thesis about mind-and-theory independent reality. He writes: *"Causation and the laws of nature are nothing over and above the pattern of events, just like a movie is nothing over and above the sequence of frames."* (Schaffer 2008). According to his view, we do not need to define precisely what kind of properties belong to the HB since this formulation of HS does not concern the subvenient base. It is important that Schaffer does not even refer to basic (microphysical) facts. His starting point is the manifest image, so it does not matter that whether we can define basic microphysical facts with the help of our best theories. However this approach relies on the presupposition that knowledge about mind independent reality can be acquired via metaphysical inquiry of the manifest image. Behind these presuppositions is the view that we have knowledge about the natural laws. This is what I doubt. For a minimalist viewpoint if it is possible to come out with a metaphysically less committed approach of nomic notions, it seems that there are no much reason to hold such a view. Thus if there is a formulation of (i) which can handle our philosophical problems about scientific laws, we can disapprove (ii). My solution relies on a distinction between natural laws (of which we does not have knowledge) and scientific laws. I think that the debate can be meaningful only if we start with the notion of a scientific law and then try to answer the question as to what philosophically relevant implications it may have.

Natural laws or scientific laws?

Metaphysical discussions of natural laws assume that science aims to discover the *laws of nature*. Consequently, a metaphysical account of the natural laws can be accepted only if it can explain the role of the laws in scientific practice. Non-humeans argue that the notion of scientific law required by science can only be explained within a non-humean metaphysical framework. Humeans (e.g. Loewer 1996) argue that HS does explain the role of the notion of scientific laws and we do not need non-Humean assumptions. Loewer argues that notion of the natural law in scientific practice does not require the governist idea that laws have to govern events so it is right to abandon governism. However, Balashov (2002) as a non-humean can also appeal to scientific practice in order to show that humean solutions are untenable. He argues that laws can be strongly supported by empirical evidence and have explanatory power only if they are necessary. It is clear that what Balashov has in mind is the concept of mind-independent natural laws which can be discovered with the help of our science.

I think that without a sharp distinction between notions of natural laws and scientific laws, it is impossible to see the problem as clearly as we need. The concept of natural laws involves universality, so “laws of nature without universality” is a contradiction while there is no such a contradiction in the sentence that “the laws of thermal expansion is not universal”. This can be demonstrated easily: if $\Delta L = k \cdot L_0 \cdot \Delta T$ were an universal truth it should be impossible to cite situations where the relation between L_0, ΔT and ΔL falls, but there is an infinite number of provisos implicit in a statement like “without hammering the heated iron rod inward at one end ...”, so it is impossible to state a genuine law-statement. Thus scientific laws cannot be universal, while implicit in the concept of natural laws we presuppose universality. scientific laws – at least most of them – certainly fail to being equal to natural laws.

A metaphysically moderate version of Humean Supervenience

I am agnostic about the feasibility of metaphysical approaches so I will focus on the question of scientific laws instead. As I said scientific laws – or at least lots of them – fail to be universal. What more can we say about scientific laws? Earman and Roberts (2005) identify laws with mathematical treatments of certain problems. They wrote: “*In modern physics (by which we mean, at least, physics since Newton), the typical form of a problem is that of solving a differential equation (or a system of such equations) subject to certain boundary conditions, which typically include initial conditions.”* (Earman & Roberts 2005, 13). The difference between nomic and non-nomic facts is merely a methodological one. We have to distinguish boundary and initial conditions and equations while we try to solve a problem, but this difference is not ontological. This definition restricts the field of scientific laws to laws which can be formulated mathematically. Thus if we hold this view as an appropriate account of scientific laws, we cannot count many of scientific generalizations as laws; e.g. laws of biology will be non-laws according to this analysis. I do not think that this would be the right move. However, implicit in the view presented by Earman and Roberts (2005) there is something resembling to Hume's original insight, that nomic relations among events is not observable. They define the Humean base as follows: “*The Humean base at a given world is the set of non-nomic facts at that world that can be the output of a reliable, spatiotemporally finite observation or measurment procedure.”* (Earman & Roberts 2005,17).

Earman and Roberts suggest a modal characterization of HB. In the light of their analysis a fact belongs to the HB if and only if there is a nomologically possible observaiton or measurement procedure to detect it. The explanation will be circular only if we interpret nomological possibility as a mind-and-theory independent feature of our world. I prefer the alternative – ontologically less committed – interpretation that something is nomologically possible if it is compatible with our scientific laws. It is an epistemological characterization of nomological possibility and it has its own problems, but I think this is the best we can offer. There is a bit obscure notion within the definition presented by Earman and Roberts, namely reliability. To disperse this obscurity I define reliability as follows: *An observation or measurement procedure can be reliable, if it is consistent with our already accepted theories and all observable fact.* In addition, reliability has a normative feature: what is and what is not reliable depends on our current theories. This normative feature can be understood with the help of Kuhn's (1996) terminology. Even if it is incorrect to talk about “Scientific revolution”, the core idea that scientific standards can change with paradigms seems plausible enough. Thus I refine my definition: *A measurement process is reliable, if it is consistent with our already accepted theories and directly observed facts, and it fits to our current scientific standards.*

A moderate version of Humean supervenience as a minimalist theory states that we can develop predictable useful scientific laws as inference rules that belongs to certain models or systems. The models we develop are useful instruments of explanation and prediction, but we can never be in an epistemological position that enables us to declare that in scientific research we can discover the mind-and-theory independent natural laws. We have an ability to perceive relevant patterns of non-nomic facts, there by creating instrumentally useful models or systems, rather than the very natural laws. The Humean Base consists of facts that we can acquire directly or with the help of any reliable observational or measurement procedure. One can argue, that the moderate version of HS is not the same as that of Lewis elaborated, and it is obviously true. Why call this thesis Humean Supervenience after all? Lewis dedicated his thesis as “*in honor of the greater denier of necessary connection”*(Lewis 1986,ix), however it is possible to come up with a supervenience thesis about laws which can express Hume's original – epistemological – motivations about nomic relations.

Literature

Balashov, Yuri 2002 "What is a Law of Nature? The Broken-Symmetry Story.", *The Southern Journal of Philosophy,* 40,459-473.

Cartwright, Nancy 1999 *The Dappled World: A Study of the Boundaries of Science.* Cambridge and New York: Cambridge University Press.

Earman, John & Roberts, John T. 2005 "Contact with the Nomic: A Challenge for deniers of Humean Supervenience about Laws of Nature Part I.", *Philosophy and Phenomenological Research* 71, 1-22.

Earman, John & Roberts, John T. 2005 "Contact with the Nomic: A Challenge for deniers of Humean Supervenience about Laws of Nature Part II.", *Philosophy and Phenomenological Research* 71, 253-286.

Kuhn, Thomas 1996 *The Structure of Scientific Revolutions.* Chicago: University of Chicago Press

Lange, Marc 1993 'Natural Laws and the Problem of Provisos", *Erkenntnis* 38, 233-248.

Lange, Marc 2000 *Natural Laws in Scientific Practice.* New York: Oxford University Press.

Lewis, David 1986 *Philosophical Papers II.* New York: Oxford University Press.

Lewis, David 1994 "Humean Supervenience Debugged", *Mind* 412, 473-490.

Loewer, Barry 1996 "Humean Supervenience" *Philosophical Topics* 101-127.

Schaffer, Jonathan 2008 "Causation and Laws of Nature: Reductionism", in J. Hawthorne, T. Sider, and D. Zimmerman, (eds.), *Contemporary Debates in Metaphysics*, Oxford: Basil Blackwell.

“In der Frage liegt ein Fehler” – Überlegungen zu *Philosophische Untersuchungen* (PU) 189A

Wilhelm Krüger, Bergen, Norway

1. “ ‘Aber sind die Übergänge durch die [...] Formel *nicht* bestimmt? ’ ” (PU 189A1).

In PU 185 bis PU 190 diskutiert Wittgenstein, inwiefern die Übergänge algebraischer Formeln bestimmt sind oder nicht. Zum Ausgangspunkt nimmt er dabei ein Verständigungsproblem über die Verwendung des Ausdrucks “+2" zwischen einem Lehrer A und einem in PU 143 bis PU 150 bereits mit der Grundzahlenreihe vertraut gemachten Schüler B. Diesem B ist jetzt auch mit guten Worten nicht verständlich zu machen, dass seine Verwendung von “+2" mit unserer Verwendungsweise dieses Ausdrucks nicht übereinstimmt. Ihn dazu zu bringen, die Reihe so fortzusetzen, "wie wir es tun" (PU 145B), gelingt dieses Mal nicht (vgl. PU 185). Diese Ausführungen Wittgensteins zum Verhalten des Schülers laufen in PU 186 erstens darauf hinaus, zu hinterfragen, was für den einzelnen zur richtigen Befolgen eines Befehls nötig ist (vgl. PU 186A1). Wittgenstein weist das intuitive Befolgen von Regeln hier zuück.[1]. Zweitens geht es jetzt darum, was entscheidend dafür ist, dass man mit Bezug auf einen Befehl (“+2!") von einer *richtigen Befolgung* sprechen kann. Der Übergang vom Verstehen des Befehls durch den Schüler zum Meinen des Befehls durch den Lehrer ist damit gemacht.

In PU 187 und PU 188[2] verdeutlicht Wittgenstein, dass nicht nur für das Verstehen (des Schülers), sondern auch für das Meinen der Übergänge (durch einen Lehrer) seelische Vorgänge u.ä. keine Rolle spielen. Wer behauptet, er habe, während er den Befehl gab, schon alle Übergänge, der von ihm gemeinten Reihe gewusst, kann damit auf jeden Fall nicht ernsthaft behaupten zu dieser Zeit an alle Übergänge gedacht zu haben. Behauptet wird damit vielmehr, so Wittgenstein, dass man auf Befragen in bestimmter Weise reagiert hätte.[3] Durch einen Vergleich bringt Wittgenstein zum Ausdruck, dass man von dem, der das und das meint, erwarten kann, dass dieser in einer bestimmten Situation in bestimmter Weise reagiert. - Das, was in PU 186A9 noch als “beinahe richtiger” bezeichnet wird, erweist sich schon vor diesem Hintergrund als (überwiegend) falsch. (Vgl. a. PU 219D.) - In PU 188 finden wir eine Art “mythologische Beschreibung” (PU 221A1) des Gebrauchs von “+2" ähnlich wie Wittgenstein dies später in PU 221A1 thematisiert. Da für denjenigen, der mit dem Ausdruck “+2" in der in PU 187 beschriebenen Weise vertraut ist, die Übergänge festliegen, unabhängig davon, ob *er* “sie schriftlich mündlich, oder in Gedanken” (PU 188B1) macht, erhält er den irrigen Eindruck, sie seien “*eigentlich* schon gemacht” (PU 188B1) und “in einer *einzigartigen* Weise vorausbestimmt” (PU 188B2). Wittgenstein weist diese "Idee" (PU 187) sowohl durch den Hinweis, dass der algebraische Ausdruck ("+2") alleine die Übergänge nicht festlegt (vgl. PU 185)[4], als auch seine Reduktio ad absurdum (an *alle* gemeinten Übergänge kann man nicht denken) und den Hinweis auf den richtigen Gebrauch des Satzes (PU 187A4) zurück. Nachdem deutlich ist, wodurch die Übergänge algebraischer Ausdrücke *nicht* bestimmt sind, setzt Wittgenstein mit der Frage aus PU 189A fort.

Wittgenstein hat die Frage bewusst formuliert. Und auch seine Antwort darauf ist, wörtlich zu nehmen: Der Fehler, der sich in der Frage befindet, besteht in dem, was die Frage auslässt, um sie beantworten zu können. Es geht hier nicht einfach darum, die bereits widerlegte Vorstellung eines "geistigen Mechanismus" (PU 689C) des Meinens noch einmal zurückzuweisen.[5] Vielmehr soll jetzt deutlich werden: Man kann nicht sinnvoll nach der Bestimmung eines Ausdrucks fragen, ohne dabei auf eine Gruppe zu referieren, in der der Ausdruck gebraucht wird. "Bestimmt" nennen wir einen Ausdruck nur dort, wo es Menschen gibt, die ihn in gleicher Weise gebrauchen, d.h. dort, wo alle auf "der gleichen Stufe den gleichen Übergang machen" (PU 189B). Wittgenstein schlägt damit zwei Fliegen mit einer Klappe: wenn alle aufgrund ihrer Abrichtung auf "+3" den gleichen Übergang machen, dann ist nicht nur bestimmt, was für einen Schüler (z.B.) als nächster Schritt zu tun ist, sondern auch wie *alle* nur denkbaren Schritte gemeint werden können, ohne an sie gedacht zu haben. Gewonnen ist damit ein Kriterium, durch das zu entscheiden (rechtfertigen) ist, welcher Schritt auf den Befehl "+3" der richtige ist oder nicht. Die Frage nach der Bestimmtheit eines Ausdrucks erweist sich so als Frage nach einem konstanten gemeinsamen Verhalten (Reagieren) innerhalb 'Gruppe'. Fügt man dem Fragetext in PU 189A1 diese hinzu, lässt sich nicht nur zwischen denen unterscheiden, für die, die Übergänge festliegen und denen, für die sie nicht festliegen, sondern innerhalb derer, für die der Ausdruck bestimmt ist, eine weitere Differenzierung der Verwendung der Formel in "bestimmt" und "unbestimmt" vornehmen (vgl. PU 189C). Wittgenstein nennt die Frage wohl deshalb in seinem MS 118 auch „zweideutig“ (118. 88v). PU 190 stellt dazu den logischen Abschluss dar: der einzelne kann mit (s)einem Zeichen nur dort etwas meinen, ihm eine Bestimmung geben, wo er zur Erklärung auf Zeichen zurückgreifen kann, die bereits bestimmt sind. Wittgensteins Wortwahl ("wie wir sie ständig gebrauchen", PU 190A) an dieser Stelle weist natürlich auf den "ständigen Gebrauch" in PU 198C hin. Unterstrichen wird damit die Nähe dieser Bemerkungen zum Regelfolgen in PU 198ff. und die Rolle, die denen, die die Zeichen gebrauchen, dabei zukommt. Im Folgenden soll anhand einiger Bemerkungen aus dem MS 109 gezeigt werden, dass diese Einsichten zur Grammatik des Ausdrucks "bestimmt" auch für Wittgenstein nicht immer selbstverständlich gewesen sind.

[1] Für Baker & Hacker (1985) beginnt mit PU 185 Wittgensteins Erörterung des Regelfolgens in PU I.

[2] Hier verläuft Wittgensteins Argumentation teilweise parallel zu der in PU 147 / 148: das Wissen der Anwendung einer Reihe ist kein seelischer Zustand.

[3] Zu dem Konditional als Kriterium des Meinens vgl. auch PU 684. Zu erinnern ist hier auch an PU 78: Wissen wie eine Klarinette klingt, heißt u.a. den Klang erkennen, *wenn* auf ihr gespielt wird.

[4] Hier gilt einmal mehr Wittgensteins "Aber da waren wir ja schon einmal. Wir können uns ja eben mehr als eine Anwendung eines algebraischen Ausdrucks denken" aus PU 146B.

[5] Wittgenstein hat die Fragestellung aus PU 189A2 so oder ähnlich in (mindestens) fünf verschiedenen Zusammenhängen benutzt (vgl. z.B. MS 113. 143, MS 131. 63, MS 135. 75; MS 136. 6a). Ausgeschlossen werden soll hier nicht, dass er diese Anmerkung auch dazu benutzt, auf irreführende Fragen hinzuweisen. Hier wird vielmehr dafür argumentiert, dass er dies in PU 189 nicht tut.

2. "Aber ist der Sachverhalt durch die genaue Beschreibung *nicht* bestimmt?" (vgl. MS 109. 298).

Am 3. Februar 1931 bringt Wittgenstein sein MS 109[6] mit einer, wie er meint, "einfache[n] Antwort auf unsere langen Schwierigkeiten" (109. 298) zum Abschluss. Die Schwierigkeit, die er hier zum Abschluss zu bringen hofft, und mit der er im gesamten MS 109 kämpft, beruht auf der später in PU 198 formulierten Einsicht, dass Deutungen allein die Bedeutung eines Ausdrucks nicht bestimmen können (vgl. 109. 281). Für ihn steht damit in Frage, wie ist es möglich, einen Satz auf bestimmte Weise zu verstehen, einen bestimmten Gedanken mit ihm auszudrücken, wenn jeder Satz auf verschiedene Weise gedeutet werden kann. Die Antwort, die er jetzt gibt, besteht auch hier in einer Anmerkung zur Verwendung des Ausdrucks "bestimmt".

> / Denn mehr bestimmt, als durch eine genaue Beschreibung, kann etwas nicht sein. Denn bestimmen kann nur heißen, es beschreiben. Und das ist sehr wichtig. (109. 298)

"Bestimmen des Satzsinnes" kann demnach nicht heißen, sich das Denken und Verstehen wie einen Mechanismus vorzustellen, "dessen äußeres wir kennen, dessen inneres Arbeiten uns aber verborgen ist" (109. 174). Die nach Wittgenstein nur scheinbare Mehrdeutigkeit des Satzes (vgl. 109. 170) kann und muss auch nicht durch verborgene physiologische oder psychologische Vorgänge kompensiert werden.[7] "Genau" und "ungenau" und "bestimmt" und "unbestimmt" sind für ihn Ausdrücke, die sich auf die Beschreibungsmöglichkeiten innerhalb eines Sprachsystems beziehen (vgl. 109. 298). Eine unbestimmte Ausdrucksweise, die innerhalb der Sprache nicht zu beseitigen ist, läst Wittgenstein nicht gelten.[8] Insoweit durch einen Regelausdruck gegeben ist, "was ich tun soll, soweit es überhaupt gegeben sein kann" (298), ist was zu tun ist, bestimmt. "Und d.h.," so Wittgenstein, "es kann der Beschreibung nur eine Handlung entsprechen (nur so können wir diesen Ausdruck ['bestimmt'] gebrauchen)" (299). - Wittgenstein folgt auch hier seiner Praxis, seine Untersuchungsergebnisse mit einem philosophischen Kommentar zu unterlegen:

> Alle Schwierigkeit der Philosophie kann nur auf Missverständnissen beruhen. Eine Entdeckung ist nie nötig, kann nie nötig sein sie aufzulösen. Es ist ein Missverständnis & kann nur als solches aufgelöst werden. (109. 298.)

In diesem Zitat wird der Ausdruck "Missverständnis" von ihm für eine bestimmte Lesart von "bestimmt" gebraucht. W. sagt hier also, dass es einem Missverständnis entspringt, die gesamte Sprache, für missverständlich zu halten. Die Forderung nach der Bestimmung eines Satzes ist nämlich unsinnig, wenn man einen Satz, der alle Möglichkeiten der Bestimmung eines Sachverhaltes ausschöpft, weiterhin "unbestimmt" nennt. Die Möglichkeit des Missverstehens, die Wittgensteins Kritiker hier gegenüber jedwedem Ausdruck der Sprache geltend machen will, fällt auf ihn selbst in Form einer Kritik seiner Verwendung des Ausdrucks "unbestimmt" ("bestimmt", "Missverständnis", etc.) zurück. Die Skeptik gegenüber jedwedem sprachlichen Ausdruck wird durch den Hinweis auf die richtige, weil einzig mögliche Verwendung des Ausdrucks "bestimmt", etc. gebrochen. Die Bedenken sind Missverständnisse, denn mehr darf man von der Bestimmung eines Sachverhaltes durch einen Satz nicht erwarten. Sich solcherart über die Bestimmtheit der Sprache klar zu werden, nennt Wittgenstein in seinem Zitat nicht "eine Entdeckung machen". Sein Ergebnis nennt er nicht "Lösung" des Problems, sondern dessen "Auflösung". Nimmt man Wittgenstein beim Wort, dann gibt es am Ende des MS 109 das Problem mit der Unbestimmtheit des Satzes, das er in diesem MS seitenweise diskutiert, für ihn nicht mehr. Es ist für Wittgenstein gegenstandslos geworden. Zur Verteidigung dieser starken These weist Wittgenstein hier ein drittes Mal auf den Zusammenhang seines Darstellungsproblems mit der Frage hin „sieht der Andere wirklich dieselbe Farbe, wenn er blau sieht, wie ich?"(299 / vgl. PU 273).[9] Können wir überhaupt wissen, welchen Farbeindruck der Andere hat, oder kann nur er selber dies wissen? Wittgensteins Antwort lautet hier nicht, dass wir mehr nicht tun können, als sein Urteil über das, was er sieht, zu akzeptieren, sondern, dass die Frage gar nicht danach fragt, ob der andere die gleiche Farbe 'in sich sieht'. Es gibt, so Wittgenstein, einen grammatisch verbrieften Rechtsanspruch das, was der andere sieht "dasselbe, was ich sehe" zu nennen, wenn sich dies "nach der gewöhnlichen Methode konstatieren" (299) lässt. Die Frage in einem anderen Sinne zu verstehen, ist hier nicht möglich. Eine scheinbar noch offene, nur schwer zu beantwortende und philosophisch überaus relevant erscheinende Frage, wird zu einem Missverständnis bzgl. des Gebrauchs des Ausdrucks "die gleiche Farbe sehen" umgedeutet; das Problem löst sich auf. - In Analogie dazu ist es Unsinn weiter nach der Bestimmtheit eines Satzsinnes zu suchen, wenn dieser Satz in dem System, zu dem er gehört, schon vollkommen bestimmt ist. Wer nach dem Sinn eines Satzes fragt, erwartet nach Wittgenstein nicht mehr als eine Erklärung; und die bekommt er ja auch. (Vgl. PU 504.)

Vor diesem Hintergrund verwundert es nicht, dass Wittgenstein nicht nur fordert, dass die ganze Sprache für sich selbst sprechen müsse (vgl. Ms 109. 280), sondern auch die Frage „Wie kann der Satz einen Sachverhalt bestimmen?" nur noch unter Vorbehalt benutzt, insofern dieser zu der Annahme verleitet, ein Satz *tue* etwas. (Vgl. dazu PU 93B.[10]) Bereits zu Beginn des MS 110 präsentiert er dazu, die aus PU 435 bekannte pejorative Form: "'Wie macht der Gedanke / Satz das, das er darstellt?'" (110. 33; "Satz" im MS ohne Streichung über "Gedanke"). Auf diese Worte wendet Wittgenstein an dieser Stelle seinen philosophischen Auftrag aus PU 116B an, "die Worte von ihrer metaphysischen wieder auf ihre richtige Verwendung in der Sprache" (110. 34) zurückzuführen.

3. Wo Wittgenstein fehlt

Benutzt man die grammatischen Hinweise, die durch PU 189 gegeben werden, wird deutlich, in welcher Schwierigkeit Wittgenstein sich 1931 befand. Einerseits hat er ja ganz recht mit dem, was er über die Bestimmung eines Satzes *innerhalb* einer Sprache sagt. Setzt man, wie Wittgenstein dies im MS 109 tut, den Gebrauch unserer Sprache voraus, ist das, was zu tun ist, durch die Ausdrücke der Sprache vollständig bestimmt. Nicht umsonst sind

[6] Die Bezeichnung aller unveröffentlichten MSS und TSS erfolgt nach v. Wright 1990, zitiert wird nach der BEE (2000).

[7] Wittgenstein wendet sich im MS 109 u.a. explizit gegen Russells (198) und Ogdens und Richards (170) Theorie der Bedeutung. Vgl. auch PU 109 mit Wittgensteins Anmerkung zur "pneumatische[n] Auffassung vom Verstehen".

[8] Hier wird wohlgemerkt nicht der Gebrauch von Ausdrücken wie "ca.", "ungefähr", "vielleicht", etc. von Wittgenstein problematisiert.

[9] Vgl. MS 109. 171, 197 und 299.

[10] Wittgenstein schreibt dort: "Durch ein *Missverständnis* erscheint es uns so, als *tue* der Satz etwas seltsames" (PU 93). Wir werden so, so Wittgenstein eine Bemerkung später "auf die Jagd nach Chimären geschickt (PU 94).

nicht nur viele der Fragen, die er zu dieser Zeit z.B. zur Grammatik von "erwarten", "wünschen" "denken" und "befehlen" stellt, bis in seine PU I gewandert, sondern auch die Antworten, die er damals dazu bereits gab.[11] „Unbestimmt“, mehrdeutig“ etc. benutzt Wittgenstein für eine Relation innerhalb einer Sprache. Dadurch wird so wenig aus der Sprache herausgetreten (vgl. 109.170), wie dadurch, dass wir “verschiedene Arten von Formeln [...] einander entgegensetzen“, wie in PU 189C beschrieben.[12] Wittgenstein operiert zu dieser Zeit ganz in Ûbereinstimmung mit seiner Ausgangsbasis in PU 204: “wie die Sachen stehen“, ist die Befolgung von Befehlen bestimmt, getreu dem Motto „Was in der Logik nicht nötig ist, ist auch nicht von Nutzen“ (109.294) Andererseits wird deutlich, dass Wittgenstein durch diese Innenansicht der Sprache, den in PU 189 vollzogenen Wechsel, der einem Wechsel der Perspektive auf seinen Untersuchungsgegenstand gleichkommt, hier versäumt. An die Stelle eines innersprachlichen Erklärungsregresses tritt in PU 189 ja der Verweis auf verschiedene Gruppen, in denen ein Zeichen infolge einer Abrichtung immer auf die gleiche Weise Verwendung findet oder eben nicht. Und damit die Einsicht, dass von einer Bestimmtheit des Sinnes dort nicht die Rede sein kann, wo es nicht, wie Wittgenstein in einer frühen Formulierung von PU 189 schreibt – „Usus“ (Fragment 178E)[13] ist, das Zeichen zu gebrauchen. Vorbereitet wird dies im Kontext von PU 189 durch die Gegenüberstellung von unserem Verhalten und denen von Marsbewohnern (vgl. PU I, S. 54.), und von unserem Verhalten und denen von abnormalen Schülern; und eben durch die Bemerkungen zum Regelfolgen in PU 198ff. konstruktiv ergänzt.[14] Solange der in PU 189A durchgeführte Wechsel der Perspektive - aus einem funktionierenden Kalkül heraus (vgl. PU 189C) – zur Verwendungspraxis unterschiedlicher Gruppen hin, nicht geleistet wird, bleibt seine Frage nach der Bestimmtheit unserer Ausdrücke zu dieser Zeit fehlerhaft.[15] Es liegt damit auf der Hand, dass das in PU 189A2ff. ausgedrückte Wissen des späten Wittgenstein dem früheren Wittgenstein (vgl. PU 189A1) gefehlt hat. Ob dieses Wissen wiederum nur zur Aufklärung von Missverständnissen (vgl. 109. 298) beitrug, oder ob es, wenn es schon keine gewöhnliche Entdeckung war, so doch eine “grammatische Entdeckung“ (111. 2) zu nennen ist, und in welcher Weise diese Entdeckung ihn überrascht haben könnte, kann im Rahmen dieser Arbeit nicht mehr thematisiert werden.[16] - In PU 189A spricht Wittgenstein auf jeden Fall erst einmal von einem “Fehler”.

Literatur

Baker, G. and Hacker, P. 1984 *Scepticism, Rules Grammar and Necessity*. Chicago: University of Chicago Press.

Hilmy, S. St. ’Tormenting questions’ in *Philosophical Investigations* section 133, in: Arrigton, L. R. and Glock, H.-J. 1991 Wittgenstein’s *Philosophical Investigations*. London and New York: Routledge.

Schulte, J.2005 *Ludwig Wittgenstein*. Frankfurt a. Main: Suhrkamp.

Wittgenstein, L. 1953 / 1997 *Philosophische Untersuchungen*, Oxford: Basil Blackwell.

Wittgenstein, L. 2000 *Wittgenstein’s Nachlass. The Bergen Electronic Edition*. Oxford: Oxford Unversity Press.

Wright v., G. H. 1982 *Wittgenstein.* Oxford: Basil Blackwell.

[11] Hier braucht man nur an seine Aussagen zur internen Beziehung zu denken, die (u.a.) aus dem MS 109 stammen. Weiterhin Befinden sich in diesem MS Vorarbeiten z.B. zu PU 429, 430, 431, 435, 438, 440, 442-446.

[12] Der Fehler, den Wittgenstein zu dieser Zeit macht, besteht scheinbar darin, dass er alle Fragen zur Bestimmtheit eines Satzes im Sinne von PU 465B versteht. Der Ausdruck der Erwartung ist für Wittgenstein dort nur insofern unbestimmt, “dass er etwa eine Disjunktion verschiedener Möglichkeiten enthält“.

[13] Dieser vermutlich erste Formulierungsversuch von PU 189 befindet sich auf einem Kalenderblatt (Fragment) vom 8. August 1937. Wann Wittgenstein den Eintrag vornahm, ist ungewiss. Zweifelsfrei ist aber, dass der dort gegebene Hinweis auf Gepflogenheiten mit zu den frühesten im gesamten Nachlass gehört.

[14] Schulte (2005, S. 39) weist darauf hin, dass Wittgenstein vermutlich von Piero Sraffa "die Anregung [erhielt] menschliches (Sprach-)Verhalten aus 'anthropologischer' - also quasi ethnologischer - Perspektive zu betrachten".

[15] Insofern das stimmt, lassen sich viele der Bemerkungen aus PU I mit Bezug auf diese Fragestellung (grob) datieren. Vgl. z.B. PU 337, PU 432 und natürlich PU 198ff.

[16] Vgl. dazu z.B. Wittgensteins Bemerkungen in PU 89 bis 133.

Problems with Psychophysical Identities

Peter Kügler, Innsbruck, Austria

1. Theories of Psychophysical Identity

The idea of psychophysical identity is sometimes expressed by saying that the mind is identical to the body, or to one of its parts, the brain. This could mean different things. If it is meant as a rejection of the Cartesian dualism of mental and physical substances, it would be better to say that minds, i.e., mental substances do not exist at all. Here, elimination seems to be more appropriate than identification. If Descartes is wrong, souls are not identical to the body, they simply do not exist.

But of course, other kinds of mind-body identification are available. *Type identity theory* says that mental types are identical to physical (neurophysiological) types. In different versions of this theory the types are conceived as types of properties, events, processes, or whatever the preferred ontology is. *Token identity theory* identifies mental tokens with physical tokens without assuming identity of types. And *functionalism*, as it is usually presented, identifies mental types with functional types which are said to be "realized" by the physical properties of the brain. It is common to note that functionalism is actually ontologically neutral, because the functional properties could also be realized by non-physical properties. But it is no less common to supplement the functionalist framework with physicalist assumptions. Functional types are defined by their causal roles, which include physical causes and physical effects. Physicalists assume that these causal relations are describable by physics. Given this additional premise, the property that realizes the functional type must be a physical property. In this way, functionalism leads to psychophysical identity; whether it is type identity or token identity does not matter in the present context.

Type identity theory, token identity theory and physicalist functionalism identify the mental with the physical. Hence they aim at *reducing* the mental to the physical in one sense of "reduction". In the following I will try to argue against these three theories. For ease of discussion, I will use the open sentence "F is identical to G" for psychophysical identity statements of all sorts. F is a psychological and G a physical term suitable for the respective theory. So the terms may either refer to types or to tokens.

2. Psychophysical Identity as Necessary

Suppose that F is identical to G. Whether F and G stand for types or tokens, there are two possibilities as to the nature of the respective identity statements. They are either meant to express a *necessary* or a *contingent* truth. Let us consider the first option first and suppose that "F is identical to G" is necessarily true. It is well known that this assumption flies in the face of anti-physicalist arguments based on modal considerations. Descartes famously argued that he is not a material substance, because whereas he cannot doubt the existence of his mind, his body might not exist. The mind could exist without the body, therefore the two are not identical. Contemporary philosophers have criticized psychophysical identity theories by assuming that in other possible worlds mental properties are correlated with different physical properties than in the real world, or by evoking a Zombie world that is physically identical to the real world but contains no consciousness at all. Arguments like these rely on the possible dissociation of the mental and the physical. It is important to keep in mind that they only work against psychophysical identity claims assumed to be *necessarily* true.

When type identity theory was put forward in the 1950s, it was meant as a theory of contingent identity; as an empirical discovery that need not necessarily be true. But today most philosophers seem to be convinced that type identity implies necessary coextension, which means that a mental and a physical type can only be identical if they are correlated in all possible worlds. If this is correct, type identity is seriously threatened by modal counterarguments. But these arguments threaten token identity too. Suppose a mental token F is necessarily identical to a physical token G. There seem to be possible worlds in which F is correlated with different physical tokens, or with none at all. Conversely, there seem to be possible worlds in which G is correlated with different or no mental tokens. For example, we can imagine a possible world that is exactly like ours except that my current visual impression is correlated with a different brain state, or my current brain state with a different visual impression. So token identity does not look like a necessary relation either.

A possible defence against these and similar objections to psychophysical identity is to challenge the underlying modal intuitions. The objections presuppose that we are able to imagine, to think about, or to consistently describe possible worlds that differ from ours in containing other psychophysical correlations, or even Zombies without minds. But our imagination is limited, and so are our thoughts and descriptions, or any other faculty that is supposed to provide epistemic access to possible worlds. Thus there are reasons to be sceptical about the reliability of possible-world arguments. But although it is good to be cautious, one must add that there is no better access to possible worlds than that provided by imagination, thought, and description. When investigating such alternatives to reality we depend on these methods of metaphysical enquiry. Without them, it would be pointless to claim that psychophysical identity exists necessarily. If you dismiss imagination, thought and description as insufficient methods for exploring possible worlds, you must also dismiss the necessity of psychophysical identity.

3. Explaining Psychophysical Identity

Since necessity seems to be a dead end, I will now turn to the assumption that psychophysical identity is a contingent truth. In this perspective, questions of *explanation* become particularly important. Suppose a psychophysical identity claim is true. Do we need to explain why it is true? And, if so, how can we explain it? Is there an answer to the question "Why is F identical to G?" And do we need such an answer to understand that identity? To preclude a possible misunderstanding: these closely related questions aim at explanation, not at epistemic justification. A justification for the claim that F is identical to G might consist in empirical evidence that F and G are regularly correlated. But even if such an observed correlation were a good reason for believing that F and G are identical, it would not *explain why* this identity exists.

With respect to explanation, there is a huge difference between contingent and necessary psychophysical identity. Philosophers who prefer necessary identity sometimes maintain that we do not need an explanation, because in their view the existence of something necessary does not require to be explained. As their argument goes, there is no sense in asking, e.g., why squares are rectangles, because the opposite cannot be the case. The same would go for psychophysical identity: if F were necessarily identical to G, it could not be otherwise; therefore we would not need to explain why F is identical to G. I think this argument is *almost* correct. It only needs to be added that there is a reasonable answer to the question why squares are rectangles, even though it is quite trivial: squares are rectangles because the words "square" and "rectangle" are used in a certain way; more precisely, because the word "square" is defined in such a way that it refers to rectangles with equal sides. In a similar vein, we could answer the question why a necessary psychophysical identity exists: because F and G (or, for that matter, the predicates these symbols stand for) are used in such a way that both refer to the same entity. In short, the existence of a necessary identity can be explained by how the language is used.

An explanation like this, however, is not sufficient in the case of contingent identity. A contingent identity cannot be explained by language use alone, although the way the words are used always contributes to the explanation of why a sentence is true. A sentence has its truth value partly because of the meanings of its components. But in the case of contingent truths this cannot be the whole explanation. That an identity statement is contingently true means that it could also be false. Why is it not false? Why doesn´t the identity not exist? If we could answer these questions solely by how the language is used, the statement "F is identical to G" would be true in all possible worlds in which the words are used as they are used in the actual world. And this would contradict the assumption that the identity statement is contingently true. Therefore we need an explanation that goes beyond language use.

Obviously, we can exclude causal models of explanation which are used in some branches of dualism but are not suitable for explaining psychophysical identity. In particular, it does not make sense to say that a mental state is *caused* by the corresponding brain state, if the two are identical. An effect must be different from its cause. So, what we look for is an explanation of psychophysical identity that goes beyond language use and does not rely on causal interaction. Which alternatives are left? It is useful to consider examples of identities and their explanations in other areas, where we find at least three different explanatory models.

4. Three Models of Explaining Contingent Identity

The first one consists in analyzing identity in terms of *nomological connections*. Suppose two different things, A and B, are united by some physical process, like two drops of water that fuse into one when touching each other. Regarding identity, this process allows for different descriptions. An interesting way of describing it is to say that A and B, not having been identical before the fusion, are *now* identical. Let us accept this description for a moment, just to have an identity that we can explain. We can do this by reference to the fact that the molecules of drop A are now connected to the molecules of drop B by physical forces which can be described by laws of nature. What we call an identity is in fact a set of nomological connections.

The second kind of identity, and the second kind of explanation, is based on *definite descriptions* that refer to the same thing. Suppose a single person wrote the *Iliad* and the *Odyssey*. If this is true, the *author of the Iliad* was identical to the *author of the Odyssey*. It is easy to understand how both descriptions can refer to the same thing: Homer wrote both books during his lifetime. We know what a human being is, and we also know how a single human being can write two books.

The third and last kind of identity explanation concerns *epistemic perspectives*. Take any of the familiar examples of the relativity of perception, say, Locke´s example of the water that feels warm to one hand and cold to the other (because the one hand has been cold and the other has been warm before immersion). The water is the same for both hands, and so is its temperature, understood as the kinetic energy of the water molecules. The sensations of warm and cold are two perceptual perspectives on the same object. Three things are involved: a sensation of warm, a sensation of cold, and the temperature itself, which is neither warm nor cold in a sensational sense.

To conclude, in order to understand identity, we may refer to nomological connections, to definite descriptions or to epistemic perspectives. When applying these three models to psychophysical identity, we realize that they correspond to well-known positions in the philosophy of mind. The first one is represented by *parallelism*: F and G, the mental and the physical, form a unity, being connected to each other by psychophysical laws. In the history of philosophy the existence of these laws has sometimes been traced back to God, e.g. by Leibniz, but this is no necessary part of the theory. We can also stop the explanatory regress at the laws themselves.

The second model suggests a *double-aspect view*, which is often confused with parallelism. What I mean, however, is the idea that F and G are but two aspects of reality which has other aspects too. As I understand the two theories, these other aspects make the difference between double-aspect theory and parallelism. Compare this to the example of Homer: if he was a real person, he was not only the author of the *Iliad* and the *Odyssey*, but also had other properties characteristic of real persons. He had a heart and a brain, was born and died, and so on. In analogy, the double-aspect theory conceives of F and G as two aspects of a larger whole with additional properties. It is often assumed that these properties are unknown or even unknowable to us, just like we do not know many of the properties of Homer. But nevertheless we may speculate that we would understand how F and G are linked to each other if we knew the other properties of the psychophysical whole.

Neutral monism, which is our third position, regards F and G as two epistemic perspectives on reality which in itself is neither mental nor physical. As its name indicates, neutral monism rejects physicalist monism. Of course, the same is true of parallelism and the double-aspect theory, but while these are varieties of dualism (or rather pluralism in the second case), neutral monism is neither physicalist nor dualist. At least this is the intention behind the theory.

These considerations suggest that the quest for an explanation of contingent psychophysical identity leads to non-physicalist models, either to a kind of non-monism (parallelism, double-aspect view) or to a non-physicalist (neutral) monism. Of course, we may still raise the question whether parallelism, double-aspect theory and neutral monism are really non-physicalist views. After all, each of them has also been interpreted as a kind of identity theory, or as a predecessor theory coming very close to psychophysical identification. However, if we view these theories from the perspective of explanation, seeing them as attempts to explain psychophysical identity, we also see that they *explain physicalism away*. What they assume is a distinction between mental and physical parts, properties or aspects, or between the mental and the physical on the one side and a neutral third on the other. None of this is physicalism.

Reducing Complexity in the Social Sciences

Meinard Kuhlmann, Bremen, Germany

1. Explanation, Reduction, and Mechanisms in the Social Sciences

Reductions are attractive since they enhance the unity of our knowledge and allow for a sparser ontology or conceptual scheme. Reduction is intimately connected with explanation since reductive relations between different theories, i.e. 'intertheory relations', figure prominently in at least two of the main accounts of explanation. According to the covering law model, explanations of phenomena or special laws are viewed as derivations from (more) general laws, or conversely, in their reduction to more general laws. In the unificationist account, the core of explanations is considered to be their unifying power. Concerning the explanation of special laws, for instance, explanation consists in the overall reduction of independent laws, says the unificationist. Eventually, certain causal theories of explanations can also be rated as reductive in the sense that a diversity of phenomena is subsumed under universal generating mechanisms.

In the social sciences (i.e. including economics) methodological individualism, first introduced by Max Weber, makes a clear and still today very influential claim about how explanations ought to proceed, namely in terms of micro reductions to (individual) human actions (see Udehn 2001 for a comprehensive account). For this reason it is often said that methodological individualism expresses a particular version of reductionism, applicable in the social sciences. Methodological individualism is a bottom-up approach since the starting point is always the bottom level of individual constituents. The main thrust of methodological individualism that matters for my concerns is the emphasis that social phenomena cannot be sufficiently understood by analysing statistical correlations between macro quantities, e. g. macroeconomic variables, but that it is necessary to refer to the micro level of actors. However, as I will show below, very often not all micro details are relevant, and moreover, the attention must not be restricted to individuals in isolation.

Friedrich August von Hayek, explicitly endorsing Weber's doctrine of methodological individualism, underlined that an understanding of economic phenomena presupposes explanations in terms of rational actions by economic agents. Hayek's ideas must not be misunderstood as a form of rationalism since he even emphasizes the limits of rationalism in the sense of social planning and control, arguing that economic phenomena often emerge as unintended consequences of the economic agent's actions, whose perspectives are always very limited (cf. Heath 2005). For this reason, systematic economic analyses should always start by considering the perspective of the economic agents, i. e. by following the doctrine of methodological individualism. In the 1980s, the advent of rational choice theory and in particular of game theory triggered new debates about methodological individualism, arguing again that "there do not exist collective desires or collective beliefs" (Elster 1986: 3). These ideas shed some light on modern agent-based models of financial markets, studied among others by physicists. In both cases it is assumed that the constituents of the system have no access to information on the level of the whole system. Nevertheless, the net effect of all the individual contributions can be large-scale structures that seem unexpected given the uncoordinated behaviour of the constituents.

Explanations in terms of social mechanisms, advocated among others by Elster, can be seen as a way to overcome the divide between methodological individualism and holism, its classical opponent, since they enable explanations in the middle between micro-sociological and macro-sociological research, which is often desirable for explanatory purposes. They "provide more fine-grained accounts of social processes than do macro structural theories, but they do not require a commitment to the strictures of methodological individualism" (Pickel 2004: 177). In particular, modeling with artificial societies combines methodological individualism on the one side and the search for mechanistic explanations on the other side. The corresponding formal manifestation is a "shift from equation-based modeling to agent-based modeling" (Sawyer 2004: 263).

In the following I want to reflect upon one particular strand of agent-based modeling. In the last decade economists and physicists investigated various so-called microscopic models of financial markets, for instance the Kim-Markowitz, the Levy-Levy-Solomon, the Cont-Bouchaud, the Solomon-Weisbuch, the Lux-Marchesi, the Donangelo-Sneppen and the Solomon-Levy-Huang model (see Samanidou et al. 2007 for a review). In the stochastic multi-agent model of Lux and Marchesi (1999), for instance, there are two types of traders, 'fundamentalists' and 'noise traders' (or 'chartists'). Fundamentalists are rational traders in the sense that their action is based on the comparison of the fundamental value of the traded asset (e.g. stocks, bonds or currencies) and the actual market price. Fundamentalists buy if the asset is undervalued, and they sell if it is overvalued. In the case of noise traders the behaviour only depends on the current price trend and the opinion of other traders. In the following I will use this approach as an exemplary agent-based model.

2. False Models as a Path towards Real Mechanisms

In non-law-based accounts of explanation, such as Woodward's (2003) causal approach, models often play a crucial role in discovering causal relations (also see Glennan 2002). I want to claim that in agent-based explanations, the identification and understanding of causal mechanisms is in fact the main function of modelling. Mechanisms are not simply very detailed descriptions of *what* is happening, but their identification is crucial for causal explanations of *why* things behave the way we observe them. The specification of mechanisms is explanatory because it abstracts from as many details as possible with respect to the explanatory target. Thus *simplicity* is a crucial characteristic of mechanisms and the best way to identify mechanisms in complex systems is by constructing simple idealized models.

It is not essential that the mechanism is completely realistic. Gibbard and Varian 1978 point out that even 'caricature models' can help to understand certain aspects of the world. It only matters that certain structural features are modelled, such as interaction between the parts of a multi-agent system or the possibility of strategy change. Once these features are incorporated, the employed microscopic models of financial markets may be surprisingly unrealistic in various other features. For instance, microscopic models of financial markets abstract from material details about traders and transactions that are considered as irrelevant for understanding the basic features of financial markets ('Aristotelian idealization'). But microscopic models of financial markets also involve certain distortions, which simplify the situation considerably ('Galilean idealization'). Moreover, playing around with numerous different more or less unrealistic models has the advantage that it is possible to single out exactly which structural mechanisms are responsible for the statistical effects one wants to explain (see Morgan 1999 for a related conclusion). In contrast, an approach with a detailed realistic model might not reveal what it is that is actually crucial for the explanation (cf. Wimsatt 1987 and Cartwright 1983). My point is that agent-based models help to explain by concentrating on significant structural features while there is hardly any pretence to realism in many other respects. Batterman (2004) has a similar point, when he argues that highly idealized and oversimplified models can sometimes be better for the explanation of the dominant phenomenon than a detailed model in terms of micro-constituents.

3. Complexity and Robust Mechanisms in Agent-Based Models

Roughly, I understand a complex system not as an object with a complicated *compositional* structure but rather as an object with highly non-trivial *dynamical* features, on the basis of a structurally simple arrangement of a large number of non-linearly interacting constituents. One example is dynamical multi-agent systems in socio-economic contexts, which deal with 'microscopic' agents in a very simple arrangement and with a very simple individual behaviour. Whereas for a classical mechanism it is usually easy to predict its behaviour once the compositional structure and the behaviour of its parts is known, this is radically different in the case of complex systems. Here the knowledge of the compositional structure, e.g. spins on a square lattice, together with the knowledge of the behaviour of its parts in isolation as well as in simple composites, allows for hardly any straightforward predictions of the dynamical behaviour of the complex system.

Although higher-level interactions and thereby higher-level mechanisms are ultimately *ontologically* determined by the underlying physics, higher-level mechanisms are *explanatorily* autonomous. For instance, if financial market crashes were described in terms of the material processes that obtain between investors and their telephones, traders and their computers, electronic processes within the computer system of the NASDAQ etc., then the mechanisms involved in a crash could never be appropriately understood. As it turns out it is sensible to abstract so much from these material manifestations that it is possible to realize that the same mechanisms obtains in other contexts, notably in statistical physics (see Kuhlmann 2006). These consideration show that it can be extremely important for explanations, in particular for explanations in terms of mechanisms, not to eliminate level-specific vocabulary, notions and methods.

For mechanistic explanations in agent-based complex systems, the occurrence of the type of dynamical higher-level pattern one wants to explain, e.g. a statistical phenomenon, must be robust. The qualification *type of* pattern is essential since in complex systems the single *tokens* of a dynamical pattern are usually *not* robust due to the high sensitivity to variations of the initial conditions. In contrast to a classical mechanism like a thermostat, from which we expect a predictable output in each single case of its working, mechanisms in complex systems mostly do not generate token outcomes that we can predict, but rather bring about a certain type of outcome. But when it comes to the explanation of statistical features, the sensitivity to variations of the initial conditions in each single case dissolves in the collective statistics, which is *not* sensitive to such perturbations, provided the explanation is successful. To put it the other way around: a mechanistic explanation of a statistical phenomenon in a complex system is only successful if the resulting collective statistics of many simulation runs is not sensitive to perturbations of the system's parameters in a reasonable range of values. If this condition were not fulfilled one would rather classify the phenomenon as an artefact of the model, which does not help to identify an explanatory mechanism.

4. Structural Mechanisms

The above considerations show that a more abstract structural conception of mechanisms is prerequisite for understanding explanations in complex systems theories. The notion of mechanisms I want to suggest applies to many cases in the social sciences but also in physics, and biology, as far as they are complex systems in the sense I specified abobe. Currently, mechanisms are often discussed on the basis of case studies about biological systems (see e.g. Machamer, Darden, and Craver 2000). Naturally, this brings about limitations in the applicability. I propose the following more general notion of a mechanism in a complex system:

> A property of the time-dependent relation between locally interacting lower-level components and a higher-level quantity of a complex system is a candidate for a mechanism if it fulfils the following requirements: (i) The dynamics of the higher-level quantity exhibits a discernible type of pattern, which we want to be explained. (ii) The occurrence of this type of higher-level pattern is robust, i.e. it remains qualitatively the same under small variations of lower-level quantities.

My emphasis on the *local* nature of interaction is meant in the following way. If the occurrence of the higher-level pattern was due to an external influence with a global effect, e.g. a coordinating external force on all constituents, then I think one should not say that the higher-level pattern was generated by a mechanism. In other words, the higher-level pattern must emerge purely out of the interaction of the system's constituents.

I want to illustrate my characterisation of mechanisms in complex systems for the above-mentioned multi-agent model by Lux and Marchesi. The higher-level quantity is the price of some asset, e.g. a stock, currency or bond, and the corresponding example for interacting lower-level components are traders in the respective financial market. An example for a discernible pattern in the dynamics of this higher-level quantity is the so-called 'volatility clustering', i.e. the tendency of quiet and turbulent (or 'volatile') periods to cluster together. The agent-based explanation of this phenomenon rests on idealized assumptions about the relation between interacting lower-level components, namely the traders, and a higher-level quantity, i.e. the price of the traded asset. The explanation is successful if it can reproduce the characteristic higher-level pattern in a robust way, i.e. if the reproduced phenomenon remains stable under perturbations in a reasonable range of initial values. If it does succeed I would say that one has identified a mechanism that brings about the observed higher-level pattern. In order for such a mechanistic explanation to be satisfactory it is often desirable to single out—by appropriate modelling—what it is in the interaction of the constituents that generates the macroscopic patterns. One example is the swapping mechanism, where traders change their strategy and swap from one camp of traders to another, which brings about the transition from quiet to volatile periods.

In conclusion, mechanistic explanations are reductive in the sense that higher-level behaviour of a complex system is explained in terms of interacting lower-level components. However, mechanistic explanations are not reductive in the sense that higher-level description and conceptualisation was dispensable. But it is the case that the irreducibility of autonomous higher-level description allows for 'reduction' in the sense of a decrease of relevant details on the micro level due to the fact that the causal mechanisms needed for explanatory purposes are of a structural nature. Thus the reference to structural mechanisms in the social sciences makes the basic ideas of methodological individualism compatible with higher-level explanations of collective social phenomena.

Literature

Cartwright, Nancy 1983 How the Laws of Physics Lie, Oxford: Clarendon Press.

Elster, Jon 1986 Introduction, in: Elster, Jon (ed.) Rational Choice, New York: New York University Press.

Gibbard, Alan, and Varian, Hal R. 1978 "Economic models", The Journal of Philosophy 75, 664-677.

Glennan, Stuart S. (2002 "Rethinking mechanistic explanation", Philosophy of Science 69, S342-S353.

Heath, Joseph 2005 "Methodological Individualism", The Stanford Encyclopedia of Philosophy (Spring 2005 Edition, http://plato.stanford.edu).

Kuhlmann, Meinard 2006 "How Do Microscopic Models of Financial Markets Explain?", Proceedings: Models and Simulations, London, 2006. URL=http://philsci-archive.pitt.edu/archive/00002788/

Lux, T., and Marchesi, M. 1999 "Scaling and criticality in a stochastic multi-agent model of a financial market", Nature 397, 498-500.

Machamer, Peter, Darden, Lindley, and Craver, Carl 2000 "Thinking about mechanisms", Philosophy of Science 67, 1-25.

Morgan, M. (1999): Learning from models, in: Morgan, Mary S., and Morrison, Margaret (eds.) 1999 Models as Mediators: Perspectives on Natural and Social Science, Cambridge: Cambridge University Press, pp. 347-388.

Pickel, Andreas 2004 "Systems and Mechanisms: A Symposium on Mario Bunge's Philosophy of Social Science" Philosophy of the Social Sciences 34, 169-181.

Samanidou, Egle, Zschischang, Elmar, Stauffer, Dietrich, and Lux, Thomas 2007 "Agent-based models of financial markets", Reports on Progress in Physics 70: 409-450.

Sawyer, R. Keith 2004 "The Mechanisms of Emergence" Philosophy of the Social Sciences 34, 260-282.

Udehn, Lars 2001 Methodological Individualism, London: Routledge.

Wimsatt, William C. 1987 "False models as means to truer theories", in: Nitecki, Matthew H., and Hoffman, Antoni (eds.): Neutral Models as a Biological Strategy, Oxford: Oxford University Press, 23-55.

Woodward, James 2003 Making Things Happen - A Theory of Causal Explanation, Oxford: Oxford University Press.

Four Anti-reductionist Dogmas in the Light of Biophysical Micro-Reduction of Mind & Body

Theo A. F. Kuipers, Groningen, The Netherlands

Introduction

There are at least four anti-reductionist dogmas: 1) reduction implies elimination, 2) multiple realizability is an obstacle to concept and theory reduction, 3) supervenience implies non-reducibility, and 4) mind-body micro-reduction amounts to neuroreduction, neglecting the embodied and embedded character of mental life.

In (Kuipers, *Structures in Science* (henceforth *SiS*), 2001, Ch. 1), I have described several goals and types of co-operation between research programs. The programs can deal with different levels of aggregation and can even belong to different disciplines. Moreover, the programs can use different styles of description and explanation: causal, functional or intentional. In the present paper, these themes are discussed in the context of mind-body research. I will focus on interlevel biophysical mind-body research, and indicate several kinds of (micro-)reduction and (micro-)correlation of concepts and laws. At the end it will turn out that none of the four dogmas is observed by *biophysical micro-reduction of mind&body*.

Global characterization

Let us concentrate on interlevel mind-body research in the form of symmetric co-operation between research programs or even disciplines, hence boundary-bridging interlevel research. More specifically, I will deal with (macro-micro) interlevel reduction and correlation of mind-body concepts and with interlevel explanation and reduction of mind-body laws. The guiding perspective is the material (more specifically, biophysical) realization of such concepts and laws in micro-states, -events and -processes. The latter not necessarily restricted to the individual, let alone, to its neural states and processes.

In what amounts to an interdisciplinary dream, I will apply the mixed strategy of moderate reductionism or moderate holism to the non-eliminative reduction that might result from the co-operation between psychology and (neuro) physiology. Justin Schwartz (1991) noted already long ago that the anti-reductionist arguments of many philosophers of psychology are motivated by the worry that successful reduction would eliminate rather than conserve the mental realm. However, as he points out in his specific terms, but in the same spirit as ours, some paradigm natural science examples of (micro-)reduction of concepts, such as water, genes and tables, are not at all eliminative. Indeed, “Philosophy of psychology needs more detailed attention to issues in natural science which serve as analogies for reduction of the mental” (Schwartz, 1991, p. 203). I am afraid this still holds to this day.

Another general point is that philosophers seem to favor discussions involving the reduction of whole disciplines. From my point of view in general, and the mixed strategy in particular, such discussions are rather pretentious. The really interesting question is whether *at least some* mental concepts and *even some* mental laws (or, more cautiously, regularities or quasi-laws) can be reduced, straightforwardly or approximately, to biophysical micro-concepts and -laws.

To state the (micro-)reductionist ambition more precisely, let us start from the common sense division of properties (events, states, dispositions, etc.) of human individuals into bodily and mental properties. That is, some individual properties are purely or largely bodily, e.g., weight, temperature, strength etc. Some others are typically mental, e.g., having memories, beliefs and desires. And there are many mixed properties and the like in between: pains, actions, including speech acts, etc. In common sense terms all these properties are causally active (or at least causally reactive) and not only within their own sphere (mind/mind and body/body interactions) but also between the spheres (mind/body interactions). Many of these properties can be described in functional and several of them in intentional terms, and these ways of description may well go together.

To describe and explain these common sense ideas, classical mind-body dualism postulated two kinds of substances, which can also interact. The modern, prevailing biophysical point of view is quite different. On lower levels of aggregation or organization everything is supposed to be of a biophysical (and partly functional) nature, such that all individual properties supervene as mental and/or bodily macro-properties of the individual conceived as an (organized) aggregate of a biophysical nature. In other words, the macro-properties are supposed to be (materially) realized by biophysical micro-states. Hence, and this is the core idea of my approach, *mental, bodily and mixed terms can be treated in the same way, and that way is such that their being causally interactive is unproblematic.*

It is important to note that actual biophysical (micro-)states of human beings have (distributed) traces of earlier states and inputs of one kind or another. This fact is not only relevant for many bodily properties, e.g., specific muscular features of sportsmen, but also for many mental properties, e.g., what one believes and, very important, how that is realized. More specifically, it is plausible to assume that the specific learning history of an individual has left its biophysical traces in such a way that the biophysical realization of the same belief may be different between different people. Moreover, the precise biophysical realization of (psychologically) the same belief of one and the same individual may change in time, with the consequence, for instance, that the strength of the belief changes. The reason is that long-term learning is nowadays conceived not only as a matter of electrical and chemical processes, within and between neurons, but also as a matter of morphological changes of the neurons (Kandel et al., 2000; Kandel, 2006).

The (biophysical micro-)reduction ambition can now be stated more clearly. I distinguish B(ody)-properties/-concepts/-terms, such as weight, M(ental)-properties, such as having a certain belief, and MB-properties, such as sexual arousal, without supposing that the division is always clear. The potential ambiguity will turn out to be of no importance. I will talk about B-laws (-regularities/quasi-laws) and M-laws, depending on whether they relate only B-properties or only M-properties, respectively. However, it remains a matter of dispute whether (pure) M-laws exist at all. MB-laws are laws that relate M- and B-properties and/or MB-properties. Note that in my approach, MB-laws are as a rule or even always correlations, MB-correlations for short. Correlations may be causal or ontological correlations, in the latter case they are incomplete, e.g., of a part-whole nature, such that they cannot be considered as ontological identities. The question of whether there are MB-identities, besides MB-correlations, is irrelevant in my micro-reductionist approach, for both concern relations on the macro-level.

In this terminology the really interesting question is whether at least some M- or MB-concepts or even some M- or MB-laws can be reduced, straightforwardly or approximately, to biophysical micro-concepts and -theories. As long as reduction of, say, an M-concept is not successful, it may well be that we can find one or more MB-correlations and 'ontological' correlations, that is, correlations between the M-concept and micro-concepts. For instance, neuro-imaging techniques may show that certain areas and processes on, relatively speaking, the macro-level of the brain are specifically related to a certain M-concept. Such MB-correlations may well be a crucial step for an ultimate reduction of the M-concept. Moreover, electrophysiological, neuro-anatomical, and neurochemical experiments, for instance, may show that certain types of neurons or neurotransmitters are involved in M. Such ontological correlations may be part of an ultimate reduction. Similarly, as long as the reduction of, say, an M-law is not successful, it may be that we can find a micro-theory that enables the explanation, together with suitable concept correlations, of the M-law.

Further Analysis

In the following analysis, I will essentially apply the analysis presented in Chapters 3 and 5 of *SiS* to individuals, humans and other animals. In order to simplify the presentation I will restrict the attention to MB-concepts and MB-laws, relating MB-concepts, since the story is essentially the same for pure M-concepts and M-laws, if any exist, for B-concepts and B-laws, and for MB-laws relating pure M- and B-concepts. Let us concentrate for a while on the macro-concepts and let us assume a finite number of families of MB-types, i.e., sets of mutually exclusive and together exhaustive, monadic MB-predicates, each family giving rise to an MB-representation of an individual at each moment. The MB-predicates at the start are supposed to be types in the sense that their application is stable and strongly intersubjective and/or because they are lawfully correlated with other types.

Let us now conceive individuals, and their environment as far as relevant, as (organized) aggregates, to be described in terms derived from 'base-terms'. It is important to distinguish between the genuine base-concepts for the constituents, e.g., molecules or cells, etc., and their mutual relations, and the representation of aggregates of these constituents, e.g., substances and tissues. The latter may be done by construing set-theoretic 'micro-structures' in terms of the (biophysical) base-concepts. The micro-structures will be indicated as CS-structures, because the micro- or base-entities are supposed to be *C*ells, in particular neurons, and (micro-pieces or -amounts of) *S*ubstances. The latter are included in order to be able to take non-biological micro-entities in the brain, the body and the environment into account. Each CS-structure represents a conceptual possibility of a state of an aggregate.

A CS-structure is a token of as many (disjunctions of) CS-types as can be meaningfully defined as sets of structures, using aggregate concepts, i.e., concepts characterizing a certain aspect of the structure, e.g., the internal and external temperature, the ratio of activated neurons in a certain layer or region of neurons involved, etc. Moreover, the idea is of course that many different, though probably in some way or other related CS-structures, may realize (approximately) the same MB-concept, e.g., the same memory or pain. This set of structures is called the *realization-class* of that MB-concept. If a structure belongs to such a class, it is also said to be a token of the *realization-type* of that concept.

Let us now formulate some possible reduction results concerning some MB-families that would be considered as successes. For this purpose we have to assume at least the *Token-Identity Hypothesis*; i.e., every state of an individual can be represented by an MB-type of each MB-family as well as by a CS-structure. Additionally the *Realization Hypothesis* has to be adopted; i.e., every CS-structure uniquely determines, as a matter of ontological fact, an MB-type for each MB-family. Note that the Realization Hypothesis implies that each structure can be represented as a member of the realization-class of precisely one MB-property of each family. In other words, that structure is a token of the realization-types of those MB-properties.

It is possible to distinguish (*SiS*, Chaper 5) three degrees of concept (micro-)reduction. A result of the first degree is the quasi-type-type reduction of some MB-type, say an MB-state. This result only presupposes that it is possible to characterize a set of CS-structures as the realization-class of that MB-state. Assuming that the CS-structure representation of a state can be experimentally established, the reduction enables the prediction of the MB-state on the basis of the CS-structure representation (quasi-reduction) and, conversely, being in an MB-state predicts that the CS-structure representation belongs to the realization-class of that MB-state. Of course, given that the realization-class is not exclusively defined in CS-terms, for it is by definition 'MB-induced', both kinds of prediction concern CS-structure representations that are (very) similar to those that were used to characterize the realization-class.

As soon as a realization-class can also be characterized independently in CS-terms, in the form of a disjunction of CS-(micro-)types, exemplifying *a multiple version* of the Type-Type Identity Hypothesis, see below, we get the possibility of a one-many or multiple type-type reduction of an MB-state: the second degree of reduction. A second degree reduction enables the deduction of the realization-type and hence the MB-state when one knows that the CS-structure of the state belongs to the disjunction of the corresponding CS-(micro-)types. The special case that the realization-type corresponds to just one CS-type, exemplifying the singular version of the Type-Type Identity Hypothesis, is a third degree result: the one-one or singular type-type reduction of an MB-state. According to the *singular Type-Type Identity Hypothesis*, belonging to a certain macro-type and to a certain micro-type amounts to an 'ontological equivalence'.

The foregoing classification immediately leads to three degrees of (micro-)reduction of laws. First, if two MB-types are lawfully connected and both can be quasi-reduced, it may be possible that the law can be reduced by a theory in a weaker sense than occurs in reduction by identification according to a general model (*SiS*, Chapter 3), viz., that, starting from one or more specific theory applications, the transformation step consists of replacing realization-type attributions by macro-type attributions. I call this reduction of a law of the first degree, or quasi-reduction. If two MB-types are one-many type-type reducible and lawfully connected, there must be a lawful connection of unions of CS-types. Formally, this must even be the case in such a way that the first can be reduced to the second, in the sense of reduction by identification, based on the appropriate multiple Type-Type Identity Hypotheses (*SiS*, Chapter 5). I call this reduction of a law of the second degree, or multiple reduction of a law. In the special case that MB-types can be reduced to proper CS-types, the law reduction is a result of the third degree, called singular or strong reduction of a law.

I have mainly restricted myself to the general perspective; for the specifications for the causal-structural, functional and intentional style I refer to *SiS*, Chapter 6.3. Paradigmatic examples of research that show progress in realizing the above described micro-reductionist ambitions concern learning and memory (Kandel et al., 2000; Kandel, 2006, Huisman, 2005). Moreover, other examples seem to fit very well. In *SiS*, Chapter 6.3, upon which the core of this paper heavily draws, indications are given to research dealing with mental disorders and 6.4 deals, more extensively, with juvenile delinquency, essentially illustrating the non-validity of the four anti-reductionist dogmas.

Conclusion

All discussed degrees of micro-reduction of concepts and laws concern straightforward *non-eliminative* reduction. Multiple reduction of concepts and laws exploits the possibilities of *multiple realizability* instead of being blocked by it. Many of the reduced concepts and laws will typically represent *supervient* properties and patterns, in particular arising from the interaction of micro-entities. Finally, wherever relevant, the micro-state goes beyond the neural state and includes the relevant parts and aspects of the (rest of) body and of the environment. Hence, none of the four anti-reductionist dogmas is observed by what I like to call *biophysical micro-reduction of mind&body*.

Acknowledgement

I like to thank Fred Keijzer for his critical comments on a draft.

Literature

Huisman, L. 2005 Neuroscientific reduction: Functional versus Phenomenal Aspects of the Mind, MA-thesis, Groningen: Faculty of Philosophy.

Kandel, E. 2006 In search of memory, New York: W.W. Norton&Cy.

Kandel, E., Schwartz, J.H., Jessel, T. M. 2000 Principles of Neural Science: International Edition, New York: McGraw-Hill Companies.

Kuipers, T. 2001 Structures in Science (SiS), Dordrecht: Kluwer.

Schwartz, J. 1991 "Reduction, elimination, and the mental", Philosophy of Science, 58, 203-220.

Two Problems for NonHumean Views of Laws of Nature

Noa Latham, Calgary, Canada

Humeans and nonHumeans offer dramatically different world-views stemming from their different accounts of the status of the laws of nature of our world. Humeans take the laws to be material conditional facts, while nonHumeans take them to be nonmaterial. And this can be shown to be equivalent to another popular way of distinguishing Humean and nonHumean views. Humeans claim, and nonHumeans deny, that the laws metaphysically supervene on the totality of particular facts—those describing the distribution of properties somewhere in spacetime—where fact A metaphysically supervenes on fact B iff A exists in all possible worlds in which B exists.

I shall assume that there is a fundamental level of reality, and I shall regard fundamental properties as those characteristic of such a level. Humeans can be characterised accordingly as claiming that the fundamental laws of our universe metaphysically supervene on the totality of its fundamental particular facts, and nonHumeans as denying this. I am also provisionally assuming a view of properties that treats them as independent of the laws in which they feature.

My discussion henceforth is thus narrower than familiar discussions of laws of nature. I shall not examine nonfundamental laws such as special science laws, ceteris paribus laws of physics and thermodynamic laws. I see it as an advantage of the account I am offering that it does not run together what I take to be very different kinds of laws in quest of a unified account.

A popular idea, which I think captures the intuitive understanding of the role of fundamental laws in our universe, is to think of them as analogous to rules of a computer programme that changes an array of pixels on a screen from one moment to the next. The way I like to put it is to say that fundamental laws of nature are derivable from a single fundamental law, analogous to the computer programme, that (i) provides an instruction, given the existence of a time-slice, as to what the next time-slice is to be and (ii) makes it the case that this next time-slice comes into existence. I call this the generative conception of fundamental laws.

Humeans typically point to the absence of the category of nomic facts on their view as a huge gain in simplicity and as greatly favouring their view. In opposition to this, it can be claimed that the advantage of simplicity goes instead to the view that the universe is generated by a fundamental law from an initial slice, as this provides a sense in which only the fundamental particular facts of the initial slice and the fundamental law are causally/explanatorily basic. The regularities observed in subsequent fundamental particular facts are all explainable, while the Humean must treat them as a cosmic coincidence. Humeans typically respond that their view attributes a causally/explanatorily basic status to the initial conditions and fundamental laws too. I think the best reply to this is to say that what is compelling about this nonHumean view is that laws and initial conditions provide a *mind-independent* causal/explanatory basis for the universe. By contrast, the laws are essentially mind-dependent on the best worked out Humean account, David Lewis's, and it does not appear that any alternative Humean account could escape this. The Humean cannot offer a mind-independent sense in which anything is causally basic, or a mind-independent sense in which anything less than the totality of fundamental particular facts is explanatorily basic.

I believe almost all nonHumean views of laws are committed to such fundamental generative laws. (Nancy Cartwright's view, if coherent, is one that is not.) These include views that take fundamental properties to be dispositional and to entail laws. Such views of properties presuppose victory over the Humean view of laws, and compete with other nonHumean views. I shall not dwell here on these differences among nonHumeans as my concern in this paper is to examine two problems that I think almost all nonHumean views face.

The idea that the observed world is driven to unfold, i.e. generated, over time by fundamental laws or fundamental dispositional properties is I think the natural way people have understood the workings of nature. This appears to be a purely metaphysical view in the sense that it is independent of scientific facts about our universe that were unavailable to our ancestors and may be unavailable to us today. In this respect the nonHumean view of laws resembles the view with which it is often compared that there is a physical reality independent of our sense perceptions. What I will be arguing in the remainder of this paper, however, is that the thesis that our universe is generated may be empirically false and thus cannot be defended as a purely metaphysical thesis. I think this empirical sensitivity detracts greatly from the plausibility of the generative view. But all the alternatives I have encountered strike me as nonstarters.

The first problem is that generative laws appear to be possible only for universes generated from a single initial slice. This precludes backwardly infinite universes such as a Steady State or oscillating universe, and backwardly finite but open universes if these are coherent. It also excludes universes for which the notion of a linear temporal ordering of events doesn't make sense in the part of the universe where one would need an initial slice to be.

Why is an initial slice required? Well, some particular fact is needed in addition to fundamental laws in order for there to be a universe. And the only alternative to the fundamental facts of a single slice appears to be the fundamental facts of an infinite sequence of slices. This reintroduces the Humean view with its problems in characterising this infinite sequence as explanatorily basic when it appears to be partly generated. Furthermore there would be an implausible arbitrariness to the cutoff point between this infinite sequence and the generated part of the universe. So an initial slice is the only plausible option from which a universe could be generated.

But why should there be a problem in generating a temporally dense backwardly finite universe? Assuming for the moment an absolute conception of time, then there is either a last moment at which there is nothing and no first moment at which there is something (a backwardly open universe), or no last moment at which there is nothing and a first moment at which there is something (a backwardly closed universe). And perhaps both open and closed options for the temporal topology of a backwardly finite universe are also coherent if one rejects an absolute notion of time and requires that something exists in order for time to exist.

The generative conception can accommodate backwardly finite but closed universes that begin at a singularity. But it is hard to conceive of our universe as such, since so little information about it could be stored at such a single spatial point. The initial conditions would have to consist in finite values of various fundamental parameters at that point. Theories of the universe based on general relativity that ignore quantum effects yield a singularity with infinite values of the fundamental parameters at that point. But such a singularity is not a candidate for an initial slice as the values of the parameters are not well defined.

Recent efforts to construct a theory of quantum gravity for our universe that integrates general relativity and quantum theory offer hope of avoiding this problem of a Big Bang singularity at which parameters cannot be defined. They mostly have the consequence that the linear temporal ordering among events breaks down, i.e. that time doesn't make sense, close to the Big Bang when the universe is smaller than the Plank scale. If universes in accordance with such theories are coherent, they too would lack an initial condition and so it would seem could not be generated.

A second problem for the view that our universe is generated concerns the direction of generation. For any generated universe there is an extrinsic direction of time given by the direction of generation (and by the temporal asymmetry in the notion of a fundamental dispositional property). And in our universe there are temporal asymmetries found in intrinsic features that mark a subjective direction of time. They are generally regarded as based on the thermodynamic condition of increasing entropy. Such asymmetries include the way actions affect the future not the past, organisms age, and scattered rubble doesn't suddenly cohere to constitute a building. A problem arises for the generative conception if there is no good reason to believe that these intrinsic and extrinsic directions coincide. For part of the appeal of the view that our universe is generated comes from its incorporating the compelling intuition that what is present was generated by the laws and what seems subjectively to be the past.

Why might the direction of generation and our subjective direction of time diverge? The direction in which time seems to be flowing in the universe might be the opposite to that in which the universe is generated. From our perspective on the direction of time we say that our universe is 13 billion years old, but if the universe has a finite duration and its fundamental laws are time-reversal invariant, then it could be generated from the other end, and we should say instead that we have another 13 billion years or so before the Big Crunch, if we take the direction of generation to be the direction of time.Given these assumptions then, it is logically possible that our universe is generated in the counterintuitive direction. However, this should worry us no more than the logical possibility that we are brains-in-vats, so long as we have good reason to believe that our universe is generated in the intuitive direction.

Tim Maudlin (2007, pp. 130-5) addresses this problem for the generative conception, adding to our assumptions of a temporally finite universe and time-reversal invariant fundamental laws the empirically well supported assumption that there is low entropy at one temporal end of the universe and high entropy at the other. Maudlin's idea, as I understand it, is that we could explain why entropy increases given the assumption of the fundamental generative laws and the condition that entropy starts off low. But we could not explain a decrease in entropy given the assumption of the laws and the condition that entropy starts off high. As we observe an entropy gradient over time in our universe, this asymmetry in explanation makes it much more reasonable to suppose that the universe is being generated from the lower entropy end than from the higher entropy end, and hence that the intuitive direction of time coincides with the direction of generation.

Where I think Maudlin goes wrong is in setting up a biased contrast that misleadingly gives the impression that for possible universes with our generative laws there are in a natural sense more of them with low to high entropy gradients than there are with high to low entropy gradients. The semblance of asymmetry comes from taking as one option that the universe has a low entropy beginning, ignoring knowledge that it has a high entropy state at its other end, and pointing out that it would be highly likely to evolve into a high entropy state. This is then compared with the option that the universe has a high entropy beginning, ignoring knowledge that it has a low entropy state at its other end, and pointing out that it would be very unlikely to evolve into a low entropy state. Each of these options embraces something and ignores something we know about our universe. But what the first option embraces—that the universe has a low entropy state—is something that would be considered highly improbable if we knew nothing about which universe was actual, while what the second option embraces—that the universe has a high entropy state—is something that would be considered highly probable if we didn't know which universe was actual.

I suggest that to determine what it is reasonable to believe about our universe we should not skew our understanding by suppressing knowledge in this asymmetrical way, but should begin with what we do know, namely that the actual universe contains an entropy gradient with extremely low entropy at one end. The salient fact is then surely that from the nature of time-reversal invariant laws, there is a natural 1-1 mapping of universes with a low to high entropy gradient onto qualitatively identical universes with a high to low entropy gradient. This leaves us without a reason for believing our universe is generated from the low entropy end rather than the high entropy end.

Nevertheless, there might be empirical reasons to reject the two assumptions we were making—of a temporally finite universe and of time-reversal invariant fundamental laws. Rejecting either of these assumptions would support belief in a coincidence of the direction of generation and the subjective direction of time. For if the universe is closed at one end and infinite at the other, it could be generated from the unique slice at one end but could not be generated in the other direction. Current estimates of the rate of expansion of the universe suggest that it is forwardly infinite, though this could be revised if evidence of a lot more "dark matter" arises. And although both general relativity and the Schroedinger equation are time-reversal invariant, a quantum theory involving a Collapse Postulate is not. Whether such a postulate is among the fundamental laws of our universe is currently a debated feature of the interpretation of quantum theory. But should it be so, and (as is widely supposed) there cannot be a time-reverse of such a postulate, then the universe couldn't be generated in the counterintuitive direction.

So science may grant both initial conditions and a secure belief that the universe is generated in the intuitive direction. However, it is disconcerting that the generative conception of fundamental laws should rest on such unsettled scientific theory. And this empirical sensitivity undermines the idea that the generative conception captures a commonsense metaphysical view. It is hard to rest content that science will bail out the generative conception. For if there are coherent epistemically possible ways our universe might be which do not meet the empirical requirements for the generative conception, we need a way to understand them.

Literature

Maudlin, Tim 2007 *The Metaphysics within Physics* Oxford University Press.

Some Remarks on Wittgenstein and Type Theory in the Light of Ramsey

Holger Leerhoff, Konstanz & Oldenburg, Germany

Paradoxes, Logicism and the Theory of Types

Russell developed his Theory of Types as an answer to a range of paradoxes he saw his logicist project confronted with. One of these paradoxes is Russell's well-known paradox about the set of all sets not containing themselves, others are such famous paradoxes as the liar, Berry's paradox and the Grelling/Nelson paradox. Henri Poincaré coined the term ›vicious-circle fallacies‹ for all of these: what they seem to have in common is, in Russell's words, that

> [i]n each contradiction something is said about *all* cases of some kind, and from what is said a new case seems to be generated, which both is and is not of the same kind as the case of which *all* were concerned in what was said. (Russell 1908, 224)

Russell presented the first draft of a solution to this class of paradoxes as an appendix to his *Principles of Mathematics* in 1903 and a full-blown solution in his ›Mathematical Logic as based on the Theory of Types‹ in 1908. The core idea behind type theory is that each propositional function has a ›range of significance‹, *i.e.*, a set of possible arguments, and the following limitation:

> This leads us to the rule: ›Whatever involves *all* of a collection, must not be one of the collection‹; or, conversely: ›If, provided a certain collection had a total, it would have members only definable in terms of that total, then the said collection has no total‹. (Russell 1908, 225)

The 1908 version of the theory played a key role in the monumental logicist project Russell and Whitehead were working on then, the *Principia Mathematica*. One serious flaw of the theory, however, was the need for the *Axiom of Reducibility*, which is everything but a *prima facie* plausible axiom of logic.

Ramsey's classification of the paradoxes

Around 1925, F. P. Ramsey, a logicist as well, was trying to find a way to dispense with the Axiom of Reducibility and to that avail examined Russell's reasons for introducing the Theory of Types in the first place: the various vicious-circle paradoxes. He introduced a nowadays generally accepted distinction between them:

> We can easily divide the contradictions according to which part of the theory is required for their solution, and when we have done this we find that these two sets of contradictions are distinguished in another way also. The ones solved by the first part of the theory [*i.e.*, the Simplified Theory of Types] are all purely logical; they involve no ideas but those of class, relation and number, could be stated in logical symbolism, and occur in the actual development of mathematics … Such are the contradictions of the greatest ordinal, and that of the class of classes which are not members of themselves. With regard to these Mr. Russell's solution seems inevitable.
>
> On the other hand, the second set of contradictions are none of them purely logical or mathematical, but all involve some psychological term, such as meaning, defining, naming or asserting. … [I]t is possible that they arise … from ambiguity in the psychological or epistemological notions of meaning and asserting. Indeed, it seems that this must be the case, because examination soon convinces one that the psychological term is in every case essential to the contradiction, which could not be constructed without introducing the relation of words to their meaning or some equivalent. (Ramsey 1926, 192)

Ramsey classified the first type as *logical*, the second type as *psychological paradoxes*, though the term ›semantical paradoxes‹ for the latter is more common today. Regarding these two types of paradoxes, different parts of Russell's 1908 type theory are responsible for their solution. For the simpler logical paradoxes, the part of Russell's theory which is akin to his first proposal from 1903 was sufficient. Ramsey distilled that part from Russell's more complex 1908 theory and coined the term *Simplified Theory of Types* (STT) for the result. The semantical paradoxes, on the other hand, proved to be consistent regarding the STT and remained a problem requiring the full-blown theory, the *Ramified Theory of Types* (RTT), for its solution. Now, according to Ramsey, the logicist project was not at all confronted with the semantical paradoxes—he claimed that those paradoxes were problems of language, not of mathematics, and so it was not mathematics' job to deal with them. If, following Ramsey, the STT was indeed sufficient for the goals of mathematics, there was no need for using the RTT—and since the STT was lacking the negative side effects of the RTT that lead to the necessity of introducing the axiom of reducibility, Ramsey's modification made the logicist project much more acceptable.

Wittgenstein's critique of the Theory of Types

In the *Tractatus*' 3.33 ff., Wittgenstein presents his arguments against Russell's Theory of Types:

> In logical syntax the meaning of a sign should never play a role. It must be possible to establish logical syntax without mentioning the *meaning* of a sign: *only* the description of expressions may be presupposed. (Wittgenstein 1921, 3.33)
> From this observation we turn to Russell's ›theory of types‹. It can be seen that Russell must be wrong, because he had to mention the meaning of signs when establishing the rules for them. (Wittgenstein 1921, 3.331)

The core element of Wittgenstein's criticism can be understood in at least three ways: (1) In formulating the Theory of Types, Russell uses terms (›truth‹, ›meaning‹, ›type‹, …) that are, according to Wittgenstein, meaningless. If one understands Wittgenstein in this way, an alternative type theory, formulated on a purely syntactical level, could escape his criticism. Church's Theory of Types (Church

1940) is constructed in such a way and, given that this is the crucial point in Wittgenstein's criticism, could be regarded as a valid alternative to Wittgenstein's approach. James Davant discussed this option in his (Davant 1975) and came to the conclusion that *any* version of type theory is incompatible with Wittgenstein's system in the *Tractatus*; I will not here repeat his arguments. (2) Wittgenstein's criticism is directed at Russell's talking about the meaning of the symbols of the ›object language‹. (3) Wittgenstein's criticism must be understood as a combination of (1) and (2)—this is the way I understand Wittgenstein.

Indeed, Russell has to classify symbols according to their type: When he says that, *e.g.*, some symbols stand for individuals of type 0 or propositional functions of type 2, Russell is *in some sense* talking about the meanings of the respective symbols. This sense is a very basic one, no more problematic than saying that the relation ›is larger than‹ has to be accompanied by exactly two terms to add up to a meaningful sentence. Nonetheless, this *is* talking about the meaning of symbols and one may very well buy Wittgenstein's arguments against this if one likes.

In my opinion, this rather fundamental difference between Russell and Wittgenstein is grounded in their different approaches to language: Wittgenstein's ideal language in the *Tractatus* is no purely artificial language but the end point of an actual analysis of ordinary language, and thus somewhere between an ideal and an ordinary language. Though we do not use the *Tractatus'* language for actual communication, according to Wittgenstein we use the language on a very fundamental level of our thinking. Its names do refer directly to the *objects* (*Gegenstände*) of the world: in the *Tractatus*, there is a very close-knit connection between language, thinking, and ontology. As a consequence of this, Wittgenstein cannot state the meanings of names, of the symbols of his language, *in* his language: The meanings only *show themselves* through their use. Russell, on the other hand, is free to do this; he may very well use a metalanguage or a hierarchy inside his language to assign meanings to his symbols, since his (much more artificial) language does not necessarily stand in a fixed relation to our thinking and hence is not subject to the restrictions holding for Wittgenstein's language.[1]

Wittgenstein's way to avoid the logical paradoxes

Since Wittgenstein has to dispense with type theory, he has to put forth an alternative way to escape the problems associated with the paradoxes mentioned above. Moreover, type theory may very well have it's origin in the solution of the paradoxes, but its benefits surpass the simple fact that it can deal with them: the theory offers some deep insights into the nature of language, *e.g.*, into ambiguity, which is a crucial element in the logical paradoxes. Wittgenstein was very well aware of that and saw the need to give an explanation of these phenomena, too:

> In order to avoid such errors [resulting from ambiguity] we must make use of a sign-language that excludes them by not using the same sign for different symbols and by not using in a superficially similar way signs that have different modes of signification: that is to say, a sign-language that is governed by *logical* grammar—by logical syntax. (Wittgenstein 1921, 3.325)

Wittgenstein's symbol/sign distinction reminds one very much of Peirce's more familiar type/token distinction. Instead of saying that the word ›count‹ has two meanings, it could be said with Wittgenstein that there are two different *symbols* (types) which have the one *sign* (token) ›count‹ in common. The connection between the symbol and its meaning is constant; it is necessary to refer to the context of the sign—its position in the sentence—to ascertain what its correct symbol is, since a sign in isolation cannot have a meaning. Hence, the analysis of the use of signs in sentences reveals their corresponding symbols and thereby their logical form (see (Wittgenstein 1921), 3.326 ff.). The first step is a kind of optional disambiguation from sign to symbol; the second step the recognition of the symbol's logical form.

Once this is established, syntactical mistakes can be recognised. This does apply to more ordinary syntactical mistakes (»table chair« is not a meaningful combination of names) as well as to the not-so-obvious logical paradoxes: In ordinary language, some sentences do occur in which there seems to be a combination of symbols leading to a kind of vicious circle. In analysis, however, these problems disappear: by regarding the sign's context one can get from the sign to the correct symbol; disambiguation takes place. Wittgenstein gives an example:

> The reason why a function cannot be its own argument is that the sign for a function already contains the prototype of its argument, and it cannot contain itself.
> For let us suppose that the function F(fx) could be its own argument: in that case there would be a proposition »F(F(fx))«, in which the outer function F and the inner function F must have different meanings, since the inner one has the form φ(fx) and the outer one has the form ψ(φ(fx)). Only the letter »F« is common to the two functions, but the letter by itself signifies nothing.
> This immediately becomes clear if instead of
>
> »F(Fu)« we write »(∃φ):F(φu) . φu=Fu«.
>
> That disposes of Russell's paradox. (Wittgenstein 1921, 3.333)

One might have strings of growing complexity, *Fu*, *F*(*Fu*), *F*(*F*(*Fu*)), … in which similar signs ›F‹ occur in different positions. Analysis reveals that, though the different symbols' signs ›F‹ are identical, every sign belongs to a different symbol. This is exactly the approach that can be found in the STT. There, similar symbols (not to be understood in Wittgenstein's sense)—*e.g.*, the relation of identity—do appear on different types, *i.e.*, are systematically ambiguous. In the example above, each step to a more complex string can be regarded as a step from one type to the next in Russell's STT. Without some explicit indicator, *e.g.*, its type attached as an index to the symbol (which would be nothing but a disambiguation of the symbol, of course), Russell would have to resort to the context of the symbol, *i.e.*, its arguments, as well, to get to know its specific type. The last sentence in the previous citation makes the whole matter clear: this kind of disambiguation is the key to the solution of the logical paradoxes (of which Russell's paradox is the most well-known and explicitly mentioned by Wittgenstein), and both Russell and Wittgenstein offer means to solve the logical paradoxes by disambiguation. In Russell's as well as in Wittgenstein's ideal language there is exactly one name for each object. So, on the most

[1] I have argued for this approach in my (Leerhoff 2008).

fundamental level, when analysis is done, there is no room for ambiguities nor, as a consequence, for the logical paradoxes, which can no longer be formulated.

The semantical paradoxes

Both the STT and Wittgenstein have similar techniques to avoid the kind of systematic ambiguity involved in the logical paradoxes. The semantical paradoxes, on the other hand, are *much* more complicated to avoid. They can still be formulated, and they are still paradoxical in an ideal language of Russell's kind with only STT-restrictions. Hence Russell developed the RTT to guard his language against them. In a nutshell, the states of affairs described in the semantical paradoxes can still be expressed in the ideal language, but the RTT enforces a non-paradoxical ›translation‹ for them. As I have stated above, there is a high price to pay for this: the RTT is extremely complicated and, at least for some areas of application, further axioms have to be postulated.

How does Wittgenstein's solution of the semantical paradoxes fare in this respect? In all these paradoxes some semantical (or, in Ramsey's word, ›psychological‹) terms play a crucial role, *e.g.*, ›truth‹, ›naming‹, ›lying‹, etc. In Wittgenstein's ideal language, there are no and can be no expressions for these ordinary-language terms, so the whole question of semantical paradoxes is a non-issue for Wittgenstein. This, of course, is a high price to pay as well, since it sets definite limits to the areas of application for the language. In Wittgenstein, these limits do not result from the threat posed by the semantical paradoxes; their ›solution‹ has to be regarded as a kind of side effect of limits that are grounded in the *Tractatus'* concept of language.

Conclusion

Ramsey's distinction of the paradoxes in logical ones on the one hand and psychological (or semantical) ones on the other proves to be valuable for an examination of Wittgenstein's alternative to Russell's Theory of Types. The logical paradoxes pose a threat for Wittgenstein's system as well as for Russell's. Since Wittgenstein cannot integrate a type theory in his system, he offers an alternative approach to the disambiguation of terms, which is the key to the solution of those paradoxes. His way of solving these problems has striking similarities to Russell's STT. The semantical paradoxes, however, do pose a threat for Russell's system, but not for Wittgenstein's. This difference is due to the diverging concepts of language in their respective variants of logical atomism.

Acknowledgements

This paper draws on results from an ongoing research project commissioned by the *Landesstiftung Baden-Württemberg*.

Literature

Church, Alonzo 1940. “A Formulation of the Simple Theory of Types”, *Journal of Symbolic Logic* 5, 56–68.

Davant, James B. 1975. “Wittgenstein on Russell's Theory of Types”, *Notre Dame Journal of Formal Logic* XVI, 102–108.

Leerhoff, Holger 2008. Logische Form und Interpretation. Eine systematisch-historische Untersuchung des Logischen Atomismus. Paderborn: mentis 2008.

Ramsey, F. P. 1926. “Mathematical Logic”, *The Mathematical Gazette* 13, 185–194.

Russell, Bertrand 1903. *The Principles of Mathematics.* (1996) New York, London: W. W. Norton.

Russell, Bertrand 1908. “Mathematical Logic as based on the Theory of Types”, *American Journal of Mathematics* 30, 222–262.

Whitehead, A. N. and Russell, Bertrand 1910–13. *Principia Mathematica*. 3 Vols., Cambridge: At the University Press.

Wittgenstein, Ludwig 1921. *Tractatus Logico-Philosophicus*. (1961) D. F. Pears and B. F. McGuinness (trans.), New York: Humanities Press.

The Tractatus and the Problem of Universals

Eric Lemaire, Paris & Nancy, France

A) The problem of objects

What we called the problem of objects covers different questions. We can distinguish the metaphysical problem and the epistemological one. The metaphysical one is concerned with 1) the existence of universals and particulars and 2) with the nature of those things. The epistemological one is "Have we some sort of knowledge of objects?" and, if yes, "What kind of knowledge?". Here we will be interested with the problem of the existence of universals.

The problem of objects received numerous answers. Without details, we can give an idea of the different points of view. The important point is that there is no consensus about the community. The difficulty with that problem is that there is no textual evidence to support one of the possible solutions. We can briefly expose some of the existing solutions. From our two principal questions about the objects, we can classify the authors. 1) The *Epistemological question*: How can we know the objects? 2) The *Ontological* question: What is their ontological status?

Those who think that we can answer to the two questions.

Jaakko and Meril Hintikka (1989) argued that the tractatian's objects are russellian's disguised objects[1] except the logical one. So objects can be known by acquaintance and ontologically they are particulars and universals.

Those who answers the ontological question and deny we can answer the epistemological one.

Peter Hacker (1972) affirmed that the objects are universals and we cannot say that they are objects of acquaintance.

Elizabeth Anscombe and Irving Copi, think that the objects are particulars Note and we cannot say that they are objects of acquaintance.

Those who think we can answer neither the ontological nor the epistemological question.

David Pears (1988), Anthony Kenny (1973) think that we cannot answer these questions because Wittgenstein does not know. They consider this as a lack.

Sebastian Gandon (2003) asserts that Wittgenstein does not know but that is not a lack. He believes the necessity to answer the questions is a delusion[2].

B) First argument

Our first argument is ground on the distinction between a proposition and a name. We will not explain in details what exactly Wittgenstein's conception of these two things is. We need not to do that. Nevertheless, we should say that by name we mean *real name*, that is a complete symbol and not a description or a logical fiction. A proposition is distinguished from a name by being in a relation of denotation with the world, whereas a name means. What is the difference? When some sign which can denote something do not actually refer to a fact it does not loose its sense. Consider the following example: Jones says that "Brandy is a nice cat". In that case, even if Brandy is not nice, the proposition Jones pronounced is perfectly intelligible or has a sense. But Wittgenstein thinks that a name is not mere noises if and only if a name has a meaning, or is related to an object in reality. It is some kind of rigid designator. It is difficult to illustrate this conception with some example for nobody, even Wittgenstein, has found real name. In fact, the only plausible candidate I can see is "this". A real name may refer to an object directly experienced. If this is true, this means that Wittgenstein endorses a russellian epistemology as Jaako and Meril Hintikka affirmed it. But we do not want to discuss the very controversial point here. The point here is that Wittgenstein needs to make this distinction in order to differentiate the symbolic behaviour of a proposition from the one of a constituent. A proposition is bipolar, which means that a proposition necessarily can be true or false otherwise it is just nonsense. In other words, there is an internal relation between a proposition and its truth-conditions. And a name needs to be related to an object in order to safe the sense of a proposition.

Another crucial distinction between a real name and a proposition is that a name is a simple symbol which means a simple object, whereas a proposition is a complex symbol which denotes a complex of objects (state of affairs or facts).

Now, suppose that Wittgenstein's ontology is nominalist, that is only concrete particulars exist. We said that a proposition is necessarily a complex otherwise the essential feature of propositions could not be bipolarity. So, at least, a proposition has two constituents. That is, according to the nominalist interpretation, a proposition is necessarily composed of two concrete particulars. But this is very doubtful. How could there be two concrete particulars in "This is red" or "Jones is nice" or "this painting is beautiful". It seems very counterintuitive. And suppose that a proposition is composed of only one constituent. In that case, we cannot say that for the simplest proposition, there is a difference between a proposition and a name. But the distinction between a proposition and a name is a crucial one in the tractarian's system. In fact, When Wittgenstein criticized Frege and Russell, he precisely insisted on this point.

C) Second argument

Our second argument focuses on the notion of concrete particulars and the notion of property. Each object, Wittgenstein says, has internal and external properties. There is no doubt that Wittgenstein thought that there are concrete particulars or individuals. Why?

In asking whether there is individuals or concrete particulars, we do not mean "Are concrete particulars reducible to universals?". If there are only universals, no

[1] Russell's thought during the two first decades of the twentieth century changed a lot. The expression "russellian's objects" refers to his posthumous book written in 1913 *Theory of knowledge*.

[2] The supporters of the New Wittgenstein did not give any direct interpretation of the problem. However, we can consider, even if he probably disagrees with this, that Sebastian Gandon's book is a speech for the defence of the Diamond-Conant's point of view.

concrete particulars can be subject of a proposition. So the only propositions which exist are proposition as “Blueness is a colour” or “Triangularity is a shape” in which no concrete particular is referred to. Moreover, we should note that if it is true, Wittgenstein’s conception of universals is Platonist because these universal exist independently of any concrete instance. In an Aristotelian conception, the existence of universal is dependent upon the existence of their instances. If there are only platonic universals, the only propositions we could make are necessary: “Triangularity is necessarily a shape”, “Blueness is necessarily a colour”, etc. Platonic universals could be constituents in contingent propositions such as “I thought to Blueness this morning”. But such propositions suppose the existence of concrete particulars (me, this morning, etc.) But it is evident that Wittgenstein’s conception of propositions is not compatible with this. Every proposition with a sense is contingent. Necessary propositions are logical and empty of sense, or metaphysical and without sense. For that reason, we believe that Wittgenstein’s conception of language and the world implies the existence of concrete particulars. One could reply that in a bundle theory of particulars, the only things that exist are universals. A proposition is for example “Blueness is co-present with Squareness”. And this is a purely contingent proposition for Blueness could be co-present with Roundness. But in that analysis, you said that concrete particulars are reducible to co-instantiated universals.

There are three conceptions of concrete particulars: 1) the bundle theory, 2) the substratum theory, and 3) the substance theory. What do they say? According to the Substratum theory, “a concrete particular is a whole made up of the various properties we associate with the particular together with an underlying subject or substratum that has an identity independent of the properties with which it found – a bare particular.[3]” According to the Bundle theory “There are no underlying substrata; ordinary particulars are constituted exclusively by the properties associated with them, there are “bundles” or “clusters” of those properties.[4]” These two theory share the common assumption that a concrete familiar particular (like a chair) is not a basic entity but is a whole made up of more basic constituents. There is a third theory: the substance theory. This theory takes the concrete particulars to be ontologically basic entities. There are not reducible to properties or to a bare substratum. Another point suggests a crucial difference between these three theories: the question of identity. The bundle theory is an essentialist one. This means that each property is essential. If a bundle looses one of its properties, it becomes another thing. The substratum theory is anti-essentialist because the very identity of the particular is assumed by the bare substratum, so each property is contingent. The identity of the particular does not depend upon its properties. In a substance theory, a particular has essential and inessential properties. Essential properties are generally thought as Kinds (Universal). A Kind term show what is the particular. For example, if you ask “What is Boby?”, the answer is “Boby is a man.”. But there are other properties which permit us to answer the question “How is it?”. For example: “Boby is beautiful”. The substance theory is generally understood as a realist one, which commits us to the existence of universal. Kinds are typically universals. The other theories are compatible with a trope-theory. And typically, an austere nominalist thinks that we cannot investigate the ontological structure of concrete particulars[5]. We do not want to discuss the merits or difficulties of each theory here. Now, come back to Wittgenstein. We just saw that the substance theory distinguish the following questions: 1) What is it?; 2) How is it?. Wittgenstein clearly distinguish these one too. For example take the remark 3.221. And in his ontology, he insists on the fact that objects have internal and external properties. The internals properties are such that it is unthinkable that the object do not possess them. And external properties can be possessed or not by the objects. The last thing is a *matter of fact*. In fact, Wittgenstein seems to avoid problems met by the Bradley and Russell in their account of relations. According to Bradley, each property of a particular is inherent to it or internally related to it. Each property is essential to the particular. According to Russell, there is no internal relation. Then, each property is externally related to a particular. That the particular possesses a property is always a matter of fact. There is no essential property. Bradley and Russell seem to be committed respectively to a bundle theory and a Substratum theory. So he seems to hold a substance theory of concrete particulars (Except for the subject of thoughts for which he seems to maintain a substratum theory.)

So we have presented two arguments in support of a realist interpretation of the *Tractatus*. Obviously, many things should be said to reply objections or to clear up our discussion. But we lack place. So, the discussion, I hope, will serve to answer questions and perplexities.

Literature

ARMSTRONG David, *Nominalism and realism, Universal and scientific realism*, volume 1 and 2, Cambridge University Press, Melbourne, 1978.

AUNE Brian, *Metaphysics The Elements*, University of Minnesota Press, Minneapolis, London, 1985.

HINTIKKA, Jaako and Méril, *Investigating Wittgenstein*, 1989, New York, Blackwell.

LOUX Michael J, *Metaphysics a contemporary introduction*, Third edition, Routledge, New York, 2006.

LOWE Jonhatan E, *The possibility of metaphysics*, Oxford University Press, New York, 1998.

Mc DONALD Cynthia, The variety of Things, foundations of contemporary metaphysics, Routledge, Victoria, 2005

Van INWAGEN Peter and ZIMMERMAN Dean (edition), *The Oxford Handbook of Metaphysics*, Oxford University Press, New York, 2003.

WITTGENSTEIN Ludvig, *Tractatus logico philosophicus*, 1961, English Translation by Pears David and Mc Guinness Brian, London, Routledge.

[3] Loux, 2006, 84.

[4] Ibid.

[5] Loux 2006, p 106-7 and Chapter 2.

A Critique of the Phenomenal Concept Strategy

Daniel Lim, Cambridge, England, UK

The Strategy

The *locus classicus* of the phenomenal concept strategy is Brian Loar's paper 'Phenomenal States' (1990). In it he claims that the Knowledge Argument relies on a dubious assumption which he dubs the Semantic Premise:

> "A statement of property identity that links conceptually independent concepts is true only if at least one concept picks out the property it refers to by connoting a contingent property of that property." (Loar 2004, 224)

He argues that the Semantic Premise, while true in standard cases of a posteriori identities like 'water is H_20', is crucially false in cases of a posteriori psychophysical identities. This is because the *phenomenal* concepts deployed in psychophysical identity statements are really type-demonstrative concepts of the form: '*this* experience'. The two features of type-demonstratives relevant for the strategy are: direct reference and conceptual independence from physical / scientific concepts. The first feature ensures that no *new* properties are introduced and the second feature ensures that the identity remains a posteriori.

The phenomenal concept strategy has the makings of a powerful response to the Knowledge Argument and defenders of the strategy claim that it satisfies three important desiderata: (i) it respects the kind of knowledge Mary gains after leaving her black and white room, (ii) it is physically explicable, and (iii) it explains why we cannot resist the illusion of ontological distinctness concerning our conscious experiences.

Phenomenal Knowledge and Physical Explicability

While there are independent reasons[1] for eschewing a type-demonstrative construal of phenomenal concepts I wish to dwell on a dilemma David Chalmers (2007) has been keen on exposing which brings features (i) and (ii) into tension. If phenomenal concepts are physically explicable they will not explain our epistemic situation. On the other hand, if phenomenal concepts can explain our epistemic situation then they will be physically inexplicable. Chalmers focuses his critique on issues of conceivability, but I will focus on the putative physical mechanisms that make phenomenal concepts possible. A nice way of bringing out the tension in this dilemma is to use Janet Levin's and David Papineau's positions as exemplars of each horn.

Levin's position exemplifies the first horn. She deliberately avoids any account of phenomenal concepts that require anything that is physically suspect: 'quotation', 'partial constitution', or 'acquaintance'. She opts for a physically *safe* version of phenomenal concepts that is limited to "causation, reliable correlation, and relations of physical inclusion or adjacency". As such she argues that phenomenal concepts should be construed as introspectively deployed demonstratives *and nothing more*. All that is needed to distinguish introspectively deployed phenomenal demonstratives from nonphenomenal ones are 'differences in what they [causally] denote'. This is worrisome because it drives, what seems to be, an unacceptable wedge between phenomenal concepts and the properties they denote. By relating the two by causation too much distance has been allowed to creep into the picture. In David Chalmers' terminology, this makes phenomenal concepts, in a sense, 'Twin Earthable'. The referents of Twin Earthable concepts will be unstable across counterfactual worlds. When Twin Oscar thinks, while on Twin Earth, that the substance in the lake looks refreshing, his thoughts about the substance will be twin water thoughts. This is because XYZ, and not H_20, causes Twin Oscar's concept of 'water' to be tokened. When Oscar thinks, while on Earth, that the substance in the lake looks refreshing, he will entertain water thoughts because they are caused by H_20.

To apply this insight to differentially caused type-demonstratives, we can imagine a scenario where I am observing a lake. The lake is partitioned into sections. Scientists, who have managed to transport a sizeable amount of XYZ from Twin Earth, have filled some partitions with XYZ and others with H_20. While looking at partition *A* I deploy a type-demonstrative '*that* liquid' and while simultaneously looking at partition *B* I deploy, what I think is, the same type-demonstrative '*that* liquid' and think in my mind: '*that = that*'. I am wrong about this since *A* is filled with XYZ and *B* is filled with H_20. However, I am wrong not because I *misapplied* one of the demonstratives, but because I unknowingly deployed two different concepts. This is because my concepts are causally individuated by the objects they denote. While attending to the liquid in *A* I may have thought that I was deploying a water concept, when in fact I was deploying a twin water concept.

Applying this scenario to our own phenomenal states[2] it is possible for a normal subject to deploy a type-demonstrative '*that* experience' while attending to the same phenomenal property twice in quick succession and yet have room to rationally doubt whether '*that* = *that*' is in fact true. Levin writes:

> "…it may seem epistemically odd that introspecting subjects can be mistaken about whether they're using the same concepts in their thoughts about their own phenomenal states. But when concept difference and identity are determined 'externally' – that is, by the features of what's denoted – this shouldn't be unexpected, even when the subject matter is one's own mental states." (Levin 2007, 108)

This makes the following scenario possible: I may mistakenly *think* that I'm deploying a phenomenal concept when in fact I am not. Let's say that brain state p_1 is my

[1] For example see Diana Raffman's (1995) critique of type-demonstratives based on the empirical fact that we can discriminate more colors than we can re-identify over time.

[2] This is a thought-experiment John Hawthorne (2007) develops against direct reference theories of phenomenal concepts used to defend property dualism.

phenomenal concept of experiencing red. This is because brain state p_2, which is my experience of red, was causally responsible for p_1. Ex hypothesi, p_1 and p_2 are distinct physical states. This would make it possible for a neuroscientist to stimulate p_1 while ensuring that p_2 is left inert. But what exactly is happening in this scenario? If it is anything like the water case, I will think that I'm entertaining a phenomenal concept that refers to p_2 but I will actually be deploying a nonphenomenal concept of the neuroscientist's electrical stimulation. This means that I will mistakenly believe that I am thinking 'phenomenally' about experiencing red when I am actually thinking nonphenomenally about something else.

If we accept Kripke's observation that there is no reality / appearance distinction when it comes to our phenomenology then this seems highly implausible. But more importantly, it is imperative that the phenomenal concept strategist develop an account of phenomenal concepts that is genuinely phenomenal and Levin, by keeping her account physically respectable, loses touch with this crucial feature. She preserves (ii) at the expense of (i). This shortcoming points us to a different account, one that makes the relationship between the phenomenal concept and the object it denotes much tighter.

David Papineau's account of quasi-quotational phenomenal concepts discharges this duty. His idea is to analyze phenomenal concepts as a species of perceptual concepts. Consider my ability to perceptually identify a flower and make judgments about it. I might be able to think '*that* flower' is beautiful when looking at a tulip. I have, at my disposal, a visual sensory template that I can use to recognize tulips. This template, however, is not only useful for thinking about external objects like tulips, but it is also useful for thinking about internal objects like my conscious awareness of the tulip. When we attend to our experiences we use the experiences themselves to think about them. The crucial feature of phenomenal concepts is that they will always deploy an instance of the experience they are about. That is, they *use* the denoted experiences in order to *mention* them. Papineau writes:

> "This means that any exercise of a phenomenal concept to think about a perceptual experience will inevitably involve either that experience itself or an imaginary recreation of that experience. If we count imaginary recreations as 'versions' of the experience being imagined, then we can say that the phenomenal thinking about a given experience will always *use* a version of that experience in order to *mention* that experience."
> (Papineau 2007, 124)

The problem with this view is that it seems to leave no room for any kind of physical explanation. The phenomenal property that the phenomenal concept is about is literally a part of the concept itself; as such the traditional distinction between representation and represented object that made it possible to specify an explanation in terms of causal / historical correlation between the two has been obliterated. It seems that the special kind of cognitive presence that Papineau's account provides must be explained by the physical presence of the experience alone. But does this really count as an explanation or is this better characterized as a stipulation?

Dualist Illusion

I believe Chalmers' dilemma is real and must be addressed. But even if a path can be cut between the horns of this dilemma it is far from clear that the phenomenal concept strategy has adequately accounted for (iii). For ease of exposition I will use the 'explanatory gap' (Levine 2001) as an instance of the dualist intuition. The existence of this gap is something defenders of the phenomenal concept strategy readily admit. Their claim is not that the gap can be bridged but that its existence can be explained.

Let's begin by thinking about a classic optical illusion based on human color constancy known as the 'checker shadow illusion'. A subject is shown what appears to be a black and white checkerboard with a cylinder on it that is casting a diagonal shadow across the middle of the board. The squares are actually different shades of gray and the image is constructed so that the 'white' squares in the shadow are the same shade as the 'black' squares outside the shadow. Despite being the same shade the squares *appear* to be very different. The standard explanation for this illusion has two parts. The first is that our visual system keys in on local contrasts – a square that is lighter than its immediate neighbors is considered 'lighter than average'. So the mere fact that the square in the shadow is surrounded by darker squares and the square outside the shadow is surrounded by lighter squares contributes to the illusion. The second part is that shadows often have soft edges while painted boundaries have hard edges. Our visual system tends to ignore gradual changes in lighting in order to determine the color of the surfaces involved without being misled by shadows. In the image, the cylinder's shadow is deliberately made fuzzy to add to this effect.

What we have here is a satisfying explanation for the existence of the checker shadow illusion. A vital feature of this explanation is that it does not *presuppose* the illusion in order to explain it. By exposing the tendencies of our visual system to key in on local contrasts and import generic information concerning boundaries, a non-circular explanation of the illusion is provided. The question is whether phenomenal concepts provide such an explanation for the explanatory gap. David Papineau claims that it does. He locates the source of the explanatory gap in the *absence* of a use / mention distinction regarding phenomenal concepts. The use / mention distinction is present in a majority of our nonphenomenal concepts but it is peculiarly absent when we think phenomenally. As already gestured at above, it is the way we *think* about phenomenal properties that makes it *seem* as though we are apprehending their essences and consequently intuiting phenomenal properties as ontologically distinct from anything physical. Phenomenal concepts literally contain the properties that they conceptualize so it is no wonder why we can disassociate the physical descriptions of the properties from the first hand experiences of the properties themselves. Papineau writes:

> "There is a sense in which material concepts do 'leave out' the feelings. Uses of them do not in any way activate the experiences in question, by contrast with uses of phenomenal concepts … After all, most concepts don't use or involve the things they refer to. When I think of being rich, say, or having measles, this doesn't in any sense make me rich or give me measles."
> (Papineau 2007, 136)

I agree that I do not have to have measles in order to think about measles. In fact, in most cases, I don't. I can think of the disease in terms of its characteristic symptoms, its scientific classification rubeola or even its distinction from smallpox among other things. But, of course, it is also possible to think about measles while actually having measles. I can think, using a type-demonstrative, '*this* disease', where I use the disease I currently have in order to mention it. If Papineau is right, juxtaposing this type-demonstrative concept with my ordinary measles concept should generate an illusion of distinctness. It is this 'special way' of thinking that creates the fallacious impression that other ways of thinking about measles fail to refer to the measles themselves. But is this at all convincing? Despite the absence of the use / mention distinction in my measles type-demonstrative I have no dualist illusions. I don't find myself thinking that I'm dealing with two ontologically distinct entities.

Since the absence of a use / mention distinction is a characteristic of certain nonphenomenal type-demonstratives that do not engender a dualist illusion, the absence of the distinction cannot be used to explain the illusion in the case of phenomenal concepts. Without offering a principled distinction between type-demonstratives like '*this* disease' as opposed to '*this* experience' there is no reason to think that a satisfying explanation for our dualist intuitions has been given. Of course one can appeal to the peculiarity of the phenomenal property itself in contrast to the property of being rubeola but this would presuppose the explanatory gap, not explain it. This move effectively eliminates the possibility of satisfying (iii) and it is a failure of the phenomenal concept strategy that is often overlooked. Most debates concerning the strategy assume that (iii) is adequately addressed, but further examination shows that this is a mistake. Referring to the absence of the use / mention distinction to explain our dualist intuitions simply does not work.

Conclusion

The phenomenal concept strategy faces some serious problems. Chalmers' dilemma is real, that is, it does not seem that (i) and (ii) can consistently be held together in a single account. However, even if this dilemma can be resolved it cannot account for (iii) because it does not provide us with a satisfying non-circular explanation for the dualist illusion.

Literature

Chalmers, David 2007 "Phenomenal Concepts and the Explanatory Gap", in: Torin Alter and Sven Walter (eds.), *Phenomenal Concepts and Phenomenal Knowledge*, Oxford: Oxford University Press, 167-194.

Hawthorne, John 2007 "Direct Reference and Dancing Qualia", in: Torin Alter and Sven Walter (eds.), *Phenomenal Concepts and Phenomenal Knowledge*, Oxford: Oxford University Press, 195-209.

Jackson, Frank 1982 "Epiphenomenal Qualia", *Philosophical Quarterly* 32, 127-136.

Kripke, Saul 1982 *Naming and Necessity*, Cambridge, Mass.: Harvard University Press.

Levin, Janet 2007 "What is a Phenomenal Concept?", in: Torin Alter and Sven Walter (eds.), *Phenomenal Concepts and Phenomenal Knowledge*, Oxford: Oxford University Press, 87-110.

Levine, Joseph 2001 *Purple Haze: The Puzzle of Consciousness*, Oxford: Oxford University Press.

Loar, Brian 1990 "Phenomenal States", *Philosophical Perspectives* 4, 81-108.

Loar, Brian 2004 "Phenomenal States (Revised)", in: Peter Ludlow, Yujin Nagasawa and Daniel Stoljar (eds.), *There's Something About Mary*, Cambridge, Mass.: MIT Press, 221-239.

Papineau, David 2007 "Phenomenal and Perceptual Concepts", in: Torin Alter and Sven Walter (eds.), *Phenomenal Concepts and Phenomenal Knowledge*, Oxford: Oxford University Press, 111-144.

Raffman, Diana 1995 "On the Persistence of Phenomenology", in: Thomas Metzinger (ed.), *Conscious Experience*, Exeter: Imprint Academic, 293-308.

Metaphorische Bedeutung als *virtus dormitiva*

Jakub Mácha, Brno, Tschechien

Die Diskussion über das Wesen der Metapher steht in der analytischen Philosophie hauptsächlich unter dem Einfluss von Arbeiten Max Blacks und Donald Davidsons. Ihre Querele betrifft vornehmlich die Problematik der *metaphorischen Bedeutung*. Black behauptet, dass der Begriff der metaphorischen Bedeutung notwendig sei, um die spezifische kognitive Kraft der Metapher zu beleuchten. Davidson leugnet das, indem er eine Reihe von Argumenten gegen die Idee der metaphorischen Bedeutung präsentiert. Im Weiteren soll ein von denen untersucht werden. Das abzuhandelnde Argument betrifft die Tauglichkeit einer solchen Idee zur Erklärung dessen, wie Metaphern funktionieren und verstanden werden. Es gehört allerdings zu einer generelleren Denkfigur, die in der Geistesgeschichte längst bekannt ist. Donald Davidson hat sie keineswegs erfunden und auch in ihrer Anwendung auf die Erklärung der Metapher behält er kein Primat. Es sei vorausgeschickt, dass der Gegenstand dieses Aufsatzes keine Beurteilung der Figur selbst, sondern nur ihre Ausprägung in der Metapherdiskusion sein soll.

Es war niemand geringerer als Friedrich Nietzsche, der in seiner berühmten Kritik an Kant schrieb:

> Wie sind synthetische Urteile *a priori möglich*? fragte sich Kant, - und was antwortete er eigentlich? *Vermöge eines Vermögens* […] – hatte er gesagt, mindestens gemeint. Aber ist denn das – eine Antwort? Eine Erklärung? Oder nicht vielmehr nur eine Wiederholung der Frage? Wie macht doch das Opium schlafen? »Vermöge eines Vermögens«, nämlich der *virtus dormitiva* – antwortet jener Arzt bei Molière
>
> *quia est in eo virtus dormitiva,*
> *cujus est natura sensus assoupire.*
>
> Aber dergleichen Antworten gehören in die Komödie […]. (Nietzsche 1954, 575f)

Hier ist nicht der Ort, um die Entscheidung zu treffen, ob Nietzsche mit diesem Tadel Recht gehabt hat. Lediglich die Struktur des Arguments soll uns interessieren: Wie funktioniert die Leistung *Y* des Systems *X*? Vermöge der Tatsache, dass *X* die Eigenschaft *Z* besitzt. Aber das ist keine Erklärung, denn *Z* stellt nur eine Umschreibung von *Y* dar oder die Eigenschaft *Z* ist genau so unbekannt wie die Leistung *Y* von *X*. Für die Variable *X* können beispielsweise die Ausdrücke „synthetische Urteile a priori", „Opium", oder „Metapher" eingesetzt werden. Für *Y* sind es „Möglichkeit", „Schlafen", „Verstehen", für *Z* daraufhin „Vermögen", „*virtus dormitiva*" oder „metaphorische Bedeutung". Also wie sind Metaphern als solche zu verstehen? – Vermöge einer spezifischen metaphorischen Bedeutung. Aber das Verstehen von Metaphern ist genau so erklärungsbedürftig wie jene metaphorische Bedeutung.[1]

Ohne den Ausdruck „Metapher" zu gebrauchen, hat Ludwig Wittgenstein in seinen *Philosophischen Untersuchungen* das Problem der Erklärung der uneigentlichen Rede aufgegriffen.

> Gegeben die beiden Begriffe 'fett' und 'mager', würdest Du eher geneigt sein, zu sagen, Mittwoch sei fett und Dienstag mager, oder das Umgekehrte? (Ich neige entschieden zum ersteren.) Haben nun hier "fett" und "mager" eine andere, als ihre gewöhnliche Bedeutung? — Sie haben eine andere Verwendung. — Hätte ich also eigentlich andere Wörter gebrauchen sollen? Doch gewiß nicht. — Ich will *diese* Wörter (mit den mir geläufigen Bedeutungen) *hier* gebrauchen. — Nun sage ich nichts über die Ursachen der Erscheinung. Sie *könnten* Assoziationen aus meinen Kindheitstagen sein. Aber das ist Hypothese. Was immer die Erklärung, — jene Neigung besteht.
>
> Gefragt, "Was meinst Du hier eigentlich mit 'fett' und 'mager'?" — könnte ich die Bedeutungen nur auf die ganz gewöhnliche Weise erklären. Ich könnte sie *nicht* an den Beispielen von Dienstag und Mittwoch zeigen. Man könnte hier von 'primärer' und 'sekundärer' Bedeutung eines Worts reden. Nur der, für den das Wort jene Bedeutung hat, verwendet es in dieser.
>
> […]
>
> Die sekundäre Bedeutung ist nicht eine 'übertragene' Bedeutung. Wenn ich sage "Der Vokal *e* ist für mich gelb", so meine ich nicht: 'gelb' in übertragener Bedeutung — denn ich könnte, was ich sagen will, gar nicht anders als mittels des Begriffs 'gelb' ausdrücken. (Wittgenstein 2000, Artikel 144, S. 79f, Reinschrift des II. Teils der Untersuchungen.)

Um die Intention des Philosophen transparent werden zu lassen, soll noch ein anderer Kommentar aus dem Nachlass vorgelegt werden:

> Könnte man hier von 'primärer' und 'sekundärer' Bedeutung eines Worts reden? — Die Worterklärung ist beide Male die der primären Bedeutung. Nur für den, der das Wort in jener Bedeutung kennt, kann es diese haben. D.h. die sekundäre Verwendung besteht darin, daß ein Wort, mit *dieser* primären Verwendung, nun in dieser neuen Umgebung gebraucht wird.
>
> Insofern könnte man die sekundäre eine 'übertragene' Bedeutung nennen wollen.
>
> Aber das Verhältnis ist hier nicht, wie das zwischen dem 'Abschneiden eines Fadens' und 'Abschneiden der Rede', denn hier *muß* man ja nicht den bildlichen Ausdruck gebrauchen. […]
>
> Man sagt nur von solchen Kindern, sie spielen Eisenbahn, die von einer wirklich Eisenbahn wissen. Und das Wort Eisenbahn im Ausdruck "Eisenbahn spielen" ist nicht bildlich gebraucht, oder im übertragenen Sinn. (Wittgenstein 2000, Artikel 138, S. 12b–13a, Band S, 31. Januar 1949)

[1] Es lassen sich viele andere Belege für diese Denkfigur erbringen: Wie werden ästhetische/moralische Urteile zustande gebracht? Vermöge eines ästhetischen/moralischen Sinns. Wie ist unmittelbares Wissen möglich? Vermöge der Intuition – d .h. vermöge einer Sehkraft (vgl. das lateinische Verbum *in-tueri* – genau auf etwas hinsehen) Oder warum waren bestimmte politische Subjekte erfolgreich? Sie besaßen eine geheimnisvolle Eigenschaft oder Kraft – die *virtù*, und man neigt dazu, auch den Willen zur Macht in diese Liste einzutragen.

Wittgenstein erbrachte mehrere Beispiele, um an ihnen die Idee der sekundären Bedeutung zu verdeutlichen. Umschreiben wir drei von ihnen in prädikativer Form:

(1) Mittwoch ist (für mich) fett.
(2) Der Vokal *e* ist (für mich) gelb.
(3) Die Rede ist abgeschnitten worden.

Alle diese Prädikate sollen eigentlich materiellen Gegenständen zukommen, was hier ersichtlich nicht der Fall ist. Daher legt sich die Auffassung nahe, dass die Wörter „fett“ oder „gelb“ in einer sekundären Bedeutung gebraucht worden sind. Wenn dies so möglich wäre, würde Wittgensteins fundamentales Argument gegen die private Sprache unterminiert werden. Denn die sekundären Bedeutungen wären in einem signifikanten Sinne *privat*, was durch die eingeklammerten Einschübe „für mich“ hervorgehoben sein solle.

Um den Anschein einer privaten Sprache abzuweisen, muss eine Erklärung geliefert werden, welche die stets hypothetischen sekundären Bedeutungen kundgeben soll. Da diese Erklärung nicht vermöge einer ostensiven Handlung erfolgen kann (die Beispiele sind so konstruiert worden), muss sie auf die ganz gewöhnliche Weise gemacht werden. – Das heißt nicht anders als mithilfe primärer Bedeutungen der involvierten Wörter: Mittwoch sei *für mich* fett, weil ich als Kind mittwochs viel Fettes zu essen pflegte; der Vokal *e* sei *für mich* gelb, weil in meiner ersten Fibel der Buchstabe *e* gelb aufgemalt worden ist. Der Gebrauch der Wörter „fett“ sowie „gelb“ ist in den kausalen Nebensätzen durchaus buchstäblich, d. h. ausschließlich ihre primären Bedeutungen werden in Anspruch genommen.

Primär oder sekundär sind demzufolge nicht Bedeutungen, sondern Verwendungen – nämlich die Stellung der Wörter in ihren spezifischen Umgebungen. In eine neue Umgebung oder in einen ungewöhnlichen Kontext wird also ein Wort samt seiner Bedeutung verlegt. Sobald die Erklärung geliefert ist, ist das Uneigentliche dieses Gebrauchs nicht mehr vorhanden.

Zum Kontrast sei die offensichtlich tote Metapher (3) gegenübergestellt. Das Verb „abschneiden“ verfügt über eine primäre Bedeutung „etwas auf eine feinere Weise abtrennen“ und über die Vielheit von bildlichen Bedeutungen wie „eine Rede abscheiden“, „einen Weg abschneiden“ oder „bei einer Prüfung abschneiden“ usw.

Das Argument wird bei Wittgenstein möglicherweise in einer nicht so expliziten Form dargestellt wie in dem obigen Nietzsche-Zitat. Die Struktur ist doch dieselbe, nur der Argumentationsgang erfolgt umgekehrt: Man neigt dazu, die Beispielsätze *X* durch sekundäre Bedeutungen bestimmter Wörter *Z* zu erklären. Aber um diese Sätze verständlich (*Y*) zu machen, werden weitere Erklärungen *Z’* erforderlich, welche sich alleinig auf buchstäbliche Bedeutungen berufen dürfen. Daher tritt die Erklärung *Z’* anstelle der Erklärung *Z* oder diese wird durch jene fixiert.

Für unsere nachfolgende Betrachtung ist wichtig: Das Konzept der sekundären Bedeutung ist nicht unsinnig, disponiert jedoch über keine Erklärungskraft. Seine Berechtigung besteht lediglich darin, dass weitere Erklärungen zu erwarten sind, welche es letztendlich ersetzten sollen. Es ist ein Provisorium.[2]

Aus Davidsons Formulierung lässt sich erkennen, er muss das Beispiel von Nietzsche bzw. von Molière im Auge gehabt haben.

> [Metaphorische Bedeutungen] erklären die Metapher nicht, sondern die Metapher erklärt sie. Sobald wir eine Metapher verstehen, können wir […] (bis zu einem gewissen Punkt) […] sagen, was die „metaphorische Bedeutung“ ist. Diese Bedeutungen aber einfach in der Metapher anzusiedeln, ist so ähnlich, als wollte man die Wirkung einer Schlaftablette durch die *vis dormitiva* erklären. Buchstäbliche Bedeutungen und buchstäbliche Wahrheitsbedingungen können Wörtern und Sätzen unabhängig von jeweiligen Verwendungskontexten zugeordnet werden. Deshalb vermag die Berufung auf sie wirklich etwas zu erklären.[3]

Die oben skizzierte Struktur lässt sich problemlos in diese Sätze hineinbringen. Erst wenn wir eine Metapher *X* verstanden haben (*Y*), können wir darüber reflektieren und teilweise herausfinden, was ihre metaphorische Bedeutung Z gewesen ist. Eine metaphorische Bedeutung *Z* vermag nicht zum Verständnis *Y* einer Metapher (token) beizutragen, sondern eher umgekehrt ist sie aus dem Verstehen der Metapher zu entnehmen. Die sekundäre/metaphorische Bedeutung war bei Wittgenstein ein Provisorium, eine Zwischenstation; Davidson hat sie ans Ende der Erklärungsfolge verbannt.

Die provisorische Annahme Davidsons ist die, dass die Metapher (token) eine metaphorische Bedeutung haben kann. Wäre diese Bedeutung dem Hörer zuvor bekannt, würde sich es um eine tote Metapher handeln. Ergo muss die metaphorische Bedeutung unbekannt sein und somit besitzt sie keine Erklärungskraft. Umgekehrt ausgedrückt heißt es, um die Rolle einer Erklärung einzunehmen, muss die metaphorische Bedeutung der Metapher als *type* zukommen, was hieße, dass es um eine tote Metapher geht.

Den Kern der Argumentation macht die Schlussfolgerung aus, dass die metaphorische Bedeutung die Metapher zu einer toten macht. Aber sehen wir uns die Struktur einer toten Metapher näher an. Sie ist eine gewesene Metapher, die gegenwärtig zwei oder mehrere buchstäbliche Bedeutungen aufweist. Sollte die Zuschreibung der metaphorischen Bedeutung die Metapher mit einer toten gleichsetzen, müsste die metaphorische Bedeutung mit der zweiten (aber schon buchstäblichen) Bedeutung ebenso gleichgesetzt werden. Folglich müsste die metaphorische Bedeutung, gegen die sich dieses Argument richtet, mit der buchstäblichen gleichartig sein.

Soweit meine Analyse des Argumentes. Max Black wehrt sich aber folgendermaßen:

> One must agree that it would be pointless and obfuscating to invoke some ad hoc “figurative” sense, not otherwise specified, to explain “how metaphor works its wonders”. Nevertheless, it would help us to understand how a particular metaphorical utterance works in its context if we could satisfy ourselves that the

[2] Man könnte zweifeln, ob die Beispiele überhaupt Metaphern sind. Denn was erst widersinnig war, und demzufolge als eine Metapher zu sein schien, ist später durch eine buchstäbliche Erklärung ersetzt, die jedoch keine systematische ist. Sie beruht auf einer durchaus inzidentellen Tatsache. Die Sätze (1) und (2) dürfen auch als eine Chiffre angesehen werden, die der Sprecher mit denen teilt, denen die nachfolgenden Erklärungen oder die in den Erklärungen ausgedrückten Tatsachen bekannt sind. Dieser unsystematische Charakter ist auch für jede Chiffre wesentlich. Ich möchte aber keineswegs bestreiten, dass man immerhin solche zufälligen Tatsachen (etwa in der Gestallt von versteckten biografischen Anspielungen) in Metaphern in der Dichtung hineinzuziehen pflegt. Die Chiffre und die Metapher können koinzidieren.

[3] Davidson 1998, S. 51f, Hervorhebung original. Der lateinische Ausdruck „*vis dormitiva*“, der die vermutliche Anspielung auf Nietzsche oder Molière evidenter macht, stammt erst vom Übersetzer ins Deutsche. Im Original steht lediglich „dormative power“.

> *speaker* is then attaching a special extended sense to the metaphorical "focus" (selecting, as I have explained elsewhere, some of the commonplaces normally associated with his secondary subject, in order to express insight into his primary subject). This view is not open to the charge of invoking fictitious entities. (Black 1979, 190f)

Man muss nun diese Sätze unter die Lupe nehmen. Was kann dem Hörer helfen, eine metaphorische Aussage (d. h. Metapher-token) zu verstehen? Black hat wohl gemeint, dass der Hörer über den Sprecher weiß, er habe einige Wörter metaphorisch benutzt und ihre neue Bedeutungen sind gemäß seiner Methode der Interaktion[4] herauszufinden. So würde die Metapher zu einer Art der Kommunikation und der Vorwurf seitens Davidson behielte seine Geltung.

Hätte jedoch Black „*the* special extended sense" anstatt des unbestimmten Artikels geschrieben, wäre der betreffende Satz einer anderen Lesart fähig. Wenn der Hörer lediglich die Tatsache weiß oder annimmt, *dass* die Aussage metaphorisch intendiert worden ist, könnte zu *der* speziellen erweiterten Bedeutung selbst die Methode der Interaktion werden. Dieser subtile Unterschied mag mithilfe von prozessualen Vokabeln folgenderweise nahegebracht werden: Die Methode der Interaktion sei eine Prozedur mit vielen Eingaben, unter denen die in der Metapher benutzten Wörter eine signifikante Rolle spielen, wobei der Kontext der metaphorischen Aussage und das Hintergrundwissen die restlichen Eingaben bilden sollen. Der Sprecher kann nun mit seiner metaphorischen Aussage entweder diese Prozedur oder ein bestimmtes Ergebnis dieser Prozedur verknüpfen. Blacks Formulierungen sprechen eher für die zweite Möglichkeit und infolgedessen sind sie ein leichtes Ziel von Davidsons Kritik. Die erste Möglichkeit hat den Vorteil, dass der Sprecher nicht alle Eingaben kundgeben und der Hörer nach ihnen nicht forschen muss. Ein anderer Vorteil ist die Tatsache, dass eine solche metaphorische Bedeutung nicht mit der buchstäblichen gleichartig ist, und somit trifft auf sie die Davidson'sche Kritik nicht zu. Dies würde jedoch heißen, die Ideen der Kommunikation und der Übereinstimmung der metaphorischen Interpretationen beider Gesprächspartner müssen preisgegeben werden, und das hat Black offensichtlich nicht tun wollen. Die bekannte These, dass die Kunst, worunter auch die Metapher gehört, zweimal gezeugt wird, muss ergänzt werden, dass daran zwei selbstständige und voneinander unabhängige Leben anschließen können. Denn wenn die metaphorische Bedeutung nur eine Methode wäre, wie Metaphern zu interpretieren sind, so gewährleistet nichts, dass zwei Durchführungen zu demselben Ergebnis kommen müssen, weil unter Eingaben dieser Methode auch der inzidentelle außersprachliche Kontext gehört.

Diese Auseinandersetzung sollte darauf aufmerksam machen, dass das Konzept der metaphorischen Bedeutung allein keine Erklärung der Funktionsweise von Metaphern sein kann. Sie kann aber als die Bezeichnung einer solchen Erklärung verstanden werden, die vorhergehen oder folgen muss. Diese Erklärung soll etwas Allgemeines aussagen und zwar über die Metapher selbst, nicht über diese oder jene konkrete Metapher. Somit muss sie der Metapher als *type* zukommen.[5]

Literatur

Black, Max 1979 „How Metaphors Work: A Reply to Donald Davidson", in: Sheldon Sacks (Hrsg.), *On Metaphor,* Chicago: University of Chicago Press, 181-192.

Davidson, Donald 1998 [^{1}1978] „Was Metaphern bedeuten", in: Anselm Haverkamp (Hrsg.), *Die paradoxe Metapher*, Frankfurt am Main: Suhrkamp, 49-75.

Nietzsche, Friedrich 1954 [^{1}1886] *Jenseits von Gut und Böse*, in: Karl Schlechta (Hrsg.), *Werke in drei Bänden*, München: Hanser, Bd. II.

Wittgenstein, Ludwig 2000 *Wittgenstein's Nachlass*, Oxford: Oxford University Press.

[4] Wie die Methode genau funktioniert, ist verhältnismäßig kompliziert und wird in dem Zitat in den Klammern angedeutet. Für unsere Überlegung ist allein wichtig, dass sie fixiert, wie aus ursprünglichen buchstäblichen Bedeutungen die metaphorische Bedeutung hervorkommt.

[5] K.-Fr. Kiesow (Hannover) hat wertvolle kritische Bemerkungen zu einer früheren Auffassung dieses Aufsatzes gemacht.

„Vom Weißdorn und vom Propheten“ – Poetische Kunstwerke und Wittgensteins „Fluß des Lebens“

Annelore Mayer, Baden, Österreich

„Das uhlandsche Gedicht ist wirklich großartig. Und es ist so: Wenn man sich nicht bemüht das Unaussprechliche auszusprechen, so geht nichts verloren. Sondern das Unaussprechliche ist, - unaussprechlich – in dem Ausgesprochenen enthalten.“ (Engelmann S 78). Das Gedicht, zu welchem Wittgenstein sich hier äußert, ist die aus 7 Strophen zu je 4 Zeilen bestehende Ballade „Graf Eberhards Weißdorn“ von Ludwig Uhlland (1787-1862). Die „Handlung“ ist rasch erzählt: Graf Eberhard „vom Württemberger Land“ schneidet sich auf „frommer Fahrt“ nach Palästina in einem dortigen Wald ein „grünes Reis von einem Weißdorn“ ab, welches er nach vollbrachten Taten mit nach Hause nimmt. Dort grünt es und wird zum stattlichen Gewächs, unter dessen „Wölbung“ der Ritter gegen Ende seines Lebens sitzt und gleichsam wie im Traum der Zeit gedenkt, als er das Reis gebrochen hatte. Der Dichter schildert hier weniger ein „Geschehen“, als vielmehr zwei in der Person des Grafen ineinander verwobene Zustände: jenen der Jugend und jenen des Greises. Die nach außenhin scheinbar volksliedhaft-schlicht erscheinende Form beruht auf einer sehr subtilen Gestaltung. Das Gedicht hat 7 Strophen. Der Mittleren, in welcher die Heimkehr Eberhards mit dem Weißdornreis geschildert wird, kommt eine besondere Funktion zu, als dort nämlich die beiden Zustände ineinander übergeführt werden. Im weiteren Verlauf bezieht der Dichter die nun folgende 5. Strophe inhaltlich auf die 3., des Weiteren die 6. auf die 2. und letztendlich die 7. auf die 1.. Konkretisiert wird diese Bezugnahme durch die Schlussworte:

> „Die Wölbung hoch und breit,
> Mit sanftem Rauschen mahnt
> Ihn an die alte Zeit
> Und an das ferne Land.“

In dem, was dem Grafen gegen Ende seines Lebens nahe ist, nämlich im mittlerweile zur „Wölbung“ ausgewachsenen Weißdorn, evidiert sich demnach für ihn das nunmehr zeitlich Ferne und dieses ist somit durch das, was ihm jetzt nahe ist, auch selbst wieder nahe.

Es kann dieses Gedicht durchaus verstanden werden als ein ganz bestimmter „Fluß des Lebens“, in welchem „die Worte ihre Bedeutung haben“ (Wittgenstein 1984, 913). Es ist diese Bedeutung als eine anzusehen, welche über eine solche auf dem den reinen lexikalischen Gehalt beruhende weit hinausgeht. Mit allen gebrauchten Wörtern – also mit dem kompletten Gedicht – zielt Uhland auf mehr denn auf ein Verständnis der einzelnen durch die Wörter ausgesprochenen Begriffe. Das Gedicht fasst diese Begriffe gewissermaßen als eine „Gesamtheit der Tatsachen“ zusammen, so dass sie in dieser Gesamtheit und als solche eine eigene Tatsache ergeben, als welche das von Uhland beschriebene ineinander Übergehen der von Eberhard gerade zum Zeitpunkt der Beschreibung erfahrenen und bedachten Lebenszustände des Greisen– und Jugendalters bezeichnet werden darf. Und so kann vielleicht doch gesagt werden, dass die Wörter nur in der durch den Dichter gegebenen Form des ganzen Gedichtes mit seinen inneren Bezügen zwischen den einzelnen Strophen ihrer Funktion für die Erreichung dieser „Tatsache als Gesamtheit der Tatsachen“ – nämlich der im Gedicht erwähnten, wie etwa „Eberhard“ oder „Weißdorn“ – gerecht werden können. In diesem Sinne ist dieses Gedicht nun auch selbst „Welt“, wenn man diese nämlich mit Wittgenstein als „die Gesamtheit der Tatsachen“ begreift (Wittgenstein 1984, 1.1., S 11). In dieser Welt gibt es das „Aussprechliche“, worunter der rein lexikalische Gehalt eines Wortes verstanden werden kann. So ist beispielsweise der im Gedicht so wesentliche Weißdorn eine genau definierte „Tatsache“ der Welt in deren Unterordnung „Pflanzenwelt“, für dessen wissenschaftliche Erkenntlichmachung und Beschreibung die Botanik das lateinische Wort „Crataegus“ verwendet. Es steht außer Zweifel, dass mit dieser rein lexikalischen Bedeutung auch im Gedicht etwas *gesagt* wird. Das „Weiße“ und das „Dornige“, durch welches sich diese Pflanze auszeichnet macht sie selbst als *die* bestimmte Tatsache „Weißdorn“ erkennbar. Gerade in diesen Bestimmtheiten mag auch Uhland den Grund gesehen haben, diese Pflanze als ein Medium seines Gedichtes erkoren zu haben. Allerdings eben nur als „Medium“, als Mittel, welches die Annäherung an etwas Anderes, ja, vielleicht sogar die Handhabung dieses „Anderen“ ermöglicht. „Es gibt allerdings Unaussprechliches. Dies *zeigt* sich, es ist das Mystische.“ (Wittgenstein 1984, 6.522, S 85). Gemäß dieser wittgenstein’schen Feststellung kann das Wort „Weißdorn“ – und zwar notwendigerweise im Kontext mit seiner lexikalisch-botanischen Bedeutung – in jenem uhland’schen Gedicht als definitorischer Teil der „Gesamtheit der Tatsachen des *an sich Unaussprechlichen*“ in jenem Gedicht angesehen werden. Das *Unaussprechliche* ist ja nicht das, was durch die Wörter „erzählt“ wird – also die Fahrt des Grafen Eberhard nach Palästina, das Ausgraben des Weißdornreises und dessen Wiederanpflanzung in der württembergischen Heimat, etc. – sondern *das*, was formal durch die Hervorhebung der 4. Strophe und der nachfolgenden Verklammerung der weiteren mit den ersten drei Strophen bewerkstelligt wird und so zur Evidenz geführt werden soll. Es ist dies tatsächlich das *Unaussprechliche*, weil es eine „Gesamtheit von Tatsachen“ –und zwar von aussprechlichen solchen - ist, die so ineinander verwoben wurden, dass sie im Einzelnen nicht mehr ausgesprochen werden können. Demnach ist auch die kunstvolle formale Gestaltung dieser Ballade ein Medium der Evidenzialisierung dieses *Unaussprechlichen*. Durch Uhlands poetisch-formale Bemühungen ist es aber auch „ausgesprochen“ und es ist ihm dabei tatsächlich nichts verloren gegangen, wie Wittgenstein doch richtig festgestellt hat. Und so gesehen ist das uhlandsche Gedicht wirklich großartig, weil es in der Gesamtheit seiner Struktur eben nichts weniger ist als ein „Fluß des Lebens“, in welchem den verwendeten Wörtern die Bedeutung zukommt, im Ausgesprochenwerden etwas Unaussprechliches handhabbar zu machen.

Einen vielleicht noch extremeren Fall als Uhlands Ballade stellt das Gedicht „Пророк“ („Der Prophet“) von Aleksandr Sergeevič Puškin (1799-1837) dar. Wittgenstein hat dieses Gedicht eigenhändig in russischer Sprache und in kyrillischer Schrift abgeschrieben (s.: Rothaupt, S 278). Der Dichter schildert darin die Berufung des Jesaja zum Propheten, genauer eigentlich die Wandlung und somit Werdung zum Propheten, ausgehend von jenem Bericht, den der Berufene selbst im 6. Kapitel seines zu den wesentlichen prophetischen Schriften des aus christlicher

Perspektive sogenannten Alten Testamentes gehörenden Buches gibt. Auch hier geht es um die Darstellung einer wesentlichen Zustandsveränderung, um die Evidenzialisierung jenes Augenblickes – oder als was sollte es bezeichnet werden – in welchem Jesaja sich vom Nicht-Propheten zum Propheten wandelt. Diese Wandlung, die im Gedicht der nunmehr Verwandeltwordende und somit Verwandeltseiende selbst beschreibt und reflektiert, geschah im göttlichen Auftrag durch einen Seraph, welcher dabei dem Jesaja u.a. „die sündige und geschwätzige Zunge aus dem Mund herauszog“ und in den durch dieses Herausziehen „verklingenden Mund“ mit der „blutbefleckten Rechten“ „den Stachel der klugen Schlange“ hineinlegte.

„И он к устам моим приник,
и вырвал грешный мой язык,
и празднословный и лукавый,
и жало мудрыя змеи
в уста замершие мои
вложил десницею кровавой.“

Bei 16 der insgesamt 30 Zeilen des ohne strophische Gliederung in einem durchlaufenden Gedichtes steht am Anfang das Wort „и“ („und“), wobei von Zeile 10 – 18 dieses Worte jeweils 9 mal hintereinander den Anfang bildet, von Zeile 21 - 23 3 mal hintereinander. Auf diese Weise bringt der Dichter die Zusammengehörigkeit, ja das gewissermaßen „Punkthafte“ all des hier vor sich Gehenden zur Sprache. Auch hier ergibt also ein Konglomerat von Einzeltatsachen die Gesamtheit einer Tatsache, auf welcher die Bestimmung des Jesaja zum Propheten und sein dadurch ein solches gewordenes „Prophetsein“ beruht.

Bestimmung, Wandlung und Werdung zum Propheten, deren Realisierung Puškins Jesaja durch den von Gott beauftragten Seraph an sich im Wahrnehmen derselben reflektiert, sie alle drei sind samt ihrem Ergebnis des „Prophet geworden Seins“ in ihrem Kontext mit dem Göttlichen etwas „Geheimnisvolles“ und somit letztlich „Unaussprechliches“. Der Dichter zeigt mithin nicht nur, dass es dieses Geheimnisvoll-Unaussprechliche gibt, sondern „dies *zeigt* sich“ – um nochmals Wittgenstein zu Wort kommen zu lassen – und es zeigt sich mithin in seiner Notwendigkeit, das „Mystische“ zu sein.

Ausgehend von Wittgensteins Wertschätzung des Gedichtes „Graf Eberhards Weißdorn“ von Ludwig Uhland und des Gedichtes „Пророк“ von Aleksandr Sergeevič Puškin könnte folgender Gedanke entwickelt werden: Ein Wort gewinnt „im Fluß des Lebens“ seine Bedeutung. Die Dichter schaffen mit ihrem ganz bestimmten Werk, also etwa mit „Graf Eberhards Weißdorn“ oder „Пророк“ einen ganz bestimmten solchen „Fluß des Lebens“. Im Gedicht kommt dank des strukturellen Vermögens des Dichters und seiner Fähigkeit, Wörter mit bestimmter lexikalischer Bedeutung zu einer neuen „Gesamtheit der Tatsachen“ zusammenzufassen und innerhalb der Struktur ihres Werkes als solche zur Geltung zu bringen das *Unaussprechliche* als *Gezeigtes* zur Evidenz.

Wenn dies aber „das Mystische“ ist, dann ist dieses „Mystische“ nichts weniger als ein vom Dichter in seiner bestimmten Art strukturierter „Fluß des Lebens“. In letzter Konsequenz könnte dies bedeuten, dass die Wörter Uhlands und Puškins aus den beiden hier angesprochenen Gedichten gleichsam in ein und denselben „Fluß des Lebens“, nämlich den „mystischen“ eingebettet sind – und dies unabhängig von der Verschiedenheit der angewendeten Sprachen Deutsch und Russisch. Das Bett des Lebensflusses ist aber dann auch gestaltgebend für die Struktur des poetischen Kunstwerkes – so es denn ein solches ist, wie im konkreten Fall von Uhland und Puškin. Die von beiden angewandte präzise Form ist demnach im den beiden Gedichten gemeinsamen „Fluß des Lebens“, nämlich jenem des „Mystischen“ grundgelegt und die beiden Künstler gestalten dieses dort Grundgelegte gemäß ihrer individuellen künstlerischen Fähigkeit und den diesbezüglichen Möglichkeiten der von ihnen verwendeten Sprachen. So gesehen haben beispielsweise die Wörter „und“ bei Uhland bzw. „и“ bei Puškin in ihrer Funktion für die Akkumulierung von Einzeltatsachen zu einer „Gesamtheit der Tatsachen“ als neuer, unaussprechlich gezeigter Tatsache trotz ihrer Zugehörigkeit zu einer jeweils anderen Sprache die selbe Bedeutung, sind also, um bei Wittgensteins Bild zu bleiben, in diesem Falle der zwei Gedichte zwei Tropfen im selben Flusse. In diesen Fluss hineinzusteigen – ist das nicht „ein Erlebnis“, vermittelt durch das Kunstwerk in seiner durch den Künstler angewandten Struktur und Wortwahl? „Aber“, so Wittgenstein, „es gibt Erlebnisse charakteristisch für den Zustand des „Sich-auskennens.“ (Wittgenstein 1984, 721, S 337).

In den beiden hier angesprochenen Gedichten darf davon ausgegangen werden, dass sich die Dichter – ganz im Sinne Wittgensteins – eben nicht bemühten, „das Unaussprechliche auszusprechen“. Und weil dadurch durch sie und in ihnen nichts verloren gegangen ist, „sondern das Unaussprechliche - unaussprechlich – in dem Ausgesprochenen enthalten“ ist, so lässt sich daraus ein letztendlicher Wert von Sprache – und zwar in der Gesamtheit der Tatsachen ihrer Erscheinungsformen, wie beispielsweise als Deutsch oder Russisch – wahrnehmen: Sinn und Funktion der Sprache ist schlechthin nicht das Aussprechen, sondern das Zeigen des Unaussprechlichen oder zumindest das Hinzeigen auf dieses. Ein sprachlich-poetisches Kunstwerk bewerkstelligt dies durch das Nichtbemühen um das Aussprechen und durch das Bemühen, die „Gesamtheit der Tatsachen“ als neue gemeinsame Tatsache zu zeigen, so dass ein „Zustand des Sich-auskennens“ herbeigeführt wird.

„Was passiert, wenn man Wittgensteins Gesamtnachlass nicht nur als Philosophie, sondern auch als Literatur, als Dichtung ... betrachtet und liest? Man würde nichts verlieren und vieles gewinnen“ (Rothaupt, S 287). Josef Rothaupts Frage mit folgernder Feststellung hat auch den hier unternommenen Inbeziehungsetzungen von wittgenstein’schen Überlegungen mit den Möglichkeiten und Erscheinungsformen poetischer Kunstwerke.

Nahrung gegeben. Äußerungen des Philosophen etwa zu Uhland, Puškin und machen anderen bedeutenden Vertretern der Literatur – so marginal sie auch fürs Erste erscheinen mögen – eröffnen vielleicht tatsächlich zusätzliche Möglichkeiten nicht nur des Verständnisses, sondern auch der Veranschaulichung von Gedankengängen, so wie es hier zu zeigen unternommen wurde. Und was spräche denn tatsächlich dagegen, Wittgenstein – unbeachtet des Gehaltes und der Originalität seines Argumentierens – auch in einem Zusammenhang zu sehen mit Denkern wie beispielsweise Johannes Tauler, Angelus Silesius oder Friedrich Nietzsche, deren Hervorbringungen ja auch durchaus anerkannten literarischen Rang besitzen oder auch mit Literaten wie Ljev Nikolajevič Tolstoj oder Rainer Maria Rilke, deren Einfluss auf die Philosophie im Allgemeinen und speziell auch auf Wittgenstein ja auch beachtlich ist? Kann eine solche Sicht nicht als durchaus angemessen erkannt werden einem Denker gegenüber, der sich selbst die Frage stellt: „O, warum ist mir zumute, als schrieb ich ein Gedicht, wenn ich Philosophie schreibe?“ (zitiert bei Rothaupt, S 288).

Literatur

Engelmann, Paul 1970 Ludwig Wittgenstein. Briefe und Begegnungen. Herausgegeben von Brian McGuinness. Wien und München: R. Oldenbourg.

Rothaupt, Josef G. F. 2006: Zu Engelmanns Buch der Erinnerung. Paul Engelmann als Dichter und Ludwig Wittgensteins diesbezügliche Wahlverwandtschaft. In: Internationale Wittgenstein Gesellschaft e. V. (Hrsg.), Wittgenstein Jahrbuch 2003/2006. München, Peter Lang, S 249 – 289.

Wittgenstein, Ludwig 1984: Tractatus logico-philosophicus. Werkausgabe Band 1. Erste Auflage, Frankfurt/M, Suhrkamp.

Wittgenstein, Ludwig 1984 Letzte Schriften über die Philosophie der Psychologie. Vorstudien zum zweiten Teil der philosophischen Untersuchungen 711. Herausgegeben von G.H. von Wright und Heikki Nyman. Werkausgabe Band 7. Erste Auflage, Frankfurt/M, Suhrkamp.

„Die Einheit hören“ – Einige Überlegungen zu Ludwig Wittgenstein und Anton Bruckner

Johannes Leopold Mayer, Baden, Österreich

Ludwig van Beethoven läßt im Finale seiner IX. Symphonie nach einem einleitenden dramatischen, fanfarenartigen Motiv die Themen der drei vorhergegangenen Sätze nochmals anklingen.

Ähnliches scheint auch in den drei großen Messen und in den Symphonien Anton Bruckners zu geschehen, wenn dort – meist gegen Ende des jeweiligen Finalsatzes – Material aus anderen Teilen des Gesamtwerkes nochmals in ein abschließendes Geschehen einbezogen wird. Vielfach wurde in der Musikgeschichtsschreibung daher festgestellt, Bruckners Verfahren beruhe auf jenem Beethovens und habe dort sein historisches, von Bruckner selbst als solches anerkanntes und daher nachgeahmtes Vorbild.

Eine solche Betrachtung lässt zwei Fragen außer acht: erstens, ob es per se das Gleiche ist, wenn zwei das Selbe tun – und zweitens, ob hier tatsächlich zwei das Selbe tun und daher Gleiches oder gar Selbes passiert.

Ludwig Wittgenstein scheint dies offenbar anders zu hören. „Die Brucknersche Neunte ist gleichsam ein *Protest* gegen die Beethovensche.“ (Wittgenstein 8/1989, S 497). In der Tat kann die Aussage des Philosophen als höchst angemessen bezeichnet werden. Eine auf Beethoven bezogene Interpretation des strukturellen Phänomens der Wiederkehr von Material aus den vorangehenden Sätzen im Finale bei Bruckner lässt nämlich das außer acht, was die beiden Komponisten genau hier – nämlich im Geschehen eines nochmaligen finalen Rekurses auf thematisches Material aus vorhergehenden Sätzen eines mehrteiligen Gesamtwerkes – geradezu fundamental unterscheidet: nachdem Beethoven die Themen der ersten drei Sätze hat Revue passieren lassen, leitet der Baßsolist mit dem Rezitativ „O Freunde nicht diese Töne. Sondern lasst uns andere anstimmen“ zur Intonation der nun tatsächlich ganz neuen, ganz anderen Melodie auf die Worte „Freude schöner Götterfunken“ über. Allein schon durch das nunmehrige Einführen menschlicher Stimmen vollzieht sich hier eine intensive Abwendung vom bisher nur instrumental geprägten Klanggeschehen. Alles was vorher war sollen die Zuhörenden, so lässt es uns der Komponist vernehmen, in radikalster Weise und ein für alle Mal hinter sich lassen.

Anders ist dies bei Bruckner, der schon in seiner ersten großen Messe d-moll von 1864 in das den Gesamtzyklus abschließende *„dona nobis pacem“* das melodische Material des „Credo“-Schlusses, dort mit den Worten *„et vitam venturi saeculi“* unterlegt, nochmals und tatsächlich abschließend-sinnfällig hineinklingen lässt. Solcherart wird auf etwas Vorausgehendes nicht nur zurückgegriffen, es wird dieses Vorausgehende durch dessen Wiederaufnahme vielmehr bestätigt, sodass dieses kompositorische Vorgehen also eine „Restitutio in integrum“ darstellt. Auch in seinen Symphonien gelangt Bruckner auf unterschiedlichen gestalterischen Wegen dazu, gegen Ende des Gesamtwerkes auf dessen Anfang zu rekursieren. Die Symphonien 2 – 8 bieten dafür jeweils ganz individuelle und exemplarische Lösungen. Für die von Wittgenstein angesprochene „IX.“ hat Bruckner wiederum ganz andersgeartete Möglichkeiten zur Zusammenführung des Gesamtmaterials gesucht. Diese sind jedoch nur aus Skizzen überliefert, weil der Tod des Komponisten diesem eine Verwirklichung nicht mehr gestattete.

Aber der von Wittgenstein gehörte „Protest“ Bruckners gegen Beethovens „IX.“ kann ja bereits aus anderen Werken als der letzten Symphonie dieses Meisters herausgehört werden, etwa auch aus dessen „VII.“, zu welcher sich der Philosoph in einem Brief an seine Schwester Helene vom 30.3.1946 äußert: „Gestern spielten mir zwei bekannte die 7te von Bruckner vor (vierhändig). Sie spielten schlecht, aber nicht ohne Verständnis. Ich hatte die Symphonie seit Jahren nicht gehört und hatte wieder einen großen Eindruck.“ (Wittgenstein 1996, S 187).

Im Finale dieser Symphonie leitet Bruckner kunstvoll die originären Motive dieses Satzes so, dass durch die Struktur eine Rekapitulation des Eröffnungsthemas des ersten Satzes geradezu evoziert wird und das Geschehen des Finales letztlich in diese Rekapitulation einmündet, sodass das Werk in seiner Gesamtheit im Erklingen seines Anfanges zum Schluss kommt. Das, was vor dem Finale war, wird in diesem Falle also nicht verworfen, um „andere Töne“ anstimmen zu können, was offenbar durch das zuvor Gehörte notwendig geworden ist, wie bei Beethoven, sondern es wird dieser Anfang vielmehr in seiner Rückkehr ins Gesamtgefüge bestätigt.

Also: Nichts von „nicht diese Töne“, sondern ganz im Gegenteil - im Sinne des Ganzen zurück zu diesen Tönen. Damit kann demnach auch die 7. Symphonie Bruckners im wittgenstein'schen Sinne als ein Protest dieses Komponisten gegen Beethoven aufgefasst werden, und mit ihr die vorangegangenen symphonischen Werke und die drei großen Messen.

Wittgenstein verweist durch seine Aussage auf eine deutliche Unterschiedlichkeit der Absichten bei einem scheinbar gleichen strukturellen Verfahren – der Hereinnahme von Material aus vorhergehenden Sätzen in den Finalsatz und er macht damit jeweils gänzlich anders geartete Gesamtkonzeptionen offenbar. Diese unterschiedlichen Gesamtkonzeptionen verlangen aber auch ein anderes Hören und Bedenken der Zusammenhänge innerhalb dieser Symphonien. Beethoven fordert nachgerade dazu auf, alles, was der Freudenhymne vorausgeht, nicht mehr aufkommen zu lassen, auch wenn es sich – wie aus dem nochmaligen Zitieren vernehmbar – aufdrängen wollte. Die anderen, letztendlich der menschlichen Stimme anvertrauten Töne, schließen alles Vorhergegangene mit aller Entschiedenheit aus. Darin liegt auch die Radikalität dieses Werkes, dass es hier zuletzt einem eigenen großen Teil von sich selbst widerspricht. Und Beethoven will ja offenbar auch, dass der letztendlich gültige Teil anders gehört wird, als alles, was ihm bevorgeht, indem er nämlich das klangliche Geschehen durch die Einführung menschlicher Stimmen völlig verändert. Das Neue, am Schlusse sich selbst Bestätigende, will, ja muss auch anders gehört werden als alles, was da vorher kommt.

Bruckners Symphonien verlangen aber ein anderes Hören: nämlich ein solches in Gesamtheiten, die akustisch einen ganzen Organismus darstellen.

„Ich könnte von einem Bild von Picasso sagen, ich *sehe* es nicht als Menschen. Das ist doch ähnlich dem: ich war lange nicht Imstande dies als Einheit zu hören, jetzt aber höre ich's so. Früher schien es mir wie lauter kurze Stücke, die immer wieder abreißen, — jetzt hör ich's als Organismus. (Bruckner).“ (Wittgenstein 7/1989, § 677, S 436).

In der Tat: *jetzt* ist etwas als Gesamtheit eines großen Organismus zu hören, dann nämlich, wenn Bruckner das Material eines Schlußssatzes kunstvoll so strukturiert, dass darauf die Wiederhinwendung zum Anfang erfolgen kann. Wie wird aber dieser in das Ende den Anfang einbeziehende Letztzustand des Werkes erreicht? Kann vielleicht davon gesprochen werden, dass der Rekurs auf den Anfang als Zusammenfassung verstanden werden soll, als Zusammenfassung nämlich von strukturellen Einzelerscheinungen, die als jeweils unterscheidbarer symphonischer Satz innerhalb des Gesamtzyklus, respective als einzelne melodisch-thematische Gebilde innerhalb eines solchen Satzes hörend wahrgenommen werden können? Durchaus verständnisfördernd ist hier folgende Bemerkung Wittgensteins: „Von einer Brucknerschen Symphonie kann man sagen, sie habe zwei Anfänge: den Anfang des ersten & den Anfang des zweiten Gedankens. Diese beiden Gedanken verhalten sich nicht wie Blutsverwandte zu einander, sondern wie Mann & Weib.“ (Wittgenstein 2000, S 110).

Es ist *tatsächlich* so, dass Bruckner die Expositionen seiner Themen als einzelne Ereignisse gestaltet. In den frühen Symphonien trennt er sie abrupt durch Pausen, später scheidet er sie durch deutlich als solche hörbare und auf etwas nun folgendes Anderes verweisende Übergänge. Das heißt, dass zuerst einmal musikalische Einzeltatsachen, nämlich individuelle thematisch-melodische Gebilde zur Darstellung kommen. Diese kommen als solche Einzelgebilde scheinbar vorerst so zu ihrer ersten Darstellung, als würde tatsächlich mit ihnen das Werk erst beginnen. Durch Wittgensteins Aussage kann man auf eigentümliche Art auf einen fundamentalen Unterschied zwischen der Exposition thematischen Materials bei Bruckner und in den Symphonien eines wiener Klassikers wie etwa Joseph Haydn hingewiesen werden. Bei Letzterem entwickelt sich die Darstellung eines Themas nämlich gewissermaßen auf ein zweites, zu jenem im Kontrast stehendes Thema hin. Diesen Prinzipien der Einführung des Materials folgen im Wesentlichen auch Mozart und Beethoven.

Was hat es nun mit der von Wittgenstein konstatierten, zum Charakter der Beziehungen des angewandten thematischen Materials bei den wiener Klassikern im Gegensatz stehenden „Nicht-Blutsverwandtschaft“ bruckner'scher Themen auf sich? Auch hier hilft ein Rekurs auf Haydn: bei ihm ist es oft der Fall, dass ein in der Haupttonart des Werkes exponiertes Thema sich so entwickelt, dass in der Folge – und zwar tonal kontrastreich auf der Dominante – dasselbe Thema nochmals erscheint. Tonale Unterscheidung und melodische Gleichheit fallen bei solchen Gegebenheiten sinnfällig zusammen.

Dergleichen gibt es bei Bruckner nicht. Im Gegenteil: der Komponist wendet die höchste Kunst an, um seine thematischen Gebilde in erkennbar unterschiedener Weise erscheinen zu lassen, etwa auch durch die Wahl der klanglichen und satztechnischen Mittel.

Es lassen sich aber darüberhinaus grundsätzliche melodische Übereinstimmungen zwischen den exponierten Themen feststellen. Dies trübt die Unterscheidbarkeit dank der angewandten Unterscheidungs*mittel* keineswegs, lässt aber für aufmerksam Zuhörende eine besondere Art von Zusammengehörigkeit evident werden. Diese Tatsache und die Art und Weise, wie der Komponist im Laufe des symphonischen Geschehens die Themen bis hin zur Gleichzeitigkeit ihres Auftretens – wie etwa am Schluss der f-moll Messe oder der 8. Symphonie – zusammenführt, all dies lässt den plastischen wittgenstein'schen Vergleich von „Mann & Weib“ als gerechtfertigt erscheinen. Einzelindividuen, die zuerst auch einzeln erscheinen, werden im Laufe von Geschehen gleichsam wie in einer ehelichen Gemeinschaft so zusammengeführt, dass sie als Gesamtheit angesprochen werden können, ohne dabei ihre erkennbare – soll also im konkreten Falle heißen: hörbare – Individualität am Ende verloren zu haben. Denn gerade auf der Bewahrung der Individualitäten und deren Erkennbarkeit beruht ja Bruckners Kunst der Zusammenführung seiner musikalischen Strukturgebilde. Die „Gesamttatsache“ Messe oder Symphonie verweist daher speziell in ihrer durch den Rekurs auf den Anfang erreichten Zusammengefasstheit nachhaltig und endgültig auf die am Ende zusammengefassten „Einzeltatsachen“ der zuerst unabhängig exponierten thematischen Gebilde. Diese „Gesamttatsache“ ist ja nur hörbar durch die Wahrnehmung der „Einzeltatsachen“, durch welche sie zu dieser endgültigen „Gesamttatsache“ wird. Es ist dann doch *so*, dass nicht nur Wittgenstein im Laufe einer langen bruckner'schen Symphonie „lange nicht Imstande ist, dies als Einheit zu hören, jetzt aber“. Erst da, wo Bruckner es unternimmt, die Einzelelemente durch deren, aufgrund von Rekursen auf Zuvorgehendes erreichtes, gleichzeitiges Auftreten in einer Gemeinsamkeit erscheinen zu lassen „höre ich es so. Früher schien es mir wie lauter kurze Stücke, die immer wieder abreißen, - jetzt hör ich's als Organismus.“

Es kann ein höchstrangiges Erlebnis sein, hörend an einem solcherart zustandekommenden Ereignis der Werdung einer symphonischen Gesamttatsache teilzunehmen. Dergleichen ist nämlich keineswegs das von Haus aus zu Erwartende. Bruckner lässt sich mit seinen Lösungen Zeit und stellt im Laufe des Geschehens mehrere Möglichkeiten solcher Lösungen zur Diskussion, indem er zuerst die Verschiedenartigkeit der einzelnen Elemente in den Vordergrund stellt. Aber er kommt damit absichtsvoll nicht ans Ende.

„Ich glaube, das gute Österreichische ist besonders schwer zu verstehen. Es ist in gewissem Sinne *subtiler*, und seine Wahrheit ist nie auf Seiten der Wahrscheinlichkeit.“ (Wittgenstein 8/1989, S 454). Das „gute Österreichische“ wird in dieser Aussage Wittgensteins u.a. durch Grillparzer und Bruckner repraesentiert.

Dass die Wahrheit hier nie auf Seiten der Wahrscheinlichkeit ist – dies mag auch noch durch etwas deutlich werden, worauf Wittgenstein im Zusammenhang mit dem Streichquartett d-moll von Franz Schubert aufmerksam macht, in dessen zweitem Satz der Komponist sein eigenes Lied „Der Tod und das Mädchen“ variiert. „Die letzten beiden Takte des ‚Tod und Mädchen' Themas“, so Wittgenstein, „das ~; man kann zuerst verstehen, daß diese Figur konventionell, gewöhnlich ist, bis man ihren tieferen Ausdruck versteht. D.h., bis man versteht, daß hier das Gewöhnliche sinnerfüllt ist.“

(Wittgenstein 8/1989, S 523). Scheinbar Konventionelles gibt es auch im melodischen Material Bruckners, zumal im Hinblick auf die Typologie der Themen in den Messen, welche vielfach auf in der katholischen Kirchenmusik traditionsreichen Topoi beruhen. Auch hier liegt die Wahrheit nicht in der leicht annehmbaren Wahrscheinlichkeit, dass es sich aufgrund der scheinbaren Konventionalität der thematischen Gebilde um ein im Ganzen konventionelles Werk handelt, sondern in der Wahrheit der Sinnerfüllung.

Es sagt diese Überlegung zur „Sinnerfüllung“ auch etwas aus über Wittgensteins Verständnis von Musik. Dieses ist mit nichten „konservativ“. Die Erkenntnis einer solchen Sinnerfüllung des Gewöhnlichen innerhalb eines bedeutenden Kunstwerkes und die Bemerkungen des Philosophen zu Bruckner lassen darauf schließen, dass es ihm nicht um etwas geht, auf das der Begriff „modern“ ad hoc anwendbar zu sein scheint, sondern um das aus sich selbst bedeutsame und etwas bedeutende Werk, das dann eben gerade auf Grund einer solchen Bedeutsamkeit „aktuell“, weil bedenkbar ist.

Auch Bruckners Musik beruht auf einer solchen Art der Aktualität, wiewohl in ihr mit recht Bezüge gehört werden können zur frühen europäischen Mehrstimmigkeit des Mittelalters gleichermaßen wie zu den Messen des Josquin des Prés, dessen Werk, geschaffen am Ende eben jener so gerne mit negativen Epitheta ornantia belegten Epoche, einen der grandiosen Höhepunkte abendländischer Musik darstellt.

Was Wittgenstein über Bruckner zu sagen hat, das kann darauf aufmerksam machen, dass seine philosophischen Überlegungen bei der gedanklichen Erschließung des bruckner'schen Werkes äußerst hilfreich sein können, ja, für das Verständnis vieler musikalischer Gegebenheiten überhaupt. Sachverhalte wie das nochmalige Aufgreifen thematischen Materiales aus den vorangegangenen Sätzen im Finale einer Symphonie erscheinen im Hinblick auf Beethoven und Bruckner unter Einbeziehung der wittgenstein'schen Überlegungen nicht mehr als Abhängigkeit des jüngeren vom älteren Komponisten, weil durch solche Überlegungen eine vermeintliche kultrugeschichtliche Kausalität, aufgrund derer es sich um den selben Sachverhalt handelt, verneint wird. Ein solcher Blick auf die jeweilige Einzeltatsache ist so geartet, dass er die Tatsachen gemeinsam vergleichend erfasst und nicht aufgrund eines historischen Hintereinander.

In Hinsicht auf das Denken über die oder mit Hilfe von Musik sollte Wittgenstein durchaus der Rang zugestanden werden, welchen diesbezüglich etwa Aurelius Augustinus, Johannes Scotus Eriugena, Mechthild von Magdeburg oder Nicolaus Cusanus zu Recht einnehmen. Durch diese Namen, denen jener Wittgensteins also begründetermaßen hinzugefügt werden darf, soll aber auch darauf aufmerksam gemacht werden, dass sowohl der Musik als ganzer als auch einem bestimmten musikalischen Kunstwerk die Qualität eines philosophischen Erkenntnismittels zukommt. Gewissermaßen ganz von Seiten der Musik hat darauf der hochbedeutende österreichische Komponist und Theoretiker Johann Joseph Fux (1660-1741) in seinem lateinisch geschriebenen Lehrwerk *„Gradus ad Parnassum“*, erschienen in Wien 1725, aufmerksam gemacht. Den zweiten Teil seines Buches gestaltet der Autor als philosophischen Dialog nach platonischem Vorbild, *„utque veritas magis elucesceret“.* Anhand dieses Buches, welches der Autor dem Anspruch verpflichtet, dass aus ihm und aufgrund seiner strukturellen Gestaltung die Wahrheit besser hervorleuchte, hat sich Joseph Haydn die theoretischen Grundlagen des Komponierens erworben. Und gemäß den von Fux formulierten Grundsätzen lernte auch ein blutjunger Schulgehilfe in einem kleinen Bauerndorf an der Nordgrenze Österreichs die Grundsätze der allem künstlerischen musikalischen Schaffen mit Notwendigkeit zugrundeliegenden Theorie - sein Name war Anton Bruckner. Und dieser war später auch als Universitätslehrer bestrebt, die Musik als philosophische Wissenschaft zu vermitteln.

Literatur

Wittgenstein, Ludwig 7/1989: *Letzte Schriften über die Philosophie der Psychologie.* In: Werkausgabe Band 7. 4. Auflage. Frankfurt/M Suhrkamp.

Wittgenstein, Ludwig 8/1989: *Vermischte Bemerkungen.* In: Werkausgabe Band 8. 3. Auflage. Frankfurt/M Suhrkamp.

Wittgenstein, Ludwig 1996: *Familienbtriefe.* Herausgegeben von Brian McGuinness, Maria Concetta Ascher, Otto Pfersmann. Wien, Hölder-Pichler-Tempsky.

Wittgenstein, Ludwig 2000: *Denkbewegungen. Tagebücher 1930 – 1932, 1936 – 1937.* Herausgegeben und kommentiert von Ilse Somavilla. 2. Auflage. Frankfurt/M. Fischer Taschenbuch-Verlag.

Counterfactuals, Ontological Commitment and Arithmetic

Paul McCallion, St Andrews, Scotland, UK

1.

On the face of it, utterances of arithmetic sentences in which number words or numerals occur as singular terms carry commitment to the existence of numbers. Yet there is no easy path from the syntax of an uttered sentence to a claim about ontological commitment. For it might be that the sentence is uttered with non-assertoric force, for example as in make-believe. Alternatively, it might be that the sentence is to be understood non-literally.

An assignment of ontological commitment to an utterance should mesh with the beliefs that may be reasonably attributed to the speaker (for a brief discussion of the case of applied arithmetic see (Rayo *forthcoming*)). It would, for instance, be implausible to attribute a commitment to the existence of the average Scot to someone who asserts

(1) The average Scot has 2.3 children,

because, for one thing, it is implausible to suppose that the speaker believes that any Scot has a non-natural number of children.

In (Yablo 2001) it is noted that the willingness of speakers to assert

(2) The number of moons of Jupiter is four

seems to turn on whether they believe that Jupiter has exactly four moons, and is apparently independent of their belief in the existence of arithmetical objects. There is therefore a prima facie case for denying that an utterance of (2) carries commitment to the existence of the number four, just as an utterance of (1) does not carry commitment to the average Scot.

Yablo suggests that sentences such as (2) are uttered with non-assertoric force, as part of the pretence that there are numbers. An alternative explanation is that utterances of (2) are to be understood non-literally. On that view, the surface syntax of (2) is misleading, and it is used to express the proposition that

(3) Jupiter has exactly four moons

An attractive feature of both Yablo’s suggestion and of the alternative just mentioned is that they promise to remove the difficulties surrounding the epistemology of mathematical objects. One might therefore ask whether these analyses can be extended to pure arithmetic, for pure arithmetic is replete with numerical singular terms. The aim of this paper is to investigate whether the latter analysis can be so extended (for an extension of the pretence view to pure arithmetic and set theory, see (Yablo 2005)).

The surface syntax of simple arithmetic sums (such as ‘5 + 7 = 12’) suggests that their assertion will carry commitment to numbers. Nevertheless, the willingness of speakers to assert simple arithmetic sums seems to turn on their belief in some sort of generalisation (evidence for which is obtained by counting), and is apparently independent of their belief in the existence of arithmetical objects. So it is prima facie plausible that simple arithmetic sums may be paraphrased by sentences whose literal assertion would not carry commitment to arithmetical objects.

2.

This may sound familiar. There have been previous attempts to paraphrase (or otherwise provide a reduction of) the sentences of pure arithmetic in terms of generalisations constructed using the adjectival (or quantificational determiner) occurrences of number-words or numerals (for an overview, see (Rayo *forthcoming*)). These have typically been undertaken in the context of a defence of nominalism or logicism (or both). The following works are broadly in defense of an adjectival strategy: (Bostock 1974); (Gottlieb 1980); (Hodes 1984); (Rayo 2002).

A standard starting point for such reductions is the paraphrasing of arithmetic sums as generalizations constructed using the material conditional. For present purposes, it is more plausible for candidate paraphrases to involve natural language conditionals. An initial suggestion might be that (schematically)

(4) $m + n = p$

may be paraphrased as the indicative conditional

(5) *for any concept F, and any concept G, if there are exactly m objects that fall under F, exactly n objects that fall under G, and no objects fall under both F and G, then there are exactly p objects which fall either under F or under G*

It is a good question whether natural language sentences involving the indicative conditional have the same truth conditions as parallel sentences constructed using the material conditional. If these particular indicative conditionals do, then that creates a potential problem for the candidate paraphrases: the standard distribution of truth values will be preserved only if there are infinitely many objects. For if there are less than *m* objects, the paraphrase of $m + n = p$ will be true even in a case where the paraphrased utterance is standardly taken to be false.

It is standardly complained (for example in the context of nominalist treatments of arithmetic) that such a reduction would be theoretically flawed because, in the absence of further assumptions, it would leave open the epistemic possibility that there is a non-standard distribution of the truth values of arithmetic sentences. Moreover, the truth of any particular arithmetical sentence would be contingent on there being enough objects around (the only necessarily true sum would be ‘0 + 0 = 0’).

A related problem for the proposed reduction is that the willingness of speakers to assert or deny arithmetic sums in line with the standard distribution of truth values should not be – but plausibly is - independent of their belief in the infinitude of the world. Might then (4) be better paraphrased as a counterfactual (or subjunctive) conditional? A good candidate would be

(6) *for any concept F, and any concept G, if there were exactly m objects that fell under F, exactly n objects that fell under G, and no objects fell under both F and G, then there would be exactly p objects which fell either under F or under G*

This is an appealing paraphrase. On limited assumptions concerning modality (firstly that each world accesses some larger world, and secondly that the modal logic is at least

S4) the necessity of arithmetic is secured. It is also well placed to explain the apriority and applicability of arithmetic sums.

It might be objected that it is still problematic to mesh this reduction with the beliefs of ordinary speakers. For suppose there could not be more than *m* objects. Then since counterfactuals with impossible antecedents are vacuously true, the paraphrase of $m + n = p$ will be true even in a case where the paraphrased utterance is standardly taken to be false. The willingness of speakers to assert or deny arithmetic sums in line with the standard distribution of truth values should therefore not be independent of their belief in the potential infinitude of the world.

There are at least two possible lines of response to this objection. It could be argued that a willingness to make standard arithmetical assertions is indeed tied to a belief in the potential infinitude of the world. Alternatively, it could be argued that counterfactuals with impossible antecedents are not uniformly (vacuously) true. A common proposal is that the counterfactual conditional *if it were the case that A, then it would be the case that B* may be analysed as *all nearby A-worlds are B-worlds*. One might then follow, for example, (Nolan 1997) in accepting impossible worlds at which not everything is the case. On that view, some impossible worlds are nearer to the actual world than others, and so it is not the case that all counterfactual conditionals with impossible antecedents are true. Someone disposed to accept that view will arguably agree that on the assumption that there could not be 100 objects, the counterfactual ‘if there were exactly 100 objects that fell under F, exactly 0 objects that fell under G, and no objects fell under both F and G, then there would be exactly 100 objects which fell either under F or under G’ is true and that the counterfactual ‘if there were exactly 100 objects that fell under F, exactly 0 objects that fell under G, and no objects fell under both F and G, then there would be exactly 999 objects which fell either under F or under G’ is not. The willingness of speakers to assert or deny arithmetic sums in line with the standard distribution of truth values can therefore be consistently held to be independent of their beliefs regarding the potential infinitude of the world.

3.

Can this approach be extended to capture more of arithmetic? One might consider directly paraphrasing quantified arithmetic sentences by introducing third-order quantifiers into the paraphrases. Alternatively, one might claim that arithmetic quantification is substitutional.

The latter option may have more philosophical appeal. It is difficult to see how to make a case for ordinary speakers’ belief in something like a third-order counterfactual paraphrase of even the simplest quantified sentence. However, the willingness of ordinary speakers to assert quantified sentences is arguably not independent of their beliefs concerning the truth values of substitution instances of those quantified sentences. One might see in this the beginnings of a case for construing arithmetic quantifiers substitutionally.

There is not enough space here to develop such a case. However, it may be objected that if arithmetic quantifiers are substitutional then arithmetic must after all involve epistemologically troublesome ontological commitment. For an explanation of the truth of sentences involving substitutional quantifiers is commonly given in terms of the existence of expression types (or the possible existence of expression tokens) and the truth of substitution instances. For one thing, it is worth noting that the problem of the epistemology of expression types (or the possibility of expression tokens) is of a different sort from the problem of the epistemology of numbers, as on a standard construal numbers are not types. But more importantly, such commitments would belong to the metatheory of arithmetic sentences, not to arithmetic sentences themselves. Substitutional quantifiers may be taken to be primitive devices of infinite conjunction/disjunction (see (Field 1984)), or to have their meaning given by their inferential role (see (Rossberg 2006) and (Wright 2007) for an inferentialist treatment of higher-order quantifiers).

4.

On the understanding of arithmetic sentences just sketched it is natural to think of syntactically singular occurrences of numerals and number-words as occurrences of mere pseudo-singular terms. Such terms do not behave semantically like genuine singular terms; they do not purport to refer to objects. It is likewise natural to think of numerical predicates as pseudo-predicates. Such predicates do not behave semantically like genuine predicates; they do not purport to ascribe properties to objects.

A consequence of this would be that second-order arithmetic quantifiers (such as those that occur in the second-order version of the induction axiom) should also be construed substitutionally. This would in turn entail that arithmetic predicates may not be impredicatively defined, on pain of circularity. The moral is that commitment to numbers as objects is therefore unexpectedly revealed not by the assertion of sentences involving numerical singular terms or quantifiers but rather by the acceptance of impredicative definitions of numerical predicates.

Literature

Bostock, David 1974 *Logic and Arithmetic. Volume 1. Natural Numbers*, The Clarendon Press, Oxford University Press, Oxford.

Gottlieb, Dale 1980 Ontological Economy: Substitutional Quantification and Mathematics, New York: Oxford University Press.

Field, Hartry 1984 Review of Dale Gottlieb, Ontological Economy: Substitutional Quantification and Mathematics, Nous 18.

Hodes, Harold 1984 “Logicism and the Ontological Commitments of Arithmetic”, *Journal of Philosophy* LXXXI, 123-49.

Nolan, Daniel 1997 “Impossible Worlds: a modest approach”, *Notre Dame Journal of Formal Logic*, 38(4).

Rayo, Agustin 2002 “Frege’s Unofficial Arithmetic”, *The Journal of Symbolic Logic* 67, 1623–1638.

Rayo, Agustin *forthcoming* “On Specifying Truth Conditions”, *The Philosophical Review*.

Rossberg, Marcus 2006 *Second-Order Logic: Ontological and Epistemological Problems*, (Ph.D dissertation, University of St. Andrews).

Wright, Crispin 2007 ‘On Quantifying into Predicate Position: Steps towards a New(tralist) Perspective’, in: Mary Leng, Alexander Paseau, and Michael Potter, (eds.) *Mathematical Knowledge*, Oxford University Press.

Yablo, Stephen 2001 “Go figure: a path through fictionalism”, *Midwest Studies in Philosophy* 25, 72-102.

Yablo, Stephen 2005 “The myth of the seven”, in: Kalderon, Mark (ed.) 2005 *Fictionalism in Metaphysics*, Oxford: Clarendon Press, 88-115.

Getting out from Inside: Why the Closure Principle cannot Support External World Scepticism

Guido Melchior, Graz, Austria

The canonical version of the argument for external world scepticism has the following structure:

Premise1:	If P does not know that she is not a brain in a vat, then P does not have knowledge of the external world.
Premise2:	P does not know that she is not a brain in a vat.
Conclusion:	Therefore, P does not have knowledge of the external world.

This version of the argument is presented by Brueckner (1994 and 2004), Byrne (2004), Pritchard (2005) and others.

If "a" is any proposition about the external world and "b" is the proposition that P is not a brain in a vat, then the sceptical argument has the following structure:

Premise1: $(\neg K(b) \rightarrow \neg K(a))/(K(a) \rightarrow K(b))$
Premise2: $\neg K(b)$
Conclusion: Therefore, $\neg K(a)$

This argument is logically valid. Therefore, any objection against the argument has to be one against one of its two premises. Subsequently, I will investigate how it can be argued for and against each of the two premises.

Argumentations for premise1: The closure principle and alternative arguments

Premise1 states that any knowledge of P about the external world implies that P knows that she is not a brain in a vat. Why should premise1 be true? Usually, in epistemological discussions, the argumentations for premise1 are grounded on one or the other version of the closure principle. I will, firstly illustrate how the closure principle can be used for arguing for premise1. Secondly, I will present an alternative argumentation.

The closure principle

The closure principle is based on our epistemic intuition that persons can gain new knowledge about facts through inference from facts they already know. The closure principle can occur in different forms. The simplest version is:

C: $(K(a) \wedge (a \rightarrow b)) \rightarrow K(b)$

According to this version, a person knows every proposition which is entailed by a known proposition no matter if the person has any knowledge about this entailment relation itself. This is obviously an implausible account of knowledge through inference. Hence, the more common version is:

CP: $(K(a) \wedge K(a \rightarrow b)) \rightarrow K(b)$

This version expresses the idea that a person, who knows a proposition a and knows that a implies another proposition b, knows b through inference from a. This version captures our intuition that a person can only gain knowledge through inference if the person also knows the entailment relation between the two propositions.

On the basis of the closure principle CP, it can be argued for premise1 in the following way:

Argument1 for premise1:

CP: $(K(a) \wedge K(a \rightarrow b)) \rightarrow K(b)$

The proposition "P is a brain in a vat" and any proposition which P believes about the external world are contradictory.

P knows this contradiction.

Therefore, P infers that she is not a brain in vat, if P has knowledge about the external world.

Therefore, P knows that she is not a brain in a vat, if P has knowledge about the external world.

I think this argument captures the common philosophical intuitions concerning premise1. The first premise of argument1 is the closure principle. One consequence of its second premise is that the term "brain in a vat" has to refer to persons, who not just have unjustified beliefs but who have totally false beliefs about the external world. The third premise implies that the person knows the contradiction between propositions about the external world which she believes and the proposition that she is a brain in a vat, i.e. the person has to be epistemologically educated. Both premises are philosophically acceptable. Argument1 explicates how it can be argued for premise1 by using the closure principle.

According to a common view in contemporary epistemology premise1 is essentially based on the closure principle. Therefore, a popular anti-sceptical strategy is to attack premise1 of the sceptical argument by denying the validity of the closure principle in the context of external world scepticism. This strategy is famously chosen e.g. by Dretske (1970) and Nozick (1981).

These philosophers regard the closure principle in the sceptical context as too strong.

Alternative argumentations for premise1

Using the closure principle is the most popular but not the only possible argumentation for premise1. Premise1 states that P knows that she is not a brain in vat, if P has knowledge of the external world. Premise1 is an implication of the form $K(a) \rightarrow K(b)$. Inferences between a and b can be drawn in two directions. According to this, any implication of the form $K(a) \rightarrow K(b)$ can be interpreted in two ways:

If P knows a, then P infers b from a.
If P knows a, then P has inferred a from b.

In the first case, the direction of inference is from a to b, in the second case it is from b to a. Both interpretations entail the implication $K(a) \rightarrow K(b)$. According to the first version, an inference is always drawn, according the second interpretation an inference has necessarily to be made.

These two interpretations can be applied analogously to premise1. Argument1 for premise1, which is based on the closure principle, obviously corresponds to the first interpretation. The argument for premise1, which corresponds to the second interpretation of K(a) → K(b), is the following:

> Argument2 for premise1:
>
> P can only have knowledge about the external world if she infers it from knowledge that she is not a brain in a vat.
>
> Therefore, P knows that she is not a brain in a vat, if P has knowledge about the external world.

To sum up, there are two possible argumentations for premise1. The common one uses a version of the closure principle for arguing that inferences from "e" to "¬BiV" can always be drawn. According to the alternative argumentation, the inference from "¬BiV" to "e" is a necessary condition for K(¬BiV).

In a next step, I will show that premise2 of external world scepticism can only be true if argument2 for premise1 holds and if argument1, which is based on the closure principle, is rejected.

Argumentations for premise2

I will now analyze possible argumentations for premise2, which states that a person cannot know that she is not a brain in a vat. If knowledge is justified, true belief, there are obviously three possible reasons why P cannot know that she is not a brain in a vat: the impossibility to believe it, the impossibility that the belief is true and the impossibility that the belief is justified. P can obviously believe that she is not a brain in a vat and it is possible that this belief is true. If knowledge is justified, true belief, the problematic aspect for P's knowledge that she is not a brain in a vat is justification.

Taking into account internalistic as well as externalistic theories of justification, the obvious candidates for methods of justification are evidence, inference and externalistic justification. If it is possible that P's belief that she is not a brain in vat can either be evident or externalistically justified, there is obviously no specific problem of justification. Therefore, premise2 can only be true, if P's belief that she is not a brain in a vat cannot be evident or externalistically justified. It is philosophically plausible that this belief is neither evident nor externalistically justified. Hence, it can be accepted that premise2 is true if P's belief that she is not a brain in vat cannot be justified through inference.

Next, I will investigate why it should be impossible that P's belief that she is not a brain in vat is justified through inference by focusing my attention on the following challenge for sceptics: Why is it impossible for person to justify that she is not a brain in vat by inferring it from knowledge about the external world? Why is the following argumentation inadequate?

> The anti-sceptical argument:
>
> K(e)
>
> P infers "¬BiV" from "e".
>
> Therefore, K(¬BiV).

This argument is at least close to the anti-sceptical argumentation proposed by Moore (1925 and 1939).

> Example:
>
> I know there is a computer in front of me.
>
> Therefore, I am not a brain in a vat.

This anti-sceptical argument is an inference from K(e) to K(¬BiV). It is based on the idea that a person can infer from any knowledge of the external world that she is not a brain in vat. The closure principle states: If K(a) and K(a → b), then K(b). Hence, the anti-sceptical argument uses the closure principle for arguing that a person can know that she is not a brain in a vat. Why should this argumentation line be wrong?

Generally speaking, there are three possible reasons why an argument respectively an inference is not a justification:

1. The argument is not preserving truth respectively justification.
2. The premise of the argument is not justified.
3. Other reasons.

I will now investigate each of these three possible reasons:

1. The inference from a person's belief about the external world to the belief that she is not a brain in a vat is truth-preserving. This follows from the fact, that premise1 can only be valid, if "brain in a vat" refers to persons with totally false beliefs about the external world. If premise1 is true, then the inference from a person's belief of the external world to the belief, that the person is not a brain in a vat is truth preserving and, therefore, preserving justification. Hence, in the context of defending external world scepticism the first objection against the anti-sceptical argument has to be rejected.

2. According to the second objection, the premise of the anti-sceptical argument is false, i.e. P's belief about the external world is not justified. The conclusion of the argument of external world scepticism is that P does have knowledge respectively justified beliefs about the external world. This is the same theory as the second objection to the anti-sceptical argument. If it is assumed that the belief about the external world is not justified in order to argue for premise2 of the argument of external world scepticism this sceptical argument itself becomes circular. Therefore, the second objection against the anti-sceptical argument is not adequate in the context of external world scepticism.

3. The first two objections against the anti-sceptical argument are inadequate. Therefore, the anti-sceptical argument only fails, if the inference is deficient for other reasons. The only reason, why a truth-preserving inference with justified premises is not a valid justification, is that the inference leads into a vicious circle. This is the case for the anti-sceptical argument, if a person can only justify beliefs about the external world by inferring it from knowledge that she is not a brain in a vat. This means that justification of beliefs about the external world is only possible through inference from "inside out". Hence, premise2 is only true if this internalistic condition for justifying beliefs about the external world is fulfilled.

Therefore, the correct argument for external world scepticism has to have the following internalistic structure:

Premise1: P can only have knowledge about the external world if she infers it from knowledge about her own sense data and from knowledge that she is not a brain in a vat.
Premise2: P does not know that she is not a brain in a vat.
Conclusion: Therefore, P does not have any knowledge of the external world.

According to this version of external world scepticism, the truth of the premises and of the conclusion essentially depends on internalistic concepts of justification. If it is possible that beliefs about the external are evident or externalistically justified, then the sceptical problem vanishes.

The inadequacy of the closure principle

It is a common view that premise1 of external world scepticism is essentially based on the closure principle which states in this context that it is at least possible for a person to justify that she is not a brain in a vat by inferring it from any knowledge about the external world. It has been shown, that premise2 is only true if a person's beliefs about the external world can only be justified through inference from her knowledge that she is not a brain in a vat. Vicious circles do not lead to justification. Therefore, justification through inference from knowledge about the external world to the proposition that P is not a brain in vat implies that premise2 is false. On the other hand, justification through inference in the other direction implies that the closure principle does not hold in the context of scepticism. If the closure principle holds, then premise2 is false. If premise2 is true, then the closure principle does not hold. Hence, premise1 of the sceptical argument can only be based on argument2, but not on argument1 which involves the closure principle.

One anti-sceptical strategy is to deny the validity of the closure principle. It is based on the intuition that the closure principle supports external world scepticism and that, therefore, arguments against this principle are objections against scepticism. Their sceptical opponents, on the other hand, defend the closure principle in order to defend external world scepticism. As it has been shown this anti-sceptical strategy as well as the sceptical replies are superfluous. If the closure principle holds, then external world scepticism does not exist.

Literature

Brueckner, Anthony 1994 "The Structure of the Skeptical Argument" *Philosophy and Phenomenological Research* 54, 827-835.

Brueckner, Anthony 2004 "Brains in a Vat", in: Edward N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy (Winter 2004 Edition)*.

Byrne, Alex 2004 "How Hard are the Sceptical Paradoxes?" *Noûs* 38, 299-325.

Dretske, Fred 1970 "Epistemic Operators "*Journal of Philosophy* 67, 1007-1023.

Moore, George Edward 1925 "A Defence of Common Sense", in: J. H. Muirhead (ed.), *Contemporary British Philosophy* (2nd series), 193-223.

Moore, George Edward 1939 "Proof of an External World", in: *Proceedings of the British Academy* 25, 273-300.

Nozick, Robert 1981 *Philosophical Explanations*, Cambridge: Harvard University Press.

Pritchard, Duncan 2005 "The Structure of Sceptical Arguments", *The Philosophical Quarterly* 55, 37-52.

Dispensing with Particulars: Understanding Reference Through Anaphora

Peter Meyer, Göttingen, Germany

1. Brandom's theory of reference

The technical notion of *particulars* or *objects* as that which singular terms purport to refer to has been known at least since Frege to be highly theory-laden. Particularly from a non-representationalist point of view it is obvious that one cannot simply take some domain of particulars as somehow simply antecedently understood or given. Instead, even language-specific syntactic patterns may have a crucial bearing on the question of what may, under which circumstances, count as an object referred to (Schneider 1992). To see how deeply this issue is entangled in foundational discussions about reference, it suffices to have a look at its repercussions in philosophical and linguistic debates, e.g., with regard to the question of Meinongian 'nonexistent objects' (cf. the defense of some of Meinong's ideas in Parsons 1980) or to the many difficulties with abstract or fictional objects in, say, 'Millian' theories of singular terms.

From an inferentialist perspective on language pragmatics and semantics, Robert Brandom has tried to explicate the notion of particular object in terms of singular term use, giving a supposedly intra-linguistic, deflationary account of reference and rejecting the idea that reference should be viewed as some sort of word-world relation (Brandom 1994; 2000). In his view, objects are given or specified by equivalence classes of symmetrically intersubstitutable terms; more precisely, reference to particulars is seen as a social practice of attributing and undertaking what Brandom calls symmetric material substitution-inferential commitments, that is, very roughly, commitments to which terms are intersubstitutable *salva veritate*.

Brandom refines this general picture by an account of *anaphora* as the mechanism that lets linguistic tokenings *inherit* the substitution-inferential commitments (and entitlements) associated with other tokenings (Brandom 1994, 455 seqq.). Explaining anaphoric mechanisms is a vital part of Brandom's enterprise since these mechanisms are needed to account for several important and interrelated aspects of his overall account of language: First, an understanding of how *unrepeatable* singular term expressions, such as indexical expressions, provide a link to extralinguistic circumstances is needed in order to explain the *empirical* contentfulness of language use. Second, inter-personal anaphoric connections account for the social, interpersonal nature of linguistic communication.

As has been noted by some commentators, Brandom's theoretical outlook on singular term reference is, on the face of it, rather strained and idealized at least from a linguist's point of view (cf. Fodor and Lepore 2001). To give but one simple example, in many cases the possibility of substituting one singular term for another is subject to grammatical constraints such as concord. In fact, apart from the use of pronominals and other strongly context-dependent proforms it is rarely the case that speakers would use different context-independent descriptions for the same object; they would rather, on different occasions, make several different assertions about an object, assertions that are linked by an *anaphoric* chain. This already suggests that, in some sense, anaphora might be the more primitive, object-constitutive relation vis-à-vis substitutional ones. Typically, it is only by using nominalization constructions that intersubstitutable singular terms come into being in the first place. As far as language acquisition is concerned, it seems plausible that mastery of anaphoric relationships between utterances (such as grasping the kind of semantic link between different utterance tokens containing the word *mama*) precedes, or is at least largely independent of, the ability to use different expressions (such as *mama* and *dad's wife*) that can be more or less always be used interchangeably. Perhaps more important in a philosophical context is the fact that there are serious problems on the semantic side, too; thus, it is possible to *understand* descriptions whose extensions are *known* to be empty, even though, in such cases, no substitutional commitments are undertaken – no other term can be taken to be coreferential with the one given. A remedy for this problem can be found, I think – but it requires looking at the anaphoric relationships empty singular terms may entertain to one another.

2. Inverting Brandom's order of explanation

In order to solve the above-mentioned and other issues, I propose to invert Brandom's order of explanation, taking the notion of *coreference* (i.e., the anaphoric relation holding between coreferential singular term *tokens*) as a theoretical basis for a deflationary account of linguistic reference and 'objecthood'. Similar to Brandom, the result is a non-representationalist view of linguistic reference that does not take some relation between (parts of) linguistic utterances and aspects or chunks of some possible or actual world as its starting point. It is possible (but not conceptually necessary) to explicate the notion of coreference as understood here in terms of Brandomian inferentialism, starting with an account of what an implicit and normative *practice* of *taking* two terms in actual discourse to be coreferential actually consists in. Of course, if one thinks along the lines pursued here, coreference cannot simply be defined as the property of different token expressions to refer to 'the same thing'. Suffice it to say here that, in prototypical cases, coreference can create a 'pragmatic link' between (aspects of) the utterance situation tokens involved – a link that cannot be created by other kinds of subsentential expressions. Two subsequently uttered sentences containing the adjective *red* are normally *not* thereby linked *qua utterance tokens*; in contrast, two subsequently uttered sentences containing the proper name *John* may give reason for a hearer to establish a pragmatically relevant connection between the two utterance tokens (e.g., upon hearing first *John is his room* and then being ordered *Summon John here* the addressee will probably go to John's room in order to carry out the order). The *kind* of 'link' created by coreferential terms is not some simple invariant but correlates, roughly speaking, with the kind of sortal that one would use in talking about phenomena of the respective kind. Deflationary though this approach is, it does allow for reconstructing ordinary 'reference talk', even in a much less artificial way than the one proposed by Brandom: An utterance like *In that utterance, the pronoun 'he' refers to John Doe* is true if and only if the singular term *John Doe* is coreferential with the pronoun *he* in the utterance token in question.

3. Consequences and applications

The account of linguistic reference briefly sketched here shares many of its strong points with Brandom's approach, since, in many cases, Brandom's 'substitutional' analyses have a 'coreferentialist' analogue. For lack of space, only some hints can be given here. Thus, the account is highly neutral with respect to ontological questions, which helps to solve some well-known problems by actually dissolving them. Thus, Quinean concerns with the 'inscrutability of reference' simply do not arise when the basic question is not *which object* is referred to on a given occasion, but which other utterances 'refer to the same thing'. Similarly, the well-known quarrels about, say, description vs. 'Millian' theories, or the Kripke belief puzzles (Kripke 1980), can be given rather down-to-earth analyses that, incidentally, bear some similarity to the line of reasoning in Katz's "new intensionalism" (Katz 2004) without sharing most of its general outlook on semantic or metaphysical questions.

Less perspicuously, the account proposed here might help to throw into relief subtle preconceptions about what may or may not be regarded as a 'proper' particular for a theory of reference. The volatile intuitions surrounding the concept of rigid designation are a case in point: In the narrowly scoped reading of the sentence *The president of France might have been bald*, the subject is usually interpreted as a quantified expression of some sort; in ordinary parlance (not Brandom's, to be sure), it is assumed to denote different individuals in different possible worlds. On the other hand, a rigid interpretation (which takes the nominal phrase to denote some kind of "generic concept") cannot simply be dismissed on *a priori* grounds. Indeed, a 'generic' reading of the nominal phrase in question would seem rather natural in the context of a legal or political discussion of the duties or obligations of 'the president of France' or even 'the present president of France', when what is at stake is not a certain person but properties or requirements concerning a political function. Positing such kinds of individuals is not as far-fetched as it might seem; for instance, the term *the mice that inhabit my kitchen every winter* might be taken to denote a particular with remarkably complex identity criteria, pertaining to a reoccurring temporary presence of a group of animals whose members are possibly different every time.

The present proposal differs, however, markedly from Brandom's in being much less committed to a picture of reference as being 'about particular objects'. On close reading, this picture still figures prominently in Brandom: First, the particulars in a representationalist conception of reference are supposed to correspond to Brandom's equivalence classes of terms; second, singular terms, on Brandom's view, can get the empirical content they have only through entry moves of the language game, specifically, non-inferential perception reports that use unrepeatable linguistic items tied up anaphorically with repeatable ones. Empirically contentful terms are prototypically linked to classes of 'external circumstances' that Brandom calls "reference classes". This way, a surprisingly direct relation between linguistic expressions and a nonlinguistic reference class sneaks in, as it were, through the back door.

In stark contrast, the approach presented here allows for a much broader understanding of the semantic and pragmatic nature of anaphoric relations, even in cases where 'syntactically conditioned' anaphoric relations clearly *cannot* be correlated with the idea of one and the same 'thing' being referred to twice. In a sense, what is proposed here is a radically deflationary attitude towards the notion of object. To take a simple example, the two sentences *John was told that his mother had left* and *Every boy was told that his mother had left* are completely analogous as to grammatical structure; yet, Brandom would, similar to many logic-based semantic approaches, be forced to assume that the semantic analyses of the two sentences differ with respect to quantification, with the consequence that the anaphoric pronoun *his* must be treated differently in these two cases. However, a philosophical account of reference should be able to say something about why this difference is so difficult to explain to a layperson. This requirement becomes more urgent in a strictly *relational* view on grammatical structure as proposed in (Meyer 2003), where I argue that assigning a grammatical structure to a sentence token supervenes on pragmatically grounded relationships between this and other utterance tokens: If such a view is on the right track, then it is difficult to explain why the two sentences can be ascribed parallel syntactic structures in spite of differing significantly in their semantic makeup.

Indeed, the alleged semantic difference between our two example sentences virtually disappears in a 'coreferentialist' perspective, where both sentences receive a parallel treatment: The two terms in question can simply be said to be coreferential in both cases, the difference residing rather in what I have called above the 'kind of link' between the coreferential expressions. In other words, traditional linguistic issues concerning quantification and scope relations can be handled successfully in the account presented here. That the idea of denying any fundamental difference between the anaphoric relations in the two sentences is not simply out of place is suggested by otherwise widely differing conceptions that try to use a unified description for quantified and non-quantified sentences; a remarkable recent example is (Shapiro 2004) who, following ideas of Kit Fine and others, assimilates a sentence such as *Every sheep is white* to the type of sentence exemplified by *Dolly is white* by proposing a logic of arbitrary and indefinite *objects* that is suitable for real-world language processing tasks via semantic networks. Surprisingly, much recent work on knowledge representation in this area is compatible with the non-denotational approach I propose (see also Helbig ^{2}2008) and deserves to be placed under more scrutiny by philosophers of language.

Literature

Brandom, Robert B. 1994 Making It Explicit. Reasoning, Representing, and Discursive Commitment, Cambridge, Mass.: Harvard University Press.

Brandom, Robert B. 2000 Articulating Reasons. An Introduction to Inferentialism, Cambridge, Mass.: Harvard University Press.

Fodor, Jerry and Lepore, Ernie 2001 “Brandom’s Burdens: Compositionality and Inferentialism”, Philosophy and Phenomenological Research 63, 465-481.

Helbig, Hermann ^{2}2008 Wissensverarbeitung und die Semantik der natürlichen Sprache: Wissenspräsentation mit MultiNet, Berlin: Springer.

Katz, Jerrold J. 2004 Sense, Reference, and Philosophy, Oxford: Oxford University Press.

Kripke, Saul A. 1980 Naming and necessity, Cambridge, Mass.: Harvard University Press.

Meyer, Peter 2003 Gebrauch und Struktur. Zu den pragmatischen Grundlagen grammatischer Beschreibung, Berlin: Logos Verlag.

Parsons, Terence 1980 Nonexistent objects, New Haven: Yale University Press.

Schneider, Hans Julius 1992 Phantasie und Kalkül. Über die Polarität von Handlung und Struktur in der Sprache, Frankfurt a. M.: Suhrkamp.

Shapiro, Stuart C. 2004 “A Logic of Arbitrary and Indefinite Objects”, in: Didier Dubois, Christopher Welty, and Mary-Anne Williams (eds.), Principles of Knowledge Representation and Reasoning: Proceedings of the Ninth International Conference (KR2004), Menlo Park, Ca.: AAAI Press, 565-575.

Reichenbach’s Concept of Logical Analysis of Science and his Lost Battle against Kant

Nikolay Milkov, Paderborn, Germany

1. Reichenbach on Logical Analysis

Reichenbach was rather negligent in both the history of philosophy and the philosophical terminology he used. In particular, he introduced a very idiosyncratic use of the concept of “logical analysis” that has little to do with the way that concept was treated by analytic philosophers like Frege or Russell. In *The Rise of Scientific Philosophy*, for example, Reichenbach simply set it against “psychological analysis”. While psychological analysis studies the “errors” of the “speculative philosophers”, logical analysis makes a “rational reconstruction” of the scientific theories. To be more specific, it sets out the principles on which the results of sciences are *really* based, and not simply the way they were set out by their originators. “It endeavours to clarify the meaning of physical theories, independently of the interpretation by their authors, and is concerned with logical relationship alone.” (Reichenbach 1949: 293) In other words, logical analysis investigates the “context of justification”, not the “context of discovery”, of this or that particular scientific theory.

Especially helpful for making logical analyses of science is Hilbert’s axiomatic. In fact, Hilbert’s axiomatic brings with itself the whole logic with the help of which Reichenbach made his “logical analysis” of science.

2. Metamorphoses of Kant’s A priori

Reichenbach set out his program for logical analysis in *Theory of Relativity and A priori Knowledge* (1920). It accepted two fundamentally different principles of knowledge: principles of coordination and principles of connection. Principles of connection are empirical laws in the usual sense, involving terms and concepts that are already well defined. Principles of coordination, in contrast, are not empirical; rather, they must be first established in order to insure such adequate empirical definitions in the first place. In other words, principles of coordination are *constitutive of the object* of every particular scientific theory. In *this* sense (A) they are a priori. Of course, these a priori principles change with every new significant theory; they are not given once and for all. In *that* sense (B), the principles of coordination are empirical. They are the result of new observations and examinations.

For taking this position, Reichenbach was severely criticized by Moritz Schlick. The latter claimed that instead of a priori in a Kantian sense, we can interpret the principles of coordination as conventions in the sense of Poincaré. Schlick’s criticism convinced the young Reichenbach: from 1922 onward, he adopted Poincaré’s terminology of “convention”. More especially, instead of principles of coordination he now spoke about definitions of coordination, underlining this way their conventional character. Unfortunately, this position totally neglected the first meaning of principles of coordination as being constitutive of the object of knowledge.

Assessing this turn in Reichenbach’s philosophy, Michael Friedman has pointed out that Reichenbach “overhastily … acquiesces in the Schlick–Poincaré terminology”. We fully agree with this judgment. We, however, cannot accept Friedman’s claim that with its acceptance, “the most important element in [Reichenbach’s] earlier conception of the relativized a priori is actually lost” (Friedman 1994: 26). In fact, Reichenbach never stopped believing that there are principles that connect the basic concepts of scientific theories with reality. This point is especially pronounced in respect of the so called principle of probability. Reichenbach already introduced it in his Dissertation (1916) and never stopped considering it necessary for any kind of knowledge (cf. Kamlah 1985: 162).

In this paper, however, we are going to track down another trace of Kantian apriorism in Reichenbach’s “new philosophy”. To be more specific, we argue that Reichenbach’s program for a logical analysis of science, which was prominent in his works after 1920 until his death in 1951, was nothing but a transformation of the idea that science contains elements that are constitutive of their objects. In other words, despite the fact that in 1920 Reichenbach officially abandoned the idea that sciences contain a priori elements, he nevertheless continued to explore this in the form of a logical analysis of sciences. How this can be?

As well-known, Kant’s position was that we can formulate all principles that make science possible once and for all through a logical deduction from pure reason: in fact, this was a task of solitary reflection. In contrast, Reichenbach believed that the definiteness of the coordination changes with every new scientific theory. Furthermore, the very idea that every significant scientific discovery brought with itself new principles of coordination posed a new task for philosophy. This was to explicate the new principles of coordination of all subsequent scientific discoveries. Among other things, this latest task led Reichenbach to set up the Berlin Group—a society for scientific philosophy with a clear interdisciplinary coloring. The Group (we shall return to it in § 4), with its most active members Kurt Grelling, Walter Dubislav and Alexander Herzberg, developed in a close relationship with the Vienna Circle.

3. Ambiguity in Reichenbach’s Program

Hartmut Hecht was the first to draw our attention to the fact that Reichenbach’s critic of Kant’s a priori and the method of logical analysis of science are but two perspectives on the one problem of human knowledge (cf. Hecht 1994: 221). Moreover, despite the fact that they were different, it is impossible to conceive of them separately.

Indeed, on the one hand, Reichenbach criticized Kant’s thesis that there is an ultimate table of the categories and principles of the scientific theoretical knowledge that is given once and for all; on the other hand, he claimed that sciences are only possible as long as they have coordination principles which are statements about the logical structure of sciences that change over time. It is exactly this way that the logical analysis of science and the criticism of the a priori made two sides of one point: the task now was not to *criticize* the *pure reason* but to *logically analyze* the *sciences* in order to find out their specific principles of coordination.

Reichenbach insisted that this is not merely a program for popularization of science. Rather, its pursue is exactly as complicated as the studies of science itself, in particular physics, are. To be sure, the philosophers must work hard in order to clarify the results that scientists achieve. This is a necessary work because the scientists themselves are more concentrated on making discoveries: "Scientific research does not leave a man time enough to do the work of logical analysis." (Reichenbach 1951: 123)

4. Reichenbach's Berlin Group and Leonard Nelson as Its Grandfather

To sum up, the task set out by Reichenbach before philosophy was a logical analysis of science: not only of physics but of *all* science. Moreover, he was deeply convinced that his program for "new philosophy" was radically anti-Kantian. In this section we are going to show the inaccuracy of Reichenbach latter claim with the help of a historical argument. To be more specific, we shall refer to the fact that a program close to that of Reichenbach was introduced, much earlier, by the Göttingen philosopher Leonard Nelson (1882–1927) who considered himself a Kantian.

In his philosophy, Nelson closely followed Jacob Friedrich Fries (1773–1843). Fries was Hegel's contemporary and also adversary and rival. He criticized Kant for his "rationalistic prejudice" that we can deduce all a priori concepts from one single principle and in one system. Fries opposed to it the program for analysing the a priori forms of knowledge by "self-observation". To be more specific, he claimed that while the subject of investigation of this program was still the a priori, the way we reach it was a posteriori, or empirical. It was a task of deduction of our a priori knowledge from our immediate knowledge, which, however, also included the scientific knowledge.

Fries' next claim was that metaphysical knowledge, which consists of a priori principles, grows; in other words, it *changes*. In particular, this is true of our knowledge of axioms of mathematics. Leonard Nelson was eager to point out that the growth of metaphysics was especially well demonstrated in its sub-discipline of philosophy of mathematics by the emergence of the non-Euclidean geometry. Indeed, it was discovered after Fries' death and introduced new axioms into it. Moreover, similarly to Reichenbach later, Fries and Nelson claimed that the task of the philosophy of mathematics is to reduce the number of the axioms to a minimum, retaining only those in it which are necessary for the logical *Aufbau* of the theory (cf. Nelson 1928: 110).

The most interesting point is that Reichenbach's two closest friends in the Berlin Group, Kurt Grelling and Walter Dubislav, were faithful followers of Nelson. Indeed, Grelling worked directly under Nelson for more than fifteen years, and while Dubislav had no direct contacts with this philosopher, he worked on Nelson and Fries for years (cf. Dubislav 1926). Apparently, this fact explains the strong theoretical integrity of the Berlin Group.

The main task of the Berlin Group was: an interdisciplinary work on sciences with the aim of establishing their specific principles of coordination. In the light of our analysis of Reichenbach's philosophy we made in §§ 2 and 3, it is clear that this program was nothing but a realization of Reichenbach's program for "logical analysis" of science. In this connection it should be pointed out that Leonard Nelson set up the Fries-Society that, in fact, was the forerunner of the Berlin Group, already before the First World War (in 1913). The Fries society was an interdisciplinary forum for discussions of philosophers, scientists and mathematicians which had its own theoretical organ: *Abhandlungen der Fries'sche Schule* (published between 1903 and 1937).

5. David Hilbert as Reichenbach's Critic

Especially intriguing is the fact that the interdisciplinary program of the neo-Kantian Leonard Nelson also inspired the top mathematician of the time David Hilbert of Göttingen—this to such a degree that the latter believed that he is a Kantian (cf. Majer 1994: 254). In particular, Hilbert claimed that mathematics is based on certain non-logical objects that are subject to our intuition. These are formal structures that have no content; Hilbert called them "ideal elements", or "implicit definitions" of thought.

This fact is puzzling for at least two reasons. (i) As already seen in § 1, Hilbert's axiomatic method played a central role in Reichenbach's program for logical analysis of science. (ii) In 1922, Schlick and Reichenbach were convinced that Hilbert's axiomatic method delivered an ultimate proof that there is no need for Kantian a priori intuition of perceptions in mathematics. How can this puzzle be explained?

Apparently, Schlick and Reichenbach treated Hilbert rather one-sidedly. Indeed, Hilbert's philosophy of mathematics can be interpreted not only as aprioristic, but also as conventionalist. What were the reasons for this oversight?

We have already noted that Reichenbach was a careless terminologist. In particular, he made a very free interpretation of the term "logical analysis". But Reichenbach's use of an inadequate terminology was even more clearly illustrated by his claim that he was an "empiricist". It seems that he had three main reasons for insisting on this point:

(i) It opposed Kant's claim that we can formulate the principles of science once and for all and in our reasoning. Instead, Reichenbach's position was that these principles change with every significant shift of science and so are a result of experience.

(ii) Further impulses to stick to this one-sided terminology came from Reichenbach's crusade in defense of Einstein's theory of relativity against idealistic philosophers of a quite different provenance, such like Hugo Dingler and Oskar Becker. Apparently, Reichenbach believed himself to be an "empiricist" because his opponents rejected empiricism.

(iii) A third reason for insisting on empiricism was the fact that "the opposition against Neo-Kantianism and other kinds of apriorism was a common bond which united logical empiricists and gave them a feeling of being part of a unique philosophical movement." (Kamlah 1985: 158)

Hilbert, who once sat in sessions of Nelson's Fries-Society, closely followed the development of the Berlin Group. Moreover, his assistant Paul Bernays actively participated in the life of the prominent offspring of the Berlin Group—"The Society for Empiric Philosophy". After the analysis we made in this section, it is no surprise that Hilbert criticized the naming of the society "empirical". Reichenbach promptly reacted to this criticism, renaming it into "Society for Scientific Philosophy" (cf. Joergensen 1951: 48). Unfortunately, Reichenbach did not realise that he must also rename his philosophy. Indeed, it was "empiricist" in a very weak sense.

Acknowledgments

The work on this paper was supported by the Fritz Thyssen Foundation.

Literature

Danneberg, Lutz *et al*. (eds.) 1994 *Hans Reichenbach und die Berliner Gruppe*, Braunschweig: Vieweg.

Dubislav, Walter 1926 Die Fries'sche Lehre von der Begründung, Dömiz: Mattig.

Friedman, Michael 1994 "Geometry, Convention, and the Relativized A Priori: Reichenbach, Schlick, and Carnap", in: Wesley Salmon and Gereon Wolters (eds.), *Logic, Language, and the Structure of Scientific Theories*, Pittsburgh: University of Pittsburgh Press, pp. 21-34.

Hecht, Hartmut 1994 "Hans Reichenbach zwischen transzendentaler und wissenschaftsanalytischer Methode", in: Danneberg *et al*. (eds.), pp. 219-227.

Joergensen, Joergen 1951 *The Development of Logical Empiricism*, Chicago: University of Chicago Press.

Kamlah, Andreas 1985 "The Neo-Kantian Origin of Hans Reichenbach's Principle of Induction", in: N. Rescher (ed.), *The Heritage of Logical Positivism*, Lahnam (MD): University Press of America, pp. 157-169.

Majer, Ulrich 1994 "Hilbert, Reichenbach und der Neu-Kantianismus", in: Danneberg *et al*. (eds.), pp. 253-273.

Nelson, Leonard 1928 "Kritische Philosophie und mathematische Axiomatik", *Unterrichtsblätter für Mathematik und Naturwissenschaften* 34, 108-115, 136-142.

Reichenbach, Hans 1920 [1965] *The Theory of Relativity and A priori Knowledge*, transl. Maria Reichenbach, Berkeley: University of California Press.

Reichenbach, Hans 1949 "The Philosophical Significance of the Theory of Relativity", in: P. A. Schilpp (ed.), *Albert Einstein: Philosopher-Scientist*, La Salle (Ill.): Open Court, pp. 289-311.

Reichenbach, Hans 1951 *The Rise of Scientific Philosophy*, Berkeley (CA): University of California Press.

Defining Ontological Naturalism

Marcin Miłkowski, Warsaw, Poland

In most discussions, “naturalism” is thought to be equivalent to “physicalism”. For example, David Papineau’s book (Papineau 1993) doesn’t even mention the word “naturalism,” and uses “physicalism” instead. The standard text-book definitions follow the same pattern (Schmitt 1995; Guttenplan 1994).

In other words, it could seem that the term “ontological naturalism” is simply reducible to “physicalism” and, therefore, can be eliminated from the philosophical vocabulary. I will argue that naturalism is not to be reduced to physicalism, and that both positions should be distinguished. Physicalism must be committed to the view that all objects are physical, and that implies that objects mentioned in special sciences, for example, are reducible to physical. Naturalism doesn’t have to embrace this view. This is not to say that naturalism is necessarily anti-reductive; on the contrary, it has to imply that all objects are natural objects, and that means that they are reducible to natural objects. The main difference is that narrow physicalism implies unity of science, and naturalism can remain neutral towards it, neither denying nor accepting it. To wit, naturalism is a broader notion and covers various other positions such as broad and narrow physicalism.

Physicalism is usually defined in two basic ways: (1) via the notion of a physical object; (2) via the notion of a physical theory (Stoljar 2001). What is notable is that it’s impossible to define the notion of a physical object without implying a theory of it (it’s not observational), and on the other hand, any physical theory will imply a notion of a physical object. So, both ways are inextricably linked to one another but the theory-based approach has two advantages: first, its ontological commitments can be analyzed the same way they are analyzed in the case of scientific theories; second, it will clearly present the theoretical background.

Even if the notion of a physical object is defined in a fashion that is deliberately non-scientific (see Strawson 2006), it will imply a theory at least in a sense that the notion of the physical cannot be taken to be purely observational. If this theory is based on *a priori* assumptions that are outright incompatible with the physics as we know it, it should be dismissed as highly objectionable example of armchair theorizing. In other words, object-based definition of physicalism must be vindicated against the objection that it is arbitrary and unjustified. Even if the definition is supposed to be based on conceptual analysis that starts with an intuitive notion of the physical (as in Strawson 2006), it should be asked which concepts were analyzed to reach this definition. If they are natural language concepts, then there is no guarantee they are correct: we still speak of the Sun rising. If they are scientific concepts, like in the case when the physical object is defined as whatever exists in timespace, it is already theory-laden. Either way, the object-based approach collapses into an implicit version of the theory-based physicalism.

Theory-based physicalism is not trouble-free, however. It cannot offer clear-cut conceptual solutions, if the theory it appeals to is scientific. For example, it is hard to stay nominalist while positing ideally black bodies or timespace points. Moreover, physical theories of the genesis of timespace can hypothesize that there were once non-timespace objects that gave rise to timespace. A theory-based physicalist will have to embrace the claim (if scientifically valid), even if it would seem counterintuitive to her.

This is a minor problem compared to an objection that if theory-based physicalism refers to scientific physics, it is false insofar as current physics is not complete, abounds in tensions between disparate theories (e.g., quantum mechanics and relativity theory) and has obvious gaps. In reply, most physicalists claim that they refer to an ideal physics. Yet, they don’t care to explicate the notion of ideal physics. Carl Hempel formulated a dilemma: physicalism is defined either with current physics, which is almost surely false, or with ideal physics, which is unknown, and therefore cannot be rationally asserted (Hempel 1970). Object-based physicalism, as it implies a theory, has to face it as well.

There are two strategies for dealing with this dilemma. First is to define ideal physics in terms of empirical accessibility (Guttenplan 1994), and the second is to specify requirements that ideal physics will have to meet. The first strategy implies either that physicalism is equivalent to empiricism (including a priori versions) or that empirical access will not be defined in terms of ontological commitments of empirical theories (possibly a posteriori). I suggest that “empiricism” and “physicalism,” are not to be equated; the second possibility boils down to adding some explicit criteria for theories that will be able to identify empirical objects. It seems that either way, physicalism will have to say something more specific about ideal theories.

One of the ways of spelling out Hempel’s objection is to say that physicalists cannot ensure that the future ideal physics won’t include the term “ectoplasm,” or “nonmaterial substance” in its vocabulary. These terms would be worrying for a physicalist because they don’t seem to be reconcilable with the current physics. Currently, referring to ectoplasm boils down to stipulating that there be a miracle happening: There is no place for any entity like that in physical laws. Should there be genuine cognitive progress in physical theories that leads to introducing the term and appropriate laws to physics just the way other entities are admitted in scientific theories, theory-based physicalists would have to accept that. In other words, as soon as ectoplasm is no longer a miracle in a theory, it is not embarrassing for physicalism. In spite of the skepticism about the cumulative nature of scientific theories, it remains relatively uncontroversial that physics remains faithful to methodological naturalism. If this feature of science is relatively constant, then we know enough about ideal physics to be able to refer it, as it won’t admit any supernatural objects. In other words, ideal theories must fulfill the Humean prescription that *explananda* cannot be more improbable than *explanantes* (famous section X of Hume 1902).

Two things might be observed. First, even ideal physics cannot be an absolutely complete theory; it will never be free from cognitive constraints, such as inability to observe infinite physical bodies at once by any finite cognitive agent. Second, the hypothesized complete physics will have limited power of expression; it could not decide logically undecidable problems, or problems that lead to combinatorial explosion.

The ideal physics doesn't have to be conceptualized as the most complete theory of the world in the sense that it would contain all the possible physical knowledge. In other words, it's not what Mary the color scientist would know (by definition she has all the possible knowledge of colors; cf. Jackson 1986); no finite cognitive agent can have all the knowledge due to cognitive limitations. It would be much more limited; by referring to this ideal physics, we mean that we are ready to accept all progress in physical theories that would enhance explanatory, predictive and descriptive powers of the current physical knowledge. To wit, theory-based physicalism defines physical objects as objects that physics is committed to, and physics is understood as the current-day physics including any future enhancements to it. So physicalism claims:

> (P) There exists everything that can be explained by ideal physical theories or observed using the best standard observational procedures in these theories, and whatever is excluded as impossible by ideal physics, doesn't exist.

The above explication of the idea of the ideal physics doesn't imply that physics will turn out to be united or unifying science at all. It just has more explanatory, descriptive and predictive power, while remaining faithful to scientific standards. It is probable that it will remain the most basic and most universal science but we can only hope that it will help us unify special sciences (interdisciplinary unification) or even physical theories (intradisciplinary unification). The claim (P) can be made stronger (or narrow) by adding an explicit condition that the ideal physics will unify the special sciences as the most basic and universal theory. Yet, such a condition is not based on any evidence and as such is simply metaphysically dogmatic and unpalatable for naturalists. Weak (or broad) physicalism doesn't have to be overoptimistic *per definitionem*.

This is one of the reasons why ontological naturalism might seem more attractive than narrow physicalism. While we might hope that physics will be the most basic science, as physical laws are known to be universal, it may turn out that special sciences that deal with objects on other level of organization and with context-dependent phenomena will remain irreducible to physics (or to one of the competing universal physical theories). Even if the microreduction should remain possible if universal laws of conservation are not undermined (the parts of complex systems as described by special sciences will remain reducible to physical processes and properties), the system-level properties, or emergent properties, could be out of the scope of physics.

There is yet a deeper reason for thinking that simple convergence to physics is not a realistic account of science. Natural kinds, and physical objects are a natural kind, are notoriously hard to define with a normal definition. They are rather determined by bundles of laws in which they are referred to. The more independent various determinations are, the more robust the objects (for more on robustness in theories, see Wimsatt 2007). Robust objects tend to appear in several clusters of laws. Real progress of science doesn't invalidate this robustness; as finite cognitive agents, we need several independent ways of confirming that objects are real, and we try to find new ways of doing that. But this also means that any kind of unification is actually detrimental to robustness of the objects we quantify over in theories: we lose ways to re-engineer and correct mistakes in theories, if we replace several theories with one. This is not to say that reduction is necessarily wrong; if successful, it shows that what was thought of as independent, is actually interrelated, and it shows unexpected features of theories.

Moreover, as there is no universal algorithm for discovering physical laws, we must use fallible heuristics instead. The biased heuristics generate different clusters of laws that operate on various levels of abstraction, and unifying them might be not only infeasible but useless as well: add as many heuristics as you might, you'll never get a universal algorithm out if it. So there is little hope for getting rid of heuristics even in the long run.

This is why it seems more appropriate to remain at least neutral towards the unification in science, and endorse a weaker naturalistic position:

> (N) There exists everything that can be explained by ideal natural science or observed using the best standard observational procedures in science, and whatever is excluded as impossible by ideal science, doesn't exist.

(N) is a paraphrase of the famous Sellars adage (Sellars 1956) that science is the measure of things. It doesn't exclude the possibility that it will be physics that will unify sciences via reduction or similar procedures but it doesn't require it. Yet, it shares a certain feature with (P) that needs to be elaborated. It could seem that it's possible that there exist some objects that are inaccessible to science because of the cognitive limitations that are specific to human beings. Though we might try to alleviate this situation by using more instruments and artificial cognitive systems, there will always remain objects that, for example, do not interact causally with anything we might possibly have access to. Doesn't (P) or (N) say that those objects do not exist? The explicit second clause states that the criteria for non-existence should be supplied by a theory. If the existence of such an isolated object X is not excluded by physics in case of (P), or any other science in case of (N), we can remain agnostic towards it. On the other hand, if anyone wants to assert that X exists, (P) and (N) will rather imply we should use standard methodological approaches, and that will include using Occam's Razor against objects with no evidence whatsoever. So, it's far from suggesting that (N) is a version of idealism where the role of the subject is played by science; it's not the science that determines what exists. It's rather other way round: science uses its procedures to see what does exist and what does not.

Ontological naturalism appreciates that we have multiple ways of access to objects on various levels of their organization. Far from denying the role of physics in contemporary science, it is able to integrate special sciences in the realistic account of human knowledge. There is no better source of knowledge than science, and there is no evidence that all special sciences will converge into ideal physics. No ideal physics will be a complete, all-inclusive theory as there are unsurmountable cognitive limitations. We will need different, independent ways of explaining, describing, and predicting the world.

Literature

Jackson, Frank 1986. “What Mary didn't know”, *Journal of Philosophy* 83 (May):291-5.

Guttenplan, Samuel 1994 “Naturalism”, in: Samuel Guttenplan, *A Companion to the Philosophy of Mind*, Oxford: Blackwell.

Hempel, Carl 1970 “Reduction: Ontological and Linguistic Facets”, in: S. Morgenbesser et al. (eds.), *Essays in Honor of Ernest Nagel*, New York: St Martin's Press.

Hume, David 1902 *Enquiries Concerning Human Understanding,* ed. by L. A. Selby-Bigge, Oxford: Clarendon Press.

Papineau, David 1993, *Philosophical Naturalism,* Oxford: Blackwell.

Schmitt, Frederick F. 1995 “Naturalism” in: Jeagwon Kim and Ernest Sosa (eds.)*, A Companion to Metaphysics,* Blackwell.

Sellars, Wilfried, 1956, "Empiricism and the Philosophy of Mind," in Herbert Feigl and Michael Scriven (eds.), Minnesota Studies in the Philosophy of Science, Volume I: The Foundations of Science and the Concepts of Psychology and Psychoanalysis, 253-329.

Stoljar, Daniel 2001 “Physicalism”, in: Edward Zalta, *Stanford Encyclopedia of Philosophy,* URL = <http://plato.stanford.edu/entries /physicalism/>

Strawson, Galen, 2006, “Realistic monism: why physicalism entails panpsychism”, *Journal of Consciousness Studies* 13, 3-31.

Wimsatt, William 2007, *Re-Engineering Philosophy for Limited Beings*, Cambridge, Mass.: Harvard University Press.

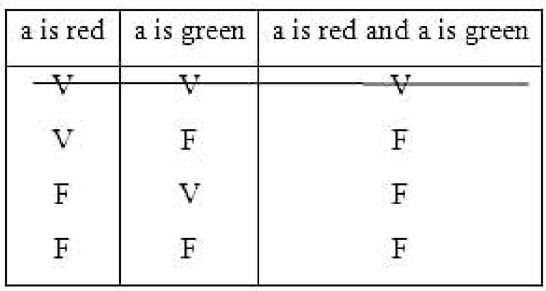
The Logic of Sensorial Propositions

Luca Modenese, Padova, Italy

1. Introduction

Propositions 6.375 and 6.3751 of the *Tractatus* state respectively that there's only a logical necessity and explain it with an example based on colours and the logical impossibility of having two different colours in the same place because of the logical structure of colour itself.

Considering an object named 'a', it seems difficult to accept that if a has two colours at the same time this generates a genuine contradiction (despite of Wittgenstein's statement), because of the absurd situation described by the first line of the truth table presented in Table 1: that needs a syntax consideration about the use of the language of colours to be rejected, and Wittgenstein himself in *Some Remarks on Logical Form* spoke of "*exclusion*" in contrast to contradiction analyzing such a type of logical product*.*



a is red	a is green	a is red and a is green
~~V~~	~~V~~	~~V~~
V	F	F
F	V	F
F	F	F

Truth table of the presumed logical contradiction indicated in proposition 6.3751

In my opinion it is not possible to conclude anything about logical necessity from a table like this, because it's not clear how to manage its first line by a logical point of view, being the "exclusion" not a significant logical operation.

This difficulty is connected to a not satisfactory logical inference theory based on tautologies that derive by independent atomic propositions. I think anyway that it's possible to overcome this problematic situation using the same principles accepted by Wittgenstein in the first phase of his philosophical development.

This paper will try to find a new point of view to consider Table 1 in order to make it clearly logically significant.

2. Sensorial Spaces

It seems reasonable to call propositions like 'This table is white' or 'This wall is rough' or 'This food is salty' sensorial propositions. These propositions describe objects using sense data and can be analyzed using 5 main classes (each related to a sense) and their combinations. This main classification is clear if a sensorial proposition is defined as a statement of a sense impression or as a logical product of statements of sense impressions[1]. 'This table is rough' can be expressed in a form that matches the definition of sense propositions if transformed in: 'This colour spot with the shape that I use to call 'table' is in my field of vision and, when I move the pink shape that I use to call 'hand' near it, I feel a sensation of roughness'.

The distinction of the contributions of the single senses in a sensorial proposition is possible also in more subtle contexts. For instance, a proposition like 'This pullover is comfortable' can be analyzed using fuzzy logic to define in a quantitative manner the grades of each sense involved.

The class of the characteristic properties related to a percipient sense can be indicated with Wittgenstein's metaphor of space.

When a proposition states a certain property of a material subject, the coordinates of a point in the space of that sense will be fixed by it. This specification will exclude other properties of that space because 'a particle […] cannot be in two places at the same time; that is to say, particles that are in different places at the same time cannot be identical.' (6.3751). This proposition of the *Tractatus* describes the main property of the structure of sensorial spaces. Is this structure logical or empirically derived? It depends on what we intend for logic: if we kept in mind proposition 5.552 ('The "experience" which we need to understand logic is not that such and such is the case, but that something *is*; but that is *no* experience. Logic *precedes* every experience that something is *so*. It is before the How, not before the What') it is probably to be considered pure logical.

3. Logic Of Sense Proposition

The logical form of the sensorial propositions will be found using the method described in proposition 3.315 of the *Tractatus*. Taken a certain proposition 'Ra', it is possible to rewrite it in the following form, where all the properties of the sensorial space are involved:

$$R_S a \wedge \sim T_S a \wedge \sim U_S a \wedge etc$$

This form is allowed by proposition 6.3751 and propositions 3.311-3.313 (see also Anscombe 1963, chap. 6, about expressions and symbols). The subscript 's' stands for a generic sensorial space. In details, the written logical product is an attempt to stress with a different propositional sign the structure of the sensorial space involved by the proposition analyzed.

The expression above can be expressed also as:

$$R_S a \wedge \sim \overline{R_S} a \quad \forall \overline{R_S} \qquad \text{with} \qquad \overline{R_S} \in S \wedge \overline{R_S} \neq R$$

where $\overline{R_S}$ could indicate every property of the space S different from R.

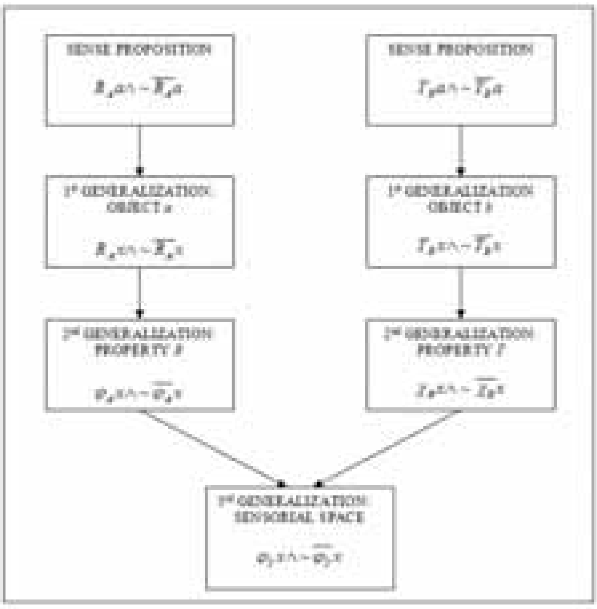
The determination of the logical form of a sensorial proposition follows three steps: 1) the generalization (i.e. the substitution with a variable) of the subject, 2) the generalization of the sense properties indicated in each component of the logical product of the sensorial proposition, 3) the generalization of the space in which the

[1] It will be not considered, at this level of the analysis, the influence if the logical nature of the subject of the proposition.

properties are located. This process transforms a sense proposition until it says nothing but it shows its logical form.

The method could be represented as in Figure 1.

Please notice that the generalization process it's much more subtle than how it seems. Especially the 3th generalization step it's critical, because assuming a *general* sensorial space, it's not enough to derive a logical form from the propositions obtained from the 2nd generalization: also the sensorial properties expressions (in Wittgenstein's use of the term 'expression', see again propositions 3.311-3.313) must be considered akin *a priori*. This is actually not guaranteed by the sensorial space structure.


SENSE PROPOSITION
$R_A a \wedge \sim \overline{R_A} a$
SENSE PROPOSITION
$T_B a \wedge \sim \overline{T_B} a$
1st GENERALIZATION OBJECT a
$R_A x \wedge \sim \overline{R_A} x$
1st GENERALIZATION OBJECT b
$T_B x \wedge \sim \overline{T_B} x$
2nd GENERALIZATION PROPERTY R
$\varphi_A x \wedge \sim \overline{\varphi_A} x$
2nd GENERALIZATION PROPERTY T
$\chi_B x \wedge \sim \overline{\chi_B} x$
3rd GENERALIZATION SENSORIAL SPACE
$\varphi_S x \wedge \sim \overline{\varphi_S} x$

The process applied to write the logical form of the sensorial propositions

At the end of the generalization process, at the level of the logical form, the sensorial space structure is still in evidence. The critical point of the whole process is once again the assumption of the structure of the sensorial spaces (derived from 6.3751) as a logic place where it is possible, defining the "position" of an object, to negate the remaining space for the same object.

4. The Contradiction

Because of the dependency between a sense proposition 'Ra' and all the propositions about 'a' related through that space properties, the strange "*exclusion*" in Table 1 is not more needed. If 'a is red' is 'Ra' and 'a is green' is 'Ga', the logical product became:

$$(R_S a \wedge G_S a) \wedge \sim (\overline{R_S} a \wedge \overline{G_S} a)$$

It is always possible to write[2]:

$$(R_S a \wedge G_S a) \wedge \sim (G_S a \wedge R_S a)$$

because $G_S \in \overline{R_S}$ and $R_S \in \overline{G_S}$.

With this method we finally obtain a genuine contradiction and also the first line in Table 1 is clearly managed and the difficulty by it generated resolved: the first line must be considered valid in the same way as the other and its apparent absurdity is resolved thanks to the second "hidden" term discovered by the analysis of the sensorial propositions.

A consequence of sensorial propositions logical form is that a logic product between propositions of the same sensorial space can be only a contradiction or a tautology (considering the same space-temporal coordinate of course). The logic product of propositions belonging to different spaces instead is never neither a contradiction nor a tautology.

The investigation here presented can be considered an example of the clarifying possibilities of logical analysis.

5. Conclusion

The method described in proposition 3.315 of the *Tractatus* was used to clarify the nature of sensorial propositions, after the assumption for all the sensorial spaces of the logic structure of colours presented in proposition 6.3751. Considering sensorial spaces that are fully involved when a single point of them is used to define an object, it is shown how truth tables derived from sensorial propositions can be expressed in a complete and clear way, useful to investigate their logical properties.

The assumptions and critical points of the process developed have been stressed in a way so that anyone could decide by himself if accept or refuse such a way of proceeding.

Literature

Anscombe, Gertrude E. M. 1959 *An introduction to Wittgenstein's Tractatus*, London: Hutchinson&Co.

Kenny, Anthony J. P. 1973 *Wittgenstein*, London: Allen Lane The Penguin Press.

Ramsey, Frank P. 1923 Critical notice of L. Wittgenstein's Tractatus Logico-Philosophicus, Mind, XXXII, 128, 465-478.

Wittgenstein, L. 1984 *Tractatus logico-philosophicus*, in: Werkausgabe in 8 Baenden, Band 1, Frankfurt am Main: Suhrkamp.

Wittgenstein, L. 1929 Some Remarks on Logical Form, Proceedings of the Aristotelian Society Suppl. vol. 9 162-171.

[2] In a slightly generalized sense, $\overline{R_S}$ and $\overline{G_S}$ indicates here the classes of the properties different from R and G in space S.

A Wittgensteinian Answer to Strawson’s Descriptive Metaphysics

Karel Mom, Amsterdam, The Netherlands

1. Introduction

In the introduction of *Individuals* Strawson expounds his idea of descriptive metaphysics (Strawson 1959 9 ff.; cf. A834 = B862). The subtitle of *Individuals*, “*An essay in descriptive metaphysics*”, indicates that Strawson is concerned with an elaboration of this idea. In this respect, Strawson's metaphysics is meant to be similar to Kant's (B24), which, with Aristotle's, equally is called descriptive. Strawson's subsequent *The Bounds of Sense* is the outcome of his decision “that (he) must try to get to grips with the work [*CPR*, k.m.] as a whole” (Strawson 2003 8). This development could, therefore, arguably be appealed to in making a case for Strawson's Kantianism.

This book, though, has a multifarious purpose. For it is a compound of Strawson's polemic intention to “give decisive reasons for rejecting some parts [of *CPR*, k.m.] altogether” and his reconstructive intention to “indicate (...) how the arguments and conclusions of other parts might be so modified or reconstructed as to be made more acceptable (...)” (Strawson 1966 11). Unsurprisingly, therefore, Strawson's alleged Kantianism became a matter–and probably also a source–of much controversy. A prompt settlement of this stubborn controversy by just qualifying it as a paradigm of *analytic* Kantianism is unlikely. For if it is to be taken as “a distinctly analytic interpretation, defence, and elaboration of Kant's ideas” (Glock 2003 16 f.), this nomenclature can hardly satisfy the sceptic about its distinctive method: connective analysis. Assessing Strawson's Kantianism might, however, clarify its systematic and historical position.

This paper attempts to do so, taking Strawson's exposition of his project as point of departure. I will argue that Strawson's project shares a similarity with the later Wittgenstein; its Kantian remnants, however, prohibit it to team up with the full potential of Wittgenstein's linguistic analysis.

2. Descriptive metaphysics

Departing from Strawson's definition of 'metaphysics' in *The Bounds of Sense,* descriptive metaphysics is the “description of the limiting framework of what we can conceive of or make intelligible to ourselves as a possible general structure of experience” (Strawson 1966 15). This definition might seem ambiguity-ridden, as it relies upon what is meant by 'description'. It thus comprises both the “important and interesting philosophical undertaking” (Strawson 1966 15) of an inquiry, which “is concerned with describing and clarifying the concepts we employ in discourse about ourselves and about the world, and in elucidating their relationships–their forms of relative priority, dependency and interdependency” (Hacker 2003 49), and its outcome.

Hacker observes that Strawson's project is continuous with traditional metaphysics in its quest for the “most general forms of connectedness that permeate our conceptual scheme (...)”, while it departs from it, insofar as it yields “insight only (...) into the forms and structures of our *thoughts* about reality” (Hacker 2003 62; my emph.). This suggests a dissociation of the epistemological element of the description–the aforementioned philosophical inquiry–from the ontological status of the description–the result of that inquiry: “the concepts we employ etc.” It is questionable, however, how this inquiry can be said to be continuous with the tradition even if its result differs categorically from the traditional one: “the (knowledge of) the primary causes and principles” (cf. Aristotle 1989 A 981b 26 – 982a 20).

If such a dissociation is defensible at all, it is improbable in Strawson's case. For, however Strawson's strategy to “develop a conception of the a priori in which pure intuition play(s) no role” justifies his assignment to the semantic tradition (cf. Coffa 1991 22), the shift from ontology to semantics that looms in Strawson's project has not cleared away all remnants of the Kantian convertibility of the epistemic and metaphysical conditions of experience (cf. Aschenberg 1978 335). Hence, the ambiguity of the definition of descriptive metaphysics somehow is inevitable. This is because the way theoretical assumptions about the subject-matter of this inquiry–the conceptual framework to be described–are intertwined with its method, without this intertwining being given due account. For however Strawson acknowledges “having been subtly and in part consciously influenced by it [*CPR*, k.m.] in (his) own independent thinking about metaphysics and epistemology (in *Individuals)*” (Strawson 2003 8), Kant's epistemological considerations regarding the possibility of metaphysics as a philosophical discipline (cf. Kant 1993 §40 ff.), are not parallelled in Strawson. As one critic puts it, Strawson's “novel merger of the virtues of cautious and piecemeal analysis with techniques of bold generalizations and systematic theorizing concerning the character of 'our conceptual scheme' (...) results (...) from a failure to attend sufficiently to the character and implications of their interconnection” (Glouberman 1976 364).

I will illustrate this by the way Strawson demarcates descriptive metaphysics, and, corollarilly, distinguishes its method from reductive analysis. It is on the interface of both components, indeed, that the aforementioned intertwining of theoretical and methodical aspects appears. This intertwining, and the ambiguity that goes with it, constitute a line of justification for Strawson's project, and thus provide, in a way, its prolegomenon. An attempt to construe a counter to relevant objections against it, with recourse to this 'prolegomenon' might show this.

To start with the demarcation, Strawson speaks of a “*limiting* framework of necessary general features of experience” (Strawson 1966 15; my emph.). To grasp the meaning of this phrase–and of its variants that occur throughout the book–I recall that Strawson demarcates descriptive metaphysics which is “content to describe the actual structure of our thought about the world”, from revisionary metaphysics “which is concerned to produce a better structure” (Strawson 1959 9). I will call this demarcation: demarcation$_1$. Strawson also draws a demarcation line (demarcation$_2$) between “gramatically permissible description(s) of a possible kind of experience”, which we could imagine, and a subclass of those: “truly intelligible descriptions” (Strawson 1966 15).

A first objection concerns demarcation$_2$. As Davidson observes, demarcation$_2$ assumes that many imagined worlds are seen from the same point of view. Thus, Strawson's Conceptual Invariance Thesis states that 'our conceptual scheme' "is constant over time and between different languages" (Haack 1979 361). Hence, demarcation$_2$ requires a linguistic dualism of concept and content, insofar it supposes that a fixed system of concepts is used to describe alternative universes. It thus rests on the idea of a distinction between theory and language; mistakenly, though, for meaning is contaminated by theory (Davidson 1984 187 f.).

The effectiveness of this objection could be weakened, however, by differently from Davidson, emphasising the methodological, rather than the semantical aspect of demarcation$_2$. Such a reading could be based on the assumption that demarcation$_2$ can be mapped on demarcation$_1$. Demarcation$_1$ should thus be understood in the aforementioned methodical sense of a particular type of inquiry. In this sense, revisionary metaphysics pertains to regimentations of our ordinary 'discourse about ourselves and the world' (cf. Hacker 2003 49); descriptive metaphysics to the employment of connective analysis as the method of analysis of this discourse. Strawson's methodological exposition of his connective model of analysis, which sets its apart from the reductive or atomist model (Strawson 1985 31 f.), viz. Quine's ontological reductionism, viewed as a consequence of his regimentation of ordinary concepts, and Moore's linguistic reductionism, which overlooks the (inter)dependency of concepts (Strawson 1985 59, 43) supports this reading.

To take full recourse to what has been labelled Strawson's prolegomenon to descriptive metaphysics, this reading should be supplemented by a reading that emphasises the semantical aspect of demarcation$_2$, and which is consonant with a plausible reading of demarcation$_1$. Here, some semantical assumptions of Strawson's logical theory can serve as point of departure. Among these assumptions, which are recurrent in Strawson's work, from *On Referring* onwards, and which are explicitly stated in *Individuals* but form implicitely in *The Bounds of Sense* a heuristic framework for the interpretation and reconstruction of *CPR*, are: (i) "the central importance of the subject-predicate distinction; (ii) "the role of particulars as objects of reference"; (iii) the conceptual "priority of particulars over universals" (Haack 1979 362). These assumptions warrant a semantic reading of demarcation$_2$ which is consonant with demarcation$_1$. It is on their account that Strawson's logical theory renders the modality of a priori necessity to the (inter)connections that make up the significance, in Strawson's sense of intelligibility, of the conceptual scheme which it is intended to describe. As such, e.g. assuptions (i) and (ii) jointly warrant in Strawson the objectivity of referring expressions in a subject-predicate sentence similarly as does the category of substance in Kant (cf. B129). By force of these assumptions, though Strawson does not state this explicitly, demarcation$_2$, of course, echoes Kant's distinction between the use of the categories in mere thinking and their application to intuitions "by which a thing is given" (B146). In this respect Strawson's logic has a transcendental aspect at it. For, although Strawson states e.g. that the performance of the referential task of certain linguistic expressions "requires no transcendental explanation", it is precisely by the use of uniquely referring expressions that "existential assertions may be implied" (Strawson 1950 335, 343).

Interestingly, though, Strawson shares his emphasis on the use of language with Wittgenstein (Wittgenstein 1958 §43), as does his suspicion of logical regimentation of ordinary language (cf. Strawson 1950 344; Wittgenstein 1958 §38). However, unlike in Wittgenstein, where 'use' extends to the grammar of expressions within a language game (cf. Hintikka 1973 55), in Strawson it seems to be restricted to their pragmatic, performative aspects.

A second objection is derived from Stroud and concerns the argument for the a priority of some basic concepts. The criterion to decide whether a concept is basic is whether it answers the demands of Strawson's austere issue (PS_S) of Kant's so-called Principle of Significance (PS_K). By this epistemic principle Kant distinguishes representations which are informed by intuition from empty representations (cf. B75, B150). PS_S "forbids, as empty, the employment of any concept for which no empirical conditions of application could possibly be specified" (Strawson 1966 192). It can be noticed that PS_S resembles the neo-positivist verificationist criterion of meaning, because it likewise is semantic in scope, as it pertains to concepts belonging to a conceptual framework. As such its employment in descriptive metaphysics could resuscitate the objection Stroud raised against the argument, in *Individuals* (Strawson 1959 38 ff.) for our knowledge of the existence of objective particulars. Stroud argues that insofar as its soundness requires the introduction of a verification principle, as he thinks it does, it is superfluous as a transcendental argument against epistemic scepticism about the existence of objective particulars (Stroud 1968 247).

However, although PS_S might resemble the verificationist criterion of meaning, its method of application, unlike the (neo-)positivist method (cf. Coffa 1991 327) is not merely verificationist in the sense Stroud would be inclined to take it, but rather transcendental. For it is applied in a test to establish the a priori status of some concepts within a conceptual framework, rather than their meaning. The scheme of this test is reductive, as opposed to deductive, and, as such, is not logically valid. It argues from a conditional assertion and its known consequent to its unknown antecedent (Bocheński 1954 101, 102 f.). The test, as Strawson conceives it, consists in the sequential performance of this scheme of reasoning in its progressive and its regressive variants, which Bocheński calls verification and explanation respectively. First, the admissibility of a concept as basic is (progressively) verified, by testing if it answers the demands of PS_S. Then its a priority is explained (regressively) by "fram(ing) a counterfactual antecedent from which we could derive (...)" the consequent that "we should have no use for this concept (Strawson 1966 115).

The degree of universality and necessity of the a priori concepts that pass this test is a function of the epistemic use of the powers of our imagination to 'frame counterfactuals'. In contrast, there is no such appeal to such powers in Wittgenstein, where the words "I can't imagine the opposite" e.g. of knowing to feel my pain, merely is a defence against a grammatical proposition being presented in the form of an empirical proposition (Wittgenstein 1958 §251). Therefore, if I can agree with Hacker's assessment of Strawson's descriptive metaphysics as being metaphysics only in an attenuated sense, "just more grammar", that is, "in Wittgenstein's extended sense of the term" (Hacker 2003 54, 59), it is only with the first part of it.

3. Conclusion

To conclude, I have shown that a defence of Strawson’s project of descriptive metaphysics can draw upon an ambiguity in Strawson’s exposition of it, by mobilizing its transcendental tendencies. This is because this ambiguity is due to the unreflective intertwining of theory and method in this exposition and the unconscious denial of a contamination of meaning by theory that goes with it; hence Strawson’s Kantianism.

Strawson’s descriptive metaphysics, as it concerns ‘our conceptual scheme’ about the world, involves a shift of focus from language to the world (cf. Strawson 1950 328f.), unlike Wittgenstein’s descriptive analysis of language, which remains within language games. Due to this digression, Strawson's analysis does not share the full potential of Wittgenstein's analysis.*

Literature

Aristotle 1989 *The Metaphysics. Books 1-9*, Harvard: HUP.

Aschenberg R. 1978, “Über transzendentale Argumente. Orientierung in einer Diskussion zu Kant und Strawson”, *Philosophisches Jahrbuch* 85, 331-358.

Bocheński, J. M.1954 *Die zeitgenössischen Denkmethoden*, Bern/München: Francke.

Coffa, J. A. 1991 The Semantic Tradition from Kant to Carnap, Cambridge: CUP.

Davidson, D. 1984 “On the very idea of a conceptual scheme”, in: *Inquiries Into Truth and Interpretation*, Oxford: OUP, 183-198.

Glock, H. 2003 “Strawson and Analytic Kantianism”, in: H.Glock (ed.), *Strawson and Kant,* Oxford: OUP, 15-43.

Glouberman, M. 1976 “Doctrine and Method in the Philosophy of P. F. Strawson”, *Philosophy and Phenomenological Research* 36, 364-383.

Haack, S. 1979 “Descriptive and Revisionary Metaphysics”, *Philosophical Studies* 35, 361-371.

Hacker, P. M. S. 2003 “On Strawson's Rehabilitation of Metaphysics”, in: *Strawson and Kant,* 43-67.

Hintikka, J. 1973, Logic, language-games and information: Kantian themes in the philosophy of logic, Oxford: Clarendon.

Kant, I. 1990 *Kritik der reinen Vernunft* (CPR; ^{1}1781 = A; ^{2}1789 = B), Hamburg: Meiner.

Kant, I. 1993 Prolegomena zu einer jeden künftigen Metaphysik, die als Wissenschaft wird auftreten können. (1783), Hamburg: Meiner.

Strawson, P. 1950, “On Referring”, *Mind* 59, 320-344.

Strawson, P. F. 1959 *Individuals*, London: Methuen.

Strawson, P. F. 1966 *The Bounds of Sense* (^{8}1993), London/New York: Routledge.

Strawson, P. F. 1985 *Analyse et métaphysique*, Paris: Vrin.

Strawson, P. F. 2003 “A Bit of Intellectual Autobiography”, in: *Strawson and Kant,* 7-15.

Stroud, B. 1968 “Transcendental Arguments”, *The Journal of Philosophy* 65, 241-256.

Wittgenstein, L.W. 1958 Philosophische Untersuchungen-Philosophical Investigations (^{3}1968),Oxford: Blackwell.

* I have benefited from a helpful conversation with Willem R. de Jong about an earlier version of this paper.

Properties and Reduction between Metaphysics and Physics

Matteo Morganti, London, England, UK

1. Tropes

Trope theory is the ontological view that reality is constituted by so-called *abstract particulars* (property-instances not derived from multiply instantiable universals) grouped together in concrete particulars (objects).

Such a view must first of all explain *similarity*, which, so to speak, 'comes for free' if one subscribes to realism about universals. Normally, trope ontologists argue that a trope *a* resembles another trope *b* exclusively in virtue of *a* and *b*, that is, of their primitively given 'causal role' in the world. This may appear more contentious than the realist's claim that similarity is determined by the numerical identity of all instances of the same universal. In fact, however, it is analogous to what the realist must accept insofar as non-exact resemblances are concerned. For, surely non-exactly-resembling entities can still be similar to various degrees, and this must be explained in terms other than numerical identity even in ontologies with universals. Hence the trope ontologist's typical claim of primitiveness appears plausible in this case.

Something must also be said with respect to the way in which tropes constitute complex particulars. Initially (Williams 1953), compresence was taken to be sufficient. However, if compresence is regarded as an external relation additional to the compresent tropes, it seems that a form of regress cannot be avoided: what connects the compresence trope and the compresent tropes? More generally, compresence does not appear sufficient for grounding the internal unity of things: what about overlapping objects?

This leads the trope theorist to account for the inner cohesion of concrete particulars in terms of internal relations of existential dependence among their constituent tropes. The first suggestion in this sense was made in (Simons 1994), who takes his clue from Husserl's foundation relations. As pointed out by (Denkel 1997), if one wants to provide room for substantial change (that is, for the type of change involving the partial or total loss of an object's essence) these relations must be regarded as holding not between specific tropes, but rather between tropes as tokens of more general trope-types (so that replacement of any trope – including essential ones belonging to what Simons calls the 'nucleus', or 'kernel', of the object - with another that acts as determinate for the same determinable is possible).

Lastly, if tropes truly are the basic 'building blocks' of reality, it seems that they had better be understood from the perspective of a sparse conception of properties, according to which not all predicates correspond to actual properties. For obviously not all meaningful predicates can plausibly be taken to refer to fundamental, non-further-analysable tropes. Following the 'scientific' approach to sparseness (Armstrong 1978), this implies that it is necessary to look at physical theory in order to identify the fundamental tropes.

2. Applying Trope Ontology

(Campbell 1990) suggests taking physical fields as the elementary tropes. However, the canonical definition of a field as an extended entity with varying intensities at various points of space seems to suggest an internal complexity and the existence of (dis)similarities between field-values, which immediately leads one to think that something more basic exists.

It seems more advisable to follow (Simons 1994) in looking for fundamental tropes at the level of particles. In fact, Simons' view, based on particles as 'kernels' of foundationally related tropes plus 'peripheral' tropes, just needs some further articulation and specification.

The hypothesis that is taken nowadays to be the best available description of the elementary constituents of reality and their interactions is the so-called Standard Model. According to it, the fundamental particles are 12 fermions constituting matter and 12 bosons mediating forces. Fermions can be either quarks (six types, or 'flavours') or leptons (six more flavours). Bosons comprise photons, W^+, W^- and Z^0 gauge bosons, and eight gluons. Fermions have antiparticles, that is, particles identical to them but with opposite electric (and, possibly, colour) charge. Each boson-type constitutes instead its own antiparticle, except for the W^+ and W^- bosons, which are each other's antiparticle. Each one of these particles has at least one of three possible properties: mass, colour and electric charge, and in most cases they have both mass and electric charge. (Photons may seem to constitute an exception to this latter claim. However, each photon possesses energy, which entails that it can in fact be attributed relativistic mass. True, the latter is distinct from the masses of the other types of particles, as those are invariant masses. Nevertheless, the difference is one of 'form' rather than 'substance': as is well-known, according to relativity theory energy and mass are two 'aspects' of the same thing. Hence, tropes from the same 'family' can be attributed to photons and to the other particles as their 'masses').

In addition to these 'fully state-independent' properties, which remain completely the same throughout the whole of a particle's existence (unless, of course, substantial change occurs), all particles have spin. However, only the absolute magnitude of spin is fixed for each particle (1/2 for fermions, 1 for bosons), while the sign can change. In fact, particles can be in a 'superposition' of the two spin values. How is this to be accounted for from the trope-theoretic perspective?

In general, in quantum mechanics a specific property can be possessed with probability p such that $0 \leq p \leq 1$. Following (Suarez 2007), I suggest a dispositional interpretation of this: whenever a quantum system does not have a specific property with probability 1, it possesses a dispositional property (or 'propensity') corresponding to a weighted sum of possible actual properties. In more technical terms, if Q is a discrete observable for the system Ψ with spectral decomposition given by $Q = \sum_n a_n P_n$, where $P_n = |v_n\rangle\langle v_n|$, and the system is in a state $\Psi = \sum c_n |v_n\rangle$ (a linear superposition of eigenstates of Q for the system), then it is possible uniquely to identify a mixed state W(Q) as the 'standard representative' of Q over the Hilbert space of Ψ. This can be taken as a representation of the dispositional property possessed by Ψ that corresponds to the observable Q. (W(Q) is, in particular, equal to $\sum_n Tr(P_\Psi P_n)W_n$, with $W_n = \frac{P_n}{Tr(P_n)}$).

An important antireductionist conclusion follows: not all properties are actual and provided with well-defined values, or reducible to 'categorical bases': some aspects of reality are in fact irreducibly dispositional.

The other, state-dependent, properties of quantum particles (e.g., position, momentum, kinetic energy etc.), I claim, are mere descriptions of the particles' dynamic 'behaviour' and/or of their relationships with the rest of reality. As such, they needn't be reified, and consequently do not require one to enlarge the set of the basic tropes. For example, space(-time) location is not a trope: it simply expresses the relation between a trope (or trope-bundle) and other tropes (or bundles) - or between tropes and space-time points. In short, the sparse view of properties is intended here as the view that the basic 'building blocks' of reality are only those properties that literally constitute things.

Given the above, the way in which the truly basic tropes give rise to the whole of material reality is readily reconstructed. For instance, it is possible that a trope of mass 0.511 MeV coexists (in a relation of existential dependence of the sort described earlier) with a +1 charge trope and a ±½ spin trope. The individual resulting from the reciprocal existential dependence between these tropes is a positron. The same applies mutatis mutandis for the other elementary particles, and for the progressively more complex levels of reality. For instance, 79 electrons, 79 protons and 118 neutrons give rise to an atom of stable gold. And many such atoms determine molecules and bigger pieces of gold. The 'new' properties of the emerging complexes, such as 'melts at a temperature of 1064.18 C' or 'is a good conductor of heat' for gold, are not tropes, but rather 'derivative' properties determined by the way in which the basic tropes get structured together. This means that they are certainly real, but also analysable in terms of simpler entities. With this, the sparse conception of properties and the 'scientific' approach to their identification find confirmation and application.

The foregoing vindicates the claim that tropes are independent and simple basic constituents of reality (incidentally, it also allows one to get rid of the so-called 'boundary problem' consisting of the fact that tropes are presented as fundamental ontological units but seem in fact arbitrarily divisible: the truly fundamental properties are not divisible, and what is is in fact just a composite trope-structure).

3. Other Properties

As is well-known, quantum mechanics allows for the possibility of many-particle systems in which the (supposed) component entities do not have well-defined values for a given property separately, and are instead mutually correlated with respect to the measurement outcomes concerning that property (even though determinate separate values will appear upon measurement).

The 'non-factorisability' of the 'entangled' states describing such systems into simpler states of the components, it is commonly agreed, points to some form of holism. Namely, to the fact that certain properties of certain physical systems cannot be analysed in terms of properties of the system's component parts, either because the system doesn't have parts (ontological holism), or because it exemplifies properties that are not reducible to the properties of its components (property holism). This entails that the relevant properties of entangled systems should be regarded as emergent tropes in the present context - either monadic and belonging to the whole system, or irreducibly relational.

Here is, then, one more antireductionist theme: the evidence just pointed at blocks all attempts at reducing all properties of physical objects to the 'truly basic' tropes. In the specific perspective of property holism, moreover, this comes together with the impossibility to reduce all relations to monadic properties. This in turn opens the way to a more general rejection of physicalism: for it is possible that non-reducible tropes emerge at levels of higher complexity than physics. (All this, however, needn't worry the trope ontologist, who is not required to commit him/herself to any of these forms of reductionism).

4. Metaphysics and Science

One last point concerns the significance of metaphysics in its relationship with science. In particular, the idea of *'experimental metaphysics'* is of obvious relevance here.

The notion of experimental metaphysics was first introduced by (Shimony 1981), who defined it in the context of a discussion of quantum mechanics, and in particular of the Einstein-Podolski-Rosen (EPR) 'paradox' and of the violations of Bell's inequalities by quantum systems. According to Shimony, a general pattern of reasoning can be individuated of the form E&H → P, where E is an accepted theory used to describe the relevant experimental setup (in the EPR/Bell case, quantum mechanics as it is employed to perform actual tests of Bell's inequalities); H a general (allegedly) metaphysical hypothesis (in the EPR/Bell case, locality as prescribed by relativity), and P a certain empirical prediction (in the EPR/Bell case, that the Bell inequalities hold). If P is disconfirmed and E is kept fixed, says Shimony, by *modus tollens* we should get to a rejection or modification of H, so bringing experiment to bear upon a metaphysical thesis. And this is exactly what happens in the case under discussion, for quantum mechanics forces us to give up locality, or at least to reformulate it in terms that allow for a 'peaceful coexistence' between quantum mechanics and relativity.

The question to ask is, though, whether experimental metaphysics is metaphysics at all. What is at stake in discussions of EPR is the status of what ultimately appears to be only a very general statement extracted from our best-established theories, and that lies *entirely* within the domain of science. Einstein's presupposition to the effect that the world must be local seems indeed to be *exclusively* a consequence of his endorsement of a specific theory (relativity); or, at any rate, of a general worldview that was the by-product of (common sense and) successful theories prior to, and including, relativity. Of course, one might call presuppositions such as locality (or, to give another relevant example, the Principle of the Identity of the Indiscernibles) 'metaphysical', on the basis that they are (among) the most general statements about reality we can make. But this would be a merely terminological choice, and would not detract from the fact that those 'principles' appear to be nothing but empirical generalizations.

Does this mean that metaphysics should be reduced to science? How does this discussion bear on what was said in this paper? Here, I want to suggest one last antireductionist idea. Another possible way of looking at metaphysics, alternative to the view of it as a mere ‘by-product’ of science, is as an autonomous discipline having to do with *hypotheses* rather than truths; and as an inquiry aiming to account for the same reality described by science but by moving at the level of the conjectural rather than at the level of the ‘verifiable’. More specifically, metaphysics is perhaps best understood as an attempt to provide general *categories* and *concepts* that transcend the empirical and yet find their best application when they are employed for interpreting what science tells us. On this construal, certainly metaphysics should not be entirely a priori but – crucially – it does not *reduce* to science either: what emerges is rather a two-way relationship of mutual support, according to which metaphysics provides the conceptual tools for the interpretation of science, and science the data to make sense of our metaphysical hypotheses.

The suggestions made in this paper, based on a prior definition and defence of trope ontology and on a subsequent implementation of it on the basis of our best science, demonstrate that this understanding of the relationship between science and metaphysics may be fruitful.

Literature

Armstrong, David 1978 *Universals and Scientific Realism*, Cambridge: Cambridge University Press.

Campbell, Keith 1990 Abstract Particulars, Oxford: Blackwell.

Denkel, Arda 1997 “On the Compresence of Tropes”, *Philosophy and Phenomenological Research*, 57, 599-606.

Shimony, Abner 1981 “Critique of the Papers of Fine and Suppes”, *Proceedings of the Philosophy of Science Association,* II, 572-580.

Simons, Peter 1994 “Particulars in Particular Clothing: Three Trope Theories of Substance”*, Philosophy and Phenomenological Research*, 54, 553-575.

Suarez, Mauricio 2007 “Quantum Propensities”, *Studies in the History and Philosophy of Modern Physics*, 38, 418-438.

Williams, Donald 1953 “On the Elements of Being” (parts I and II), *Review of Metaphysics*, 7, 3-18 and 171-192.

Functional Reduction and the Subset View of Realization

Kevin Morris, Providence, Rhode Island, USA

1. Introduction

Functional reduction is the view that functional, realized properties are reducible to realizer properties. One way to challenge this reductionism is to develop an account of realization under which realized properties cannot be so reduced. Thus Sydney Shoemaker has argued, and others have concurred, that it follows from the "subset view" of realization that realized properties are typically irreducible. In what follows, I argue that the reductionist can adequately address the challenges raised by this account of realization.

2. Reductionism and the Subset View of Realization

A functional property is one that can be exhaustively characterized, or defined, in terms of a causal role. Arguably at least some mental properties are functional in this sense and it may be that most nonbasic (for instance, biological, mental, economic) properties can be understood in this way (Lewis 1972, Chalmers 1996, Kim 1998 and 2005). Such a property is said to be *realized* by another in virtue of the latter play the role individuative of the former. The reductionist contends that a functional property can be reduced, in a given system, to its realizer in that system (Lewis 1972, Kim 1998 and 2005). There are at least two reasons we might draw this conclusion. First, if a property M is "second order," such that having M is defined in terms of having some other property that plays a certain causal role, it seems that a system's having M cannot be anything beyond its having whatever property P realizes M. Second, it seems that the causal powers of M—the effects that the instantiation of M is apt to bring about—will be identical with those of M's realizer P in a system. If we adopt even a weak causal theory of properties under which different properties cannot have the same powers in the actual world, we are thereby compelled to identify M in S with P (Kim 1998).

While there are a number of issues that will determine whether realized properties can be reduced in this manner, perhaps the most crucial concerns the nature of realization. Of the accounts of realization that have been developed in recent years, Shoemaker's subset view arguably presents the most serious challenge to functional reduction. As on the view just sketched, functional properties are again defined in terms of causal roles and again realization consists in a certain relationship between the role that individuates a functional property and the role played by some other property in a system. Under the subset view, however, P is a realizer of M just in case the effects that the instantiation of P is apt to bring about include as a *subset* those that M is apt to bring about (Shoemaker 2001 and 2007).[1] Shoemaker and Jessica Wilson have argued that realized properties will typically be irreducible under this account (Shoemaker 2001 and 2007, Wilson 1999 and 2002). This will be the case whenever the powers of M are a *proper subset* of those of P. Since M and P have nonidentical powers, they cannot be identified; thus M cannot be reduced to P (Shoemaker 2001 and 2007, Wilson 1999 and 2002). Shoemaker suggests that paradigmatic cases of realization (for instance, the mental by the neurophysiological) are like this. Thus we have the following argument:

> S1. Where P realizes M, the powers contributed by M are typically a proper subset of those contributed by P; thus, in these cases,
>
> S2. M ≠ P; thus, in these cases,
>
> S3. M is irreducible.

I believe that the reductionist can adequately respond to this argument. First, S1 can be rejected by maintaining that a realized property inherits whatever powers are contributed by its realizer to a system and that Shoemaker does not provide reason to think otherwise. Second, the inference from S2 to S3 can be challenged by appealing to the possibility of nonconservative "eliminativist" reduction. Finally, even if S2 follows from S1, this does not entail the failure of conservative reductionism about functional properties.

3. The Subset View and Causal Inheritance

The arguments advanced by Shoemaker and others in favor of S1 are inconclusive at best. Moreover, the reductionist can explain why the powers of a realized property might seem to be a proper subset of the powers of its realizer even if this is not the case.

First, the reductionist need not claim that the powers we ordinarily or *apriori* associate with realized properties correspond exactly to the powers of any realizer. But it does not follow from this that every realized property does not inherit the powers of its realizer in a system (Kim 1998). The picture here is one in which realized properties are understood in terms of a limited set of powers but in which we further reason that given that P realizes M in S, the powers of M in S are identical with those of P, and thus that M "inherits" the powers of its realizer. This reasoning is legitimate in *at least* some cases, as it amounts to the possibility of discovering powers of realized properties in addition to those typically associated with such properties. Moreover, that realized properties are understood in terms of a limited set of powers can explain why it might seem that the powers of a realized property will be a proper subset of those contributed by its realizer even if this is not the case.

Shoemaker considers several cases in which it seems that a realizer has powers beyond those of a realized property. For instance, consider a mental property M, say pain, and its neurophysiological realizer P in humans. While the instantiation of both M and P are apt to cause the subject to wince, it seems that P will have powers beyond those of M: P might contribute the power of producing a "P" reading on a cerebroscope attached to a person's head (Shoemaker 2001). Now, it is true that we

[1] In Shoemaker 2007, realization is officially defined not only in terms of causal powers, but also in terms of "backward looking causal features," what brings about the instantiation of the properties in question. However, Shoemaker suggests that the issues here of interest can be addressed by considering the simpler formulation in terms of powers. Because of this my focus in what follows will be on causal powers, what Shoemaker calls the "forward looking causal features" of the properties.

do not pick out pain in a system by considering whether a property causes "P" to appear on a cerebroscope. Further, instances of pain with a realizer other than P may not cause "P" to appear on a cerebroscope. But we cannot conclude from this that pain in humans does not have this power. For example, presumably pain in humans causes aspirin seeking behavior. Yet it is doubtful that this will feature in a functional characterization utilized to pick out pain in a system; likewise, pain does not contribute this power to nonhuman systems. But just as we cannot conclude from this that pain in humans does not cause aspirin seeking behavior, we should not conclude from like considerations that pain in humans does not cause "P" to appear on a cerebroscope. Analogous considerations can be advanced in response to the other cases put forward in favor of S1, [2] and it thus remains open for the reductionist to contend that he has not been given adequate reason to relinquish his thesis that a realized property inherits the powers of its realizer in a system.

4. An Eliminativist Response

Suppose that in at least some cases a realized property only has a proper subset of the powers of its realizer in a system. Contra Shoemaker, this does not entail S3. This is because S3 does not follow from S2, since S2 does not rule out nonconservative eliminativist reduction. That is, the nonreductive import of the subset view comes *entirely* from the claim that where the powers of M are a proper subset of those of P, M cannot be identified with P. However, the reductionist can reject the assumption that reduction requires identities and maintain that if we cannot identify realized properties with realizers, we should consider the possibility that there are *only* realizers. While this is not ontologically conservative, it can be considered a form of reduction, given its contrast with more radical versions of eliminativism.

This eliminativism is motivated by noting that under the subset view, realized properties are superfluous in that all effects brought about by realized properties are redundant, since they are also brought about by realizers. Now, Shoemaker and Wilson have argued that this should not be regarded as objectionable overdetermination (Shoemaker 2001 and 2007, Wilson 1999). But whether this overdetermination is objectionable is beside the point, since the reductionist can appeal to parsimony to support his commitment to exclusively realizer properties.[3] This amounts to the suggestion that we cease to take those proper subsets of the powers of realizer properties associated with realized properties to determine any such properties.

Further, this eliminativism does not require denying that systems have the powers associated with realized properties. For example, we may have to deny that there is a property of human pain determined by a subset of the powers of pain's realizer in humans. But this does not require denying that the relevant systems have the powers associated with pain; rather, the claim is just that these powers do not determine a property. That this eliminativism does not require denying that systems have the powers associated with realized properties arguably provides a basis for taking such systems to satisfy functional concepts even if we cease to regard the powers in question as determining functional properties. [4] While this is not conservative, it is not the more radical sort of eliminativism under which systems just do not have the powers associated with eliminated properties (Churchland 1979). Nor does it entail that the concept of an eliminated property cannot be useful, since certain powers could be of interest for epistemological and pragmatic reasons without these powers determining a property.

5. Conservative Reduction under the Subset View

While nonconservative eliminativist reduction is a viable response for the reductionist, the subset view does *not* entail that realized properties cannot be conservatively reduced via identities. Even if S1 entails S2, it does not entail the following:

> S4. There is no physical property Q such that Q = M.

This is because even if S1 entails that M cannot be reduced to its realizer P, it does not entail that M cannot be identified with some physical property Q determined by a proper subset of the powers of P. To get S4 from S1, we need the following:

> S5. For any realized property M, if there is a physical property P such that M = P, it must be the physical property that realizes M.

This says that if a realized property is reducible via identities at all, it is reducible to its realizer. Given that S1 entails S2, S1 and S5 together entail S4. But if we assume, in contrast to the eliminativism just sketched, that some proper subset of the powers contributed to a system by a physical property determines some other property, the question is why this latter property should not itself be regarded as physical. Thus we can consider two theses:

> R1. Possibly, M is realized by a physical property P but is identified with Q, where Q is a physical property determined by a proper subset of the powers of P.
>
> R2. If a proper subset of the powers of a physical property P determines a property M, there must be a physical property Q determined by this set such that M = Q.

If R1 holds, then S5 fails and so we do not have a valid argument for S4. The truth of R2 entails the falsity of S4. While both R1 and R2 are in need of an argument, nothing in the subset view entails even the falsity of R1, which is to say that the subset view does not entail S5 and so S4. This means that the subset view at most implies that realized properties cannot be reduced via identities *to realizers*.

[2] For example, Shoemaker and Wilson, following Yablo 1992, consider the case of Alice, a pigeon conditioned to peck at scarlet things but not at shades of red other than scarlet, and argue that scarlet thus has at least one power not possessed by red: the power to produce a pecking response in Alice (Shoemaker 2001 and 2007, Wilson 1999). But even if this power is not ordinarily associated with red and shades of red other than scarlet do not have this power, this does not rule out taking red to have this power in virtue of being realized by scarlet in the system in question.

[3] This is similar to the reductionist argument recently presented in Gillett 2007.

[4] This is similar to the eliminativism that reductionists have advanced in response to multiple realization: given the reduction of M in S to P, M in S* to P*, and so on, we should consider the possibility that there is no structure unrestricted property corresponding to our concept of M (Kim 1998, Lewis 1980). However, given that there is M in S, M in S*, and so on, we can arguably take such systems to satisfy the concept of M.

An initial worry with this line of thought is that it seems to express a sort of “mysterianism” about physical reducer properties. What is needed at this point is a principled account of just what subset of the powers of a physical realizer property we are interested in and whether it is plausible to regard these powers as determining a physical property insofar as they determine any property at all. One nonarbitrary possibility is that the powers in question will be those shared by all of the realizers of a given realized property. The question is then whether we should hold that these powers determine a physical property insofar as they determine any property. While I will not attempt to answer this question here, it should be noted that the worry here under consideration does not threaten the minimal claim that given that the subset view does not entail the falsity of R1, it does not entail the failure of conservative reductionism.

6. Conclusion

For these reasons, then, it seems that the subset view does not mandate the rejection of reductionism about functional, realized properties. First, the reductionist can insist on causal inheritance, and so reject the first premise in Shoemaker’s argument. Second, the reductionist can appeal to nonconservative reductionism about such properties. Finally, the reductionist can contend that without an additional premise, the subset view does not entail the failure of conservative reductionism about realized properties.

Literature

Chalmers, David 1996 *The Conscious Mind*, New York: Oxford University Press.

Churchland, Paul 1979 *Scientific Realism and the Plasticity of Mind*, Cambridge: Cambridge University Press.

Gillett, Carl 2007 “Understanding the New Reductionism: The Metaphysics of Science and Compositional Reduction”, *The Journal of Philosophy* 104, 193 – 216.

Kim, Jaegwon 1998 *Mind in a Physical World*, Cambridge, MA: The MIT Press.

Kim, Jaegwon 2005 *Physicalism, Or Something Near Enough*, Princeton: Princeton University Press.

Lewis, David 1972 "Psychophysical and Theoretical Identifications", *Australasian Journal of Philosophy* 50: 249 – 58.

Lewis, David. 1980. “Mad Pain and Martian Pain”, in: Ned Block (ed.), *Readings in the Philosophy of Psychology Vol. 1*, Cambridge, MA: Harvard University Press.

Shoemaker, Sydney 2001 “Realization and Mental Causation”, in: Carl Gillett and Barry Loewer (eds.), *Physicalism and Its Discontents*, Cambridge: Cambridge University Press.

Shoemaker, Sydney 2007 *Physical Realization*, New York: Oxford University Press.

Wilson, Jessica 1999 “How Superduper Does a Physicalist Supervenience Need To Be?”, *The Philosophical Quarterly* 49: 33 – 52.

Wilson, Jessica 2002 “Causal Powers, Forces, and Superdupervenience”, *Grazer Philosophische Studien* 63: 53 – 78.

Yablo, Stephen 1992 “Mental Causation”, *The Philosophical Review* 101: 245 – 280.

The Writing of Nietzsche and Wittgenstein

Elena Nájera, Alicante, Spain

1. Fragmentary philosophical writing

There is no doubt that we are faced with two writers who are interested in making their thoughts take on a certain form. Nietzsche is sure that "better writing means better thinking" (Nietzsche 1999 2-592). In a similar sense, Wittgenstein insists that the value of his thoughts will be all the greater, the better expressed they are, although he feels obliged to grant them a margin of imperfection: "of all of the sentences that I write here", he points out, "only one or the other will make any kind of progress" (Wittgenstein 1980 §384). And in this respect, precisely with regard to our other writer, he explains:

> "Nietzsche wrote somewhere that even the best poets and thinkers have written things that are mediocre or bad, yet they have separated them from what is good. But it's not exactly like that. Of course, in his garden a gardener keeps roses alongside manure, rubbish and straw; but it is not only their goodness which makes them stand out, but, above all, their function in the garden" (Wittgenstein 1980 §338)[1].

Thus relinquishing an entirely elaborate way of writing, Wittgenstein also compares his philosophical observations with "raisins", which may be the best part of a cake, although adding them does not ensure a perfect, complete form of expression (Wittgenstein 1980 §386). This is why, although he acknowledges that he is captivated by his way of guiding his thoughts towards philosophy, he says that he is not captivated by his own style (Wittgenstein 1997 100). In the prologue of *Philosophical Investigations,* he confesses in that respect his inability to make his thoughts progress in a natural seamless sequence: "After several unsuccessful attempts to weld my results together into such a whole, I realized that I should never succeed". His reflections tend, on the contrary, to "jump all around the subject", finding themselves spread around on "loose notes" and breaking themselves up into "countless pieces" which are impossible to piece back together, like "excerpts from an enormous landscape" in which it is difficult to find one's way (Wittgenstein 1980, §§ 156, 317 & 452). In short, the Wittgensteinian essay submits to the juxtaposition and incompleteness typical of an "album" (Wittgenstein, 1958).

Nietzsche also shares this tendency towards fragmentation and criticises philosophical systems. He introduces himself as "master" of the aphorism and the sentence that guides thought along an unhindered path which only a particularly conscientious reader could follow (Nietzsche 1999 6-153). He acknowledges that the aphoristic form creates difficulties and insists on the great hermeneutic effort which it requires. He in fact claims, "not to write more than that which could plunge «hurried» men into despair", therefore transforming good reading into an art by which nothing is achieved unless it is done "*slowly*" (Nietzsche 1999 5-256). Wittgenstein appreciates calm in intellectual work too and urges the reader to take their time:

> "I really want my copious punctuation marks to slow down the speed of reading. Because I should like to be read slowly (As I myself read.)" (Wittgenstein 1980 §393).

And this is, indeed, the pace set by his writing in spite of its brevity. In this sense, Wittgenstein is aware of the difficulty and obscurity of the extremely short observations which make up his work and, therefore, of the fact that only few readers will be able to understand it, this perhaps being the desired effect as it is possible that, for our two writers, style may be best justified as a discriminatory measure. To this respect, Nietzsche wrote that "all the nobler spirits select their audience when they wish to communicate; and choosing that, one at the same time erects barriers against *the others*" (Nietzsche 1999 3-633). And, with regard to *Philosophical Investigations*, Wittgenstein writes along the same lines that:

> "The book must automatically separate those who understand it from those who do not. [...] If you have a room which you do not want certain people to get into, put a lock on it for which they do not have the key" (Wittgenstein 1980 §34).

But, who has the key to style? In a rough draft of the prologue to *Philosophical Investigations*, Wittgenstein dedicates the book to those who are closest to him in a cultural sense: "my fellow citizens as it were, in contrast to the rest who are *foreign* to me" (Wittgenstein 1980 §495). However, he regrets that:

> "It will fall into hands which are not for the most part those in which I would like to imagine it. May it soon – this is what I wish for it – be completely forgotten by the philosophical journalists, and so be preserved perhaps for a better sort of reader." (Wittgenstein 1980 §384).

The text is aimed, without a doubt, at a close circle of people and requires an interpreter who knows how to handle the language of philosophy in such a way that is *neither journalistic* nor academic, who, perhaps, instead is sensitive to literature and poetry.

2. The limits of writing

Nietzsche and Wittgenstein's styles make an effort to express their thoughts which seems to bring them close to the imaginative or suggestive register typical of poetry. And it may well be said that the former wrote all of his works in the same hand as the creative *poetic reasoning*, as well as composing actual poems. Wittgenstein, on the other hand, confesses to an inability in that respect, which is, however, very significant when determining what is to be expected from his writing: "Just as I cannot write verse", he points out, "so too my ability to write prose extends only so far, and no farther." (Wittgenstein 1980 §336). So his style seems to admit a limit which nonetheless manages to highlight his firm poetic vocation. He writes:

> "I think I summed up my position on philosophy when I said: philosophy ought really to be written only as a form of poetry […] For with this assertion I have also revealed myself as someone who cannot

[1] Wittgenstein refers to *Human, All Too Human I,* § 155.

> quite do what he would like to do" (Wittgenstein 1980 §129).

The fact that it is impossible to give philosophical writing a completely poetic form perhaps justifies its inadequacy. Wittgenstein in fact acknowledges that he perhaps expresses only a tenth of what he wants to express, which make his texts seem like "mumbling" (Wittgenstein 1980 §§100 & 145). In this sense he ends up admitting that not all that one thinks should be written on paper:

> "Really all that can be written —that is, without doing something stupid and inappropriate— is that which emerges in the form of writing. All the rest is comical and comparable to rubbish, so to speak" (Wittgenstein 1997 27).

Nietzsche also seems to number his words and reserves them to tell of some experiences, warning that "one should only speak where one cannot remain silent, and only speak of what one has *conquered"*. The rest is all "chatter", "literature", bad breeding" (Nietzsche 1999 2-369).

In the same way, the proposals of *Tractatus* rule that "whereof one cannot speak, thereof one must be silent", drawing a precise line between the sphere of the *speakable*, the scientific description of the world, and that which can only be *shown*, the mystic (Wittgenstein 1961 §7). Decades later, Wittgenstein continues to insist that "the indescribable (that which seems mysterious to me and which I don't dare to express)" is the background upon which the thoughts that he wants to express acquire their meaning (Wittgenstein 1980 §83). In any case, the question which we are interested in raising is that his literary style favours the unspeakable. The laconic proposals of *Tractatus* create the effect of a certain dogmatism —not in vain did their author intend to convey an untouchable and definitive truth through them—, indicating a road to the mystic which suggests, precisely through the obscurity of his writing, an indisputable clarity.

Turning to the very terminology of *Philosophical Investigations*, it can be affirmed that the aphoristic form which Wittgenstein's writing tends to take on facilitates the *synoptic vision* which provokes understanding, an understanding that consists of "seeing connections" and depends on "finding and inventing intermediate cases" (Wittgenstein 1958 §122). The hermeneutic key to aphorism is, in fact, the capacity to provide examples which forsake an explanation in favour of a merely *descriptive* illustration[2]. And this, without a doubt, forces philosophy to adapt its writing not to a chain of inferences, but to a collection of images which intends to appeal to the personal point of view.

In this sense, Wittgenstein warns his reader that he merely intends to be the "mirror" where he can see his own thoughts with all of their errors, so helping him to correct them (Wittgenstein 1980 §93). In the same way, he seems to abandon discursive reasoning when he affirms that philosophy purely and simply places everything in front of us and does not conclude anything. For this reason, he emphasises that:

> "Writing in the right style is setting the carriage straight on the rails. [...] All we want to do is straighten you up on the track if your carriage is crooked on the rails. But then we'll let you travel alone" (Wittgenstein 1980 §§212-213).

So, the literary way of thinking is in itself significant from a philosophical point of view and reveals something which words cannot say. "Style" is the "expression of a general human necessity [...] seen *sub specie aeterni*" (Wittgenstein 1997 28). With this it is acknowledged that an author's way of writing allows for the understanding of their own particular circumstances and their aspirations to be placed in perspective, seen from outside the ordinary logic of words, reaching a compromise with the undescribable: with the sphere of values, with the mystic.

In accordance with this idea, Nietzsche and Wittgenstein's works may well be an attempt to *show* a cultural situation from a critical point of view, their styles suggesting something more than that which the language of the time —an egalitarian and scientistic era— allows, because that which has been said up to now leads us to suspect that our two thinkers did not have too much faith in their present nor in what their present had to offer, in short, good readers.

3. Where are the good readers?

In the case of Nietzsche, he would actually be contradicting himself if he was to expect to find "ears and *hands*" for *his* truths in life: "that today one doesn't hear me and doesn't accept my ideas is not only understandable, it even seems right to me" (Nietzsche 1999 6-298). In the same sense, Wittgenstein, in the correspondence surrounding the publication of *Tractatus*, proves to be equally resigned to the idea that "nobody will understand it"[3]. And with regard to the "spirit" of *Philosophical Investigations* he regrets the same lack of understanding during that era:

> "This book is written for those who are in sympathy with the spirit in which it is written. This is not, I believe, the spirit of the main current of European and American civilization"
> (Wittgenstein 1980 §§29 & 34).

As we insinuated a few lines ago, the philosophies of our two writers contain, more or less explicitly, a criticism of civilisation which brings them together and in which they collaborate and converge their styles. In a text from 1930, Wittgenstein points out that there are "problems in the western intellectual world" which he has not come up against and which no philosopher has ever confronted, although he specifies in brackets that "perhaps Nietzsche passed them by". To have done so would mean having known how to predict and describe the "odyssey" of the west before its end, something reserved for certain poets, for which reason it should not seem strange that it is written "in the obscure knowledge of premonition and it may only be understandable to a few" (Wittgenstein 1980 §41). That same year, confirming the wisdom of the Nietzschean cultural diagnosis, Wittgenstein wrote about the decadence of the contemporary world:

> "Our age is really an age of the transvaluation of all values. (The procession of humankind turns a corner & what used to be the way up is now the way down, etc.) Did Nietzsche have in mind what now is happening & does his achievement consist in anticipating it & finding a word for it?"
> (Wittgenstein 1997 53).

According to these passages, Wittgenstein seems certainly to have read Nietzsche and to have made use of some of

[2] Cfr. Cavell, Stanley 2004 "The *Investigations´* everyday aesthetics of itself", in: *The Literary Wittgenstein*, New York-London: Routledge.

[3] Letter to Russell, 13.03.1919.

his teachings. He coincides with him in the moral censorship of a world which is united around science, industry and progress and which, because of this, suffers acute nihilism. We are speaking about a world which is impervious to value and to feeling, in which the light has gone out: “it is as if the shine were erased from everything, everything is dead” and “one suddenly realizes that one's mere existence is still completely empty, deserted” (Wittgenstein 1997 198-199). For this reason, in these dark and desolate coordinates, authenticity, the value of the individual, becomes an arduous task: “For in times like these, genuine strong characters simply leave the arts aside and turn to other things and somehow the worth of the individual man finds expression” (Wittgenstein 1980 §29).

For Wittgenstein, cultural disappointment prevails, but he believes that the individual may still have the chance to express himself. It is a question of raising oneself to the higher and undescribable perspective of the mystic, touching upon aesthetic and religious hope. The price to be paid, however, is the creation of something from this feeling which cannot be communicated, cannot be said in the everyday common language, which is the language of argumentation and criticism. So a victory, which could almost be described a Pyrrhic victory, is celebrated of authenticity over the nihilism of the western civilisation.

Literature

Nietzsche, Friedrich 1999 *Sämtliche Werke. Kritische Studienausgabe in 15 Banden,* Munich – Berlin: Walter de Gruyter.

Wittgenstein, Ludwig 1961 *Tractatus Logico-Philosophicus*, London: Routledge &Kegan Paul Ltd.

Wittgenstein, Ludwig 1980 *Vermischte Bemerkungen*, *Culture and Value*, Oxford: Basil Blackwell.

Wittgenstein, Ludwig 1997 *Denkbewegungen. Tagebücher 1930-32/1936/37*, Innsbruck: Haymon-Verlag.

Wittgenstein, Ludwig 1958 Philosophische Untersuchungen, Philosophical Investigations, Oxford: Basil Blackwell.

Word-Meaning and the Context Principle in the *Investigations*

Jaime Nester, Blacksburg, Virginia, USA

In the *Investigations*, Wittgenstein suggests we should, "let the use of words teach (us) their meaning" (Wittgenstein, 2002, p.187). By drawing our attention to *use*, Wittgenstein believes we will see how our linguistic practices confer meaning on words. Though this line of thought seems promising, there may yet be an issue concerning how we come to understand word-meaning. To clarify how word-meaning can derive from use, I will tie Wittgenstein's notion of meaning-as-use to Frege's context principle; in doing this, I will show how Wittgenstein attributes a broader scope to the context principle that extends beyond mere propositions. I intend to argue that Wittgenstein's meaning-as-use shows how Frege's context principle is open to circularity, while his transformation of it is not. To make this argument, it will be necessary to explain what Frege's context principle is and to show how it operates in conjunction with his other two guiding principles. This explanation will enable me to show how Wittgenstein's transformation of the context principle allows him to claim that our linguistic practices confer meaning upon words without opening himself to circularity.

Frege's first guiding principle is, "Always separate sharply the psychological from the logical, the subjective from the objective" (Frege, p.X). Frege believed arithmetic fell under the laws of logic and that the laws of logic govern all thought. Frege is not concerned with the subjective mechanics of thinking, but only with what is essential to thought in order that judgments have truth-values; judgments have truth-values regardless of whether they are ever thought by individuals (Frege, p.36-38*)*. In contrast with psychology, logic is essentially a subject matter concerned with truth. Frege's first guiding principle is aimed at showing how logic furnishes the laws of thought, which makes possible the claims of truth in any other discipline, including psychology (Frege, p.21).

Frege's second guiding principle (the context principle), enjoins us to 'look for' the meaning of a word only in the context of a proposition (Frege, p.X). While this formulation suggests the possibility that words may have meaning in isolation, Frege nonetheless holds one cannot identify or judge the meaning of a word unless it is in the context of a proposition. At this point, the importance attributed to Frege's context principle is that it helps one avoid violating his first guiding principle. Frege holds that if one takes a word in isolation, one may be tempted to take the meaning of that word to be some idea ('*Vorstellung*') one associates with it. Later in the *Grundlagen*, Frege gives a much stronger formulation of the context principle when he states words do not have a meaning when taken in isolation (Frege, p.71). So, it is not that words have a meaning outside the context of a proposition, but rather, the proposition confers meaning on words. Why is a proposition essential to word-meaning? Why is it inessential that we have intuitions associated with words?

When one takes a proper name in isolation (e.g., 'Tolstoy'), it states nothing; it has no truth-value. Likewise, predicates (e.g., 'wrote *War and Peace*) have no truth-value by themselves. In combination, however, a name and predicate express a proposition that necessarily has a truth-value (e.g., 'Tolstoy wrote *War and Peace*). The meaning of the components goes back to the contribution they make to the truth-value of the proposition as a whole. So, the name 'Tolstoy' gains its meaning from the fact that it occurs in a proposition with a truth-value (i.e., the proposition has a sense); whatever subjective impressions I have of Tolstoy are irrelevant to the meaning of 'Tolstoy;' and, 'Tolstoy' cannot be placed together with just any words to produce a proposition. For example, 'Tolstoy Gottlob Frege' does not express anything. Rather, a name must be coupled with a predicate in order to produce a proposition. Why is this the case? Is Frege only drawing on our grammatical knowledge of natural language in assessing what is requisite for a proposition with a sense?

The grammatical categories of names and predicates correspond to features of propositions that make a systematic contribution to the truth-value of a proposition. These features are then logical categories that divide the essential logical components of a proposition. Names correspond to the logical category of 'object,' predicates to 'concept;' the meaning of a name is the object to which it refers, the meaning of a predicate the concept it picks out. The crucial point, however, is that 'reference' in both cases is derivative from the sense of the proposition (Dummett, p.5). The reference of 'Tolstoy' to Tolstoy stems from the sign making a contribution to the sense of a proposition; this contribution shows the meaning of 'Tolstoy.' If 'Tolstoy' did not do that, it would be logically inert, meaningless. The name 'Tolstoy' contributes to the meaning of a proposition by picking out an object; the predicate names a concept and thus contributes to the proposition by picking out a property to be asserted of that object. The object-concept coupling yields a full proposition; this calls us to Frege's third guiding principle: "Always distinguish between concept and object" (Frege, p.X). In some sense, this principle is an outgrowth of the second because it tells us what, within the context of a proposition, is essential to its having a truth-value. On my interpretation, Frege's three guiding principles work in concert to protect the logical values of propositions.

Frege builds word-meaning out of a linguistic calculus that focuses on the truth-value of propositions, and this shows how truth-values derive from a proposition's component parts. Wittgenstein takes issue with this view of meaning because the components of propositions are words, and if words are to be used correctly, we must have some knowledge of their meaning if we are to use them correctly. Frege hints at how this could be a plausible conception of meaning when he claims, "the definition of an object does not really assert anything about the object, but only lays down the meaning of a symbol" (Frege, p.78). Since words operate as symbols of objects for Frege, it seems we could grasp the meaning of words by simply looking at their definitions. This calls our attention to Frege's problem of circularity: the meanings of words are just more words that stand in for them (Wittgenstein, 2002, p.12). Frege distinguishes between sense and meaning, but it not possible for us to grasp the sense of a proposition without first knowing the meaning of its constituent words. For Frege, grasping the sense of a proposition is something we ought not to question because it is a psychological matter; this is problematic because it suggests that the logic of grammar itself provides us with word-meaning. For Frege, grasping the sense of a proposition is supposed to lead us to the meaning of that proposition; but we cannot grasp the sense of a proposition without first knowing how to use words in a meaningful way. Wittgenstein grounds our meaning in use because he realizes no proposition can be understood without some mastery of language. Baker and Hacker claim Wittgenstein turned from Frege's conception of meaning because the various uses a proposition may have cannot be depicted as a mere function of the meanings of its component parts and structure (Baker and Hacker, p.281). To understand the truth-value of a proposition requires that we first know how to use words.

Wittgenstein's focus on use challenges Frege's formulation of the context principle that insists the meaning of a word is tied to the sense of a proposition. For Wittgenstein, "the meaning of a word is its use in language" (Wittgenstein, 2002, p.18). Wittgenstein calls us to "*look and see*" (Wittgenstein, 2002, p.27) how words are used. When we think of actual cases in which we use words, the problem of their meaning disappears; we then can see how words operate in the context of propositions that in turn only operate in a larger linguistic context. Frege's theory of meaning makes it seem as though propositions are intelligible in isolation from the rest of language, but Wittgenstein argues, "There is no such thing as an isolated proposition. For what I call a 'proposition' is a position in the game of language" (Wittgenstein, 1995, p.5) The meaning of a proposition is not to be thought of as something independent of the rest of language; rather, propositions can only be understood in the context of linguistic practices. Our linguistic practices show how we use words, and word-use is directed by the rules of language. Though rules help us understand how to use words, words do not have a fixed meaning or application; words maintain 'family resemblances.' The notion of family resemblances make clear that our application of a word 'resembles' other ways in which we use that word (Wittgenstein, 2002, p.27). Since linguistic practices operate according to the rules of certain linguistic contexts guiding us toward the meaning of words, Wittgenstein describes our use of language in terms of language-games. What are language-games? What do language-games show us about our linguistic practices?

Wittgenstein's use of 'language-game' is not his attempt to offer a systematic account of language, as Frege had done. Rather, Wittgenstein uses language-games to look more carefully at what we do in our linguistic practices while drawing our attention to the limitations of systematic analyses (Stern, p.21). Language-games illuminate the similarities between language and games by calling our attention to the role that rules play in these practices. Though language-games function as heuristic tools, they should not be considered only in this way; 'language-game' employs the use of language itself. Language-games are practices of language, and they exemplify our use of language in certain contexts. By comparing language with games, Wittgenstein underscores the importance of rule-following. How do rules function in language-games, and what impact do they have on word-meaning?

Wittgenstein claims, "A rule stands there like a sign-post" (Wittgenstein, 2002, p.34) This claim draws our attention to signs, which help us understand the role that rules play in word-meaning. When dealing with a sign, we need not interpret the rules of that sign in order to obey it; rather, "Obeying a rule is a practice" (Wittgenstein, 2002, p.69). Following a rule is not a matter of guessing at the intended meaning of a sign; rather, our use of a sign that is in accordance with a certain rule involves explicitly formulating the rule one is following (Wittgenstein, 2002, p.69). For example, when one points a finger, it operates as a sign showing others to look at whatever it may be pointing and not at the finger itself. This elucidates how rule-following operates without inciting widespread ambiguity of a sign's meaning. This is not to say ambiguity never arises; if it does, one must raise questions and provide explanations, but there is no need to explain ambiguity that may arise unless some ambiguity actually does arise (Stern, p.125). Wittgenstein claims: "One may say: an explanation serves to remove or to avert a misunderstanding – one, that is, that would occur but for the explanation; not every one that I can imagine. The sign-post is in order – if, under normal circumstances, it fulfills its purpose" (Wittgenstein, 2002, p.35). The importance that I am attributing to signs is that their meaning is unambiguous because of the role of rule; this point can be made by looking at the role rule play in games.

When rules of a game are taught, one learns a practice that assures obedience to those rules. These practices do not need further explanation because rules guide the moves we make. By following the rules of our language, we can unreflectively understand new propositions and use words without having to raise questions about how we understand their meaning. Wittgenstein argues we show that we understand the meaning of words if we can use them in meaningfully ways; issues concerning word-meaning do not arise in our linguistic practices because rules govern how we use words in context. The contexts of our linguistic practices in which our words have meaning are language-games. Whether a word is in accord with or conflicts with the rules of a language-game stems from the more fundamental concept of obeying a rule. If a word is to have a meaning, it must be used in agreement with the rules of a language-game. I take rule-following to be central to the question of word-meaning for Wittgenstein because rules determine what count as valid moves in language-games. Thus, the context of language-games is that which confers meaning upon our various uses of words for Wittgenstein. It is clear Wittgenstein extends the scope of Frege's context principle to consider our use of words in language-games rather than focusing on the logical role words play in individual propositions. This draws our attention to how linguistic practices are similar to games, which underscore the importance attributed to rule-following and the way in which word-meaning is tied to use. Wittgenstein's transformation of the context principle does not open him to Frege's problem of circularity. Frege sought to ground the meaning of words in the logic of our grammar, but Wittgenstein focuses on our use of language. The move from propositions to use enables Wittgenstein to highlight the conventionality of how our words gain meaning. Since Wittgenstein ties word-meaning to our conventional practices, he avoids Frege's problem about how it is that we can grasp the sense of a proposition without first knowing the meaning and use of its constituent parts (words).

Frege's three guiding principles offer an account of word-meaning that stands open to the objection of circularity. I have argued Wittgenstein's meaning-as-use can best be understood as a transformation of Frege's context principle. By focusing on use, Wittgenstein shows that words have meaning because our use of them follows from the rules of particular language-games. Frege sought to establish the context principle to protect the truth-values of propositions and their components parts from the psychological. Frege's focus on the internal logical relations between the words of a proposition to fix their meaning led him into problems of circularity. Wittgenstein, however, extended the scope of Frege's context principle to underscore the importance that rule-following plays in our linguistic practices, and to situate the meaning of words and propositions within the larger framework of language; "A proposition is a sign in a system of signs. To understand a proposition is to understand a language" (Wittgenstein, 1995, p.131). We understand propositions when we understand the role that words play in a language-game. To formulate propositions, we need to understand the meaning of our words, and we need to know how to use them. Thus, Wittgenstein looks not to propositions for the meaning of words as Frege had, but he focuses on use.

Literature

Baker, G.P., Hacker, P.M.S. 1980 *Wittgenstein: Understanding and Meaning Vol.1*, Chicago: The University of Chicago Press.

Dummett, Michael 1981 *Frege: Philosophy of Language*, Cambridge: Harvard University Press.

Frege, Gottlob 1980 *The Foundations of Arithmetic*, Trans. J.L. Austin. Evanston: Northwest University Press.

Stern, David 1995 *Wittgenstein on Mind and Language*, New York: Oxford University Press.

Wittgenstein, Ludwig 1995 *Philosophical Grammar*, Trans. Anthony Kenny. Berkeley: University of California Press.

Wittgenstein, Ludwig 2002 *Philosophical Investigations*, Trans. G.E.M. Anscombe. Malden: Blackwell Publishers.

Naturalistic Ethics: A Logical Positivistic Approach

Sibel Oktar, Istanbul, Turkey

The view that ethical words, such as good, correspond to a natural object is generally categorised as ethical naturalism. In a wider perspective, it is a view that abandons any link to the supersensible account of ethics. Alternatively, if we recall some of the definitions of G.E. Moore's naturalistic fallacy, like 'reduction of the ethical to the non ethical' and say that this is mostly committed by the naturalists; we arrive at a narrower sense of ethical naturalism. In this sense of ethical naturalism, ethical knowledge, if any, needs to be acquired by experience. Thus, statements of ethical value judgements could be examined in the same way as empirical propositions. The logical positivists' refutation of metaphysics is based on the fundamental idea that any meaningful statement should be capable of being empirically verified. This includes value judgements.

Schlick's position, no doubt, represents such a view. He openly states, in Wittgensteinian sense, that any ethical question that has meaning can be answered, thus if there is a meaningful question then "ethics is a science" (Schlick 1959, p.247). So, before deciding whether ethics is a science or not, we need to answer the question, 'are there ethical questions that have meaning?'

First of all, looking at the fundamental nature of ethics Schlick defines it as 'theory or knowledge'. He puts it clearly that ethics 'seeks knowledge' and it 'seeks to understand' its subject matter. For Schlick, the subject matter of ethics – if we think that it is a science – must be known as clearly as the subject matter of 'biology' or 'optics.'

Schlick thinks that, as we are talking about ethics in ordinary life without difficulty, as we know the word 'light' even before there was such a science of optics, therefore we must know the meaning of the names of the objects in ethics. So restricting the subject matter of ethics to the definition of 'good' is not reasonable, it might have started with defining 'good,' but it should not end when we define it. Although Schlick allows the idea of inventing the concept of good 'quite arbitrarily', he does not accept defining the concept 'completely arbitrarily'; the person who is defining the word 'good' will be limited by some norm as a guiding principle (Schlick 1959, p.250). In this line of argument, R.M. Hare's main criticism of naturalistic ethics is that defining the word 'good' arbitrarily becomes meaningful. Hare points out that this is different from a logician's arbitrary definition of 'his own technical words' to provide clarity. Considering the nature of the study, this way of defining concepts is not acceptable for the word 'good'. As the word 'good' has a function in language, while investigating we should let it function as it is. For Hare, if we change the function of the word 'good' by an 'arbitrary definition' then we are no longer studying the same thing (Hare 2003, p.92). Schlick escapes this criticism by saying that the concept of good is already determined by norms, but whether these norms let the language function as it is depends on what he understands by these norms.

It is difficult, if not impossible, to point at 'good.' At this point, says Schlick, most philosophers develop a false hypothesis that taking the fundamental concept of good given, we possess a special 'moral sense' that point out the 'presence of good.' So we are able to say that good has an objective character. But this hypothesis falls short in explaining the variations in moral judgement. So how would 'ethics' take its place, if it could, in the realm of facts?

Schlick's mention of norms does not presuppose a normative ethics. His method is somewhat similar to Wittgenstein's, he introduces 'normative ethics' as one of the approaches that draws a connection between facts and values, but it is not what takes ethics to be. Having known that Wittgenstein has a great influence on him, it is not surprising that he follows Wittgenstein's steps.

Introducing 'norms' and/or 'standards' to define (to fix) the meaning of the word 'good', reminds us of Wittgenstein's relative sense of value. In "A Lecture on Ethics" (LE), Wittgenstein makes a distinction between relative and absolute senses of value judgements and he says that it is the absolute sense of value judgements that cannot be said. We can express value judgements if they are relative value judgements, i.e., if they correspond to a fact or predetermined standards. However, absolute value judgements do not correspond to facts thus they cannot be expressed.

Schlick's emphasis on norms does not suggest that ethics is a 'normative science,' rather it is the starting point of his quest to define whether it is a 'normative science' or a 'factual science'. As the characteristics of good must be able to point certain facts you could distinguishing the 'formal' and 'material' characteristics of 'good' (Schlick 1959, p.252). Schlick says that in the external or formal characteristic of good, "the good always appears as something that is demanded, or commanded" (Schlick 1959, p.252); as seen in Kant's moral philosophy, in which the formal characteristic is displayed in 'the categorical imperative'. This formal characteristic of good is not only seen in Kantian ethics, but also in others, as theological ethics taken to rest on God's command. Schlick appreciates the formal characteristics of good as a preliminary step, the mistake, he thinks, is considering it as the only characteristic of good. On the other hand, there are material characteristics of good which, for him, need to be considered.

The way Schlick formulates his ideas of what could be the material characteristics of good is very similar to Wittgenstein where in LE he compares substitutes for good to the Galton's composite photographs method, in which I think lies the germ of the idea of family resemblance. Schlick's suggested procedure is looking at the individual cases where we used the word good and search for common features of each case. For Schlick, within the common features of the word 'good' "must lie the reason why one and the same word, "good," is used for the several cases" (Schlick 1959, p.253). This procedure almost echoes Wittgenstein who, in *Philosophical Investigations* (*PI*), says that when we are searching for the meaning of the word 'good' we must look at the language-games where the word 'good' is used (*PI*, 77).

The critical question at this point is, 'are there any such common features?' At first sight it seems that there are more incompatibilities than similarities in various actual cases. Here the question is the universal validity of these common features. Schlick gives the example of polygamy to point out that a discrepancy in ethical judgements is only ‘apparent and not final.’ He states that what is morally judged is not polygamy or monogamy rather what is morally valued is the ‘peace of family’ or ‘order of sexual relationships’. One culture believes that these can be attained by polygamy whereas the other believes that they can be attained by monogamy. Both are trying to attain the same end by different means. What is different is the “virtue of their insight, capacity of judgement or experience” (Schlick 1959, p.254).

Applying this procedure of common features we end up having ‘norms’ as mentioned above. When we apply the procedure to norms it leads us to ‘moral principles’. If we think that the aim of ethics is to determine the concept of good and find out that this can be accomplished by providing moral principles through norms, then we could conclude that ethics is a ‘normative science’. But, for Schlick, positioning ethics as a normative science makes ethics seem something completely different from ‘factual sciences’ and this position is fundamentally false.

The main reason is this: Even if we accept ethics as a normative science, it does not matter whether it is normative or factual; a science can only ‘explain’ and cannot “establish a norm”. He thinks that, if we explain ‘what is good?’ using norms we can only tell what it ‘actually’ means rather than what it should mean. For him, the search for an ‘absolute justification’ of ‘ultimate value’ is senseless. Echoing *Tractatus* (*TLP*) 6.4, which says, “All propositions are of equal value”, Schlick says: “there is nothing higher to which this could be referred” (Schlick 1959, p.257). So, the justification process ends at the highest rule, on which the justification of others depends. What are we trying to attain? Absolute certainty? Schlick states that “[a]ll important attempts at establishing a theory of knowledge grow out of the problem concerning the certainty of human knowledge. And this problem in turn originates in the wish for absolute certainty” (Schlick 1959a, p.209).

Similarly for ethics, the problem turns out to be the certainty of ethical knowledge. For Schlick, theory of norms is not an answer for “the validity of valuation” (Schlick 1959, p.257).Schlick thinks that even if ethics is a normative science, you cannot escape its connection to the factual sciences, because “[t]he ultimate valuations are facts existing in human consciousness” and for him this is “the most important of the propositions which determine its task” (Schlick 1959, p.258).

Although the attempt of the theory of norms does not go beyond trying to find the ‘meaning of the concept of good’, Schlick appreciates it as a preliminary step into the main concern of ethics. But, he immediately adds that “only where the theory of norms ends does ethical explanation begin” (Schlick 1959, p.260).

A system of norms provides “a relative justification of the lower moral rules by the higher” (Schlick 1959, p.261). When it comes to the justification of moral rules and the universality of them, Schlick’s conception of the theory of norms almost brings us to Kant’s hypothetical imperative. Only through the hypothetical imperatives can we talk about the relative sense of values (cf. AK 4:428[1]).

For Kant, hypothetical imperatives do not provide strict universal validity. Here we come across with the idea of causality, which is important both for Kant and Schlick, although they reach totally different conclusions. Schlick says that scientific knowledge “refers to the cause, concerns not the justification but the explanation of moral judgements” (Schlick 1959, p.261). Kant says that moral law is a law of a special causality “just as the metaphysical law of events in the sensible world was a law of causality of sensible nature” (AK 5:47). But this is quite different from what Schlick has in mind when he says that ‘ethics seeks causal explanation’. The difference is that Kant asserts that moral law is “a law of causality through freedom and hence a law of possibility of a supersensible nature” (AK 5:47), Schlick in no way could accept this. I believe, Schlick could sacrifice the idea of strict universality and he could live with the universality that experience provides. Hence his attention turns from justification to explanation.

The explanation of moral judgements takes us into the realm of observable causes and effects. For Schlick, the explanation of moral judgement and conduct is inseparable (Schlick 1959, p.261). So the question becomes ‘why is it *a* standard of conduct?’ rather than ‘what is *the* standard of conduct?’ We need to look at the behaviour of people to understand and explain because a person’s “valuations must somehow appear among the motives of his acts” (Schlick 1959, p.262). Considering that language is also a kind of action Schlick states that: “What a man values, approves, and desires is finally inferred from his actions” (Schlick 1959, p.262).

Schlick suggests that instead of just focusing on moral conduct, it is better to study ‘motives of conduct in general.’ So first we must study the ‘natural law governed behaviour’ and then study moral behaviour, and find what it is that is special in moral action. And this brings us to the conclusion that “moral behaviour is purely a psychological affair” (Schlick 1959, p.263). This does not mean that ‘there is no ethics’ but that ethics belongs to the realm of psychology because its method is psychological.

Separating ‘value judgements’ into two categories, as ‘relative’ and ‘absolute’ is enforced by the fact-value distinction. We could explain certain uses of ‘good’ with the help of facts whereas other uses of the word good could hardly be explained by facts. That is why Schlick and Wittgenstein had the urge to introduce the relative and absolute sense of value judgements. But, at this point, the main difference between Wittgenstein and Schlick is that Wittgenstein was aware that relative value judgments are not problematic, the real issue was in the absolute sense. Schlick never attempted to approach absolute value judgements and tried to explain only relative value judgements. So we can ask, is it really only relative value judgements that we are concerned with ethics? I suppose, this is not what Wittgenstein understands by ethics. Thus, saying that ethics is psychology, is only answering the questions related to relative value judgements. Since ethical discourse related to predetermined standards was never a problematic concept in terms of fact-value distinction, Schlick’s scientific approach to ethics still leaves the absolute sense of ethics as inexpressible.

[1] References to Kant give the pages in German Academy of Sciences (AK) edition of Kant’s collective works.

Literature

Kant, Immanuel 2006 *Critique of Practical Reason*, trans. and ed. Mary Gregor, Cambridge: Cambridge University Press.

Kant, Immanuel 2006 *Groundwork of the Metaphysics of Morals,* trans. and ed. Mary Gregor, Cambridge: Cambridge University Press.

Hare, R.M. 2003 *The Language of Morals,* Oxford: Oxford University Press.

Schlick, Moritz 1959 “What is the Aim of Ethics?”, in *Logical Positivism*, ed. Ayer, A.J, New York: The Free Press, 247-265.

Schlick, Moritz 1959a “The Foundation of Knowledge”, in *Logical Positivism,* ed. A.J. Ayer, New York: The Free Press, 209-227.

Wittgenstein, Ludwig 2005 *Tractatus Logico-Philosophicus*, trans. D.F Pears and B.F. McGuinness, London: Routledge Classics.

Wittgenstein, Ludwig 1965 “A Lecture on Ethics”, *The Philosophical Review*, 74, 3-12.

Wittgenstein, Ludwig 2005 *Philosophical Investigations,* trans. G.E.M. Anscombe, Oxford: Blackwell.

The Evolution of Morals

Andrew Oldenquist, Columbus, Ohio, USA

> "Any animal with social instincts
> would inevitably acquire a moral sense
> as soon as its intellectual powers
> became like those of humans."
>
> Charles Darwin, *The Descent of Man*, Ch. 4

We have ancestors, 100,000 years ago, I'll guess, who had no morality–no moral concepts, moral beliefs or moral codes. We have more recent ancestors who did have moral beliefs and moral codes. What happened in between? By what describable changes did our earlier ancestors' anger at theft become moral disapproval? There are two parts to my explanation of this change: an account of how most of the content of current morality resulted from the evolution of love and human sociality, and second, bridge theories, which are lists of word usage descriptions that tell us when a positive or negative feeling turns into a moral belief. From facts about innate sociality and language I shall derive "S believes A is wrong," but not "A is wrong." Moreover, unlike most definitions of "good" or "morally wrong," a description of usage can convey the function of moral language without designating anything that is morally right or wrong.

The consensus of paleoethnologists is that humans evolved biologically to be social animals, which included the evolution of certain wants, fears and anxieties required for social living and which then were culturally reinforced. Even in pre-linguistic societies some behavior had to be taboo and deterred by fear of punishment or banishment.

Philosophers and scientists have long tried to explain altruistic motives, given that they appear to diminish likelihood of survival and therefore ought to be selected against in evolution (Hamilton, 1964). It is widely believed either that only perceived self interest can move us to act, or that both morality and self-interest are effective motives for action. Both alternatives depend on a false dichotomy that gets its plausibility from the distinction between particulars and kinds. The object of self-interest is a particular, not a kind of thing: my self-interest attends to me but even in the same circumstances not necessarily to my clone or identical twin. But morality, it is said, may judge a person only by qualities other people can have too such as cruelty, kindness or unfairness. No rule of social morality can refer to me and consequently moral terms designate qualities, not particulars.

However, group egoism generates moral judgments that combine descriptions and egocentric particulars: "Because it's mine" is as fundamental as "because it's me." Group egoism explains a large part of social morality including obligations based on love and loyalty to my mate, my children, my clan or country. It can conflict with egoism as well as with impartial principles.

It will not do simply to say that if I may do something everyone may, for the natural response is, "Every what?" Every fellow club member, fellow American, fellow Christian, fellow human being, rational being, suffering being? These nested and overlapping domains of course make morals complicated. Social morality's constraint within domains defined by group loyalties and social identities shows there is no sharp line between self-interest and altruism and that the possibility of altruism is not the fundamental question of moral psychology. The neo-Darwinian explanation of group loyalties as well as kin selection is that they are non-universalizable outside of a designated group because they fix on the physical coordinates of where one's DNA type is likely to be located, or where protectors or caretakers of it such as one's clan or country are located.

1. Kin selection, as developed by William D. Hamilton (Hamilton, 1964), is caring for relatives according to their degree of relatedness and it evolved independently of motives or understanding, as in the clear case of the social insects. Human parents value their child, who has one half of each parent's DNA, more than their grandchild, who has one fourth, and their grandchildren more than mere friends.

2. Increasingly prolonged infancy and the dependence of young children were made safer by the evolution of parental love, loss of estrus and sexual romantic love. Each of these increased the likelihood that dependent young children would have both parents around long enough to survive on their own. Love, like loyalty, makes certain behavior feel necessary independently of considerations of self interest. It is our strongest passion, explains our strongest feelings of obligation because they most directly protect our DNA, and shape our world. Love is directed to a particular and not to a kind of thing because it evolved to protect one's children, who have a particular location.

3. A number of mutually reinforcing things evolved to make us innately social, including kin selection, love, group loyalty, the felt need to belong, and fear of banishment. Feelings of security when living amongst familiar people with familiar social practices and in familiar spaces, fear of being outcast, and the world-wide development of ritual and ceremony, are all constitutive of human sociality.

Kin selection cannot explain altruism on the broader level of the clan. What was selected for was clan loyalty and other varieties of group egoism which do not depend at all on how close one is genetically to fellow clan members. Group loyalty was selected for because people in clans were safer than those who lived alone or just with immediate family.

Evolved emotional predispositions include our need to belong to groups and acquire social identities and loyalties, all of which makes the group fare better and thereby protect us better than if there were no group loyalties and social identities. Love, kin selection and innate sociality constitute the evolutionary basis of social morality and explain actions felt to be necessary independently of self interest. Arriving more recently than kin selection and love, loyalty made individuals emotionally dependent on clans and willing to sacrifice for them. This is in our DNA because those who clove to their clan were more likely to survive and pass on this disposition, whereas those who lacked such an attachment were more likely to wander off and starve, be killed by an enemy tribe or be dinner for a big cat. Another way to view a clan is as an advantageous environment to which individuals adapted.

What we now require is a bridge theory–something which, based on the preceding, says how ordinary likes and dislikes differ from moral beliefs. The bridge theory lists conditions I call marks of the moral. Satisfying the marks of the moral tells us that S believes or at least asserts that A is immoral, but it does not tell us that A is immoral. Our ancestors had moral beliefs and moral codes when their aversions, hates and likings came to satisfy the marks of the moral. If their beliefs only partly satisfied these conditions they would have had borderline cases of moral beliefs. We want to think a judgment is either moral or non moral, but in human affairs almost everything shades off into what it isn't. A dislike or negative attitude toward something turns into a moral belief or moral judgment when enough of the following features characterize it:

1. It concerns benefit and harm to humans and the higher animals.
2. It is communicated by special words.
3. It appeals to reasons that have a general appeal in the community.
4. It is universalizable, that is, a person is willing to judge similar cases similarly, even when one of these cases concerns oneself.
5. It can require actions contrary to self-interest.
6. It is taught to the young.
7. It is all things considered, that is, it judges an action in the light of self-interest, effects on others, and anything else thought relevant.
8. It often is promulgated ritually and ceremonially, as a way of indicating that the community and not just an individual is speaking.
9. It expresses a positive or negative attitude toward the object of the judgment.
10. It is urged upon the listener and rejection of what is urged is answered with anger or argument.
11. It is preached in formal religious and political settings.

If a clan spoke, reasoned and acted in ways 1-11 they had morals, but if in not enough of these ways they did not. The list aims to describe the contexts and conditions under which reasonably educated English speakers use "morally wrong," etc., and is what I suggest should replace definitions of moral words. This is in the spirit of Ludwig Wittgenstein's admonition, in Philosophical Investigations, to consult the use, not the meaning. These eleven conditions, singly or together, carry neither moral realist nor emotivist theoretical implications.

Before they had language our ancestors had to be social. The emotional predispositions for sociality had to evolve before the evolution of language, the latter requiring the evolution of the brain's speech center, the voice box, infant babbling and, of course, people to talk to. Could they have moral beliefs? My suggestion is they could not if they couldn't talk, and therefore couldn't give reasons and argue. The evolutionary sequence had to be sociality first, then language, and finally morals.

We can conjecture how particular moral ideas arose. For example, sense of unfairness, a moral idea, very likely can be deconstructed into clan rejection anxiety. The mechanics of unfairness is relatively straightforward; it is being denied benefits others receive in similar circumstances. A clue to understanding this is the outrage and anger perceived unfairness elicits, typically more than from equally harmful illnesses, accidents or combat. Unfairness has little to do with degree of perceived harm and everything to do with actual or symbolic exclusion, with being treated as an outsider or non-member when one is not an outsider.

When young people are not shamed or blamed for behavior for which others are shamed and blamed, they are being treated like outsiders or non-members, that is, like invading Huns or wild animals. The result is alienation, a loss of sense of belonging and hence loss of one's social identity. Given that these young people evolved to be innately social animals like the rest of us, they seek substitute social identities in gangs or counter-culture groups. Alienation kills sense of belonging, and hence pride and shame on which traditional social control largely depends.

Another bridge theory provides an explanation of retributive justice. It is often said that retribution is revenge and therefore has no moral status. Retributivists explain retribution in terms of desert, reciprocity, or making things even again, so as to distinguish the moral idea of retributive justice from the non-moral (or immoral) idea of revenge. I accept that retribution is a moral idea and revenge is not. However, explaining retributive justice without incorporating revenge is hopeless. Revenge turns into retributive justice when the desire to harm wrongdoers is constrained by the following empirical conditions (or by an improved version of them):

1. Those who decide how, if at all, to punish A are neither A's relatives or friends nor stand to gain or lose from the decision.
2. Similar punishments are given for similar offenses.
3. The punishment is decreed in a setting of formality and ritual, which conveys the idea that the community and not just an individual is speaking.
4. Punishments are not secret but are codified and promulgated by an appropriate official body.
5. Criminals must be believed to have actually done the deed for which they are being punished.

Retributive justice thus is sanitized revenge. Vengeance, the idea of a person being owed something bad, is fundamental to humans, showing itself not just in criminal justice but also in countless informal interactions such as ignoring or snubbing someone, cursing them, ignoring them, refusing to invite or to help someone, assaulting them and so on. Personal accountability is a primary way societies distinguish members from non-members. The anthropologist Christoph von Furer-Haimendorf (Furer-Haimendorf, 1971) explained criminal justice as the institutionalization and ritualization of retaliation as societies became sufficiently secure and complex.

But are moral judgments true or false, do they assert moral facts? These moral realist claims are logically compatible with the explanation of morality I have laid out. But must genuine moral judgments assert moral facts or be literally true or false? Many people, philosophers as well as non-philosophers, believe this is part of what moral words mean and they would feel that morality is an illusion or a fraud if moral judgments were never true or factual.

Suppose there is an antiquated community where shepherds tell time, direction and the seasons by watching the stars and planets. When asked what stars are they say the stars are gods. One of them is persuaded, with the aid of telescopes and a little schooling, that the stars are not gods. He might respond, "Rats, stars don't exist" and stop looking at them, on the ground that part of what he means by “stars” is “gods.” Or he might be persuaded that what he called stars are still useful for telling time, etc., and that he might as well keep calling them stars. If he does, we need not conclude that “gods” isn’t part of what he meant by “stars.” Rather, he was persuaded to give up part of what he meant by “star” in the light of plausible empirical claims. We chiseled off part of what he meant by “star” but the ways he used the stars were not affected. So too, even if part of what people meant by “morally right” and “morally wrong” were moral facts, and moral facts do not exist, might we not chisel that off without their needing to conclude that nothing is right or wrong?

The salient truths are the empirical ones: Our society and our security depend on honesty, fairness, and keeping unwanted hands off other people’s bodies and property. The differences between our morals and premoral clans that just yelled and banished people for violating taboos are smaller and more enlightening than some people might think.

Literature

Boyd, Robert, 2006, “The Puzzle of Human Sociality,” Science, December 8

von Fuerer-Haimendorf, Christoph, 1964, Morals And Merit, Chicago, University of Chicago Press

Hamilton, William D., 1964, “The Genetical Evolution of Social Behavior, I,” Journal of Theoretical Biology,

Oldenquist, Andrew, 1984, “Loyalties,” Journal of Philosophy;

Oldenquist, Andrew, 1988, “An Explanation of Retribution,” Journal of Philosophy

Species, Variability, and Integration

Makmiller Pedroso, Calgary, Canada

1. Introduction

Different from the visible spectrum, the variation among living organisms does not form a continuum. Because of evolution, life comes discretely organized in clusters called **species**. We refer to these clusters via **species names** such as "*Homo sapiens*" or "*Drosophila pseudoobscura*".

One can pick a particular species name and ask what it refers to. A different kind of question is to wonder what is common between the referents of every species name. That is, one may ask what every single species has in common. This paper is about the latter question. In particular, the main aim of this paper is to assess the answer provided by Boyd (1991; 1999).

Clearly, a satisfactory account of the referent of species names has to capture what is characteristic about species. So before presenting Boyd's view, I discuss two important features of biological species. Firstly, I discuss the extent to which the individual traits within a single species can vary (Section 2). In Section 3 I discuss how this variability is balanced with some integration. Based on those two sections, I derive two desiderata that a satisfactory conception of species should satisfy. Section 4 motivates Boyd's position in face of these desiderata. The last section offers some objections against Boyd's view.

2. Biological species and variability

According to **essentialism** concerning species, all and only members of a species necessarily share an intrinsic property.[1] That is, an organism cannot be a member of a species without sharing a certain intrinsic property. This property is commonly known as an **essence**. So, for the essentialist picture, the living world comes in "packages" because (1) every member of a certain species necessarily shares an essence; and (2) different species have different essences. However, despite its explanatory power, essentialism appears to be incompatible with contemporary biology.

As Okasha (2002) points out, the incompatibility between essentialism and biology has empirical and conceptual grounds. On the empirical side, we find examples of species that exhibit intra-specific variability which rules out the possibility of species essences. On the conceptual side, even if all and only the members of a certain species share some intrinsic property, this property does not count as necessary for membership to the species. I now turn to these criticisms.

The essence of a species can be either phenotypic or genotypic features of its members. Let us first consider the case in which essences are taken to be phenotypic.

As mentioned earlier, essentialism can be understood as comprising two assumptions: (1) every member of a species shares the same essence; and (2) different species have different essences. Assuming that essences are phenotypic, there are two sorts of examples of species in biology that go against both (1) and (2). As to (1), there are examples of polytypic species that are immensely diverse in terms of phenotypic traits. One example is the butterfly species *Heliconius erato*. Concerning (2) there are sibling species that are phenotypically alike but are considered as different species because they cannot interbreed among themselves. The fruit flies species *Drosophila pseudoobscura* and *Drosophila persimilis* form such a case.

Now consider the case in which essences are genetic. As before:

> Intra-specific genetic variation is extremely wide – meiosis, genetic recombination and random mutation together ensure an almost unlimited variety in the range of possible genotypes that the members of a sexually reproducing species can exemplify (Okasha 2002, p. 196).

Furthermore, we can have distinct species sharing a considerable array of genes. Thereby, the assumption that essences are genetic is empirically problematic because it fails to single out individual species.

The argument presented above against essentialism is strictly empirical. In face of this, one may argue on behalf of essentialism that the empirical arguments presented above are not sufficient to show that it is *impossible* to find a common intrinsic property among the members of a species. Maybe the failure in finding species essences is just an empirical limitation. Okasha's conceptual argument aims to rule out this possibility. The argument runs as follows. Suppose that every member of a species shares some intrinsic property. In Okasha's view, this shared property still does count as an essence

> For if a member of the species produced an offspring which lacked one of the characteristics, say because of mutation, it would very likely be classed as con-specific with its parents. So even if intra-specific phenotypic and genetic variation were not the norm, this would not automatically vindicate the essentialist (Okasha 2002: p. 197).

To sum up, the argument against essentialism has the following format. First, if we look at species studied in biology, essentialism has no empirical support. It is not the case that, for any species, we can find some trait – be it phenetic or genotypic – that is shared by all and only the members of the species in question. In addition, even if we find a trait shared by all and only members of a species, it does not follow that this trait is an essence because it is not necessary for the species to instantiate it. That is, we may have a circumstance in which a member of the species does not possess the trait in question.

3. Biological species and integration

As argued in the previous section, to appeal to intrinsic properties of organisms is not sufficient to demarcate biological species. Hence the individuation of species has to be based on the relational properties of its members and the environment. Contemporary biology provides an array

[1] From now on, I will use the word "essentialism" as short for "essentialism concerning species".

of species definitions in terms of relational properties. Such definitions are known as **species concepts**. In what follows, I consider two examples of species concepts: the *Biological Species Concept* and the *Ecological Species Concept*.

According to the Biological Species Concept (henceforth, BSC), species are groups of natural populations that are reproductively isolated from other such groups.[2] An important feature of BSC is its connection with population genetics, because a "reproductively isolated" population forms a gene pool in which gene frequencies vary through gene transfer within the population. Thereby, according to BSC, the stability of a species depends on **isolating barriers** "that would favor breeding with conspecific individuals and that would inhibit mating with nonconspecific individuals" (Mayr 2004, p. 178). Examples of such barriers include habitat isolation or reduced viability of hybrid zygotes.

For the Ecological Species Concept (henceforth, ESC), species are understood as a set of organisms adapted to a particular set of resources – or, a niche.[3] So, according to ESC, species are formed because of how resources are made available in face of selective pressures. The parasite-host relations illustrate this fact (Ridley 2004: 353). Suppose a parasite exploits two host species that have different characteristics such as morphology. In such a situation, the parasites will have different ecological resources and, consequently, they will tend to develop different adaptations that will in turn cause them to form different species.

ESC and BSC are related because gene flow within a reproductively isolated population may develop shared adaptations to a certain niche (Ridley 2004: 353-54). However, there are situations in which these two species concepts conflict. The North American oaks form distinct species despite gene flow among these different species (Van Valen 1976). Furthermore, there are cases of single species that do not exhibit gene flow among its members (Ehrlich & Raven 1969). Another point of conflict is that, in contrast to BSC, ESC permits species with asexual organisms as members.

The point of this section is not to solve the conflicts between BSC and ESC but rather to illustrate what makes species concepts distinct from essentialism. In particular, species concepts do not invoke any intrinsic property to define what a species is. Rather they show how a species is integrated via relational properties of its members like *interbreed with* or *occupy the same niche as*. Unlike essentialism, species concepts are both compatible with widespread variability of both phenotypic as well as genotypic characteristics of the members of a species.

4. Boyd's proposal

The goal of this section is to present Boyd's view about species known as *Homeostatic Property Cluster* theory. Based on the two previous sections, I describe two desiderata for a satisfactory account of species. After this, I describe how Boyd's view accommodates these two desiderata.

The incompatibility between essentialism and contemporary biology (Section 2) makes room for the following desideratum:

> *Desideratum I*: It is not necessary that members of a biological species share any intrinsic property.

The species concepts (Section 3) grounds in turn the additional desideratum:

> *Desideratum II*: Every biological species is somehow integrated (e.g., it forms a gene pool, it shares the same niche, etc).

In face of these desiderata, a satisfactory conception of what a species is has to ensure a big range of variability within a species (Desideratum I) despite the fact that species are interconnected via relational properties (Desideratum II). In what follows, I present Boyd's view in face of these two desiderata.

According to Boyd, biological species are natural kinds. Boyd's conception of natural kinds is called Homeostatic Property Cluster (henceforth, HPC) theory. HPC theory comprises, *inter alia*, the following two claims:

> **(C_1)** There is a family (F) of properties that are contingently clustered in nature in the sense that they co-occur in an important number of cases.
>
> **(C_2)** Their co-occurrence is, at least typically, the result of what may be metaphorically (sometimes literally) described as a sort of *homeostasis*. Either the presence of some of the properties in F tends (under appropriate conditions) to favor the presence of others, or there are underlying mechanisms or processes that tend to maintain the presence of the properties in F, or both (Boyd 1999: 143).

Different from essentialism, HPC theory does not assume that there is a property that is both necessary and sufficient for membership in a species. For HPC theory allows the existence of species with members that do not share the same single property. Because of this, HPC theory seems to permit enough variability within a species to satisfy the first desideratum.

HPC theory also ensures that species have some integration (Desideratum II) because species are coupled with some homeostasis. A species forms a unit because there is a certain set of properties that tend to co-occur among the species' members.

5. Assessing Boyd's position

As stated by (C_2), homeostasis may occur in two forms:

> *Homeostasis-I*: some properties of the cluster F induce the instantiation of other properties in F.
>
> *Homeostasis-II*: mechanisms are present that induce the instantiation of properties in F.

In the case of species individuated by BSC, we seem to have these two sorts of homeostasis. The fact that organisms within a species interbreed may cause the species to share some phenotypic trait. Thus the property of *interbreeding with conspecific organisms* would induce the instantiation of the property *to share some phenotypic trait* (Homeostasis-I). As an instance of Homeostasis-II, one

[2] See Mayr (1970).

[3] For a more detailed definition of the ecological species concept see Van Valen (1975).

could mention the many sorts of isolating barriers (Section 3) that prevent gene flow between distinct species (Wilson *et al.* forthcoming). Despite its apparent plausibility, the goal of this section is to present an objection against HPC theory.

The properties in the cluster F can be either intrinsic or relational. According to the conceptual argument against essentialism (Section 2), no intrinsic property may count as necessary for membership in a species. But the same argument is also effective against Boyd's finite disjunction of intrinsic properties.[4] For if Okasha is right, there is no boundary on the variation of individual traits among the members of a species. Thus, if the cluster F contains intrinsic properties, then given the rejection of essentialism F could not exclude any phenotypic or genotypic trait. If it did exclude such a trait, F would entail that not having a certain property is necessary for membership in a species. But this consequence cannot be right because F would then not single out individual species. To make this clear, consider an example. Suppose an unbounded disjunction of intrinsic properties P_1 v P_2 v … and objects *a* and *b* both containing P_1 and P_2. Are they both members of the same species? As the disjunction is unbounded, the disjunction by itself cannot decide this question. Therefore, if we accept the conceptual argument against essentialism, F cannot contain intrinsic properties and so *F can only contain relational properties.*

Now let us move to the second kind of homeostasis (Homeostasis-II). As mentioned earlier, isolating barriers between species seem to count as Homeostasis-II. An important feature about isolating barriers is that they are evolved characters between two species – e.g., courtship. Since isolating barriers are evolved characters, they are analogous to phenotypic/genotypic traits: they not only can vary through time, but also there is no boundary to such variation. In contrast, non-interbreeding caused by geographic separation is not considered as an isolating barrier because it is not an evolved character (Ridley 2004: 355).

But if the previous paragraph is correct, we can extend the argument used above against Homeostasis-II. A finite disjunction of isolating barriers cannot single out a species. Otherwise, we would have to accept that there is some a priori impediment to how isolating barriers between two species can evolve. Hence, *species cannot be distinguished via Homeostasis-II*.

I have drawn two conclusions so far: (i) F can only contain relational properties; and (ii) species cannot be distinguished via Homeostasis-II. Because of (i) and (ii) it follows that, if species are HPC kinds, then they are a cluster of relational properties where some of these properties induce the presence of others (thereby, intrinsic properties and Homeostasis-II are both excluded). If I am right about this, when applied to species, HPC theory collapses into a theory that is no more explanatory than the species concepts themselves. Boyd's theory can only state that some relational property such as *interbreeding with con-specifics* induces other relational property like *belonging to the same gene pool as*. So, although the notions of homeostatic mechanisms or cluster of co-occurring properties seem to carry some additional explanatory power, I tried to show above that these notions are irrelevant to comprehending what a species is.*

Literature

Boyd, R. 1991 "Realism, Anti-foundationalism, and the Enthusiasm for Natural Kinds" *Philosophical Studies* 61: 127-48.

Boyd, R. 1999 "Homeostasis, Species, and Higher Taxa", R. A. Wilson (ed.), *Species: New Interdisciplinary Essays*. Cambridge: MIT Press, 141-85.

Ehrlich, P. and P. Raven (1969) "Differentiation of Populations" *Science* 165: 1228-32.

Ereshefsky, M. 2001 *The Poverty of Linnaean Hierarchy* Cambridge: Cambridge University Press.

[4] See Ereshefsky and Matthen (2005), p. 9.

* Special thanks to Travis Dumsday for his helpful comments.

Limiting Frequencies in Scientific Reductions

Wolfgang Pietsch, Munich, Germany

1. Introduction

Limiting frequencies have been largely discredited in the philosophical discussions on the interpretation of probability. A crucial reason is that frequency interpretations are usually classified as objective accounts of probability which stands in obvious contrast to the non-empirical notion of infinite frequencies. In this essay I will argue that notwithstanding these conceptual difficulties, limiting frequencies are an indispensable tool in theory reductions.

Section two will be concerned with a classification of reduction according to its methodological role: This role can be ontological, i.e. concerned with unifying phenomena that were originally thought to be of different nature. Or reduction can have an epistemological function in making a theory simpler and thereby better applicable – e.g. when the general theory of relativity is simplified to Newton's theory of gravitation, which suffices to treat most phenomena concerned with the motion of planets.

In Section three an important case of such an epistemological reduction is examined, which will be termed *statistical reduction*.. In statistical reductions the concepts of the higher-level theory necessarily involve a large amount of entities of the lower-level theory. Therefore the probability calculus is essential for this type of reduction. Furthermore, it turns out that such reductions usually require the limit of an infinite number of lower-level entities, i.e. limiting or infinite frequencies.

Finally, Section four will discuss how probability in statistical reductions should be interpreted. The described use of limiting frequencies generally calls for further research into the conceptual difficulties of this notion: Can the limit be empirically justified by means of certain averaging procedures, e.g. time or ensemble averages? What is the relation between such different types of averages? Are there subjective and epistemic elements involved and if so, which role do they play? Due to their indispensable role in theory reduction, infinite frequencies cannot simply be dismissed, as often happens in philosophical discussions on probability.

2. Reduction as an Epistemological Enterprise

There has long been the sense that reduction comes in two different kinds, although the dividing line has been drawn in several quite different ways. Ernest Nagel (1974) distinguished *homogeneous* from *inhomogeneous* reductions: In the former, all concepts of the reduced theory are already contained in the reducing theory. In the later, the reduced theory employs additional concepts beyond those of the reducing theory, which leads to the need to establish bridge laws between these new concepts and the reducing theory. Another way, in which a distinction has been made, is between *interlevel* and *intralevel* reductions. Interlevel reductions concern theories on different levels of fundamentality – for example chemistry and physics – while intralevel reductions refer to the same level, for example the relation between Newton's theory of gravity and the general theory of relativity. When the reduced theory is not eliminated in favor of the reducing theory – which is mostly the case in interlevel reductions – one speaks of *synchronic* reductions. When the reduced theory is given up after a successful reduction, this is called a *diachronic* reduction.

In what follows we will adopt a classification that somewhat differs from these suggestions. We will distinguish two types according to the function of the reduction: (i) reduction as an ontological enterprise and (ii) reduction as an epistemological enterprise[1]. As with the other distinctions this is not meant to suggest, that one can always unambiguously classify a reduction as one of these cases. Rather, we always deal with a mixture, where each of the types is more or less clearly present.

(i) In ontological reductions, phenomena that were originally thought to be quite different in nature are traced back to the same mechanism. Examples for this type are the derivation of optics from electrodynamics or the unification of Galileo's law of falling bodies and Kepler's laws for the planetary motion within the general framework of Newton's theory of gravity. Usually, these are intralevel and diachronic reductions, where the original theories do not survive.

(ii) Reduction as an epistemological enterprise deals with those cases where considerable simplification is required in order to apply theories in a specific context. Examples for this type are the reduction of thermodynamics to many-particle mechanics or the reduction of macro- to microeconomics. These epistemological reductions need not always be fully worked out. In many cases only certain elements of the higher-level theory can actually be reduced to concepts of the lower-level theory, i.e. there are emergent aspects in the higher-level theory – at least for the time being. Most epistemological reductions are 'imperfect' in this way: the reduction of psychology to neuroscience, of biology to chemistry, or of chemistry to physics. Usually these reductions are interlevel reductions, but not always: a counterexample is the reduction of geometrical optics to wave optics. Since the different theories in an epistemological reduction each retain their significance in particular contexts, mostly both theories are kept, i.e. the reduction is synchronic.

For the rest of the paper, the focus will be on reduction as an epistemological enterprise. Whenever a theory is simplified in order to make it better applicable, this can in principle be interpreted as a reduction of a simpler framework to a more complicated and general one. Two important types can be further distinguished, that I will call (a) parametric and (b) statistical reduction. This classification is not necessarily exhaustive, one can well imagine other types of simplification.

(a) In *parametric* reduction the limit of a parameter of the reducing theory is taken in order to make the equations mathematically better manageable. This type is often said to be typical for reduction in physics (Nickles 1973, Batterman 2002) and indeed many physical reductions seem

[1] The distinction between the ontological and epistemological role of reduction is certainly not novel (compare e.g. Hoyningen-Huene 2007, 181-183). However, there are subtle differences between our treatment and that in the literature.

to work this way. Classic examples are the reduction of Newtonian mechanics to the special theory of relativity in the case of small velocities (velocity of light c → ∞) and the reduction of quantum mechanics to classical mechanics in the case of large quantum numbers (Planck's quantum of action ħ → 0).

(b) We speak of *statistical* reductions whenever the higher-level theory deals with a large amount of entities of the lower-level theory, which are treated by means of statistical methods. The reducing theory is normally concerned only with the behavior of a single one or at most a small amount of these entities while the reduced theory deals with the collective behavior of a huge amount of them. In much of the literature these types of reduction (a) and (b) are often lumped together – statistical reduction is considered a parametric reduction in terms of the number of particles *N*. But for several reasons that is itself not a good reduction of type (b) to type (a).

First, statistical reduction is so abundant, that it merits to be considered separately. It can be found across all boundaries in the sciences, whenever large numbers of similar entities are involved: humans, neurons, atoms, goods etc. Second, there are important conceptual aspects in which statistical reduction differs from parametric reduction. The calculus of probability plays an essential role. Also, the reduced theory deals with a peculiar kind of entities and laws: They are statistical in nature as we will see in the next section. For the rest of this essay we will concentrate on statistical reductions.

3. Statistical Reduction

Not very surprisingly the decisive property, that distinguishes statistical reductions from other kinds, is that on the level of the reduced theory we deal in all aspects with statistical phenomena: statistical entities, statistical properties of the entities and statistical laws connecting these properties. We will illustrate this in the following by means of the reduction of thermodynamics to statistical mechanics. Only at the end of this section other examples will be quickly addressed.

Thermodynamics deals with statistical entities, e.g. gases, fluids, or solids. These macroscopic entities are composed of a large amount of the fundamental entities that mechanics is concerned with – namely a large amount of point-like masses. The properties of the thermodynamic entities are also statistical in nature: Quantities like volume V, pressure p, temperature T, or entropy S require a large amount of mechanical point masses in order to be adequately defined. Finally, statistical laws relate these properties with each other – examples are the ideal gas law $pV \sim T$ or the second law of thermodynamics describing the increase of entropy.

Starting from the lower-level theory, which is mechanics in the considered example, it turns out that most of the elements of the higher-level theory can only be defined in terms of continuous and differentiable probability distributions of the lower-level entities. On the basis of a frequency view of probability, these continuous distributions must necessarily refer to an infinite number of lower-level entities. A finite number of instances, e.g. a finite number of atoms, can only yield a discrete distribution function corresponding to a weighted sum of *δ*-functions. Thus, in order to establish the concepts of the higher-level theory the transition to the infinite limit is essential.

Microscopically, matter is not continuously distributed in space – a gas actually consists of many point-like masses and a lot of empty space. To define simple properties like volume or pressure on the basis of the actual discrete distribution function for the atoms turns out quite difficult: How for example should the boundaries of the volume along the edges of the gas be determined? There is just no unambiguous way to decide which parts of the empty space belong to the gas and which not. The question is answered on the basis of pragmatic and in particular symmetry considerations. The actual probability distribution is smoothed out everywhere and is chosen in such a way that the edges are as geometrically simple as possible. In this manner, we are led from the actual discrete distribution of the atoms to a continuous distribution which determines the macroscopic concept of volume.

Similarly, the concept of thermodynamic pressure involves an infinite limit in order to arrive from the discrete concept of microscopic collisions to the continuous macroscopic concept of force per area. Again, symmetry considerations lead the way from the actual microscopic events to a continuous probability distribution which presupposes an infinite number of those microscopic events.

The laws that connect macroscopic quantities – as the ideal gas law connects pressure, volume, and temperature – presuppose the infinite limit as well, simply because properties like pressure and volume are otherwise ill-defined. Sometimes, the macroscopic laws crucially depend on the way the limit is taken. A good example is the second law of thermodynamics which describes the increase in entropy during the approach to equilibrium. Depending on the way the infinite limit is taken, one either derives from statistical mechanics a deterministic second law which allows no entropy decrease at all – as in Ludwig Boltzmann's *H*-theorem (Boltzmann 1872). In other cases one may obtain entropy fluctuations. These results are not contradictory – they just correspond to different extents of coarse-graining.

When other examples of statistical reductions are examined one encounters the same need for continuous distributions and the infinite limit. Statistical reductions can be found across all boundaries in natural as well as social sciences, whenever one deals with a great amount of similar entities: with neurons in neuroscience, goods in economy, human beings in population science, or errors in error theory. Considering the abundance of error theoretic methods in all kinds of scientific enterprises, the last example once more underlines the ubiquity of statistical reductions. Whenever Gaussian distributions for measurement values are assumed, one has implicitly taken the limit of an infinite number of equally distributed, miniature errors.

4. Conclusion: The Need for Infinite Frequencies

The indispensable role of limiting frequencies for the macroscopic concept formation in statistical reductions provides an interesting test case for the different interpretations of probability. Lately, limiting frequencies have not enjoyed a good reputation both among scientists and philosophers of science, e.g. illustrated in the following quote by Alan Hájek (2007): “To be sure, science has much interest in finite frequencies, and indeed working with them is much of the business of statistics. Whether it has any interest in highly idealized, hypothetical extensions of actual sequences, and relative frequencies therein, is another matter." H. Cramér takes a similar stance: "[Limiting frequencies] involve a mixture of empirical and theoretical elements, which is usually avoided in modern axiomatic theories." (quoted in Gillies 2000, 100)

As the quote by Cramér suggests, the main problem of limiting frequencies seems the inextricable mixture of ontic and epistemic elements. This squares badly with the usual distinction between epistemic and ontic accounts of probability[2]. Subjective interpretations of probability – in terms of degrees of belief – can readily account for the smoothing out of the discrete probability distributions. However, it seems odd to interpret probabilities as degrees of belief within objective sciences like physics. But thus far, nobody has yet managed to justify the infinite limit on the basis of a purely objective account. Notwithstanding these difficulties, the scientific practice in statistical reductions forces us to make sense of the notion of limiting frequencies. Maybe the lesson is to abandon the hard distinction between ontic and epistemic accounts altogether.

Literature

Batterman, Robert 2002 *The Devil in the Details: Asymptotic Reasoning in Explanation, Reduction, and Emergence,* New York: Oxford University Press.

Boltzmann, Ludwig 1872 "Weitere Studien über das Wärmegleichgewicht", Wiener Berichte 66, 275-370.

Galavotti, Maria C. 2005 Philosophical Introduction to Probability, Stanford: CSLI.

Gillies, Donald 2000 Philosophical Theories of Probability, London: Routledge.

Hajek, Alan 2007 "Interpretations of Probability", in: Edward N. Zalta (ed.) Stanford Encyclopedia of Philosophy, http://plato.stanford.edu.

Hoyningen-Huene, Paul 2007 "Reduktion und Emergenz", in: Andreas Bartels and Manfred Stöckler (eds.) Wissenschaftstheorie. Ein Studienbuch, Paderborn: mentis, 177-197.

Nagel, Ernest 1974 "Issues in the Logic of Reductive Explanations", reprinted in: Martin Curd and J.A. Cover (eds.), Philosophy of Science. The Central Issues, London: Norton, 905-921.

Nickles, Thomas 1973 "Two concepts of intertheoretic reduction", *The Journal of Philosophy*, 70/7: 181–201.

[2] Excellent overviews of the different accounts of probability are Galavotti 2005 and Gillies 2000.

The Key Problems of KC

Matteo Plebani, Venice, Italy

1. The key problem of KC

According to Floyd and Putnam, we can extrapolate from Wittgenstein's 'notorius' remarks on Gödel's theorem some philosophically insightful remarks. Let "P" be the Godelian sentence for the logical system of *Principia Mathematica* (PM). Wittgenstein's key claim (KC) runs as follows:

> **KC**: If one assumes that ¬P is provable in PM, then one should give up the "translation" of P by the English sentence "P is not provable"

The key problem of KC is the following simple remark (call it KP): KC is compatible with realism. In this context, "realism" is the claim that mathematics is the study of a well–defined domain of abstract objects that exist independently of our thought, language or experience; it implies the view that arithmetic is the study of the standard model N of natural numbers.

However, this problem is considered in strict relation to the goal of providing arguments to clarify and support Wittgenstein's stance concerning Gödel's theorem. Otherwise KC would be an interesting remark. It is true that it provides genuine insight into the philosophical meaning of Gödel's theorem, but it certainly throws little light on Wittgenstein's thought. In order to understand why this is the case, it is important to follow Timothy Bays's (Bays 2006, p.6) recollection of the three uncontroversial mathematical results upon which KC is based:

> 1. If PM ⊢ ¬P, then PM is ω-inconsistent.
>
> 2. If PM is consistent but ω-inconsistent, then all of the models of PM contain non-standard natural numbers—i.e., elements which the model treats as natural numbers but which do not correspond to any of the ordinary natural numbers.
>
> 3. The translation of P as "P is not provable" depends on interpreting P at the "natural numbers alone." If we interpret P at a non-standard model—i.e., at one of the models described in 2—then there is no reason to think that this will lead to a translation of P as "P is not provable."

Bays goes on to criticize the passage between 1-3 and KC. He maintains that we shouldn't give up our translation of P as "P is not provable", in the case PM turns out to be ω-inconsistent, because "there's no reason to constrain our translation of P to the class of models which happen to satisfy PM" (Bays 2006, p.6). I agree with him on this latter point, but still think that the merit of KC is to underline that the equivalence between P and "P is not provable" holds only in the standard model. And we can think of cases in which this does matter[1]. But the issue is that the existence of non-standard model could perhaps pose a problem on choosing of how to translate P, but it is perfectly compatible with the fact that P is true in N iff P is not provable in PM and that if PM is consistent, than P is true in N (and not provable in PM).

In other words, not one of these results provides any ground for scepticism concerning the existence of the standard model N, because all the results are obtained using model-theoretic machinery, and, as has been argued by many, model theory is the realist framework *par exellence.*

There is only a way in which Floyd and Putnam's suggestion might be saved: it is possible (although questionable, see Rodych 2003) that Wittgenstein claimed something like KC; if so, he certainly had a great insight into Gödel's results. But he might have made such remarks only in order to highlight an important fact that could be acknowledged also from a realistic viewpoint. This is tantamount to claiming that KC does not help us understand what Wittgenstein's stance about Gödel's theorem was.

2. The myth of prose

Why do Foyd and Putnam think that KC is a philosophical claim of "great interest" (p. 624)? Because they believe it helps to avoid a misinterpretation of Gödel's result:

> That the Gödel theorem shows that (1) there is a well defined notion of "mathematical truth" applicable to every formula of PM; and (2) that if PM is consistent, then some "mathematical truths" in that sense are undecidable in PM, is not a mathematical result but a metaphysical claim. But that if P is provable in PM then PM is inconsistent and if ¬P is provable in PM then PM is ω-inconsistent is precisely the mathematical claim that Gödel proved. What Wittgenstein is criticizing is the philosophical naiveté involved in confusing the two, or thinking that the former follows from the latter. But not because Wittgenstein want to simply deny the metaphysical claim; rather he wants us to see how little sense we have succeeded in giving it.

That's an application to the case of Gödel's theorem of a general way of reading Wittgenstein's remarks on the Foundations of Mathematics: I will call it "the myth of prose". According to the myth of prose, the task of philosophical investigation of Mathematics is to distinguish between the real mathematical content of a theorem and some philosophical thesis often associated to it from mathematicians when they expose it informally. This apparently sensible approach leads to an implausible result. As has been argued by many, there is a perfectly legitimate mathematical sense in stating that Gödel's theorem shows that if PM is consistent, than there are sentence that are both true and undecidable in PM. Certainly, in Gödel's original paper (Gödel 1931) the theorem is formulated in syntactical terms, using the notions of consistency

[1] We won't discuss this point here, but we can give a sketch of our argument. Following an hint from Martino 2006, we think this could play an important role in the formulation of Godel's second incompleteness theorem. The problem, roughly stated, is this: an idealised mathematician without any spatial or temporal limitation, could acknowledge the consistency of a system s as logical consequence of its axioms (that means: in every model – if there are – of the axioms of S, it is true that S is consistent). But not in every model for S the arithmetic sentence that should express the consistency of S (call it Con) is true. This explains why the mathematician doesn't draw it as a conclusion from the axioms of S. This could give an idea of a contest in which the translation of the predicate "Proof", involved in the construction of P and Con, as "provable" should be given up.

and ω – consistency, but in currently available semantic proofs of the theorem the notion of truth is explicitly used, thus providing a more simple and clearer demonstration than the original one (e.g. Smullyan 1992). Gödel himself proposed this concept in the introduction to his 1931 paper and, after Tarski, a precise mathematical sense can be ascribed to the notion of truth, thus leading to the central point: there is no *mathematical* reason to prefer the syntactic formulation of the Gödel's theorem to the semantic one. This does not mean that there are no reasons whatsoever: there are *philosophical* reasons, the main one being that using the notion of truth may suggest a Platonist reading of the theorem and, of course, Wittgenstein, among others, would not allow such a reading. But the problem now is to account philosophically for this rejection, and to do so, a philosophy of mathematics alternative that of Platonism would be called for. In this enterprise it is not helpful to assert that the Platonist's favourite way of stating the theorem is misleading: it is misleading only from an anti-realist view-point; this move thus merely begs the question.

I want to make a simple point: maybe Wittgenstein is really a quasi revisionist in Frascolla's sense (see Frascolla 1994), that means that he may only want to show that, without the metaphysical interpretation which it is usually accompanied by, the notion of a true but not provable proposition loses all its charm. But this is not the same as claiming that the notion of true but not provable sentence is a metaphysical one and, this is the central point: Wittgenstein must justify his position by giving philosophical reasons for it. Wittgenstein and his friend had to face the burden of the proof: the myth of prose could not help them.

The issue also becomes problematic if we contemplate that which Floyd and Putnam consider the mathematical theorem proved by Gödel:

> that if P is provable in PM then PM is inconsistent and if ¬P is provable in PM then PM is ω-inconsistent is precisely the mathematical claim that Gödel proved

Is the above, an apparently an uncontroversial mathematical result, really metaphysically neutral? I argue that it is no more neutral than the supposed "metaphysical thesis" (see Martino 2006).

What does it mean to say that a formal system is inconsistent? In textbooks on logics the usual explanation runs along the following lines:

> A System S is called inconsistent iff for some well formed formula of the language L of the system α, both α and its negation $\neg\alpha$ are theorems of the system S.

On making such a claim, we are considering the well formed formulas as a whole; we are considering all of them, and the same holds for the theorems of the system. This is tantamount to considering the well-formed formulas as a recursively enumerable set, a set isomorphic to the standard model N of the natural numbers. If there is a well defined notion of well-formed formula, as much as of theorem of a formal system or of numeral, there is a well defined notion of a structure that has the same structure as the standard model, N. Hilbert's notion of a formula as a *finite* sequences of signs is unintelligible if we do not grasp the notion of *finite*. But grasping this notion amounts to grasping the notion of natural number.

In short: if there is a well-defined notion of consistency for a formal system, there is a well-defined notion of a numeral, well-formed formula, theorem, and so forth, and there is a well-defined notion of a structure isomorphic to N. If this holds, there is a well-defined notion of mathematical truth applicable to every formula of PM, which is what we obtain when we interpret our formal language using this structure. So the supposed mathematical theorem collapses into the metaphysical thesis. The conclusion is that either the two formulations of Gödel's theorem are both metaphysical theses or they are both mathematical results: there is no room for the prose *versus* proof distinction.

Other factors make it extremely difficult to give an account of Gödel's first theorem, which avoids make reference to the model N: for example, natural numbers are used in Gödel numbering. Of course, even if we accept the semantic version of Gödel's theorem, many philosophical options alternative to Platonism are left open: we could be fictionalists, or nominalists, or intuitionists, although we could hardly be strict finitists. We might wonder whether we might be Wittgensteinians, and this issue is dealt with in the next paragraph.

3. Wittgenstein and revisionism

An important feature of Wittgenstein's philosophical reflection is his constant claim that it should not interfere with the work of mathematicians: he maintained that the clarification of the content of a mathematical theorem would never amount to giving up this very theorem. No mathematical acquisition should come under attack from philosophical analysis (the polemical target is the attempt made by intuitionists to reform classical mathematics by ruling out all non–constructive proof). This is another aspect of what I previously referred to as the myth of prose. It is acknowledged that Wittgenstein hated Set Theory and made serious efforts to contrast it, as he also did on referring to "curse of the invasion of mathematics by mathematical logic" (Wittgenstein 1956, p.19). This stance appears to contradict Wittgenstein's claim to non–revisionism. The usual reply to this objection is to state that, in discussing set theoretical topics (e.g. Cantor's diagonal proof), Wittgenstein's concern was only to make us look at them in the right way: he believed that, without all the metaphysical smoke that they are usually surrounded with, they would lose all their charm; however, this would not mean abandoning set theory as a calculus, as piece of mathematics. Herein lies the sense of Wittgenstein's claim that he didn't want not drive us out of Cantor's Paradise; he just wanted to make us realise that it is not a paradise.

It is beyond the scope of the present study to discuss whether this interpretation works for Wittgenstein's view of set theory; however, I do not believe that it works for the remarks made by Wittgenstein concerning Gödel's theorem. Although it is a controversial issue among Wittgenstein's scholars, many authoritative commentators (e.g. Rodych 2003 or Shanker 1988b) have pointed out that, in discussing Gödel's result, Wittgenstein's main concern was to show that in Mathematics the notion of truth must be identified with that of provability. This was in order to avoid a referential picture of mathematics: Wittgenstein rejected the idea that mathematics is *about* something (whether it consisted of mental, non- mental or even concrete sequences of signs is immaterial). It is not easy to see how this concept, if taken seriously, could fail to affect mathematical practice. For example, what sense could we give to a subject like model theory if we adopted Wittgenstein's picture?

Any attempt to defend Wittgenstein's claims is thus a hard job. This probably explains why so many authors have embraced the myth of prose: it saves us the trouble of doing such a job. The same advantage, as Russell said in another context, “of theft against honest toil” (Russell 1919, p. 71).

Literature

Bays, Timothy 2006 Floyd*, Putnam, Bays, Steiner, Wittgenstein, Gödel, Etc*, unpublished.

Bays, Timothy 2004 *On Floyd and Putnam on Wittgenstein on Gödel*, The Journal of Philosophy CI.4, 197 – 210.

Floyd, Juliet, Putnam, Hilary 2001: *A note on Wittgesntein “Notorius Paragraph”* about the Gödel Theorem, The Journal of Philosophy, XCVII, 11, 624-632.

Frascolla, Pasquale 1994 *Wittgenstein’ s Philosophy of Mathematics*, New York: Routledge.

Gödel, Kurt 1986, *On formally undecidable proposition of Principia Mathematica and related systems, in Collected Works*, New York: Oxford University Press.

Martino, Enrico 2006, *Ragionamento astratto e teoremi di incompletezza di Gödel*, unpublished.

Rodych, Victor 1999, *Wittgenstein’s inversion of Gödel’s Theorem*, Erkentniss 51 (2/3), 173-206.

Rodych, Victor 2003 *Misunderstanding Gödel: New Arguments about Wittgenstein and New Remarks by Wittgenstein*, Dialectica 57, 3, pp. 279-313.

Russell, Bertrand 1919 *Introduction to Mathematical Philosophy*, London: Allen & Unwin.

Shanker, Stuart G. (ed.) 1988 a *Gödel’s Theorem in focus*, London: Croom Helm.

Shanker, Stuart G. 1988b *Wittgenstein’s Remarks on the significance of Gödel’s Theorem*, in Shanker 1988a, pp. 155-256.

Smullyan, Raymond 1992 *Gödel incompleteness Theorems*, New York: Oxford University Press.

Steiner, Mark 2001 *Wittgenstein as His Own Worst Enemy: The case of Gödel’s Theorem*, Philosophia Mathematica, IX (2001), 257-79.

Wittgenstein, Ludwig 1956 *Bemerkungen über die Grundlagen der Mathematik*, Basil Blackwell: Oxford.

The Metaphysical Relevance of Metric and Hybrid Logic

Martin Pleitz, Münster, Germany

1. Temporal Reasoning

To most of our temporal statements, a quantitative element is essential. The statement that there will be rain says more than that, in the infinitely long stretch of future time, there exists a rainy moment. Rather, it conveys that rain will come *a reasonable temporal interval hence*. And in reasoning, our temporal reference often is much more precise. When recording events, we often reason in the following way:

P1: Presently, it is windy.

P2: Presently, it is 4 p.m.

C1: It is* windy at 4 p.m.

In planning for the future, we often reason like this:

P3: The talk starts* at 5 p.m.

P4: Presently, it is 4 p.m.

C2: The talk will start in one hour.

(An asterisk (*) indicates that the verb is used tenselessly.) There are four kinds of temporal statements involved here. *Present statements* are made by temporally indefinite sentences in the present tense which do not refer to a date (P1). *Clock statements* are made by tenseless sentences of the form “Presently, it is t” (P2, P4). *Diary statements* are made by tensed sentences that give a date (C1, P3). *Metric tense statements* are made by temporally indefinite sentences which give the duration from the present moment to a past or future event (C2).

A temporal logic that captures our temporal reasoning must have the resources to express all four kinds of temporal statements. *Standard logic* (propositional and predicate logic) cannot formalize any of these statements in a way which preserves their temporal characteristics. *Date logic* states relations between events and dates and nothing else. Thus it can express diary statements (“$@_t p$” for “At t, it is* the case that p”), but neither present, clock nor metric tense statements. *Standard tense logic* allows temporally indefinite statements and therefore can formalize present statements, but none of the others, because its operators F and P cannot cope with quantities.

2. Metric Tense Logic and the Reduction of Dates

Metric tense logic with its operators F(t) for “It will be the case t time-units hence that” and P(t) for “It was the case t time-units ago that” can express metric tense statements and present statements. But it allows neither clock nor diary statements because it cannot deal with dates. Dates can only be dealt with by a metric tense logic that is *grounded*, i.e. that is supplemented by a *unique* proposition, i.e. a proposition c guaranteed to be true at exactly one moment. The clock statement “Presently, it is t” then can be translated as “P(t)c” and the diary statement “At t, it is* the case that p” as “Sometimes: $p \wedge P(t)c$”.

For a realistic example of a grounding unique proposition c, we only need “some standard event which is presumed to be unique” (Prior 1957, 19). For the current system of time-keeping, c is the proposition that presently, Christ is born. “Presently, it is 2008” can be translated as “It was the case 2008 years ago that, presently, Christ is born.” Note that in metric tense logic, *one* unique proposition helps to reduce the whole system of dates (cf. section 4).

If temporal instants are nothing more than dates, then grounded metric tense logic and Ockham’s Razor lead naturally to a tensed metaphysics of time, albeit one where past and future are graded.

3. Hybrid Tense Logic and the Reduction of Instants

Although Arthur Prior described metric tense logic in detail (Prior 1957, 18-28; 1967, 95-112; 2003, 159-171) and used it to translate date statements (Prior 1957, 19; 1967, 103ff.), he took another way to reach the metaphysics free of temporal instants that his tense-theoretical intuitions made him look for: He invented hybrid modal logic, that provides, for each point of the frame, a unique proposition, which Prior called “world-proposition” (Prior 1967, 89) and nowadays is known as a “nominal”, as in a sense it *names* an instant (Blackburn 2006, 343ff.). Nominals (i, j …) allow the translation of date logic (Prior’s “logic of earlier and later”) into hybrid tense logic. “$@_i p$” becomes “Sometimes: $i \wedge p$”, “i is earlier than j” becomes “Always: $i \rightarrow Fj$”, etc. (Prior 1967, 88ff. and 187ff.; 2003, 124ff.).

Prior took this logical result to be of metaphysical importance: “A world-state proposition in the tense-logical sense is simply an *index of an instant*; indeed, I would like to say that it *is* an instant, in the only sense in which ‘instants’ are not highly fictitious entities.” (Prior 1967, 188f.)

4. Metric or Hybrid Tense Logic?

For those sharing Prior’s metaphysical tense-theoretical convictions, there are reasons to prefer metric tense logic to its hybrid alternative. (I) Hybrid tense logic does not capture the quantitative side of many temporal statements (section 1). (II) Nominals, though formally innocent (e.g. Blackburn 2006, 343ff.; Øhrstrøm et al. 1995, 221ff.), are suspect from a natural language stance. As only some times can be characterized by unique events that are publicly known, the only natural candidates we have for nominals are (a) complete descriptions of instants and (b) clock statements. But (a) leads to modal problems, because it makes impossible that something else could have happened at a certain time than what actually did. And (b) translates hybrid tense logic into metric tense logic. (III) The reduction of dates to grounded metric tense logic is better suited than hybrid tense logic to capture the epistemic side of time-keeping. Not only can a person forget the date (cf. Müller 2002, 193ff.), but we can also imagine a whole time-keeping community losing track of their grounding event, but still knowing the truth of many metric tense statements.

A fourth reason to prefer metric to hybrid logic concerns its generalizability from time to other dimensions of logical space (sections 5-8).

5. Prior's Problem

As Patrick Blackburn has pointed out, Prior ran into deep problems because his tool for the reduction of temporal instants, hybrid logic, worked rather too well and allowed to reduce objects of other categories, as well – especially troublesome in the case of persons (Prior 2003, 213ff.; cf. Blackburn 2006, 362ff.). If we interpret the points of the frame of our modal logic as persons and personally indefinite propositions to express predicates, we can translate statements of standard predicate logic to a "personal" analogue of date logic: "Socrates is wise" becomes "$@_{\text{Socrates}}$ be-wise". Nominals used to name persons then dissolve the points of personal space; we go over to "Some-personally: be-Socrates $\wedge$ be-wise" and have lost Socrates (cf. Prior 2003, 215-219).

Prior was aware of this problem. Yet he found no solution but to simply proclaim that "persons just *are* genuine individuals […] whereas […] instants are *not* genuine individuals" (Prior 2003, 219f.; cf. Blackburn 2006, 364). Does metric tense logic, as a tool of metaphysical reduction, run into similar difficulties? I will argue in sections 6-8 that it does not.

6. Putting Time and Space into Perspective: Standpoint Logic

Thomas Müller has shown how to generalize metric tense logic to cover physical space and relativistic frames of reference (Müller 2002, 268-279): Atomic propositions are indefinite concerning time, space and frame, and there are operators changing the temporal or spatial perspective or which lead from one frame to another. All operators are metrical, corresponding to temporal intervals, spatial vectors and Lorentz transformations. Müller points out that each set of operators constitutes a group in the algebraic sense.

Now, the reduction of dates (section 3) generalizes to the reduction of all context-free coordinates (Müller 2002, 276ff.). We only need to add to Müller's standpoint logic one proposition uniquely characterizing a place ("Here is Greenwich") and another proposition uniquely characterizing a frame ("At-this-velocity is Earth"). The metaphysical upshot is that places and frames are no more genuine objects than instants; our spatiotemporal world ultimately is perspectival.

7. General Perspectival Identification

But why is there no metric logic for other dimensions of logical space, like persons and possible worlds? To answer this question, we have to isolate those features of metric logic and standpoint logic that allow the reduction of instants, places and frames. When is a modal logic suited for a similar reduction of the points of its frame?

It is not the numbers: The metrical operators of a distance logic do not allow the reduction of spatial coordinates. E.g., "Two meters from here it is the case that" takes us to a sphere of two meters radius and thus is hopelessly ambiguous. We need each operator to transfer us to exactly one point. This *condition of identification* must be *general* in the sense that it is satisfied at every point of the model. But general identification is only a necessary condition, because it is met by the satisfaction operator $@_t$ of date logic, which of course does not allow a reduction of dates. For genuine reduction, we need our operators to be *perspectival*, i.e. the point the operator takes us to must depend on the point where it is employed. I argue that every general perspectival identificatory modal logic can reduce the points of its frame.

To be honest to the metaphysical project of reducing the objects these points purport to be, we have to move from a model-theoretical to a syntactical (i.e. proof-theoretical) characterization of general perspectival identification. Otherwise, the objects we want to get rid of will recur on the meta-level. I will discuss the following meta-theorems: (M1) A modal logic is *generally identificatory* just in case for all its operators V(n), $V(n)\neg p \leftrightarrow \neg V(n)p$ is a theorem, i.e. iff all operators are self-dual. (M2) A modal logic is *perspectival* just in case its operators form a group. The group axioms rule out satisfaction operators, because none of them has an inverse element. In sum, I suggest that a modal logic can reduce the points of its frame iff its operators are self-dual and form a group.

8. Solving Prior's Problem

To solve Prior's problem, we thus have to find reasons why, for persons and possible worlds, we cannot construct modal logics that allow general perspectival identification, i.e. whose operators are self-dual and form a group. For possible worlds, self-duality clearly fails, because it is essential to alethic modality that the necessary and the possible do not fall together. For persons, the obvious candidate for a relation of accessibility allowing perspectival identification is the system of family ties (everyone can perspectivally identify her mother). Here generality fails: To construct a modal logic that would reduce persons, we would need a relation that allows *everyone* to perspectivally identify *everyone*. But there is no relation that holds among *all* persons that is systematic and static enough for this purpose.

So, where hybrid logic is too powerful a tool of metaphysical reduction, generalized metric logic is just right. It can translate context-free temporal and spatial statements to perspectival statements, but not statements about persons and possible worlds. This fits Prior's metaphysical intuition that while times and places are logical fictions, persons are genuine objects.

9. Worlds and Selves: What is the Role for Hybrid Logic?

But what about possible worlds? If we replace hybrid logic by metric logic *tout court*, we are left with possible worlds seeming to be genuine objects just as much as persons. But while the reasons counting against hybrid tense logic (section 4) generalize easily to space and frame, this is not so for possible worlds. There is no quantitative element to modality (reason I), and (as shown in section 8) we cannot identify possible worlds by their relation to a unique world (reason III). What we actually do to identify a possible world is to *describe* it. Therefore, in the case of alethic modality, we *do* a have a natural interpretation for nominals, namely for each world, the maximally compatible proposition that uniquely describes it (reason II). In sum, what counts against the employment of hybrid logic concerning time and space, does not count against Prior's project of reducing possible worlds to world-propositions.

So, from a natural language stance, hybrid logic is well-suited to deal with possible worlds. But not with persons (or other enduring things): Here, the only natural candidates for nominals contain proper names. "I am

Martin Pleitz”, e.g., is true at the unique point of personal space constituted by me. But proper names already presuppose the objects they name. So, a hybrid logic of persons, when founded on natural language, cannot reduce persons.

10. Logic and Metaphysical Commitment

Today, modal logics generally are seen as tools not for the decision of metaphysical questions, but to give an internal characterization of given structures. In this “Amsterdam perspective” (Blackburn 2006, 330ff.), it is left entirely unspecified what these structures are. It may therefore seem strange that the preceding arguments about metric and hybrid logic have led to the following substantial metaphysical theses:

> Instants, places and frames can be reduced to a perspectival metaphysics of space-time (by metric logic).
>
> Possible worlds can be reduced to maximally compatible propositions (by hybrid logic).
>
> Persons (and other enduring things) are the only genuine objects we are left with.

But these metaphysical commitments do not rest on formal considerations alone, but on reasons concerning the applicability of logical systems to our ordinary use of natural language. The reasons for perspectivist spatiotemporal metaphysics lie in our practices of time-keeping and referring to places. The reason for linguistic ersatzism about possible worlds is given by our practices of describing possible situations. And the existence of persons and things is connected to the practice of naming. Only in cooperation with the philosophy of language can logic answer the question of what there is.

Literature

Blackburn, Patrick 2006 “Arthur Prior and Hybrid Logic”, Synthese 150, 329-372.

Müller, Thomas 2002 Arthur Priors Zeitlogik, Paderborn: Mentis.

Prior, Arthur N. 1957 Time and Modality, Oxford: Clarendon.

Prior, Arthur N. 1967 Past, Present and Future, Oxford: Clarendon.

Prior, Arthur N. 2003 Papers on Time and Tense. New Edition [eds. Per Hasle, Peter Øhrstrøm, Torben Braüner and Jack Copeland], Oxford: Oxford University Press.

Øhrstrøm, Peter, and Hasle, Per F.V. 1995 Temporal Logic, Dordrecht: Kluwer.

Reductionism in Axiology: the Case of Utilitarianism

Dorota Probucka, Cracow, Poland

Utilitarianism claims that any axiology and psychology of valuation is grounded in individual experience. Values arise in primordial cognitive acts which form any axiology's elementary *datum*. Hence, the range of values peculiar to a given axiology depends on the kind and content of individual experience. The indissoluble connection between axiology and epistemology prompts the impossibility to build an independent theory of value. Thus, epistemological (as well as ontological) assumptions must become integral part of any general axiological theory.

Under the utilitarian conception of axiology as a theory based on individual experience, axiology, epistemology and psychology should be integrated into one theoretical framework. The aim of this paper is to analyze this framework, unveil its internal structure and show that reductionism, which pervades contemporary philosophy and axiology, is an essential feature of the utilitarian theory of value. Subjectivism and relativism are main components of reductionist way of thinking.

The doctrine of utilitarianism was created as a secular ethics which, in accordance with the assumptions of British empiricism, refused to acknowledge the existence of empirically unrecognizable beings. Empirically recognizable world is the only one on which affirmative judgment can be made. Therefore, any attempts to go beyond human experiences are unwarranted as being cognitively meaningless. What remains to be done in these circumstances - the utilitarians claim - is to search for values as somehow existing in the world accessible to our senses.

Once axiology is to be grounded in individual human experience, there is no other choice for a utilitarian but to reduce values to some facts conceived of as empirical, psychosomatic data. In other words, the recognition of values as values amounts to the recognition of facts. Does the reducing of values to facts mean that every fact is a value? Certainly not, for not all facts have anything in common with values. Nevertheless, those facts which are recognized as values still remain no more than facts. That's how the idea to derive axiology from individual experience leads to reductionism (typical of any naturalistic stance in philosophy), reduction being understood as a special relation between values and facts.

What are facts? On The whole, they are states of affairs given to us in external and internal experience. Which states of affairs can be referred to as values? According to the utilitarians, those which are preferred, valuated, desired by subjects who experience these states. One must not attach values with states of affairs without considering human valuating, preferring, desiring. Values do not exist independently of human beings, they can't be traced outside their external or internal experience. According to H. Sidgwick, a leading utilitarian philosopher, material things have no intrinsic value. We should estimate their worth and consider the importance of their existence solely with reference to human beings. Values are related solely to human beings. Such is the consequence of reductionist thinking in axiology and ethics.

In utilitarian view, factual values do not exist outside sensual experience. They are important component of this kind of experience, but still belong to its empirical dimension as a part of natural order. Let us repeat that under this kind of reductionism facts form an ultimate reality which is the only reality. According to R. Brandt's views expressed in his book *Ethical Theory,* utilitarianism is far from the view that values exist independently of the world and unlike variable things are changeless.

Moreover, if values are entirely connected with individual experience, subjectivism is the logical consequence of that thesis. Indeed, individual and empirical knowledge is conceived as the sum of the subject's psychosomatic experiences. The factual values are one of possible contents of these experiences.

Utilitarianism is a philosophical trend which has evolved for 200 years and it has many variants. I take into account mainly its traditional interpretation represented by J. Bentham and J. S. Mill (earlier by C.A. Helvetius and D. Hume). J. Bentham was first to clearly define its principles later developed by J. S. Mill, J. Austin, H. Spencer and H. Sidgwick. Contemporary utilitarianism is represented by G. E. Moore, C.D. Broad, S. Toulmin, H. Rasdall, J. Narveson, J.J.C. Smart, R. B. Brandt, W.T. Stace, D. Parfit, D. Lyons, J.D. Harsanyi, D. Brink, R.E. Bayles. In my paper I omit intuitivism (so called ideal utilitarianism) initiated at the second half of XIX c. by H. Sidgwick and developed at the first half of XX c. by G.E. Moore. I analyze mainly hedonistic branch of utilitarianism in its standard version. The adherents of ideal utilitarianism assume the existence of the moral sense which provides the abilities for the recognition of values existing beyond the subject.

Coming back to the subject matter I would like to notice that the arising of utilitarian thinking was connected with the evolution of empirical psychology (turn of the XIX and XX c.) and its primordial current – psychological hedonism. We should underline that the classical utilitarianism was founded on the psychological hedonism. That theory, having a long tradition and many adherents (d'Alembert, Helvetius, Hartley, Priestley, Tucker, Locke), assumed that the escape from pain and desire for pleasure were the main motives of human action. Positive values have been reduced to pleasurable states of subject. Utilitarian axiology underlines this conclusion and treats it as a result of inquiries made by empirical psychology. For these psychological hedonists "pleasure" is the primary, undefined notion. Thus, experience is the only way of describing it. According to C.D. Broad, there is possible only ostensive definition of that term, made on the basis of introspection. All of us know what pleasure is, one way or another, we pursue it and avoid pain and frustration caused by the lack of satisfaction. The experience of pain and pleasure must be reduced to subjective feelings. But, the terms, corresponding to them, will be connected with the individual expression of positive or negative sentiments evoked by their empirical referents.

There are many variants of psychological hedonism. L. T. Troland in his book entitled "*Fundamentals of Human Motivation*" describes its three versions: - psychological hedonism oriented at the future (the thought about the future pleasure as a consequence of some course of action is a necessary condition of that action and the main reason of it), psychological hedonism oriented at the present (the pleasure or annoyance experienced at the present determine our course of action),

psychological hedonism oriented at the past (our current decision are determined by the pleasures and annoyances experienced in the past).

W. Tatarkiewicz, the Polish philosopher, describes the discussed theory in that manner:

everything what we do, we do with a view on the future,

we need solely a pleasure,

if we need to do something, we do it solely for the sake of pleasure,

the desire of pleasure is one of the main human motives and it directs us in our decisions,

pleasure is the aim of all human efforts.

Independently of all these versions, for all adherents of psychological hedonism, the principle "people avoid pain and desire pleasure" is a descriptive one, which covers some psychological facts and characterizes human motives, reasons and behavior. Such is its epistemological status. I don't inquire into the truth of this principle. The problem involves another question. - What will happen if we give normative character to the above rule? Then, a judgment about facts will change its meaning. It will turn into a judgment about duty. But, the sentence "people experience some pleasure and pain" is different from "people ought to aim at pleasure and to avoid pain". If we assume the identity of these two sentences, then we will pass from epistemological and psychological considerations to the normative level of our reflection. According to the descriptive principle people organize their lives. According to the normative principle people should organize their lives. For the utilitarians a transition (from the psychological level to the normative one) will be possible, provided that the factual values are perceived by people as superior aims of human conduct and are actually respected by them.

The world of factual values will be identified with the world of human purposes. But, to do it, we have to transcend individual, empirical experience and to extend the human knowledge by a contents which exceeds empirical data. This content must be derived from another source. If so, the empirical approach is insufficient. But utilitarian theory assumed that empirical data was the only possible one. Thus, we should reject that thesis, because it is incompatible.

We have approached the basic problem of utilitarian reduction called *the naturalistic fallacy*. What does this fallacy mean (in the case of utilitarian thinking)? Let's appeal to J.S. Mill and his argumentations included in the fourth chapter of "*Utilitarianism*". According to him, the only proof of that some object is visible is that it is viewed by people. Thus the only proof that an object is desirable is that it is desired. The above argumentation is based on an analogy between epistemology and ethics, an analogy assumed by J.S. Mill. The epistemology, based on the sense of perception, would be impossible if it predicated on things which have never been seen by anybody. Ethics would be in practice unacceptable if it pertained to such desirable aims, which nobody has never desired. But, the judgment "the visual perception of some object confirms its capability of being visible" is not equivalent to the judgment "the desire of something confirms its desirability". Because from the fact that something is desired doesn't follow that it is worthy of being desired. From the fact, that something is desired follows solely that it is capable to being desired. J.S. Mill's intention was to identify the word "desirable" with the phrase **"**worthy of being desired". In this case, *the naturalistic fallacy* rests on indistinguishing what *is* from what *ought to be* and it is based on ambiguity of the word "desirable". Duties do not follow from facts. An attribute of things worthy of being desired is that there some people who desired them. But, this does not imply that all things desired by people are worthy of being desired. The fact that something is desired is not sufficient to prove that it is worthy of being desired. From the judgment "people desire pleasure" does not follow the opinion that pleasure ought to be the aim of human action and, secondly, that the pleasure ought to be desired every time, everywhere and by everybody.

Let's recall the opinion of J. Dewey. According to him, desires are merely desires and that's all what we can say about them. None of psychological theories is sufficient to build the general theory of values. In utilitarian axiology, the conclusion drawn from the psychological inquiries and the description of the human desires are identified with the evaluation of these states of affairs. Let's repeat, description is identified with evaluation. That is a primary mistake of all utilitarian hedonistic thinking.

Another important issue involves the possibility of experience of sadistic pleasures, such which are the source of the human pain. According to J. Bentham, there are people who have done something with ill intention. Let's call it malevolence, jealousy, cruelty. Their motive of action will be always some kind of pleasure. It will be the pleasure which people experience when they think of somebody's pain. That unworthy pleasure is good in itself, as well. It may be undignified but it is as good as any other. There are no better or worse pleasures. The value of pleasures is connected entirely with their intensity. Thus, if sadism is a source of our pleasure it ought to be recommended. If we are pleased with the sadistic acting we ought to prefer it. Such conclusion ensues from the assumption that pleasure could be the only basis of axiology and ethics.

The next problem connected with reduction in axiology is presented by R. Nozick. Generally, this problem concerns what happens if we reduce all values to pleasurable states. R. Nozick describes some hypothetical situation. Let's assume, that we are in a special constructed machinery which provides abilities for experience of many pleasures. Neurologists stimulate, by means of electrodes, our brains in such way that we feel again and again sensual pleasure. Do we want to live in such constructed machinery? R. Nozick gives a negative answer to such a question. According to him, all of us wish to be persons, somebody more than only the object of positive sensations. Therefore, a valuable life is something more that the sum of pleasurable experiences. If we agree with the philosopher in this matter we will have to admit that axiology, based on the primacy of individual experience, makes the axiological sphere very poor. That way of thinking reduces possibilities of men's evolution because it does not invite to philosophical inquiries. Poor axiology makes human life shallow and degrades a human being. The rule "people ought to pursue of pleasurable states" reduces the human seeking for sense of life solely to one sensual sphere.

To sum up, hedonistic utilitarianism has reduced human life to the pursuit of pleasurable states. Such thinking has treated one value as the only one and the most important one. Thus, pleasure has gained a monopolistic position in the realm of values at the costs of other values. N. Hartmann has compared it to the developing of cancer tissue in human body, underlying pathological nature of such way of thinking about values. Reduction in axiology leads to tyranny of one value over the others and makes it impossible to reach the emotional, psychical and intellectual maturity of man.

Literature

Bentham, Jeremy 1948, *An Introduction to the Principle of Morals and Legislation*, Oxford.

Brandt, Richard 1959, *Ethical Ttheory*, Englewood Cliffs.

Broad, Charlie 1930, *Five Types of Ethical Theory*, London.

Brink, David 1986, *Utilitarian and the Personal Point of View*, Journal of Philosophy 83.

Brink, David 1989, *Moral Realism and the Foundation of Ethics*, New York.

Fritzhand, Marek, 1982, Wartości i fakty, Warszawa.

Mill, John, Stuart 1861, *Utilitarianism*, CW v.10. Toronto.

Sidgwick, Henry 1907, *The Methods of Ethics*, London.

Tatarkiewicz, Władysław 1985, *O szczęściu*, Warszawa.

Troland, Leonard, Thompson 1928, *Fundamentals of Human Motivation*, New York.

The Return of Reductive Physicalism

Panu Raatikainen, Helsinki, Finland

1. Motivating Physicalism

As Papineau (2001) argues, the popularity of physicalism among contemporary philosophers is not just a result of arbitrary fashion: it has rather been characteristically motivated by a certain line of reasoning, which is based on the apparently plausible assumption of "the completeness of physics" and the worry that the mental would end up being causally epiphenomenal. By "the completeness of physics" one means here the assumption that all *physical* effects are due to physical causes. The exclusion argument is then, roughly, that anything that has a physical effect must itself be physical. Thus, if the mental is capable of causing physical effects, it must be itself physical.

Smart (1959) therefore proposed that we should identify mental states with brain states, for otherwise those mental states would be "nomological danglers" which play no role in the explanation of behaviour. Armstrong (1968) and Lewis (1966, 1972) argued that, since mental states are picked out by their causal roles, and since we know that physical states play these roles, mental states must be identical with those physical states. Lewis (1966) made the completeness assumption explicit. More recently, Kim (1989, 1992) and Papineau (1993, 2001) in particular have pressed the exclusion argument in defence of physicalism. In sum, it is fair to say the exclusion argument, or something like it, is essential for contemporary physicalism.

2. The Causal Exclusion Argument

Let us look a bit closer at the argument. Assume first that reductive physicalism is false:

> *Assumption:* (distinctness)
> Mental properties are distinct from physical properties.

However, the following three premises are – so the argument goes – apparently indisputable:

> *Premise 1* (the completeness of physics):
> Every physical occurrence has a sufficient physical cause.
>
> *Premise 2* (causal efficacy):
> Mental events sometimes cause physical events, and sometimes do so by virtue of their mental properties.
>
> *Premise 3* (no universal overdetermination):
> The physical effects of mental causes are not all overdetermined.

However, these four claims together are arguably inconsistent. Therefore, physicalists conclude, the assumption must be rejected.

> Conclusion:
> Mental properties must be identical to physical properties.

On can surely dispute this argument (as I, for one, would), but the point here is to emphasize just how vital this argumentation strategy is for contemporary physicalism.

3. Multiple Realizability

The (type) identity theory is, of course, today disbelieved by many because of the so-called multiple realizability argument: it is suggested that a particular mental kind can be realized by many distinct physical kinds (see e.g. Putnam 1967, Fodor 1968, 1974, Block and Fodor 1972). Consider, for example, pain. It seems plausible that various different animals – humans, primates, other mammals, perhaps even birds and reptiles – are all capable of having (the same kind of) pain. However, it is also clear that these different animals must have radically different physical-chemical build-up. Therefore, it would be a mistake to identify the property of having (a certain kind of) pain with any particular underlying physical-chemical property, for the latter must vary greatly between different species. Further, it has been argued that the underlying physical state which realizes a certain mental state may be different even in the same individual at different times (see below).

4. Reductionism Strikes Back: Attacking the Nagelian Model

It would be, however, premature to proclaim the death of reductionism. Several philosophers have accepted the claim that psychological kinds are multiply realized while nevertheless denying such non-reductionist conclusions.

In particular, it has been often suggested that the anti-reductionists' master argument (based on multiple realizability) presupposes the classical Nagelian picture of reduction (cf. Nagel 1961), but that this theory is arguably outdated and problematic; and that if one leans instead on a more accurate picture of reduction, one might be able to vindicate reductionism. Such arguments are certainly worthy of serious consideration.

(*1*) *Mere biconditionals are too weak.* It has been argued that biconditional bridge laws of the Nagelian model are too weak, and must be strengthened into *identities* if they are to yield genuine reductions (Sklar 1967; Causey 1977; Wimsatt 1976, 1979; cf. Kim 1998). What is really needed is, for example, the identity "temperature = mean molecular kinetic energy", not just a law that merely affirms the covariance of the two magnitudes. In Kim's words, "as long as the reduction falls short of identifying them, there would be temperatures as properties of physical systems 'over and above' their microstructural properties." Furthermore, such correlations are badly in need of an explanation: why does temperature covary in just this way with mean molecular kinetic energy? Identifying the two magnitudes answers this nicely: because they are in fact one and the same (see Kim 1998).

Such a view certainly has much to be said for it, but the relevant point here is that this stronger requirement hardly makes life any easier for reductionists: if multiple realizability ever was an obstacle to Nagelian reducibility via biconditionals, how much more so would it be if genuine identities are required? Therefore, apparently pressing this point is in itself little help to reductionism.

(*2*) *One-way Nagelian bridge principles.* It has also been suggested that the common anti-reductionist argument involves a misunderstanding of the Nagelian model: although Nagel's examples involve biconditional bridge laws, one-way conditional connections expressing sufficient conditions at the reducing level are all that his "principle of derivability" requires. Multiple realizability, on the other hand, only challenges necessity reducing conditions, and so is not a challenge to even a projected Nagelian reduction of psychology to the physical sciences (Richardson 1979; cf. Bickle 2000). However, this approach is in considerable tension with the first objection to the Nagelian model, which requires that the connections between the reduced theory T_1 and reducing theory T_2 must be strengthened, not weakened. Also, this strategy is in conflict with the exclusion argument: mere one-way conditionals simply won't do; only genuine identities suffice. Otherwise the mental (even if it may be "reducible") may end up being causally epiphenomenal.

(*3*) *Reduced theories are often false.* Still another problem for the Nagelian model is presented by the undeniable historical fact that often the reduced theory T_1 is, strictly speaking, false, and hence its laws can be derived from the reducing theory T_2 (which is assumed, for the time being, to be true) only with some false auxiliary assumptions, for example. This holds even for many classical examples of reduction in science. Consequently, it has been proposed that what one really reduces (e.g., derives from T_2) is a revised version of T_1 (or its "image" in T_2) rather than T_1 itself (Schaffner 1967; Churchland 1986).

Yet, this observation poses no serious problems for the anti-reductionist; she need not commit herself to any specific theory involving mental states (indeed, it is not even clear that there exists a definite theory of the mental that would be at stake here). She can well allow extensive correcting of theories. The real question here is about ontological reduction of properties, not about epistemological reduction of any particular theory. (For this distinction, see Silberstein 2002.)

(*4*) *Local reductions.* Finally, it has been argued that the Nagelian condition of connectibility is "unrealistic and can seldom be satisfied" (Kim 1998), and that even in many paradigmatic examples of reduction in the history of science the reduced concept is multiply realizable; e.g., that temperature cannot be uniformly identified with a single micro-based property; it may be mean kinetic energy of molecules for gases, but something else for solids or in vacuums; but one may still have "local reductions" which are specific to domain, structure or species – for example of temperature in gas, or, pain in humans (Hooker 1981, Enc 1983, Churchland 1986, Ch. 7, Kim 1998, Bickle 2000).

However, this strategy too faces problems in the presence of the exclusion argument: together they threaten to make the general property of pain but also of temperature causally epiphenomenal (see below). And in any case, this kind of reply may well be insufficient: one may argue for a much more radical type of multiple realizability, that the underlying physical realizations of a certain psychological property may differ even in the one and the same token individual over time (Block 1978, Horgan 1993). Arguably there is some empirical evidence about such plasticity (Endicott 1993). Yet, in the case of such a massive kind of multiple realizability, the idea of local reduction leads to absurdity. At least one key advocate of new wave reductionism, John Bickle, grants this much: "The more radical type of multiple realizability seems to force increasingly narrower domains for reductions to be relativized – at the extreme, to individuals at times. This much 'local reduction' seems inconsistent with the assumed generality of science" (Bickle 2006).

5. A Reductionist Reply to Radical Multiple Realizability

Bickle – following Hooker (1981) and Enc (1983) – argues that the radical type of multiple realizability is a feature of some "textbook" cases of reduction, such as the one of classical thermodynamics to statistical mechanics: "For any token aggregate of gas molecules, there is an indefinite number of realizations of a given temperature – a given mean molecular kinetic energy. Microphysically, the most fine-grained theoretical specification of a gas is its microcanonical ensemble, in which the momentum and location (and thus the kinetic energy) of each molecule are specified. Indefinitely many distinct microcanonical ensembles of a token volume of gas molecules can yield the same mean molecular kinetic energy. Thus at the lowest level of microphysical description, a given temperature is vastly multiply realizable in the same token system over times … So this type of multiple realizability is not by itself a barrier to reducibility" (Bickle 2006).

To begin with, a "textbook case" or not, one should note that some distinguished philosophers of physics have expressed serious doubts about this alleged reduction of classical thermodynamics. The locality of it mentioned above is actually one of the reasons, but there are many more: for example, statistical mechanics is time symmetric whereas thermodynamics possesses time asymmetry (Primas 1991, 1998, Sklar 1993, 1999; cf. Silberstein 2002). So perhaps one should not make too much out of the contingent fact that Nagel and some others took this as a good example of reduction.

Moreover, this strategy has its price; now the exclusion argument enters, and one must conclude again that temperature is a causally epiphenomenal property. For, temperature clearly is not *identical* to any particular token microlevel arrangement or such.[1] But without genuine identities, one cannot circumvent the exclusion problem. And surely, the conclusion is implausible. For example, consider water boiling in a test tube. Now the difference-making cause of the boiling is seemingly the temperature (> 100°C) of its surroundings, no matter whether it was solid or gas, and whatever its specific arrangement of microparticles is. Therefore, something must have gone wrong.

6. Conclusions

Advances in understanding reduction in the philosophy of science must certainly be taken into account. However, arguably they are not as much help to reductive physicalism as is sometimes suggested. In particular, contemporary physicalism which essentially depends on the exclusion problem does not harmonize well with certain new views of reductionism.

[1] As Tim Crane pointed out in a discussion, essentially the same worry is already presented (in a slightly different context) in Crane (1992).

Literature

Armstrong, D. 1968 *A Materialist Theory of the Mind*, London: Routledge.

Bickle, J. 1998 *Psychoneural Reduction: The New Wave,* Cambridge, MA: MIT Press.

Bickle, J. 2000 “Concepts of Intertheoretic Reduction in the Contemporary Philosophy of Mind”, *A Field Guide to the Philosophy of Mind* (online): http://host.uniroma3.it/progetti/kant/field/cir.htm

Bickle, J. 2006 “Multiple Realizability”, in: Ed Zalta (ed.), *Stanford Encyclopedia of Philosophy,* http://plato.stanford.edu/entries/ multiple-realizability/

Causey, R.L. 1977 *Unity of Science*, Dordrecht: Reidel.

Churchland, P.S. (1986) *Neurophilosophy*, Cambridge, MA: MIT Press.

Crane, T. 1992 “Mental Causation and Mental Reality”, *Proceedings of the Aristotelian Society* 92, 185–202.

Block, N. 1978 “Troubles With Functionalism”, in: C.W. Savage (ed.), *Perception and Cognition: Issues in the Foundations of Psychology,* Minnesota Studies in the Philosophy of Science, vol. 9. Minneapolis: University of Minnesota Press, 261–325.

Block, N. and J. Fodor 1972 “What Psychological States Are Not”, *Philosophical Review* 81, 159–181.

Enç, B. 1983 “In Defense of the Identity Theory”, *Journal of Philosophy* 80, 279–298.

Endicott, R. 1993 “Species-Specific Properties and More Narrow Reductive Strategies”, *Erkenntnis* 38, 303–321.

Fodor, J. 1968 *Psychological Explanation,* New York: Random House.

Fodor, J. 1974 “Special Sciences: Or the Disunity of Science as a Working Hypothesis”, *Synthese* 28, 97–115.

Hooker, C. 1981 “Toward a General Theory of Reduction”, *Dialogue* 20, 38–60, 201–35, 496–529.

Horgan, T. 1993 “Nonreductive Materialism and the Explanatory Autonomy of Psychology”, in: S. Wagner and R. Warner (eds.), *Naturalism: A Critical Appraisal,* Notre Dame: University of Notre Dame Press, 295–320.

Kim, J. 1989 “The Myth of Nonreductive Physicalism”, reprinted (1993) in Kim, *Supervenience and Mind*, Cambridge: Cambridge University Press, 265–284.

Kim, J. 1992 “Multiple Realization and the Metaphysics of Reduction”, *Philosophy and Phenomenological Research* 52, 1–26.

Kim, J. 1998 “Reduction, Problems of”, in: E. Craig (ed.) *Routledge Encyclopedia of Philosophy*, London: Routledge.

Lewis, D. 1966 “An Argument for the Identity Theory”, *Journal of Philosophy* 66, 23–35.

Lewis, D. 1972 “Psychophysical and Theoretical Identifications”, *Australasian Journal of Philosophy* 50, 249–258.

Nagel, E. 1961 *The Structure of Science*, New York: Harcourt.

Papineau, D. 1993 *Philosophical Naturalism*, Oxford: Blackwell.

Papineau, D. 2001 “The Rise of Physicalism,” in: Loewer and Gillett (eds.), *Physicalism and Its Discontents*, Cambridge: Cambridge University Press.

Primas, H. 1991 “Reductionism: Palaver without Precedent”, in: E. Agazzi (ed.) *The Problem of Reduction in Science*, Dordrecht: Kluwer, 161–72.

Primas, H. 1998 “Emergence in the Exact Sciences”, *Acta Polytechnica Scandinavica* 91, 83–98.

Putnam, H. 1967 “Psychological Predicates”, in: W.H. Capitan and D.D. Merrill (eds.), *Art, Mind, and Religion*, Pittsburgh: University of Pittsburgh Press, 37–48.

Richardson, R. 1979 “Functionalism and Reductionism”, *Philosophy of Science* 46, 533-558.

Schaffner, K.F. 1967 “Approaches to Reduction”, *Philosophy of Science* 34, 137–47.

Silberstein, M. 2002 “Reduction, Emergence and Explanation”, in: P. Machamer and M. Silberstein (eds.) *The Blackwell Guide to the Philosophy of Science,* Oxford: Blackwell, 80–107.

Sklar, L. 1967 “Types of Intertheoretic Reduction”, *British Journal for the Philosophy of Science* 18, 109–24.

Sklar, L. 1993 *Physics and Chance: Philosophical Issues in the Foundations of Statistical Mechanics,* Cambridge*:* C.U.P.

Sklar, L. 1999 “The Reduction (?) of Thermodynamics to Statistical Mechanics”, *Philosophical Studies* 95, 187-99.

Smart, J.J.C. 1959 “Sensations and Brain Processes”, *Philosophical Review* 68, 141–156.

Wimsatt, W. 1976 “Reductive Explanation: a Functional Account”, in: *PSA 1974*, Dordrecht: Reidel, 671–710.

Wimsatt, W. (1979) “Reduction and Reductionism”, in: P. Asquith and H. Kyberg (eds.), *Current Research in the Philosophy of Science*, East Lansing: PSA, 352–377.

Rethinking the Modal Argument against Nominal Description Theory

Jiří Raclavský, Brno, Czech Republic

Taking advantage of Kripke's famous claim – "If the name means the same as that description or cluster of descriptions, it will not be a rigid designator." (Kripke 1972, 276) –one can complete a *Kripkean modal argument* (MA) against description theory of proper names. It is sufficient to use Kripke's key thesis − proper names are rigid designators − as a second premise and formulate an appropriate conclusion.[1] I will confine myself to the version of MA directed against the *Nominal description theory* (NDT). According to NDT, the meaning of a proper name *N* is the same as the meaning of a *nominal description ND*, 'the only individual named *N*' (or 'the only bearer of *N*'). It is an interesting proposal going against the meaning reduction advocated by Saul Kripke. NDT as a semantic theory has been recently defended mainly by Kent Bach (1987).[2] Here is my MA:

> If a proper name has the same meaning as
> a nominal description, it is not a rigid designator.
>
> A proper name is a rigid designator.
>
> Thus, a proper name does not have the same
> meaning as a nominal description.

MA is an instance of *modus tollens,* which shows its validity.[3] Nevertheless, Bach's early critical comments on the use of the argument (Bach 1981, 152) made me rethink the formulation of MA and re-examine its soundness. Here are several objections to MA (only the first one was expressed by Bach):

> i. *ND* is a rigid designator. For there is no possible world *W* in which an individual is named *N* and it is not picked out by the respective *ND*.
>
> ii. If *ND* is a non-rigid designator due to the fact of baptizing of an individual by *N*, thus not singling out N before the baptism by *N*, then *N* is a non-rigid designator too. For despite *N* behaves rigidly after the baptism, it does not designate N before the baptism (or in those worlds in which N is not baptized by *N*).
>
> iii. When we identify the meanings of *N* and *ND* (or certain other suitable rigid description), the Kripkean MA does not force us to accept its conclusion. Although MA is valid, its soundness is doubtful.

The evident contingency of baptism is in fact the main issue. The question of rigidity or non-rigidity of nominal descriptions is closely connected to it. To resolve the puzzles arising from the above objections, I will expose a comprehensive picture of language and its explication.

Language and its explication within an intensional framework

A natural language was developed by our predecessors as a medium serving for transmission of information about facts. An elementary fact consists basically in an individual possessing certain property. Suppose that a subject S confronted with an external reality recognizes various individuals and manages a set of pre-theoretically given attributes (properties, relations) denoted by predicates of a language L used by S. Modelling attributes as mere classes of individuals, as it is common within extensionalistic construal of language, is inadequate because a class is given by objects belonging to it. It is thus necessary for an individual to be a member of a given class. Empirical facts, however, are undoubtedly contingent. It is clear that there is a plentitude of conceivable distributions – not just a single one – of the same attributes over the same individuals. Each of (realizable) distribution is a *possible world*. Possible worlds serve then as an index enabling us to model contingency, a *modal variability*. Attributes are explicated as certain intensions, namely as mappings from possible worlds to classes of individuals (or classes of *n*-tuples of individuals). Propositions are intensions having truth-values as their values. Etc. It seems reasonable to adopt also a *temporal parameter* because possession of a property by an individual changes through the passing of time. Then *intensions* are mappings from possible worlds to chronologies of objects; they will be spoken simply as mappings from worlds and times (to certain objects). Now I wish to stress the most important point implied by the above consideration. The language L is underlaid by a definite range of possibilities which I will call *intensional ground*. Realize that another language, L', can be underlaid by a different ground because when it names different individuals than L, the thinkable distributions of attributes are naturally different from those related to L.

There is also another important point. The language L is construed as 'synchronically given'. L is explicated as a (fixed) mapping associating expressions with meanings, or more simply: expressions with denoted objects (intensions or non-intensions). Such mappings may be conveniently called (linguistic) *codes*. One might doubt whether our natural languages are such codes. But it seems quite inadequate to suppose that languages do not at least contain something as a code, i.e. a vehicle for communication of meanings. According to another possible objection, it is more natural to conceive language as a grammatical system. However, the idea that there is a finite system of grammatical rules generating (or operating on) an infinite list of expression-meaning pairs, i.e. a code L, does not contradict the idea of a code. Another objection against L explicated as a code is based on the assumption that expressions belonging to L 'change' their meanings: an expression can, for example, mean C_1 at one time and it can mean C_2 at another time or in another world. Within our intensional framework, this fact has a natural explanation. I will borrow the term 'diachronically given language' from linguistics intending to name a language modally and temporally conditioned. One should thus distinguish a natural language construed as

[1] Scott Soames (1998) formulated another MA. I will not examine here the differences between my and Soames' version. Btw. Soames did not take into account rigid descriptions (including nominal ones).

[2] The first proponent of NDT was Bertrand Russell followed by William Kneale. Recently, François Récanati and also Stefano Predelli proposed a 'formal-pragmatic' version of NDT (the meaning of a nominal description in not the meaning of proper name but it occurs somehow in its 'content').

[3] The attack against its formal validity raised by Brian Loar (and followed by Bach) is wrong. I have not enough space to prove it here.

a code, L, from that understood in the sense of a mapping from possible worlds and times to codes, sign it "L".

Lt me state a *semantical scheme* I presuppose. An expression E *expresses* in L a *meaning* C which determines the *denotatum* D of E in L. The denotatum of E in L is an intension / a non-intension / nothing. I believe that the adoption of an 'hyperintensional' level is reasonable, differentiating thus 'structured meanings' from unstructured denotata.[4] But since I will suppress this semantical consideration in the present paper, the meanings of expressions will be identified with their denotata. *Non-empirical expressions* – e.g. genuine proper names of individuals, names (or descriptions) of numbers or mathematical/logical functions – are expressions whose reference is stable across the possible worlds and times. On the other hand, a typical *empirical expression* has a stable denotation but a varying reference. For instance, the word 'horse' denotes a property but it refers to various classes of individuals in distinct worlds and times. The *reference* of an empirical expression E is the value of the intension denoted by E in a particular possible world *W*, time-moment *T*. Examples of empirical expressions: 'the U.S. president' (which denotes an 'individual concept' but it refers in some *W*'s, *T*'s to G.W. Bush, in other *W*'s, *T*'s to J. Ratzinger), 'It rains in Austria'. To know the reference of a typical empirical expression in the actual world and the present time one has to examine the contingent state of reality; it is not deducible by means of pure logic. On the other hand, to know the reference of a non-empirical expression is in principle an *a priori* matter. Note also that the *name relation* is best identified with the denotation relation, not with the reference relation. For instance, 'G.W. Bush' names what it denotes, i.e. G.W. Bush; 'the U.S. president' names certain denoted individual concept (it is quite futile to insist on its naming G.W. Bush).

Refinements for MA

I have to elucidate some very important distinctions Kripke took for granted. One of the simplest is concerned with the rigidity or non-rigidity and expressions. An expression is rigid iff its reference is stable across possible worlds and times; otherwise it is non-rigid. But since rigidity is a semantical property of expressions, it is *language-relative*. For instance, when 'the U.S. president' denotes in L' the number 7, it is a rigid designator, despite it is a non-rigid designator in L. Correcting thus the above definition: an expression is *rigid* in L iff its reference in L is stable across the possible worlds and times. (A variant of this definition has "L" instead of both occurrences of L.)

When discussing semantical properties of expressions in L, Kripke has in fact *used another language*, call it M. Assume that the code M is a language of our considerations too. (M works as a certain metalanguage in which we grasp L.) As I have discussed above, L is underlaid by a specific intensional ground, IG_{L}. Also M is underlaid by a specific intensional ground, IG_{M}. Within IG_{M}, there is a conceivable circumstance that S speaks in W_k, T_k by means of L, whereas there is also a thinkable circumstance that S enjoys rather L' in the same W_k but at T_{k+1}, or a circumstance that S uses L also in $W_{k'}$ *at* T_{k+1}. Remember, therefore, that M is underlaid by a specific IG_{M} which enables us to discuss various contingencies, e.g. those about the uses of L.

It is quite clear that a genuine proper name such as *N* is a rigid designator (of L). Thus the individual N − named in L by *N* − figures in the intensional ground IG_{L}. However, not every proper name syntactically possible within L names a particular individual. For the reasons of simplicity I will assume that a proper name not naming an individual in L is meaningless in L; it may be also spoken as a non-designator of L. For instance, an individual N' cannot be directly referred to by a proper name *N'* of L when *N'* was not endowed in L by a meaning (denotation). Now when users of L encounter N', they can baptize it by the expression *N'*. After the successful baptism, the users of L cease to use L in which *N'* is meaningless − they begin to use L' in which *N'* is a genuine proper name. Needless to say, *N'* is a rigid designator of L', thus the individual N' figures in $\mathrm{IG}_{\mathrm{L'}}$. The changes of codes are not usually noticeable because we do not name codes by L or L'; we use rather "L", i.e. a description singling out particular codes. Briefly, a baptism of an individual amounts to the replacement of L by L' in W_k within one time-interval, a passage from T_k to T_{k+1}. The description "L" picks out L in W_k, T_k but it picks out L' in W_k, T_{k+1}. *N'* is a non-designator of L but it is a rigid designator of L'.[5] A *baptism is a contingent matter* figuring inside IG_{M}; when users of (a value of) "L" baptize certain individuals, "L" changes its value – L is replaced by L'.

Now we are ready to distinguish two kinds of nominal descriptions. The description *NDL* (or *NDL'*), i.e. 'the only individual named in L by *N*', is *a rigid nominal description* denoting an intension which picks out the very same individual N in all possible worlds and times. The relation "named" mentioned in it links an individual with *N* and the code L. To know which individual is picked out by *NDL* one need not examine worlds and check time – it is sufficient to find out which individual is named in L by *N*.[6] However, the description *ND"L"*, i.e. 'the only individual named *N* in "L" ', is (typically) *a non-rigid nominal description* denoting an intension (an individual concept) which is not constant. The relation "named" links an individual with *N* and a code which is a contingent value of "L". When the value of "L" such as L' contains *N* as a meaningful proper name of N, then *ND"L"* picks out N. When the value of "L" such as L does not contain the proper name *N* as meaningful, then *ND"L"* picks out nothing. When the value of "L" is L'', in which *N* means horsiness, then *ND"L"* picks out nothing because no individual is identical with horsiness. Notice also that the above disputed circumstances belong to IG_{M} and that *NDL* and *ND"L"* are meaningful parts of M (not of L or any other value of "L").

Soundness of two versions of MA

The original MA should be properly refined according to the above considerations. There arise thereby two versions of MA: MAL containing rigid nominal descriptions and MA"L" containing non-rigid nominal descriptions. It is easy to conclude that MA"L" is a sound argument. As NDTians prefer rather rigid nominal descriptions, the

[4] Nearly all ideas from the present section are adapted from the work of Pavel Tichý (e.g., 1988). As structured meanings, Tichý introduced so-called constructions − abstract (structured) procedures that may be seen as objectual pendants of λ-terms.

[5] Note that *N'* can be a non-rigid designator of L'' or that *N* − originally a rigid designator of L − can become a non-designator in L'', when users of L'' have forgotten what *N* meant in the preceding values of "L".

[6] Rigid nominal descriptions split into two kinds: with or without a reference. For instance, *N'DL* refers to no individual because *N'* does not name in L anything at all. I classify such descriptions as rigid because their reference (that is null) is stable, non-varying.

soundness of MA“L” does not disquiet them. However, the truth of the MAL’s consequence − that a proper name does not have the same meaning as a nominal description − is a disputable matter: Kripkeans consider it false whereas NDTians consider it true. Hence, MAL as such is insufficient for the change of opinion on the part of NDTians.

Presumably, both groups of theoreticians share the belief that proper names are rigid designators. Realize, however, that the respective Kripke’s *semantical thesis* about the meaning of proper names is in fact weak because it cannot distinguish proper names from rigid descriptions. Therefore, we need to suggest another, more provident, semantical thesis. My suggestion of ST is as follows: *a proper name is an expression whose denotation is the same as its reference* (language-relativity should be added, of course). Rigid descriptions – including the nominal ones – are thus not allowed to be proper names. (Notice also that Kripke’s key thesis follows from my ST.). NDTians may still disagree with this proposal. Consequently, NDTians construe MAL as not sound. Now we should argue that NDT is a materially less adequate explication of proper names’ meaning than ST because most of competent language users do not think that the meaning of *N* contains the meaning of ‘the’, ‘bearer’, etc. Hence, ST is a more preferable proposal and MAL becomes sound.

Literature

Bach, Kent 1981 “What’s in a Name”, *Australasian Journal of Philosophy* 59, 371-386.

Bach, Kent 1987 *Thought and Reference*, Oxford-New York: Oxford UP.

Kripke, Saul 1972 “Naming and Necessity”, in: Gilbert Harman and Donald Davidson (eds.), *Semantics of Natural Language*, Dordrecht-Boston: D. Riedel, 253-355, 763-769.

Soames, Scott 1998 “The Modal Argument: Wide Scope and Rigidified Descriptions”, *Noûs* 32, 1-22.

Tichý, Pavel 1988 *The Foundations of Frege’s Logic*, Berlin-New York: Walter de Gruyter.

Different Ways to Follow Rules? The Case of Ethics

Olga Ramírez Calle, Granada, Spain

1. The Proposal

The suggestion I want to put forward is that reflection on a correct account of rule-following in ethics invites to consider what I will call a 'three fold model' of conceptual application, that varies not just from the very basic cases but also from what has been called 'modus ponens' cases of rule following. Or at least may be considered a specific variation of such a model. The working out of this possibility may help us gain a better understanding of what is at stake in competent use of a given group of concepts and, maybe, to sort out competing interpretations of rule-following in a more case-specific and less general account.

If conceptual acquisition and understanding is to be adequately understood according to the rule-following model, this model must be capable of accommodating very different kinds of concepts (classificatory, relational, functional, evaluative, mathematical concepts). It must give account of the different ways concepts relate to, and characterize, experience. We may put aside the relevance of this question either per impossibility – we cannot really separate experience from conceptual understanding – or simply acknowledge without further interest the obvious existence of differing instructions we grasp by grasping rules, and otherwise go on giving a uniform account.

But if it should be possible, and I believe it is, to attend to differences at this level it could turn out that we require differing interpretations of the very idea of rule following for different cases.

That there should be a difference between very simple, basic cases, of rule following and more complex ones, is widely acknowledged and supported by Wittgenstein's own writings. Basic cases are those where no further non-redundant linguistic specifications or reasons can be given, besides direct illustration of how concept application goes. In other cases, while acquisition may succeed without linguistic aid by the participation in further practices, some already linguistically trained subjects could on demand give some clearing explanations. And at higher levels of complexity in language acquisition, some language users at least may be able to articulate more or less sophisticated reasons to justify conceptual application. This need not amount to an exhaustive definitive definition but just sufficiently articulated necessary conditions – all these specifications resting surely at the end in basic concepts whose meanings cannot be put down in any fixed formula –. However, when rules are to some extent linguistically articulated in such a way, we get what has been characterized as the 'modus ponens model' of rule following.

These distinctions allude to the grade of specification of the rules given. Although this may not be unrelated to the topic, what I was questioning before was the different ways rules connect our concepts to experience. The focus is here on the specific conceptual contents involved. Some related proposals are made, for example, by Crispin Wright's (1992), (2002) distinction between *extension-determining* and *extension-reflecting* concepts. Extension-reflecting concepts would register "self standing properties" and therefore the possibility of getting it wrong makes sense. Even if what it is to fall under the concept is epistemically constrained (weather something falls under the concept subject to human considerations) its existence is not constitutively dependent on human responses or considerations. By extension determining cases, though, there is no sense in which truth could transcend what we would ourselves say, as what seems to us, our own impressions or responses to some experience, are part of the content we are registering. And it is the very conceptual content we have to do with that demands it to be so. In Wright words: it is *a priori* that best opinion determines truth. Concepts of primary qualities would fall under the extension-reflecting and concepts of secondary qualities, under the second. This last distinction, however, is supposed to have a wider application and extend to avowals, for example, and, maybe, to moral concepts.

If we try, however, to figure out an adequate understanding of the rule-following considerations in ethics, trying to specify the rules governing conceptual application, we become (as a result of considerations of content) what appears to be a variation of the modus ponens model. Weather this model fits into the extension determining schema will depend on how this is exactly formulated, but some additional distinctions will be called for on its regard.

2. Rule-Following in Ethics

According to McDowell (1981) non-cognitivist disentangled explanations of thick ethical concepts could not explain their consistent rule guided use if it were not on the basis of some value neutral feature we should be capable of recognizing and to which we would be responding to. And the problem is that it should not be possible to sort out what feature this is, what all the members of the extension have in common, without the aid of corresponding evaluative considerations.

Carefully considered there are actually two different assertions in McDowell's just cited claim:

(i) It should not even make sense to pick up a value neutral class equivalent to the one thick ethical concepts sort out without taking into account evaluative considerations.

(ii) It should not be possible, once the class is constituted, to see value neutral common features among the members of the extension of a thick ethical concept.

McDowell, I believe, wants to assert both (i) and (ii), but these two claims do not necessarily have to go together. Let's call the first the 'generation argument' (GA) and the second the 'application argument' (AA).

McDowell illustrates his general claim arguing that given a list of items (individuals, actions, etc. let us suppose) that belong to the extension of a thick concept we most probably won't be able to tell what such items have in common. The class consider in abstraction of evaluative aspects need have "no shape", form no kind.

The argument would reach a non-cognitivist claiming that it is at some such recognizable level of appraisal that we are to find the features we respond to. And surely if what is at issue is a model of causal reactions to an independent world, we would situate ourselves at some such level. But this does not preclude us from finding some common factor at a higher order level. There is nothing in the class of 'communists' or 'lawyers' or items used 'to hold the door' that could be distinguished at such a level either. The argument, thus, does not necessarily refute disentangled explanations. We may perfectly well have some such (non-recognizable) morally neutral class to which a moral value is added and a thick ethical concept applied.

This would allow an answer to AA: there is no reason why it should not be plausible to assume that it is some specific class of behaviour, say – identifiable separately and morally neutral on the first place – that guides or application of a thick ethical concept. Once the class is identified, the value is added and the concept applied.

But McDowell could still counter that an explanation saying *why* we should at all pick up such a class as morally relevant is missing. The apprentice who is suppose to understand how to apply the concept may this way learn to apply the term on the basis of this independently discernable class, but would not understand why such behaviours or persons are to be called morally good. He would be like a child sticking red labels to all square things, without making any more sense of this than following orders. Moreover such behaviours in virtue of which the moral evaluation is done, may not be paid much attention to if it were not for our interest in the moral classification: and this, without having to embrace McDowell's position, supports GA. In order to pick up some morally neutral class that is to be evaluated, we need some reason to fix special attention on it.

This seems right, without contradicting AA, that the relevant common characteristics are there all the way long. But it is not any more open to give a disentangled account insisting on the possibility of AA, this bringing us back to a modus ponens case –always when C(x), apply MV(x); application proceeding as in our child example. If this is not convincing, what kind of further specification would a disentangling supporter need in order to explain why some type of behaviour is to be morally evaluated such and so?

What fails is an explanation telling us why this or that behaviour is to be called 'good'. But behaviours may be good for satisfying very different goals: to be healthy, becomes acceptance in some sect, favour the Gods, the clouds or whatever. What we are looking for is something specific, what a thick moral concept is expressing is that some behaviour is *morally* good or bad, so what we look for is not just good but *morally good.* What needs explanation is

(i) What makes some kind of behaviour *morally good*?

(ii) What do we mean first of all with *morally*?

A somehow standard explanation would be to say that morality has to do with *the relation of men to each other (and their surroundings) we want to expect from all.* Substituting we obtain:

(iii) What kind of behaviour is *good* with respect to *what we want to expect from all in their relation to one another and their surroundings*?

Following Kantian proposals some will conclude, for example, that good relative to (iii) is what would *equally protect the needs and interests of the affected*. It is not my purpose to enforce this particular conclusion right now, the point is that whichever conclusion we may arrive at as an answer, it will deliver the *measure,* call it FM, relative to whose fulfilment some morally neutral behaviour is to be called 'morally good'.

(iv) A behavioural type is *morally good* if they fulfil FM

This binding engine is what would mediate between the antecedent and the consequent of the modus pones model: between the class of neutral behaviours on which basis we apply a moral value and the moral evaluation. We would actually have to do with a function that working on some descriptive level yields the evaluative as a result:

(I.) Behavioural type input b → MF → Evaluative (+/-) output MV

Conceptual application C

Applying this rule we obtain:

(1) b_1 fulfils MF
(2) MV+(b_1)
(3) C_1(b_1)

However by our rule guided application of a moral concept what we follow is the derived more simple modus ponens rule:

(II.) Behavioural input b_1 (assumed MV+)→ C_1(b)

Because it has already been calculated that characteristics b_1 fulfil MF, it is now a priori that whichever token falls under type b_1 it is MV. In this specific rule it remains implicit that FM is fulfilled and therefore MV+. This would account for the fact of children and traditional people finding no problem in blindly applying some such thick concepts on the basis of b_1, b_2, etc. without being capable of giving further explanations of why this behaviour is morally blameable, for example. The concept includes the explanation on itself by having established as *a priori* the relation between b_1 and MV by means of MF.

3. The Three Fold Model and its Implications

The presented model is what I call the *three fold model*. It is obtained by trying to give a more explicit and satisfactory disentangled account of our rule guided use of thick ethical concepts. The result being that there is some function such that, when it is fulfilled by some type behaviour, it qualifies it as morally good. So the content we have to do with is established by some operation that assigns by each ongoing input a given output. Thick ethical concepts result out of synthesizing some such result in a concept. Therefore, given some type of behaviour for which it has being established that it fulfils FM it is a priori that a given moral value applies to it. It is a priori determined, that whichever extension b_1 has, all its members (in virtue of some given operation) become a given value and fall as a result under some new class.

Do we have to do with *extension-determined* concepts then? The idea is here not that the decision of weather the concept applies depends on our best opinion because our own reactions or impressions to some behaviour should be *directly* decisive of the case. This would apply to response-dependent models. Here we have two different questions actually. a) Weather, in deciding if a thick concept applies, best opinion is all there is to truth.

And in presuming that given b_n some value applies, the question is if we do or don’t have to do with b_n. b) Weather the calculation required to assign a value to some b_n depends on best opinion. This will depend lately on weather human needs and interests, for example, can be determinable independently of our own responses –. Both questions I shall leave open here.

On the considerations made, however, some other distinction appears to be relevant. Contrary to concepts such as ‘red’, ‘tiger’, ‘cup’ or ‘tree’ whose meaning is open to development *on the way,* so to speak.. Some other concepts are such that their extension is dependent upon prefixed operations and to this extent there is no development of meaning on application. Any change would require going backwards and proving the correctness of the calculations made in its establishment. If this is right, we may distinguish between *open-ended* and *invariably prefixed* rules. That this distinction is not to be put together with that between *extension-determined* and *extension-reflecting* concepts can be seen as ‘red’, for example, would be a *extension-determined* but *open-ended* (susceptible of refinement or development). The distinction does not depend on weather best opinion determines of truth, but on the determinateness of meaning itself. Three fold concepts would fall under the second category but the distinction is not necessarily restricted to them.

Literature

Blackburn, Simon 1981 in: S. Holtzman and C. Leich (eds) *Wittgenstein: To Follow a Rule*, London: Routledge

McDowell, John 1981 “Non-Cognitivism and Rule Following” see above.

Wittgenstein, L 1967 *Philosophical Investigations*, 3rd edition Oxford: Basil Blackwell

Wright, Crispin 2002 ‘What is Wittgenstein’s point in the rule-following discussion?’ online in Boghossian/Horwich *Language and Mind* seminar, NYU

Wright, Crispin 1992 *Truth and Objectivity* Cambridge MA: Harvard University Press

Atypical Rational Agency

Paul Raymont, Toronto, Canada

1. The Capacity for Autonomous Decision-Making

In respecting one's autonomy I acknowledge her capacity to determine her own values and shape her life in accordance with them. To have this capacity, one must have some stable grasp of one's values and be able to articulate projects on the basis of them. Those who lack such skills are deemed to be 'incompetent', or incapable of truly autonomous agency.

In order to determine whether one possesses the requisite abilities, we are to focus not simply on a decision but, instead, on how it is reached, for we want to see whether the cognitive prerequisites of autonomy are evident in this decision-making process. We need to determine whether this person has the cognitive skills necessary for autonomous decision-making. We are not supposed to be addressing the quite different question of whether the choice itself is a good one. While such evaluation of the choice is relevant to moral and legal questions, it is not germane if our task is to determine whether the choice expresses genuine autonomy. Otherwise, there could be no bad autonomous decisions.

2. Reason in Action

Our aim is thus to achieve an understanding of the agent's decision, an account that will show her choice to be rational in the light of her values and projects. We need not agree with the choice, for we may not share these ideals. What is required, rather, is that her choice should appear to be rational *if* one starts from her ideals.

This sort of understanding is not supplied by a purely causal account of her choice; for the beliefs and values that make sense of her choice do so by means of prescriptive, normative standards rather than simply by means of the descriptive, nomological principles that sustain a mere causal explanation. To elaborate, when I give a mere causal explanation of an event, I subsume it under law-like generalizations, the implication being that the event occurred because things like it just do typically follow from those initial conditions. As John McDowell puts it, in this sort of causal account, "one makes things intelligible by representing their coming into being as a particular instance of how things generally tend to happen." (McDowell 1985, 389) By contrast, when I make sense of an action by rationalizing it, my objective is not to portray the act as how people just do tend to behave in such conditions. Rather, I aim to portray the action as what the agent rationally *ought* to do given her values and other attitudes. As McDowell says of such normative accounts, they are "explanations in which things are made intelligible by being revealed to be, or to approximate to being, as they rationally ought to be." (McDowell 1985, 389)

The distinctive nature of rationalizing accounts can be appreciated by juxtaposing them with mere causal explanations. Thus, suppose I want juice and believe that I can most readily satisfy this desire by getting the drink from the fridge. It is then rational for me to get the juice from the fridge, since the statement that I ought so to act is the conclusion of a practical syllogism in which the premises express the contents of the belief and desire in question. (Anscombe 1957) So the account of my action by appeal to this belief-desire pair does double duty as both an explanation and a rational *justification* that presents the action as being rationally appropriate.

By contrast, suppose that this same belief-desire pair were regularly followed by the motion of one's left hand one millimeter to the right. In that case, one could causally explain this hand motion by appealing to my desire for juice and my belief about how best to obtain that drink, together with the (*ceteris paribus*) law that links these attitudes to such a motion. Here, the *contents* of my belief and desire play no central role in accounting for the explained behaviour; after all, one can imagine the same sort of nomic link connecting that hand motion to different beliefs and desires, and (unlike in the case of rationalizing explanations) this variation in the attitudes' contents would subtract nothing from the explanatory work that is accomplished by appeal to such states

In this second, mere causal account, the explanation works because of the described nomic pattern, a relation that leads us to expect that the hand motion just will typically follow the onset of that belief-desire pair, without any implication that it is rationally appropriate for it to do so.

3. Further Distinctive Features of Rationalizing Explanation

It has long been recognized in psychiatry that there are two such distinct modes of explanation. This is due largely to the influence of Karl Jaspers. Jaspers adopted from Max Weber and others the distinction between understanding an action from the agent's perspective (*Verstehen*) and giving a causal account of the bodily motions that constitute the action (*Erklaren*). In his version of this distinction, Jaspers stressed that unlike the laws of nature, the rational principles that help to make sense of an action do not require confirmation by supporting cases in order to do their explanatory work. Whatever explanatory work is to be achieved by such rationalizing explanations does not await the discovery of a nomic pattern connecting the reasons to the action that they rationalize but is, instead, already there to be grasped just by understanding the belief-desire contents and their rational connection to the action.[1]

In his *Blue Book*, Wittgenstein likewise contrasts mere causal explanations with rationalizations. He notes that in the former case, the claim that an action resulted from a particular cause is a hypothesis, and adds that this hypothesis relies upon confirming instances which show "that your action is the regular sequel of certain conditions which we then call causes of the action." (Wittgenstein 1933/1965, 15) He contrasts this way of explaining an action with an account of the act in terms of the agent's reasons, where "no number of agreeing experiences is

[1] In Jaspers' words, "Frequency in no way enlarges the evidence for the connection. Induction only establishes the frequency, not the reality of the connection itself…. A poet, for instance, might present convincing connections that we understand immediately though they have never yet occurred." (Jaspers 1923/1963, 304)

necessary." (Wittgenstein 1933/1965, 15) Here, again, rationalizing explanations are independent of empirical confirmation.

Suppose that the general principles at work in a rationalizing account do not await empirical confirmation. Is it the case that they also retain their explanatory force in the face of disconfirmation? Yes, for they purport only to be normative principles, not true descriptions of actual patterns. Thus, for example, consider the case of Joan of Arc. We can explain (by rationalizing) her heroic actions against the English by appealing to her ideals even if many of her compatriots shared her ideals *without* acting on them as she did. This can be so even if Joan of Arc herself had not previously shown any greater tendency towards heroic deeds than her contemporaries. In this case, the statement that one who holds such ideals really ought to 'stand up for them' and oppose the enemy is not generally followed in the relevant population, but this is no obstacle to explaining or making sense of Joan of Arc's actions in terms of those such ideals.

4. Starson's Capacity for Rational Autonomy

Let us now examine issues concerning rational autonomy in the context of individuals who suffer from psychiatric illness.

Some of these people continue to exhibit rational patterns in their decision-making to greater or lesser degrees. It is difficult in such cases to determine to what extent such patterns must be present in order for one to be capable of exercising genuine autonomy in determining the course of her own health care.

A case of this nature was recently heard by the Supreme Court of Canada. (Supreme Court of Canada [SCC] 2003) The case involves Scott Starson, who was charged with issuing death threats to his neighbours. He was found to be not guilty by reason of his mental illness but was detained in a psychiatric hospital on the grounds that he posed a threat to others.

Starson refused to take medications that had been prescribed by his doctors, who then claimed that Starson was not capable of making his own treatment decisions and should therefore be required to follow the prescribed treatment. Rejecting this determination, Starson appealed to the courts. After appeals to various courts, the Supreme Court ruled that Starson was competent to make his own treatment decisions.

Starson's case attracted much attention because of his intellectual accomplishments, which include co-authored publications in physics. Indeed, one prominent physicist (Pierre Noyes of Stanford University) says that Starson has done "exciting" work that has stimulated some of his own thinking about the theory of relativity. While Starson has not published a scientific paper since the 1980's, he believes that his thinking about physics is a central source of meaning in his life. It is this dimension of his life that would, he believes, be extinguished by the medication. He bases this concern on previous experience with another anti-psychotic medication (Haldol), which dulled his mind to the point where he could no longer pursue his intellectual work.

In explaining the Court's ruling, Justice John Major did not deny that it may well be in Starson's best interests to take the medication. He adds that respect for capacity derives not from the concern for another person's best interests but, rather, from the duty to respect autonomy. Says Major, "The right to refuse unwanted medical treatment is fundamental to a person's dignity and autonomy." (SCC 2003, para. 75) According to Major, one's autonomy must be respected even at the cost of one's well-being.

Granted, but was Starson competent to make autonomous decisions? In the relevant jurisdiction, Ontario, the legal standard for competency, or 'capacity', is as follows:

> A person is 'capable' with respect to a treatment … if the person is *able to understand* the information that is relevant to making a decision about the treatment … and *able to appreciate* the reasonably foreseeable consequences of a decision or lack of decision. (SCC 2003, para. 12)

According to Major, the first part of this standard, the 'understanding' condition, "requires the cognitive ability to process, retain and understand the relevant information." (SCC 2003, para. 78) The second part, the 'appreciation' condition, requires that "the patient be able to apply the relevant information to his or her circumstances and to be able to weigh the reasonably foreseeable risks and benefits of a decision or lack thereof." (SCC 2003, para. 78)

Starson met these conditions, as is evident from his reasons for his choice. To wit, he knew that the medications were intended to slow his 'racing thoughts', and it was for that very reason that he rejected them. He rejected the risk of having his mind dulled to the point where he would be unable to pursue the central project in his life, his physics research.

As in the above example of Joan of Arc, to make sense of Starson's choice in this way is not to regard it as a typical choice, or as one that is statistically normal. More specifically, when we see his choice as being rationally motivated by his projects we do not thereby assume that most people would make the same choice as he did. Hence, we can take his choice to have issued from reasons that support it, and we can thereby regard his choice as an expression of rational autonomy, while at the same time seeing it as an atypical choice. Indeed, we can take Starson to be quite unlike most rational agents, to be quite odd in comparison to them, without this compromising our view of him as a rational agent who is capable of exercising genuine autonomy. It is not even required that we see Starson's choice as one that most people would make if they shared his goals and values, just as we need not take Joan of Arc's choices to be the most likely ones for someone who shared her ideals. We can, in other words, allow for disagreement among rational people.

This is because a rational agent's perspective typically encompasses a host of competing interests and convictions. Thus, Starson, while wanting to pursue his work in physics, at the same time recognized that his symptoms led him into conflict with others, and also desired to be released from the hospital in which he was detained. These countervailing concerns could equally rationalize a decision to comply with the prescribed treatment (just as a concern for self-preservation could rationalize a decision by Joan of Arc not to confront the English). It is for this reason that opposing choices can equally be seen as expressions of a rational, autonomous self, the implication being that we should not see just one choice, the 'normal' choice, as the sole candidate for being an expression of rational, autonomous agency.

Literature

Anscombe, G. E. M. 1957 *Intention*, Oxford: Basil Blackwell.

Jaspers, Karl 1923/1963 *General Psychopathology* (2nd ed.), trans. J. Hoenig and Marian W. Hamilton, Chicago: The University of Chicago Press.

Supreme Court of Canada 2003 *Starson v. Swayze*, 1 S.C.R. 722; 2003 SCC 32 (June 6, 2003), Available at http://www.lexum.umontreal.ca/csc-scc/en/pub/2003/vol1/.

Wittgenstein, Ludwig 1933/1965 *The Blue and Brown Books*, New York: Harper Collins.

Indexwörter und wahrheitskonditionale Semantik

Štefan Riegelnik, Wien, Österreich

Als die bestimmende Eigenschaft von Indexwörtern wird gemeinhin die Kontextsensitivität angegeben. Gemeint ist damit, dass das, worauf man sich mit einem Indexwort bezieht, nur im Kontext einer Äußerung bestimmt werden kann und der Bezug eines Indexwortes je nach Kontext variiert:

> „The ‚context-dependence' of indexicals is often taken as their defining feature: what an indexicals designates, *shifts* from context to context." (Perry 1997)

> „[...] it is nowadays a triviality to claim that indexical expressions are context sensitive, that is, their reference depends on the context in which they are used." (Corazza 2004)

Indexwörter sind strikt von singulären Termini und Namen abzugrenzen, denn während der singuläre Terminus „Wien" einen bestimmten Ort bezeichnet, der Name „James Joyce" eine bestimmte Person und „23:55" einen bestimmten Zeitpunkt, ist dies bei den Ausdrücken „hier", „er", oder „jetzt" nicht der Fall. Deren Bezug ergibt sich daraus, wo der Ausdruck von wem und wann gebraucht wird. Der Kontext einer Äußerung, also die Bestimmung des „hier", „er" oder „jetzt", ist für das Verständnis einer Äußerung zentral, wenn und weil es eine Bedingung der Wahrheit einer Äußerung ist, worauf sich die Wörter als Teil einer Äußerung beziehen. Da der Kontext variiert und Äußerungen auch dann verstanden werden, wenn der bestimmte Kontext nicht bekannt ist, stellt sich die Frage, wie Kontextbedingungen mit einer wahrheitskonditionalen Semantik vereinbart werden können.

Mit Indexwörtern kann eine wahrheitskonditionale Semantik scheinbar nicht derart umgehen, dass sowohl die Besonderheiten von Indexwörtern berücksichtigt werden und zugleich auch die Interdependenz von Wahrheit und Bedeutung nicht durchbrochen wird. Um die Schwierigkeiten anhand eines Beispieles zu skizzieren: „Dies ist weiß' ist wahr genau dann, wenn dies weiß ist" gibt zwar die Wahrheitsbedingungen des Satzes „Dies ist weiß" an und der Grundannahme zufolge, wonach man einen Satz versteht, wenn man weiß, unter welchen Bedingungen er wahr ist, müsste damit auch das Verstehen geklärt sein. Aber worauf man sich mit dem Ausdruck „dies" bezieht, wird durch die Wahrheitsbedingungen des Satzes „Dies ist weiß" nicht erklärt. Kurz: die besondere Rolle des Indexwortes „dies" kommt dabei nicht zum Ausdruck. Eine wahrheitskonditionale Semantik ist nun mit folgendem Dilemma konfrontiert: *Entweder* begnügt man sich damit, dass ein Ausdruck wie „dies" auf einen Gegenstand derart Bezug nimmt, dass damit in einer bestimmten Situation ein Gegenstand herausgegriffen wird. Das hätte aber zur Folge, dass die Wahrheitsbedingungen eines Satzes wie „Dies ist weiß" *situationsabhängig* sind, was wiederum die Folge hätte, dass dasjenige, was man versteht, wenn man diesen Satz versteht, gar nicht mehr situationsunabhängig geklärt werden könnte. *Oder*, um auf die andere Seite des Dilemmas zu verweisen, man beharrt darauf, dass das Verstehen von Äußerungen, die Indexwörter enthalten, *situationsunabhängig* ist, dann müsste man daraus aber schließen, dass die Bedeutung einer Äußerung *nicht* durch Wahrheitsbedingungen angegeben werden kann. In beiden Fällen ist die Interdependenz von Wahrheit und Bedeutung durchbrochen. Donald Davidson geht in „Truth and Meaning" (Davidson 1967) auf dieses Dilemma ein:

> „No logical errors result if we simply treat demonstratives as constants; neither do any problems arise for giving a semantic truth-definition. '"I am wise" is true if and only if I am wise', with its bland ignoring of the demonstrative element in 'I' comes off the assembly line along with '"Socrates is wise" is true if and only if Socrates is wise' with *its* bland indifference to the demonstrative element in 'is wise' (the tense)." (Davidson 1967)

Davidson relativiert die Wahrheitsbedingungen auf Person, Zeit und Ort der Äußerung, um so der Kontextsensitivität Rechnung zu tragen:

> „We could take truth to be a property, not of sentences, but of utterances, or speech acts, or ordered triples of sentences, times, and persons; but it is simplest just to view truth as a relation between a sentence, a person, and a time." (Davidson 1967)

Aber auch wenn die Wahrheitsbedingungen eines Satzes auf Zeit, Ort und Person relativiert werden, stellt sich die Frage, welchen Beitrag Indexwörter zur Bedeutung einer Äußerung leisten. Da die Wahrheit oder Falschheit einer Äußerung nicht nur davon abhängt, wie die Welt ist, sondern auch davon, welchen Beitrag die in einer Äußerung verwendeten Ausdrücke leisten, ist die Beantwortung dieser Frage für die wahrheitskonditionale Semantik von besonderer Relevanz.

Um auf das Dilemma zurückzukommen: im Sinne der ersten Alternative, also der situationsabhängigen Ausprägung, sind in letzter Zeit auch Versuche unternommen worden, die Bedeutung einer Äußerung gänzlich als vom Kontext abhängig zu verstehen, und zwar nicht nur derart, dass Person, Zeit und Ort für die Interpretation relevant sind, sondern dass der Kontext *allein* den Inhalt einer Äußerung bestimmt:

> „As the ordinary language philosophers rightly argued, literalism can't be right, because sentences are context-sensitive: in vacuo, the do not carry content, but do so only ‚in context.'" (Recanati 2006)

Die Strategie der Vertreter des Kontextualismus – so der Name dieses Ansatzes – besteht hauptsächlich darin, mit einer Reihe von Beispielen zu zeigen, dass die „richtige" Interpretation nur abhängig vom Kontext entschieden werden kann. Die bloße Angabe von Wahrheitsbedingungen, so die Vertreter dieser Richtung, reiche nicht aus, um alle für die Interpretation einer Äußerung relevanten Faktoren zu erfassen. Gerade in Hinblick auf die Klärung von Indexwörtern scheint dies auch eine attraktive Position zu sein, denn wie ich einleitend festgehalten habe, kann der Bezug eines Indexwortes nur im Kontext einer Äußerung bestimmt werden. Wenn aber der Beitrag zur Wahrheit oder Falschheit eines Satzes, in dem ein Ausdruck wie „dies" verwendet wird, vom unterschiedlichen Kontext bestimmt wird, dann kann ein Ausdruck gar nicht unabhängig vom Gebrauch im jeweiligen Kontext verstanden werden. Offen bleibt folglich auch, wie Indexwörter von Eigennamen unterschieden werden, denn in einer bloßen Kontext-situation bieten sich keine Kriterien an, die eine Unterscheidung zulassen würden. Erst die Beobachtung, dass auch andere Personen in einer anderen Situation ein Indexwort derart verwenden, dass sie auf unterschiedliche Gegenstände Bezug nehmen, erlaubt es, ein Indexwort von einem Eigennamen zu trennen. Um hier die Kritik am Kontextualismus allgemeiner zu formulieren: wenn Äußerungen nur situationsabhängig verstanden werden, wie kann dann überhaupt auf etwas Bezug genommen werden, was als Gegenstand einer semantischen Betrachtungen fungieren soll? Die Korrektheit dieser Theorie hätte schlicht das Abhandenkommen des Untersuchungsgegenstandes zur Folge.

Kritik am Ansatz, die Wahrheitsbedingungen auf Sprecher, Zeit und Ort zu relativieren, übt auch Stefano Predelli. Er führt eine Reihe von Äußerungen mit Indexwörtern an, in denen die Angabe von Sprecher, Zeit und Ort nicht ausreicht, um das intuitive Verständnis theoretisch zu erklären. Um hier ein Beispiel zu skizzieren (Predelli 2005): eine Person schreibt im Büro auf einen Zettel „Ich bin hier“ und trägt ebendiesen mit der Absicht nach Hause, jemanden über ihren Aufenthalt (zu Hause) zu informieren. Findet nun jemand im Haus diesen Zettel, wird er den Satz, so Predelli, *intuitiv* derart interpretieren, dass die Person, die diesen Zettel geschrieben hat, zu Hause sei. Die relativierten Wahrheitsbedingungen beziehen sich jedoch auf den Zeitpunkt der Äußerung und da befindet sich die Person im Büro. Predelli führt noch weitere, ähnliche Beispiele an, etwa historische Erzählungen und Beispiele aus der Literatur. Es sind aber vor allem geschriebene oder anders aufgezeichnete Äußerungen mit Indexwörtern, die als „Problemfälle“ gelten und dies hauptsächlich dann, wenn sie zu einem anderen Zeitpunkt interpretiert werden als sie geäußert oder aufgezeichnet wurden.

Wie soll man mit der Kritik Predellis umgehen? Ich möchte zunächst darauf hinweisen, dass die Beispiele von Predelli kein Problem für eine Interpretationstheorie darstellen, wenn man die Notwendigkeit holistischer Betrachtungen anerkennt, denn für die Interpretation einer Äußerung ist das Haben einer Reihe von Überzeugungen notwendig. Ein Interpret, der die Äußerung auf dem Zettel findet, wird auch der Überzeugung sein, dass es sich um eine geschriebene Mitteilung handelt, dass die Mitteilung vielleicht an einem anderen Ort geschrieben wurde und der Interpret wird auch wissen, wie er eventuelle Zweifel bezüglich der Nachricht beseitigt, etc. Die Liste an Überzeugungen lässt sich zwar nicht endgültig festlegen, aber wenn ein Interpret nicht über ein Netz von Überzeugungen verfügen würde, wäre die Interpretation *prinzipiell* unmöglich. Wird dieser Umstand bei der Formulierung einer Theorie miteinbezogen, bereiten Beispiele wie das von Predelli skizzierte keine Probleme mehr.

Weist man diese Kritiken am „traditionellen“ Ansatz der wahrheitskonditionalen Semantik zurück, bleibt trotzdem noch die Frage offen, wie die Besonderheiten von Indexwörtern mit einer wahrheitskonditionalen Semantik in Einklang gebracht werden. Zunächst möchte ich festhalten, dass Theorien, die die Erklärung des Gebrauchs von Indexwörtern auf den Akt des Zeigens zu reduzieren versuchen, keine befriedigende Erklärung bieten können, denn eine Theorie des Verstehens kann die Frage des Bezugs von Indexwörtern nicht unabhängig von der Bedeutung der ganzen Äußerung klären. Dennoch wurde eine Reihe von Versuchen dieser Art unternommen. Versucht wurde etwa, bestimmte Indexwörter durch andere zu ersetzen, etwa „heute“ mit „der Tag, der jetzt ist“. Versucht wurde auch, den Bezug von Indexwörtern als kausale Verbindung zwischen Sprecher und Gegenstand aufzufassen oder auf eine Geste des Zeigens zu reduzieren. Diesen Versuchen ist gemein, den Bezug von Zeichen auf Gegenstände durch den Akt der Deixis zu erklären. Aber weder die Geste des Zeigens noch das Aussprechen eines Wortes wie „dies“ oder „jener“ sind ausreichend, um etwas im Unterschied zu etwas anderem zu *meinen.* Denn man braucht sich nur vorzustellen, dass dem Interpreten der Akt des Zeigens nicht verständlich ist oder der Sprecher nur einen Teil des Gegenstandes meint, um zu sehen, dass eine Verwendung eines Indexwortes allein keine Kriterien für die Referenz zu Verfügung stellt. Würde dennoch versucht werden, wäre das, wonach gefragt werden würde, unsere grundsätzliche Möglichkeit sein, sich mit Zeichen auf Gegenstände der externen Welt zu beziehen. Ein solcher Wandel einer *semantischen* zu einer *erkenntnistheoretischen* Frage hätte aber die Konsequenz, die Wende zur Sprache zurückzunehmen, die eine Grundvoraussetzung der Akzeptanz der Disziplin „Semantik“ überhaupt ist. Anerkennt man hingegen, dass Semantik durch Anliegen gekennzeichnet ist, die mittels Reflexion auf *rein* sprachliche Phänomene erfolgt werden können, dann verbietet sich die Frage nach der Bedeutung von „dies“ als Frage nach der Möglichkeit von Bezug überhaupt. Diese Zurückweisung bedeutet auch, Fragen wie „Was bedeutet ‚dies'?“ unabhängig vom Gebrauch im Gesamtzusammenhang als Frage nach der Teilbedeutung eines isolierten Satzes zurückzuweisen. Da die Funktion von „dies“ scheinbar die ist, eine Worterklärung einzuführen, die sich gerade nicht auf das Mittel, dies zu tun, bezieht – mit „Dies ist weiß“ soll eben das angegeben werden, was weiß ist, aber nicht dasjenige, was die Bedeutung von „dies“ ausmacht – so zeigen diese Überlegungen auch, dass all unserem Spreche ein deiktisches Element zukommt, welches nicht in Analogie zur Bedeutung von Indexwörtern erklärt werden kann. Fragen, wie mit Indexwörtern Gegenstände bezeichnet werden, wie durch Indexwörter etwas als Gegenstand der Rede herausgegriffen wird, was die Bedingungen der Verwendung von Indexwörtern sind, sind folglich einzuordnen in die allgemeine Frage nach dem Satzverständnis und der Referenz von Ausdrücken auf Gegenstände. Ein Versuch einer reduktionistische Erklärung der Bezugnahme auf Gegenstände ist nicht zielführend und hätte, wie oben ausgeführt, Konsequenzen für die Disziplin der Semantik:

> „If the name of ‚Kilimanjaro' refers to Kilimanjaro, then no doubt there is *some* relation between English (or Swahili) speakers, the word, and the mountain. But it is inconceivable that one should be able to explain this relation without first explaining the role of the word in sentences; and if this is so, there is no chance of explaining reference directly in non-linguistic terms.” (Davidson 1977)

Die Frage nach der Funktion von Indexwörtern ist daher *zweitrangig* zu behandeln. Indexwörter stellen gerade deswegen kein Problem für eine wahrheitskonditionale Semantik dar, weil eine wahrheitskonditionale Semantik vielmehr den Gebrauch von Indexwörtern zur Voraussetzung hat.

Der Umstand, dass eine Äußerung unter bestimmten Umständen wahr ist und unter anderen falsch, stellt für die wahrheitskonditionale Semantik keine Probleme dar, wenn die Wahrheitsbedingungen auf die Person, die Zeit und den Ort der Äußerung relativiert werden, was aber nicht die Relativierung des Wahrheitsprädikats einschließt, und grundlegende Überlegungen zur Formulierung einer Interpretationstheorie nicht ignoriert werden. Fragt man nun dennoch nach der

Bedeutung von Indexwörter, so kann nur noch auf Verwendungsweisen von Indexwörtern (in Äußerungen) verwiesen werden.

Literatur

Corazza, Eros 2004 „Kinds of Context: A Wittgensteinian Appraoch to Proper Names and Indexicals“, *Philosophical Investigations* 27:2 April 2004, 158-188.

Davidson, Donald 1967 „Truth and Meaning“, in Donald Davidson, *Inquiries into Truth and Interpretation*, Oxford: Clarendon Press, 2001, 17-54.

Davidson, Donald 1967 „Reality Without Reference“, in Donald Davidson, *Inquiries into Truth and Interpretation*, Oxford: Oxford University Press, 2001, 215-225.

Perry, John 1997 „Indexicals and demonstratives“, in Bob Hale and Crispin Wright (eds.), *A Companion to the Philosophy of Language*, Oxford: Blackwell, 1997, 586-612.

Predelli, Stefano 2005 *Contexts. Meaning, Truth, and the Use of Language*, Oxford: Oxford University Press.

Recanati, François 2006 „Crazy Minimalism“, *Mind & Language* 21, No. 1, 21-30.

Two Reductions of ‘rule’

Dana Riesenfeld, Tel Aviv, Israel

1. First reduction: The reduction of rules to conventions[1]

Kripke (1982) sees Wittgenstein’s discussion on rule following as presenting a paradox which brings about a skeptical conclusion: “..there is no fact about me that distinguishes my meaning a definite function by ‘+’ …and my meaning nothing at all” (1982: 21). But as this “skeptical [*sic.*][2] conclusion is insane and intolerable” (1982: 60), Kripke attempts to solve the it by offering a skeptical conclusion that involves an appeal to community, and which has later been entitled ‘the community view’. The outcome of the paradox is that words do not have pre-existing meanings, however, he argues, they do gain their meaning by the rule’s tendency to match in their usage within the linguistic community. Meanings are not derivative of their conforming to rules but rather derive from their compliance with a social consensus regarding their proper or correct use: “the community must be able to judge whether an individual is indeed following a given rule in particular applications, i.e. whether his responses agree with our own” (1982: 109).

The first chapter of Kripke’s book (“The Wittgensteinian paradox” pp. 22-37) consists in the attempts of the interlocutor to oppose the skeptic’s doubts concerning the *meaning* of ‘+’, the addition sign. The skeptic casts doubt on whether anything at all (not exclusively a rule) can establish meaning. Having been persuaded by the skeptic’s arguments that there is no fact about me which ensures my meaning ‘plus’ rather than ‘quus’, Kripke examines a few plausible answers to the skeptic, devoting most of his efforts to reject the dispositional account as an inadequate account of rule following.

Dispositionalism in this context is the idea that “To mean addition by ‘+’ is to be disposed to, when asked for any sum ‘x+y’, to give the sum of x and y as the answer … to mean quus is to be disposed, when queried about any arguments, to respond with their *quum*” (1982: 22-23). For Kripke it is very important to distinguish between his position and the dispositionalist account. Kripke raises a few difficulties for the dispositionalist account, the most important of which is the claim that it aims to shed light on the normative practice of rule following in descriptive terms.

Dispositions tell us how we *will* answer not how we *ought* to answer. This point is stressed by Kripke throughout his attempts to dismiss the dispositional account as an adequate candidate to answer the skeptic:

> Suppose I do mean addition by ‘+’. What is the relation of this supposition to the question how I will respond to the problem ’68 + 57’? The dispositionalist gives a *descriptive* account of this relation: if ‘+’ meant addition, then I *will* answer ‘125’. But this is not the proper account of the relation which is *normative*, not descriptive. The point is *not* that, if I meant addition by ‘+’, I will answer ‘125’ but that, if I intend to accord with my past meaning of ‘+’ I *should* answer ‘125’… The relation of meaning and intention to future action is *normative*, not *descriptive* (1982: 37).

Kripke’s discussion of the dispositionalist account and his critique of it make it clear that any adequate explanation of ‘rules’ must retain the normative aspect of rules. But does Kripke’s own account give an explanation of rule following in normative terms? In what follows, I claim it does not.

Kripke’s skeptical solution is achieved by a ‘widening of the gaze’, “…from consideration of the rule follower alone and allow ourselves to consider him as interacting with a wider community” (1982: 89). So in answer to the question ‘what are rules?’ Kripke replies that they are in themselves and by themselves nothing at all. Only when thought about against a background of the community of speakers within which they operate they have meaning. The concept of rule means nothing when abstracted from the use of rules by a community, and does not have any objective meaning outside of how rules operate in actual speech. Rules of meaning are thus rules of use, and correct, justified and guided use of rules is something that is in principle dependent upon the pronouncement of the community of speakers.

This idea is presented by the emphasis Kripke puts on the procedure he calls public checkability: “Wittgenstein’s sceptical solution to his problem depends on agreement, and on checkability – on one person’s ability to test whether another uses a term as he does” (1982: 99). Public checkability, or agreement, now sets the standard of correctness and replaces the appeal to the rule as providing such standard.

‘Agreement’ for Kripke is not an objective fact but it is a brute fact. It is not objective in a Platonic sense, for agreement can change and shift and what we agree on today may not be what we will agree upon tomorrow. It is in this sense that agreement lacks the necessitation of e.g., a law of nature. However, agreement is a brute fact, a descriptive notion, capturing the conventions the linguistic community abides by at a certain point in time. In fact, Kripke claims, we do agree (e.g., on the meaning of the addition sign) generally. But what does it mean to generally agree? Does it mean that most of us agree most of the time? That most of us agree part of the time? Perhaps that part of us agree all of the time? In other words, how do the concepts of agreement and public checkability differentiate between correct and incorrect rule following, and between those and not following any rule at all?

According to Kripke’s solution, when asked to justify an action (e.g., giving the answer 125 to the addition problem: 68+57) one should answer: “I gave this answer because everyone else does”. Following a rule is found to be no more (and no less) than acting in conformity to others. The maxim “act like everyone else does” replaces the appeal to rules when one searches for justification. An important merit of this solution is that it maintains one of the most basic intuitions about rules: their fallibility. Rules

[1] Throughout, I refer to Kripke’s views, ignoring the question of whether the view Kripke is assigning to Wittgenstein is really his, whether it is Kripke’s position or whether it is a hybrid view sometimes assigned in the literature to ‘Kripkenstein’. Kripke himself is unclear about this issue, (1982: 5).

[2] Throughout Kripke’s text ‘sceptic’ and ‘scepticism’ are spelled with a ‘C’. When quoting, I left the spelling as it appears in the original text.

of meaning lack necessitation, logical or physical (in this respect they are different from logical rules and physical laws) in the sense that we can perfectly well imagine that things might have been otherwise. This is what I call the conventionality of rules, the fact that rules are arbitrary. Another merit of Kripke’s solution is that by preserving the conventionality of rules it also preserves the descriptive aspect of rules. The way people speak is all the data we need to consider in order to know whether they are rule followers. However, by putting emphasis on the conventionality of rules, Kripke’s solution neglects the normativity of rules in that it seems to suggest that the way people talk is all we need to know in order to know how they ought to talk. Kripke’s community view suggest that we read off correctness in our following the rules from the actions of the majority. It is in this sense close to a naturalistic fallacy: the futile attempt to derive *ought* from *is*. Kripke’s suggests that agreement on what it is to follow a rule constitutes rule following. I claim that this solution, like the dispositional account rejected by Kripke, cannot account for the normativity of rules.

Kripke’s solution reduces rule following to the super-rule: “act/respond as everyone else does”. Generally, he claims, there is agreement within a society regarding what it is to follow a certain rule. In this case, no one in the society will be likely to correct it, and no one outside it has a right to correct it. This is the strength of this position but also its weakness; it does not allow for a case wherein the community agrees upon a certain usage but is at the same time wrong in so agreeing. In the realm of ethics this situation is quite plausible; the fact that almost everyone thinks a war is justified does not make it so. This is an ethical problem, and of course claiming that justification depends solely on agreement in actions or opinions would be wrong. In the realm of language the problem is not ethical but conceptual. The problem with Kripke’s solution is that it assumes that there is a way of explaining the normativity of rules merely by taking into account their conventional application.

2. Second reduction: The reduction of rules to norms

The question concerning the (im)possibility of solitary rule following is an important test case for the community view. Kripke’s position is that if “…we think of Crusoe as following rules, we are taking him into our community and applying our criteria for rule following to him” (1982: 110). Kripke differentiates between an individual being physically isolated and his being considered in isolation (1982: 110). So Robinson Crusoe, in spite of his physical isolation can be said to be a rule follower as long as he is not *considered in isolation*. In my opinion, the distinction is an elegant vent for what could have posed a problem for the community view: on the one hand, the community view by definition is incompatible with solitary rule following, on the other, some of Crusoe’s actions indeed seem fit to be described as rule following activities. The distinction between physical and epistemological isolation explains that the community view bans the later but not the former.

Baker and Hacker, on the other hand, consider Kripke’s distinction ‘a muddle’ (1984: 39) that conceals the real question at hand: had, or had not Crusoe been following rules? The distinction blurs out the fact that both the physically isolated person and the one considered in isolation can manifest a rule-following behavior. The fact that this is so, they claim, is shown by Crusoe’s *regulative practice*; he uses the rule as “a canon or norm of correctness” (1984: 39) and is able to correct his own mistakes when and if they occur. Thus rule following for Baker and Hacker is performing an activity which is regular and which can be corrected by the rule follower. To them, then, the question whether anyone else observing or considering the isolated individual can or cannot detect the rule following behavior is distinct from the question whether the individual has in fact been

Baker and Hacker’s critique of Kripke is encapsulated in their different interpretation of the term ‘private’. While Kripke interprets ‘private’ as the opposite of ‘social’ (hence the distinction between physical solitude and epistemic solitude and the *prima facie* impossibility of solitary rule following), for Baker and Hacker ‘private’ means the opposite of ‘public’. Consequently, rule-following is essentially public in nature *but not necessarily social* (see also 1985: 161-165). That is why the question of whether the outside observer may or may not detect the rule-following taking place, will find it hard (perhaps even impossible!) or easy to learn the rules he is observing becomes secondary. What is crucial is that the practice is essentially public. Baker and Hacker, then, defend a position opting for solitary-public rule following.[3] Following rules, they emphasize, “…is not a matter of collective dispositions, but of a normative practice, which may be collective, but need not be” (1984: 74). Kripke, they claim, is deriving the answer to the normative question ‘how ought we to follow rules?’ from the empirical question ‘how do most people follow rules?’ because ‘rule’ is a normative concept, it cannot be claimed that the following of rules can be the object of observation. It is impossible to derive the norms of a given society by looking at the society’s conduct. Without stipulation of a rule, actions are but ‘empty vehicles’ of what would have been considered a norm, had we a rule. By considering the rule as distinct from its application Kripke’s analysis not only ignores the internal relation between the rule and its application, but also creates a situation wherein *agreement* becomes an internal property of the rule. This is nonsense, Baker and Hacker claim, for then the test of the rule’s correct application becomes correspondence with community agreement, the action of the majority.

However, this position, I argue, is highly problematic by their own assumptions. The alternative position presented by Baker and Hacker, I claim, only succeeds in solving those problems by creating others, no less resounding, namely that of reducing rules to norms. Such reduction, i.e., a normative explanation of rule is at the same time devoid of conventionality, of the capacity to describe how in fact a linguistic community follows its rules. Although following rules is a practice, it is wrong, they say, to think of it as essentially a *social practice* (1985:164). Wittgenstein’s conception of practice demands it to be *shareable,* not *shared*. For a practice to be considered rule following “it must be possible to teach a technique of applying rules to others, by grasping the criteria of correctness, to determine whether a given act is a correct application of the rule” (ibid). Baker and Hacker characterize the practice of rule following as essentially possible but not as essentially occurring, a rule following practice which is possibly learnable, teachable correctable and regular. None of these conditions are sufficient by themselves. But a rule following practice is one which, when meeting these conditions, could possibly occur. Baker and Hacker characterize a rule following practice as

[3] Baker and Hacker claim that this is Wittgenstein’s position in unpublished manuscripts (see, e.g., 1985: 172).

one that *could take place, but does not necessarily take place*. Their criticism of the community view alternative compels them not only to defend solitary rule following (that allegedly eliminates the role of society in rule following practice) but also opens the way to the possibility that *no one in the society follows its own rules.*

Rule following is essentially a normative practice, one “*which may be collective, but need not be*” (1984: 74, my italics). This means that it is possible that the normativity of rule following may not prevail, i.e., that not everyone, indeed, that no one, will actually abide by the normative rules. Baker and Hacker’s suggestion thus neglects the descriptive aspect of rule following, making it an inappropriate candidate for a picture of the role of rules in language. This creates a tension within their own position. For on the one hand Baker and Hacker demand that rule following be a learnable, teachable, correctable practice, but on the other, it may happen, that it is not actually learned, taught, corrected or practiced by anyone. Nothing in their suggestion, I claim, prevents such a possibility.

Baker and Hacker’s suggestion allows therefore for a hypothetical situation in which no one in the linguistic community follows it’s rules. I would like to claim at this point, that even if we assign to them the weaker position that some of the people, but not necessarily all or most, follow the normative rules, my argument against them remains valid. If indeed Baker and Hacker’s argument amounts ‘merely’ to allowing the possibility of partial rule following, the consequence (like those of the stronger version) is a disregard of the empirical fact of the percentage of the population that do follow the rules, thus rendering it irrelevant for the purpose of this discussion. This consequence, I wish to claim, is intolerable. I agree with Baker and Hacker that a reduction of rule following to consensus is wrong in that it cannot account for the normative aspect of rules, however, their position ignores the descriptive, empirical, conventional aspect of rules. For where there are normative rules, actual rule following, the actual way people use their language becomes, in Baker and Hacker’s picture quite immaterial. They accuse Kripke for divorcing the rule from its applications, whereas they are guilty of separating between the normative notion of how we *ought* to follow rules from the conventional way in which we speak *in fact*. We are in a place where, at least possibly, perhaps actually, the picture of rules and rule following presented by Baker and Hacker is at best contingently related to the way people act linguistically, speak their language.

3. Conclusion

Both Kripke and Baker and Hacker agree that a full account of ‘rule’ must accommodate for the rule’s normative dimension as well as its descriptive, conventional one. In this paper I have attempted to show that both the reduction of rules to conventions as well as the reduction of rules to norms does not succeed in providing the sought-after account of the concept. In so doing, I hope to have shed light on two concepts; conceptual reduction and rule.

Literature

Baker, G.P. and Haker, P.M.S. 1984. *Skepticism, Rules and Language.* Oxford: Blackwell.

Baker, G.P. and Hacker, P.M.S. 1985. *Wittgenstein: Rules, Grammar, and Necessity - Volume 2 of an Analytical Commentary on the Philosophical Investigations*. Oxford, UK and Cambridge, Mass.: Blackwell.

Kripke, S.A. 1982. *Wittgenstein on Rules and Private Language*. Oxford: Blackwell.

Wittgenstein, L. 1953. *Philosophical Investigations*. Oxford: Blackwell.

Wittgenstein, L. 1956. *Remarks on the Foundations of Mathematics*. Oxford: Blackwell.

Scientific Pragmatic Abstractions

Christian Sachse, Lausanne, Switzerland

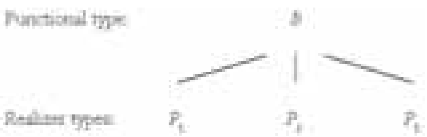
I. Starting point

As Kim argues (Kim 1998, Kim 2005), the causally efficacious property tokens considered by the special sciences are identical with tokens of physical property configurations. Thus, ontologically speaking, biological property tokens are identical with configurations of physical properties. Taking for granted token identity in what follows, one may wonder what the relationship between biological and physical property *types* is like.

Fodor and Putnam developed a famous argument in the late sixties and early seventies that hinged on what they called the possibility of 'multiple realization of property types' in order to exclude a bi-conditional connection or identity between property types of the special sciences, such as biological property types, and physical property types (Fodor 1974 and Putnam 1967/1975). In contrast to physical property types, they argued, it is possible that one and the same biological functional property type may be realized by configuration tokens coming under different physical types. In the case of biology, the possibility of multiple realization is ultimately based on natural selection, in accordance with the Paul Ehrlich's dictum – summarizing the work of biologist W.D. Hamilton - that "Selection operates when carriers of some genes out-reproduce carriers of other genes." (Ehrlich 2000, p. 38). In other words, the evolutionary salience of phenotypic effects of genes is defined by it contribution to the fitness of the organism in question in a given environment insofar as this has a positive effect on their proliferation. This is the essential point of biological evolution by natural selection – even though it is of course quite more complicated than illustrated here.

The following analysis presents a way of accounting for this evolutionary context within the functional definition of biological property types, which is a first step in sorting through the problems facing a reductionist theory that wants to grant biology scientific standing, yet ultimately seeks to defend the principle of ontological reduction.

A biological property is a functional property that is characterized in terms of fitness contribution or contribution to reproduction (for more details of the debate, cf. Weber 2005, especially pp. 38-41; for the argument to consider biological properties always in the light of evolution, cf. Dobzhansky 1973). Using this working definition, we can understand multiple realization as follows: let us say, for instance, that there is a functionally defined gene type (*B*) that is realized by different physical configurations (of type P_1, P_2, P_3, etc.). This multiple realization is possible since it is the phenotypic effect of the genes that characterises the gene type in question, whereas the different possible ways in which this phenotypic effect is physically produced – such that there are different physical types – is generally not important:



Based on the possibility of multiple realization, theory reduction of biology is by and large supposed to fail since such a reductive approach to the special sciences is generally taken to require nomological bi-conditional connections (Endicott 1998, section 8). Therefore, the special sciences such as biology are generally taken to be *scientifically indispensable* in providing explanations of certain parts of the world – namely, those having to do with living systems.

II. The dilemma of a non-reductionist framework for biology

The multiple realization argument poses a fundamental challenge to the anti-reductionist position: if one takes the MR argument to be an ontological one, it leads to an epiphenomenalism as regards the properties of the special sciences. Alternately, one may take the multiple realization argument to operate purely on the epistemological level as an argument against theory reduction. However, this, too, is not satisfying.

Taking multiple realization as an ontological argument, it gives us the following asymmetry: on the one hand, we have tokens of one and the same functional biological property type, *B*. On the other hand, the possible realizer tokens of *B* may be of different physical types. Thus, *B* is not identical with any of these physical realizer types. From this it follows that there is also an ontological difference between each token of *B* and the respective physical realizer token because *B* is taken to be something ontological (for a contrary position see MacDonald & MacDonald 1986). However, to claim that there is a causal power of the property tokens of *B over and above* the causal power of its respective physical realizer tokens contradicts ontological reductionism. If, then, we insist on an ontological difference between property tokens of *B* and their physical realizer tokens, we must conclude that this ontological difference is causally impotent. At this point, the whole scientific status of law-like generalizations comes into question, insofar as they are couched in terms of concepts referring to *epiphenomena*.

The other approach to the problem is to take multiple realization merely as an epistemological issue consisting in multiple *reference*. On the one hand, there are property tokens that are *described* by the same functional concept *B* (capital letters will be taken as concepts in what follows). On the other hand, these property tokens are differently described in terms of physics (P_1, P_2, P_3, etc.).

Let us keep in mind that, ontologically speaking, the similarities homogenously brought out by the functional concept *B* are nothing that physics can't explain, since every token coming under *B* is identical with something physical and can be, because of the completeness of physics, described and explained in physical terms (Cf. Chalmers 1996, pp. 44). In considering a single property token, physics always provides more detailed causal explanations than biology does. However, abstracting from physical differences, only the functional concept *B* seizes salient similarities among the entities in question. Biology may thus provide explanations in an unificationist manner physics is not able to make, since physics does not dispose of the conceptual means to carry out such abstractions (cf. Kitcher 1981).

Yet this unification by abstraction from physical details remains opaque so long as we lack a systematic link to physics. If we adhere to ontological reductionism and the completeness of physics, everything causally efficacious can be considered in terms of physics; therefore, the inability to generate a *systematic link* between biological concepts (law-like generalizations) and physics is a major epistemological blow to biology. What this means is that biological concepts are fundamentally unintelligible from the physics standpoint. This, in turn, casts doubt upon the scientific credibility of biological concepts. In other words, even if we cast our problem in epistemological terms, in the end, we can't coherently construct a "soft" autonomy for biology without introducing conceptual incoherence into ontological reductionism.

Given this way of stating the problem, it is obvious that, in order to save our ontologically reductionist program, we are going to have to find a conceptual schema that allows for making systematic links between biology and physics. The two main reductionist approaches do just this, but, as I will show, they are both vulnerable to criticism. The two main approaches are: a, that suggested by Lewis and Kim, entailing the construction of concepts that are semi-physical-semi-functional ones, coextensive with physical concepts (see Lewis 1980, Kim 1998, 93-95); and, b., that suggested by Bickle, the construction of physical theories that are partly coextensive with the special science theory in question (Bickle 1998). For instance, one may construct a gene concept that includes physical criteria in order to be coextensive with the physical concept in question or one may construct a physical genetic theory that refers to all and only the entities described by genetics within a certain species. Evidently, the two approaches contain enough overlaps to be combined.

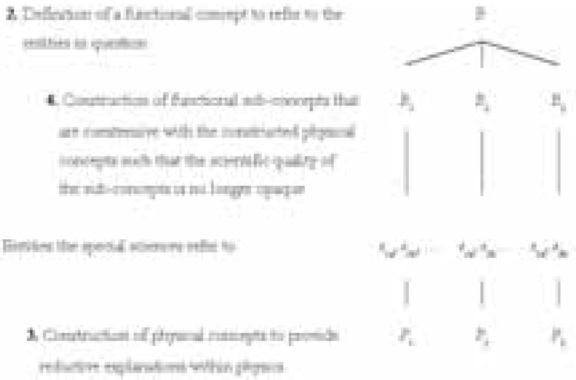
Let us take for granted for the purposes of argument that one can ascribe a scientific quality to the semi-physical-semi-functional concepts (something not trivial). Does this get us from the abstract concept *B* to terms of physics (P_1, P_2, P_3, etc.)? Since biology only works with functional concepts, but not with concepts specified by physical criteria, it is puzzling how Kim's semi-physical-semi-functional concepts could serve as bridge principles, since it only seems to repeat the problem in other terms. This is why it remains unintelligible, from the biological point of view, how the salient similarities brought out by *B* can be brought out by the semi-physical-semi-functional concepts without this resulting in a conflict with ontological reductionism and the completeness of physics. Kim does not give us a mechanism whereby it is possible to abstract from the physical part of the semi-physical-semi-functional concepts, and hence it remains unclear how *B* can causally explain something.

Against this background, we might want to make the radical move of replacing biology *tout court* with physical Ersatz theories. This is in fact Bickle's solution: the construction of physical theories that are, taking together, co-extensional with the biological theory in question. In other terms, one constructs several physical theories with applicable physical concepts (applicable in the sense that they cover target objects) that are co-extensional with biological concepts (which target the same objects). Since bridge-principles are still missing, this approach as well does not make intelligible how abstract biological concepts and law-like generalizations could be vindicated. This approach by Bickle (and Hooker) is more general than the approach of Kim (and Lewis), but it also ends up in suggesting the elimination of biology.

III. Reductionist framework without elimination

Because of multiple reference, the starting point is that tokens of physical property configurations that come under the functionally defined biological concept *B* may be described by different physical concepts (P_1, P_2, P_3, etc.). This implies that there is a causal difference among the physical configuration tokens coming under a single functional concept *B* (Kim 1999). In other words, there are different ways to bring about the effects on which the functional concept *B* focuses (Esfeld & Sachse 2007), as, for instance, in the difference created by a phylogenetic effect that elevates the rate of the reproduction of one gene over another. From the physical point of view, there is thus a difference in the production of side effects that are systematically linked with the main effects (characterizing *B*) in question.

The differences resulting from side effects can be detected from the biological point of view in a given physical environment, thus giving it standing as a scientific fact. This can be illustrated by, for instance, the empirical data with which genetics deals, which is often cited as a classical case of multiple realization (reference). It can be shown that differences between DNA sequences that come under a single gene concept (multiple reference) are linked to different molecular ways to bring about the effect on which the gene concept in question focuses (causal implication of compositional differences). These different ways to produce characteristic effects in question are systematically linked with side effects such as the speed or the accuracy of the protein production (see Bulmer 1991) that can be salient for selection. To put it in other terms, it is possible to construct purely *functionally* defined biological concepts that are nonetheless *coextensive* with the physical concepts. This means, for any concept *B*, it is possible to construct functional sub-concepts B_1, B_2, B_3, etc. coextensive with the physical concepts P_1, P_2, P_3, etc. Since any physical difference accounting for multiple reference leads to side effects that can be in principle detected from the biological perspective, there thus is a *nomological* coextensionality (Sachse 2007, 138-152). The following figure may help to illustrate the most important steps in this argument.


2. Definition of a functional concept to refer to the entities in question
B
4. Construction of functional sub-concepts that are coextensive with the constructed physical concepts such that the scientific quality of the sub-concepts is no longer opaque
B_1
B_2
B_3
1. Entities the special sciences refer to
3. Construction of physical concepts to provide reductive explanations within physics
P_1
P_2
P_3

Note that the construction of such functional sub-concepts is first and foremost an intermediate step in order to establish bridge-principles. The important thing here is that we can show that the biological concepts have a non-opaque scientific status in that the sub-concepts are coextensive with physical concepts, even though all of them may not be of any particular biological interest. Let me thus call this their *possible* scientific status. By this means, we can bootstrap upwards to establish the scientific status of the more abstract concept *B*. To put it in other terms, since any token coming under the abstract biological functional concept *B* also comes under a functional sub-concept whose scientific quality would prima facie not be opaque, *B* cannot be opaque either in as much as the only difference between *B* and one of its sub-concepts (B_1, B_2, B_3, etc.) is the degree of abstraction *within a purely functional* theory. A sub-concept brings out the same salient similarity as does its more abstract concept (its relevance here being defined in the context of selection under normal conditions) while also adding a functional detail (side effect that is salient for selection under special conditions) that is linked to this outlined salient similarity (B_1 = "B + B_{minor}"). Since the matter is so crucial, let me stress here that both the abstract concept and its sub-concepts are constructed in terms of *one* single theory, such that the abstraction from side effects is a purely theory-immanent matter with a conceptual linkage. Thus, under this schema, we clarify the assumed scientificity of the abstract unifying concepts of biology (for instance, a certain gene concept that accommodates the fact that the gene tokens are physically different), as we cannot do in the other reductionist approaches. There are now bridge-principles sufficient to make the abstraction step intelligible.

This philosophical foundation will help to normalize the undoubted pragmatic advantage of biology as a special science within a unified conceptual schema that retains the completeness of physics and ontological reductionism. Biology is scientific because of the systematic link to physics, and objective because the outlined biological salient similarities are those that exist in our world as they depend on biological evolution by means of natural selection. Its abstract functional concepts, integrated within the proposed reductionist framework of constructing functional sub-concepts, counter the twin threats of epiphenomenalism and eliminativism. Abstract biological concepts can be systematically linked with physics. This does not ratify the claim of the indispensable character of biology, since that does not seem to be compatible with the completeness of physics and ontological reductionism, but it does give us pragmatic wiggle room – one can now argue that the pragmatic value of biology is scientific and objective. Biological concepts and the abstract law-like generalizations governing them bring out salient similarities among entities that are physically different. This is the epistemological power belonging to biology alone: its ability to explain biological evolution in *homogeneous* terms that can't be selected from a wholly physics-based point of view. Hence, there is no positive argument left for the eliminativist approach to biology. Which gives us what we want: biology is the more unifying theory about a certain ensemble of entities (the living beings) while physics is the more unifying theory in general but not as concerns the living beings.

To sum up and conclude: there is a strong causal argument in favour of ontological reduction. Based on this argument and the completeness of physics, the standard anti-reductionist argument of multiple realization faces the dilemma that it apparently leads to either epiphenomenalism or eliminativism with regard to biology, that is, in respective to its status as a science. In order to avoid these consequences, we show that a systematic link between biology (and other special sciences) and physics is philosophically and empirically possible by means of the construction of functional sub-concepts that are coextensive with (in the last resort) constructed physical concepts. Based on this systematic link to physics, the scientific quality of biology and its abstract concepts is no longer opaque. This should not be taken as a warrant to regard biology as indispensable, given the principles of the completeness of physics and ontological reductionism, but it does show that, within our proposed reductionist framework, biology accrues standing as a objective, pragmatic science, which conceptualizes parts of the world (living systems) with abstract unificationary concepts that have no equivalent in physics.

Literature

Bickle, John 1998 *Psychoneural reduction: the new wave*. Cambridge (Massachusetts): MIT Press.

Bulmer, Michael 1991 "The selection-mutation-drift theory of synonymous codon usage". *Genetics* 129, pp. 897-907.

Chalmers, David J. 1996 *The conscious mind. In search of a fundamental theory.* Oxford: Oxford University Press.

Dobzhansky, Thedosius 1973 "Nothing in biology make sense except in the light of evolution". *American Biology Teacher, 35: pp. 125-129.*

Ehrlich, Paul 2000 Human Natures: Genes, Cultures, and the Human Prospect. Washington DC: Island Press.

Endicott, Ronald P. 1998 "Collapse of the new wave". *Journal of Philosophy* 95, pp. 53-72.

Esfeld, Michael & Sachse, Christian 2007 "Theory reduction by means of functional sub-types". *International Studies in the Philosophy of Science* 21, pp. 1-17.

Fodor, Jerry A. 1974 "Special sciences (or: the disunity science as a working hypothesis)". *Synthese*, 28, pp. 97-115.

Hooker, Clifford A. 2004 "Asymptotics, reduction and emergence". *British Journal for the Philosophy of Science* 55, pp. 435-479.

Kim, Jaegwon 1998 *Mind in a physical world. An essay on the mind-body problem and mental causation*. Cambridge (Massachusetts): MIT Press.

Kim, Jaegwon 1999 "Making sense of emergence". Philosophical Studies 95, pp. 3-36.

Kim, Jaegwon 2005 *Physicalism, or something near enough*. Princeton: Princeton University Press.

Kitcher, Philip 1981 "Explanatory unification". *Philosophy of Science* 48, pp. 507-531.

Lewis, David 1980 "Mad pain and Martian pain". In: N. Block (ed.): *Readings in the philosophy of psychology. Volume 1*. London: Methuen. Pp. 216-222.

MacDonald, Cynthia, MacDonald, Graham 1986 "Mental causes and explanation of action". *Philosohical Quarterly* 36, pp. 145-58.

Putnam, Hilary 1967 / 1975 "The nature of mental states". In: H. Putnam (ed.) 1975 *Mind, language and reality. Philosophical papers. Volume 2*. Cambridge: Cambridge University Press. Pp. 429-440. First published as "Psychological predicates" in W. H. Capitan and D. D. Merrill (eds.) 1967 *Art, mind and religion*. Pittsburgh: University of Pittsburgh Press.

Sachse, Christian 2007 *Reductionism in the philosophy of science*. Franfurt: Ontos Verlag.

Weber, Marcel 2005 *Philosophy of experimental biology*. Cambridge: Cambridge University Press.

Wittgenstein’s Attitudes

Fabien Schang, Nancy, France

1. Between language and mind: a logic of propositional attitudes ...

The statement in §5.542 concerns the logical form of peculiar propositional attitudes, viz. belief-statements:

> But it is clear that "A believes that *p*", "A thinks *p*", "A says *p*", are of the form " '*p*' says *p*": and here we have no co-ordination of a fact and an object, but a co-ordination of facts by means of a co- ordination of their objects.

This formulation sounds queer, and we will attempt to see why Wittgenstein did state it so before considering Hintikka's replies in favor of epistemic modal logic.

The core problem concerns *truth-functions theory*: is any meaning-function a truth-function? (Russell 1923) made a distinction between two sorts of occurrence for a proposition, namely: *meaning*-functions that contain propositions as a member are also *truth*-functions whenever the component proposition occurs as expressing a fact (i.e. an ontological entity); they are not so whenever the proposition occurs as a fact in its own right, given that the whole sentence then talks about the component proposition itself. It is precisely the case with propositional attitudes, where the fact in consideration is the form of words uttered by the speaker. It thus seems that not every meaning-function is truth-functional and, in this respect, Russell's position is to be compared with what Frege argued about the change of denotation in a context of indirect discourse.

Nevertheless, (Wittgenstein 1922) does not accept any other meaning-functions than the truth-functional ones: not only “Propositions are truth-functions of elementary propositions” (§5), but also “There is one and only one complete analysis of the proposition” (§3.25). If so, the preceding logical analyses as suggested by Russell and Frege cannot be accepted because they go beyond truth-function theory, the only one for Wittgenstein (“In the general propositional form, propositions occur in a proposition only as bases of the truth-operations”, §5.54). Therefore, the point is not to delimit one context of application for truth-functional propositions while ruling out some propositions of an intensional sort; rather, the point is to streamline *every* meaningful proposition within the unique pattern of truth-functions. There cannot be any exception to the theory of extensionality, from a Tractarian perspective.

For one thing, the analysis of "A believes p" excludes the subject A from its logical form while replacing it by a *mention* of the proposition within single brackets, 'p'. The result seems to be counterintuitive, reducing belief to an impersonal relation between a linguistic expression (i.e. the propositional sign) and that what it designates (i.e. the propositional fact that constitutes a thought). Why such an exclusion of the thinking subject, and how to analyze a belief while eliminating the psychological side of an attitude? (Russell 1923) did not reject it from his own analysis, given that he conceived the believer as a sequence of psychological facts expressed by means of sentences. But those beliefs were then associated with a single subject; now Wittgenstein's account definitely cancels this particular subject and talks instead about some arbitrary sentence in the form 'p'.

In order to understand such a mysterious statement as §5.542, several writers accounted for it in two steps, namely: Wittgenstein's theory of object and his subsequent distinction between an empirical and a metaphysical subject.

2. … is not a problem of mind (no psychologism!) ...

In (Russell 1923)'s account, each proposition was treated as a class of psychological facts that introduce A's mind through the analysis of propositional attitudes. The logical form of "A believes that p" thus corresponds to the correlation of a fact, i.e. the propositional fact that p, and an object, i.e. A's mind. However, any object is simple, Wittgenstein claims (“The object is simple”, §2.02), whereas A's mind is complex (as a sequence of psychological facts), so that the logical form assigned to propositional attitudes is not correct. The logical form required for any states of affairs (“An atomic fact is a combination of objects (entities, things)”, §2.01) thus leads Wittgenstein to discard propositional attitudes as states of affairs, in their current reading as a co-ordination of a fact and an object. Such a position leads him to the equally queer statement: “This shows that there is no such thing as the soul – the subject, etc. – as it is conceived in superficial psychology. A composite soul would not be a soul any longer” (§5.5421).

Isn't the price to pay for accepting Wittgenstein's logical analysis too expensive, if the rejection of propositional attitudes apparently leads one to a rejection of psychology? (Favrholdt 1964, 559) notes that this result directly follows from the Tractarian theory of objects:

> For the superficial psychologists that maintain this it would be reasonable to say that "A says p" is a co-ordination of a fact in the Wittgensteinian sense, namely a propositional sign, and an object, namely the thinking, presenting soul, which being simple is to be called an object. This view Wittgenstein is bound to reject. According to the picture theory in the *Tractatus* no co-ordination could ever be established between a fact and an object. The two entities in question have to be equally articulated in order to be co-ordinated. Objects can be co-ordinated with objects (because they are simple) and facts can be co-ordinated with facts in so far as they can be analysed into the same amounts of elements. (559)

This prevents Wittgenstein from viewing propositional attitudes in the usual way, to be found in epistemic modal logic. Hence his second argument that accounts for §5.542: the *distinction between an empirical and a metaphysical subject*.

(Hintikka 1958) puts such a distinction to avoid some misunderstanding in Wittgenstein's language theory, namely: his thesis of *solipsism*, ordinarily considered as an argument for private language. In order to clarify the following passage: “That the world is *my* world, shows itself in the fact that the limits of the language (*the* language which I understand) mean the limits of *my* world" (§5.62), (Hintikka 1958) argues that Wittgenstein's concern

> is not the empirical subject but the "metaphysical" subject discussed in philosophy. In other words, he is interested only in what can be said to be mine *necessarily*; for otherwise he would only be doing empirical psychology. But the only necessity there is, according to the other doctrines of the *Tractatus*, is the empty tautological necessity of logic. (89)

As a matter of fact, solipsism suggests the private character of our current thoughts: "*The limits of my language* mean the limits of my world" (§5.6). Now Wittgenstein doesn't support the view of a private language altogether. As a way to disentangle this wrong connection, Hintikka argues that the Tractarian "I" is not a psychological ego or single thinker. It is not the agent of Hintikka's later epistemic logic, but an abstract subject embodying the whole set of propositions: "The subject does not belong to the world but it is a limit of the world" (§5.632). Wittgenstein's picture theory of language should recall us that the limits of the world are determined as the limits of language, where the projective relation between both stands for the correspondence between a pictured fact and a picturing proposition.

Moreover, the metaphysical subject cannot talk about itself within the very language it embodies, contrary to the case of propositional attitudes: "No proposition can say anything about itself, because the propositional sign cannot be contained in itself (that is the "whole theory of types")" (§3.332). In virtue of such an impossible self-reference for the Wittgensteinian subject, believing that p is the case is thus confined to the impersonal relation between a propositional sign and a proposition: 'p' says p, meaning that the propositional sign expresses p's being the case. Consequently, solipsism means the obvious impossibility for the metaphysical subject to go beyond the limits of language, given that the latter is a precondition to the former; but solipsism does not mean the impossibility for a psychological subject to express her own thoughts. On the contrary, Wittgenstein's thoughts are as public as Frege's ones (the *Gedanken*) and his solipsism does not mean at all that thoughts are private representations (*Vorstellungen*). Nevertheless, such public thoughts are separated from the psychological subject that grasps them in the *Tractatus*, hence the resulting logical form in §5.542.

Now (Favrholdt 1964) recalls in the same time that the think*er* implicitly occurs in the relation expressed in §5.542 between 'p' and p:

> "'p' says p" says nothing more than p. It states that the propositional sign *is being thought*, and this is the same as *asserting* the proposition p. Therefore, according to Wittgenstein, in sentences as 'A says p', p is not occurring in a proposition in a special way which is in conflict with his general theory of truth-functions. (560).

3. ... but a problem of language (no metatheory!)

It will be attempted to show in the following that epistemic logic amounts to some compromise between both topics: it introduces belief into logic while presenting it as the public occurrence of a statement, or assertion. But the Tractarian view of logic excluded to do so.

Assuming that *assertion* refers to the occurrence of a belief by means of a statement, it does not add anything to propositions that serve to make it explicit and is to be located in the domain of psychological events. The very project of a "doxastic logic" is therefore absurd, in the light of the Tractarian language theory:

> The thought makes a une proposition out of the propositional sign p and this is the same as asserting p. If p is not thought it remains a propositinal sign and the expression " ├p" in this case is absurd; you cannot at the same time assert, that is to say think p, and not think p. Hence the assertion sign is logically altogether meaningless (see 4.442). (560)

As to the *rules of logic*, they specify the limits within which subjects do and can express themselves: inferences and tautologies don't say anything but embed propositional forms that subjects cannot think of, because either these forms don't depict any particular image (excluded middle) or cancel any of them (non-contradiction). The projective nature of language according to Wittgenstein makes his logical theory appear as a sort of *transcendental* frame for thinking. Such a view could be interpreted as reminiscent of Kant's transcendental logic, to be defined as an inquiry into the conditions of *a priori* possibility for experience according to the categories of understanding. Apart from the notion of understanding, Kant's criticism is found again here in the impossibility for any empirical subject to know the limits of language; empirical subjects think *within* language, and they cannot depart from it in order to contemplate outwardly what makes a distinction between logical and illogical thoughts.

Logic is thus characterized as a method of projecting true or false propositions of a language into states of affairs (*Tatsachen*) or mere situations (*Sachverhalte*), respectively; but these methods are inexpressible by themselves. Formal semantics cannot be described in the latter: the rules for applying a set of formulas into some given model, as depicted in every metalanguage from a model-theoretical semantics, couldn't be conceived in a Tractarian line. If any subject A believes in a contradiction, for example, (Favrholdt 1964) recalls that the distinction between a propositional sign and a thought makes such a belief meaningless (its projection is impossible, as pointing out no plausible direction):

> Wouldn't it be possible for A to say "p . ~p" thus violating the laws of logic? The answer is no (...) A can think "p" or he can think "~p". In the first case the first link of p . ~p will become a proposition but the last part (~p) will remain a propositional sign, because it is not thought (…) Therefore if one cannot think anything unlogical he cannot present anything in language which "contradicts logic" either. For language is not the physical facts that we call propositional signs, but these facts in their projective relation to other facts. (561-2)

Epistemic modal logic is in total agreement with this, when forbidding any two contradictory propositions to be embedded into one and the same "model". The metaproperty of *consistency* says no more than Wittgenstein did here; it does the same thing but in saying it with terms, that is, within a construed formal semantics.

4. Conclusion: metatheory as a precondition for modal logics

Universality of language and ineffability of semantics are two preconditions that Hintikka will rule out from his very view of logic, in accordance with his distinction between logic as a universal language and logic as a calculus; the same does for other modal logics than epistemic logic, given that any judgment about a proposition was made impossible by Wittgenstein. Such a Tractarian impossibility came from ineffability as a unknowable relation between language and reality. It also follows from this a crucial nexus between *symbolism* and *formalism*: language symbolizes the world, nothing else, and any formalized language should yield a genuine picture of reality.

The point with epistemic logic is that it becomes acceptable only when the preceding preconditions have been qualified. Such a qualification is allowed only within a model-theoretical framework that Wittgenstein refused for philosophical reasons, so that Suszko's initial objection pointed to the right direction while assuming uncharitably something justifiably refused by the *Tractatus*.

In a nutshell: only God can rule in logic, for Wittgenstein; but Suszko was an atheist and God is (officially) dead with Tarski, so to say.

Literature

Favrholdt, David 1964 "Tractatus 5.542", *Mind* 73, 557-562

Hintikka, Jaakko 1958 "On Wittgenstein's 'Solipsism'", *Mind* 67, 88-91

Russell, Bertrand 1923 "What is meant by 'A believes p'?", reedited in *The Collected Papers of Bertrand Russell*, Vol. 9: "Language, Mind and Matter: 1919-26", 159

Suszko, Roman 1968: "Ontology in the *Tractatus* of L. Wittgenstein", *Notre Dame Journal of Formal Logic* 9, 7-33

von Wright, Georg Henrik 1986: "La logique modale et le *Tractatus*", in *Wittgenstein*, trad. E. Rigal, TER, 195-213

Wittgenstein, Ludwig 1922: *Tractatus Logico-Philosophicus*, C.K. Ogden (trans.), London: Routledge & Kegan Paul

Warum man auf transzendentalphilosophische Argumente nicht verzichten kann

Benedikt Schick, Berlin, Deutschland

1. Lebenswelt, Reduktion, Elimination

Der Ausgangspunkt aller philosophischen Bemühungen wie auch wissenschaftlicher Forschung ist die Lebenswelt. Wir finden uns „immer schon" – wie manchmal gesagt wird – in einer natürlichen und sozialen Welt vor, haben „immer schon" Meinungen, Überzeugungen, Erwartungen, Deutungsmuster und Vorurteile von dieser Welt, von uns selbst und von unseren Mitmenschen. Dieses Meinungs-Geflecht mit seinen vielen einzelnen Komponenten und größeren und kleineren Zusammenhängen scheint uns nur teilweise und nur zeitweilig wirklich bewusst zu sein. Dass die Lebenswelt den notwendigen Ausgangspunkt aller menschlichen Tätigkeiten bietet, ist kaum zu bestreiten. Zu offensichtlich ist, dass die „immer schon" akzeptierten lebensweltlichen Annahmen in ihrer je nachdem eher theoretischen oder praktischen Ausrichtung eine notwendige Voraussetzung allen Handelns sind. Das scheint nicht nur für das alltägliche Handeln eines bestimmten Individuums zu gelten, sondern auch für das Funktionieren aller möglichen kulturellen und sozialen Institutionen und Projekte, und selbstverständlich eben auch für den Wissenschaftsbetrieb. Die Lebenswelt ist in einem bestimmten Sinn *primär*, das meint, sie ist Voraussetzung und Grundlage allen Forschens und Fragens.

Dieser Zusammenhang mag einleuchten, es drängt sich aber dann eine Frage auf: Folgt aus der unbestrittenen Vorrangigkeit der Lebenswelt auch ihre Unkritisierbarkeit? Kann das Geflecht von theoretischen und praktischen Annahmen, das von jeder wissenschaftlichen oder philosophischen Kritik vorausgesetzt wird, überhaupt von eben dieser Kritik in Frage gestellt werden? Zwingt uns die Tatsache, dass die Lebenswelt primär ist, dazu, auch alle ihre Elemente für sakrosankt zu halten?

Wer hier mit ja antworten möchte – so scheint es –, treibt die Transzendentalphilosophie entschieden zu weit. Lebensweltliche Annahmen sind manchmal begründet, manchmal sogar gut begründet, häufig aber bloß mehr oder weniger gut bewährt. Die Lebenswelt steckt voller Vorurteile, Inkohärenzen, Widersprüche und Irrtümer. Das Gesamte lebensweltlicher Annahmen ist auch keineswegs stabil, sondern sowohl im Hinblick auf einen bestimmten Menschen als auch im Hinblick auf die historische Entwicklung von Gemeinschaften sehr dynamisch, unter Umständen geradezu revolutionär. De facto werden lebensweltliche Annahmen also häufig kritisiert und korrigiert. Außerdem, was ist denn *die* Lebenswelt? Unterscheidet sich nicht die Lebenswelt eines Menschen zumindest teilweise von der eines anderen, die der einen Kultur von der einer anderen?

Werfen wir einen Blick in die Wissenschaftsgeschichte, so fehlt es nicht an Fällen, wo vermeintlich unaufgebbare lebensweltlich verwurzelte „Wahrheiten" sich als Täuschung erwiesen haben. Als Beispiel mag der Verweis auf den Wechsel vom geo- zum heliozentrischen Weltbild, der heute gern als „erste Kränkung der Menschheit" gezählt wird, genügen.

Es kann kein Zweifel bestehen, dass viele lebensweltliche Annahmen kritisierbar und korrekturbedürftig sind, und dass hierbei den Wissenschaften eine zentrale Rolle zukommt. Das Verhältnis, in dem Ergebnisse einzelwissenschaftlicher Forschung zu lebensweltlichen Annahmen stehen, muss allerdings in unterschiedlichen Fällen auch unterschiedlich beurteilt werden. Nicht in jedem Fall ist eine Elimination der lebensweltlichen Deutung nötig. Vielleicht können in manchen Fällen eine wissenschaftliche Beschreibung und eine lebensweltliche Beschreibung desselben Phänomens ohne Widerspruch nebeneinander stehen. In anderen Fällen kann es sein, dass die Wissenschaft eine Erklärung „von unten" für eine lebensweltliche Gegebenheit liefert und so eine lebensweltliche Beschreibung auf eine wissenschaftliche reduziert. Elimination und reduktive Erklärung müssen unterschieden werden. Im Gegensatz zur Elimination führt die reduktive Erklärung ein Phänomen nur auf eine zugrunde liegende Ebene zurück. Das Phänomen wird dabei erklärt, aber nicht wegerklärt.

Aufgrund der Vorrangigkeit der Lebenswelt scheint es jedenfalls nicht möglich zu sein, die Lebenswelt als Ganze in Frage zu stellen, noch die für unser Welt- und Selbstverständnis grundlegenden Voraussetzungen revisionär zu eliminieren. In eine ähnliche Richtung geht Julian Nida-Rümelin, wenn er schreibt: „In unserer Lebenswelt sind abrupte Veränderungen […] nicht zu erwarten. Sie [die Lebenswelt] sperrt sich gegenüber rationalistischen Konstruktionen, da sie zu tief mit den je individuellen und gesellschaftlichen Lebensformen verbunden ist, um *in toto* zur Disposition gestellt zu werden." (Nida-Rümelin 2005, 40). Wenn die Lebenswelt als Ganze nicht revisionär zu überholen ist, dann könnte es auch *konstitutive Elemente* der Lebenswelt geben, die sich einer Elimination oder auch einer reduktiven Erklärung entziehen.[1]

Vor diesem Hintergrund stellt sich die Frage nach den Kriterien. Wie kann man denn entscheiden, welcher Fall jeweils vorliegt. Wann gebietet es sich von Elimination zu sprechen, wann eher von reduktiver Erklärung? Anhand eines Beispiels soll im Folgenden geprüft werden, ob hierbei nicht eine Art transzendentaler Test weiterhelfen kann – ja ob ein solcher Test nicht vielleicht sogar unverzichtbar ist.

2. Ursachen und Gründe

In unserer Lebenswelt ist die Praxis des Begründens fest etabliert und bewährt. Gründe spielen nicht nur im Bereich des Handelns eine Rolle sondern ebenso im Bereich des Wissens, und selbstverständlich in der Wissenschaft. Das Gründe-Sprachspiel scheint zu denjenigen lebensweltlichen Phänomenen zu gehören, die eine nicht eliminierbare Voraussetzung eines jeden möglichen Welt- und Selbstverständnisses bilden. Auf der anderen Seite findet empirische Forschung bei ihrem Bemühen, kognitive Prozesse und menschliches Verhalten zu erklären, immer nur

[1] Auch eine reduktive wissenschaftliche Erklärung kann in manchen Fällen unmöglich sein. Der des Anti-naturalismus unverdächtige Ansgar Beckermann rechnet beispielsweise mit der Möglichkeit, dass phänomenales Erleben echt emergent gegenüber körperlichen Prozessen ist, und wir es daher hier mit einem unlösbaren „Welträtsel" zu tun haben. (Beckermann 1999, 34).

Ursachen. Neuronale Prozesse etwa sind offenbar durchgängig von Ursachen bestimmt. Ist daher die antinaturalistische Schlussfolgerung zu ziehen, dass, wenn wir an der Möglichkeit festhalten wollen, dass Menschen ihr Handeln zumindest prinzipiell an Gründen ausrichten können, dass wir dann einen ontologischen Physikalismus zurückweisen müssen?

In seinem Beitrag *Ursachen und Gründe: Zu Ihrer Unterscheidung in der Debatte um Physikalismus und Willensfreiheit* bestreitet Michael Pauen genau das (Pauen2005, 7). Die Unterscheidung von Ursachen und Gründen kann seiner Ansicht nach auch dann aufrechterhalten werden, wenn der Physikalismus wahr wäre. Gemeint ist damit, dass selbst wenn menschliches Verhalten vollständig durch physische Prozesse bestimmt wäre, es dennoch unter dem Einfluss von Gründen stehen könnte. Wie argumentiert Pauen nun für diese Ansicht? Zunächst einmal verzichtet er darauf „psycho-physische Identitätsbehauptungen“ (Ebd. 8) aufzustellen, ihn beschäftigt lediglich das Verhältnis der verschiedenen Beschreibungsebenen: der lebensweltlichen, in der Gründe vorkommen und der neurobiologischen, in der Ursachen vorkommen. Was dieses Verhältnis betrifft, so sind „Reduktionsbehauptungen“ durchaus denkbar, so Pauen. Mit Reduktion ist dabei gemeint die „erklärende Zurückführung z.B. einer aus dem Alltag bekannten Phänomenbeschreibung auf eine wissenschaftliche Theorie, die es erlaubt, das Alltagsphänomen zu erklären“ (Ebd. 8). Entscheidend ist der Zusatz: „Selbstverständlich wird die Existenz des Alltagsphänomens damit nicht angetastet.“ (Ebd. 8). Pauen scheint hier einen Mittelweg gehen zu wollen. Auf der einen Seite wehrt er sich gegen eine „abenteuerliche Identifikation von Gründen und Neuronen“ (Ebd. 10), andererseits hält er daran fest, dass Überlegungsprozesse, die sich auf Gründe stützen, neuronal realisiert sind. Als Analogie dient ihm das Verhältnis von Computern zu ihren Programmen. Nur weil alles, was sich in Computern ereignet vollständig physikalisch beschreibbar ist, heißt das nicht, dass nicht Programme die Funktion des Computers bestimmen. Die Software existiert und ist wirksam, allerdings nicht als etwas Eigenständiges sondern als etwas materiell – in der Hardware – Realisiertes.

3. Ein Argumentationsschritt fehlt

Michael Pauen möchte in dem genannten Beitrag nicht für die Wahrheit des Physikalismus argumentieren, er möchte vielmehr die grundsätzliche Vereinbarkeit des Physikalismus mit der Möglichkeit, dass Gründe wirksam sind, aufzeigen. Er bilanziert daher: „Es mag viele Einwände gegen den Physikalismus geben – die Präsumtion, dass unser Handeln von Gründen geleitet wird, gehört jedoch offenbar nicht dazu“ (Ebd. 11).

Mit Blick auf diesen Ansatz Pauens ließe sich nun manches in Frage stellen. Man könnte etwa den vorausgesetzten Kompatibilismus kritisieren (vgl. dazu die Replik auf Pauen: Nida-Rümelin 2005), und man könnte zweitens bestreiten, dass eine vollständige Reduktion der mentalen Ebene auf die neurobiologische Ebene gelingen kann (Pauen selbst äußert diesen Zweifel deutlich. Pauen 2005, 10). Stellen wir diese beiden Einwände aber einmal zurück und nehmen um des Arguments willen an, das Verhältnis der Redeweise von Gründen zu neurobiologischen Beschreibungen sei treffend beschrieben. Es ergibt sich dann – so meine ich – ein Problem mit Pauens realistischer, nicht-eliminativer Deutung von Gründen. Warum geht Pauen so selbstverständlich davon aus, dass selbst wenn die lebensweltliche Beschreibungsebene auf die neurobiologische reduziert werden kann, die lebensweltliche Beschreibung dennoch nicht ihren ontologisch verbindlichen Charakter verliert? Dazu noch einmal Pauen: „Selbstverständlich wird die Existenz des Alltagsphänomens damit [durch die reduktive Erklärung] nicht angetastet“ (Ebd. 8).

Ein Beispiel und eine theoretische Überlegung dazu sollen im Folgenden zeigen, dass das alles andere als selbstverständlich, vielmehr begründungsbedürftig ist.

4. Die Dämonen-Theorie als Beispiel

Das folgende Beispiel hat in ähnlicher Form Richard Rorty im Rahmen seiner Argumentation für den *Eliminativen Materialismus* verwendet (Rorty ^{2}1993, 96-99).

Denken wir uns eine Gesellschaft, in der Krankheiten dadurch erklärt werden, dass man sagt, die Kranken seien von Dämonen oder bösen Geistern besessen. Es gibt in dieser Gesellschaft bestimmte Arten der Behandlung, die in der Regel auch wirken, und die Wirkweise wird natürlich ebenfalls mithilfe des Dämonensprachspiels erklärt. Diese Gesellschaft kommt irgendwann in Kontakt mit der westlichen Welt, lernt die moderne Medizin kennen und ersetzt nach und nach die alte Beschreibung von Krankheiten durch eine Erklärung mithilfe von Viren und Bakterien. Ist das alte Sprachspiel damit nicht völlig obsolet geworden, und erzwingt nicht die erfolgreiche Reduktion damit auch das Zugeständnis, dass es Dämonen nicht gibt? Niemand würde doch die Behauptung stehen lassen wollen: Dämonen sind bakteriell realisiert. Die Redeweise von Viren und Bakterien erklärt nicht, wie Dämonen Krankheiten bewirken, sie ersetzt vielmehr das Dämonensprachspiel. Es gebietet sich hier von Elimination, nicht von reduktiver Erklärung zu sprechen. Offensichtlich haben wir es beim Dämonenglauben mit dem Fall einer in einer bestimmten Gesellschaft etablierten lebensweltlichen Annahme zu tun, die aber wissenschaftlich kritisiert und eliminiert werden kann.

Der entscheidende Grund dafür, dass eine realistische Deutung des Dämonensprachspiels nicht sinnvoll ist, scheint darin zu liegen, dass im Fall einer wirklich vollständigen Reduktion eines lebensweltlichen Sprachspiels auf eine zugrunde liegende wissenschaftlich beschreibbare Ebene dieses Sprachspiel theoretisch überflüssig wird, und daher auch jede ontologische Verbindlichkeit einbüßt. Die Möglichkeit, dass das überholte Sprachspiel aus pragmatischen Gründen beibehalten wird, ist damit natürlich nicht ausgeschlossen. Eine solche pragmatische Beibehaltung verpflichtet aber nicht ontologisch.

Man könnte das hier Gemeinte auch deutlich machen, indem man sich auf das „Prinzip der wissenschaftlichen Eleganz“ (Schmidt-Salomon 2007, 183)[2] beruft. Es besagt, dass man zur Erklärung eines Phänomens nicht mehr Annahmen investieren soll, als unbedingt nötig sind. Oder man greift zu Ockhams Rasiermesser und verlangt, dass man die Menge der Entitäten nicht ohne Not vermehren soll. Die Redeweise von Dämonen ist nach der erfolgreichen Etablierung wissenschaftlicher Medizin nicht notwendig für die Erklärung irgendeines Phänomens. Daher sollte man diese Redeweise aufgeben bzw. zumindest nicht mehr

[2] Michael Schmidt-Salomon wendet dieses Prinzip in der Frage nach Willensfreiheit an – allerdings wohl auf problematische Weise.

ontologisch ernst nehmen. Zu diesem Schluss kommt auch Rorty wenn er schreibt: „Der Fall der Dämonen macht jedoch klar, dass die Entdeckung einer neuen Möglichkeit, Phänomene zu erklären, die früher durch Bezugnahme auf eine bestimmte Sorte von Entitäten erklärt worden sind […], einen guten Grund für die Behauptung abgeben kann, dass es keine Entitäten dieser Sorte gibt."

Was ergibt sich nun aus diesen Überlegungen für das Problem der Unterscheidung von Ursachen und Gründen? Es soll hier nicht der Eindruck erweckt werden, dass dieses Problem genauso zu handhaben ist, wie der Fall des Dämonenglaubens.[3] Das Problem besteht vielmehr darin, dass es einer Begründung für die Ungleichbehandlung beider Fälle bedarf. Warum ersetzt eine vollständige neurobiologische Beschreibung nicht das Gründe-Sprachspiel, sondern erklärt nur, *wie* Gründe realisiert sind? Mit anderen Worten, warum haben wir es beim Gründe-Sprachspiel anders als bei der Dämonentheorie mit einem lebensweltlichen Phänomen zu tun, das resistent gegenüber wissenschaftlicher Elimination ist, obwohl es ebenfalls – zumindest nach der von Pauen versuchsweise eingeführten Voraussetzung – vollständig reduzierbar auf eine zugrunde liegende Ebene ist? Wenn jedes menschliche Verhalten, das wir gewöhnlich unter Berufung auf Gründe zu erklären versuchen, vollständig mithilfe neurobiologischer Ursachen erklärt werden kann, dann scheint das „Prinzip der wissenschaftlichen Eleganz" doch den Schluss zu erzwingen: Die Erklärung durch Gründe erklärt nichts, was wir nicht auch durch neurophysiologische Ursachen erklären können, also sind lebensweltliche Erklärungen, die sich auf Gründe berufen überflüssig (nicht notwendig) und daher ontologisch zu eliminieren.

5. Transzendentale Argumentation

Mir scheint es notwendig zu sein, auf irgendeine Art transzendentalphilosophischer Argumentation zurückzugreifen, um den Unterschied zwischen Phänomenen wie dem Geben und Nehmen von Gründen einerseits und Phänomenen, die durch wissenschaftliche Erklärungen eliminiert werden können, andererseits, begründen zu können. Zumindest sehe ich nicht, wie diese Begründungsleistung sonst erbracht werden könnte. Auch könnte die Selbstverständlichkeit, mit der Pauen an einer realistischen Deutung des Gründe-Sprachspiels festhält, dafür sprechen, dass ihm diese Realität intuitiv unkritisierbar erscheint.

Die Beeinflussbarkeit unseres Handelns und Denkens durch Gründe ist eine nicht aufgebbare Voraussetzung für unsere lebensweltliche Praxis insgesamt, auch für Wissenschaft und Philosophie. Die Praxis des Begründens ist so grundlegend mit unserem Welt- und Selbstverständnis verflochten, dass es als ein zelnes Phänomen gar nicht aus diesem Gesamtzusammenhang herausgelöst und zum Gegenstand einer kritischen Überprüfung gemacht werden kann. Hierin ist der wesentliche Unterschied zu sehen im Vergleich zu Annahmen, wie der Dämonentheorie oder auch des schon erwähnten Wechsels vom geo- zum heliozentrischen Weltbild.

Eine transzendentalphilosophische Prüfung der Rolle, die das Gründe-Sprachspiel in unserer Lebenswelt spielt, zeigt, dass es sich bei diesem Sprachspiel um ein konstitutives Element handelt, dass so wie die Lebenswelt als Ganze nicht als Täuschung entlarvt werden kann. Erst mit dieser transzendentalphilosophischen Hintergrundannahme wird die Argumentation Michael Pauens schlüssig. Sie erklärt, warum selbst unter der Voraussetzung des Physikalismus die mentale Beschreibungsebene – und mit ihr die Rede von Gründen – nicht obsolet wird.

Will man hier noch mir Pauen von *Physikalismus* sprechen, so scheint mir diese Bezeichnung allerdings irreführend zu sein. Physikalismus heißt in diesem Zusammenhang ja nur, dass eine vollständige reduktive neurophysiologische Erklärung mentaler Phänomene möglich ist. Das Explanandum wird hierbei aber nicht etwa wegerklärt, sondern es bleibt Ausgangs- und bleibender Bezugspunkt für die wissenschaftliche Erklärung. Die Erklärung setzt das zu Erklärende voraus, nicht umgekehrt. Die transzendentalphilosophisch abgesicherte Lebenswelt erweist sich damit als die eigentliche ontologische Basis.

Literatur

Beckermann, Ansgar1999 „Bewusstsein – neurobiologisch erklärbar?", *Forschung an der Universität Bielefeld* 20, 29-34.

Nida-Rümelin, Julian 2005 *Über menschliche Freiheit*, Stuttgart: Reclam.

Nida-Rümelin, Julian 2006 „Ursachen und Gründe: Replik auf: Michael Pauen, Ursachen und Gründe", *Information Philosophie* 1/2006, 32-36.

Pauen, Michael 2005 „Ursachen und Gründe: Zu ihrer Unterscheidung in der Debatte um Physikalismus und Willensfreiheit", *Information Philosophie* 5/2005, 7-16.

Rorty, Richard ^{2}1993 „Leib-Seele-Identität, Privatheit und Kategorien", in: Bieri, Peter (ed.) *Analytische Philosophie des Geistes*, Bodenheim: Athenäum Hain Hanstein, 93-120. (Original erschienen 1965 „Mind-Body Identity, Privacy, and Categories", Review of Metaphysics 19, 24-54.)

Schmidt-Salomon, Michael 2007 „Von der illusorischen zur realen Freiheit: Autonome Humanität jenseits von Schuld und Sühne", in: Liessmann, Konrad Paul (ed.) *Die Freiheit des Denkens*, Wien: Paul Zsolnay, 179-218.

[3] Genau das ist die These, die Rorty mit Hilfe des Dämonen-Beispiels verteidigen will: „Gerade so, wie wir jetzt die Existenz von Dämonen leugnen wollen, könnte die zukünftige Wissenschaft die Existenz von Empfindungen leugnen wollen." (Rorty 21993, 98).

Making the Mind Higher-Level

Elizabeth Schier, Sydney, Australia

Central to Kim's position is a distinction between levels and orders. Levels are understood in mereological terms. Lower-levels contain parts which are organised into structured wholes at higher-levels. Orders are ways of conceptualising properties without providing details of their various realisers. So water is at a higher-level to the H_20 molecule because water is made up of H_20 molecules. Liquidity is at a higher-order to water because water is one way to realise liquidity. But liquidity is not at a higher mereological level to water because water is not a part of liquidity in the same way that H_20 molecules are a part of water.

Crucially Kim is a realist and a physicalist about higher-level phenomena. He is a realist because higher-level properties have causal powers that their parts lack. He is a physicalist because the properties of the higher-level wholes are determined by their mirco-physics constituents. A brief reflection on a simple example shows that this is obviously correct. Water can do things that no individual H_20 molecule can: you would not want to dive into a pool containing only one H_20 molecule. This difference in causal powers is due to the arrangement of the molecules: you also would not want to dive into a pool containing ice or steam. The fact that the isolated constituents or the same constituents in different structures can have different types of causal powers means that the different structures can not be identical to their constituents. There must be more to being water, snow, or ice than being composed of H_20 molecules. But the extra thing is simply the arrangement of the parts. Importantly, the ways in which the parts can be arranged depends on the nature of the parts. It is the because oxygen is a small highly electronegative atom that H_20 is a polar molecule and it is because H_20 is a highly polar molecule that hydrogen bonds form between the molecules and it is because hydrogen bonds form between the molecules that water has unusual properties (e.g. liquid water is denser than ice so ice floats on water). The nature of the H_20 molecule determines how it can be arranged and therefore determines the causal powers of the structures that are composed of it. So it is the nature of the constituents that determines the causal powers of the whole. According to Kim, although physicalism is true, a strongly eliminative micro-physicalism is clearly false. Although everything is physical, not everything is identical to the entities of micro-physics because not all the causal powers are possessed by the entities of micro-physics.

While there are controversies regarding Kim's picture, they do not concern his claim that higher-level phenomena are real and physical. No one doubts that water can do things that an individual H_20 molecule can't and no one doubts that water is physical. This suggests that if we can show that the mind and brain are in a mereological relationship in the same way as water and H_20 then we can show, to everyone's satisfaction, that the mind is as real and physical as water. I want to suggest that this option has not been explored because the standard, classical digital account of the mind does not make the mind higher-level. Fortunately, a connectionist, analog account of the mind does.

What might distinguish the mind from its lower-level constituents? What might the mind be able to do that its parts can't? One obvious answer is that the mind can compute, but its parts can't. So what would it take for computation to be a higher-level property? What is doing the causing in computation? Following O'Brien and Opie (2006) I suggest that computation is a causal process involving representing vehicles where the trajectory of that process is shaped by the meaning of the vehicles. That is, computation involves representing vehicles causing each other in a way that depends on the meaning of those vehicles. Meaning is doing the causing in computation. So for computation to be a higher-level process we need meaning to be higher-level property. That is, we need the meaning of a representing vehicle to depend on the organisation of its parts. Although we can get this for analog computation, we do not get this for digital computation.[1]

Digital computation is the rule-governed manipulation of symbols (Haugeland 1985). A symbol is a representing vehicle that bears an arbitrary relation to its represented object (Copeland 2005). The content of a symbol does not depend on the intrinsic nature of the symbol. For example there is nothing about the numeral '3' as a physical object that means it must be about the number three. Another way to put this is that there is an arbitrary relation between the syntactic and semantic properties of a symbol. This means that, for symbols, meaning is not a higher-level property of the symbol. The meaning of a symbol does not depend on the organisation of the parts of a symbol. You can change the structure of your symbol, and as long as the rules are changed to recognise the new structure, the system will still function correctly. This is why it is possible to build a digital computer out of anything, even beer cans or bricks.

But there is an alternate analog account of computation according to which computation is the manipulation of analogs, representing vehicles that resemble what they are about. Consider for example an architect's scale model, which is being used to calculate the shadows that will be cast by a building.[2] Like water, the organisation and nature of the parts of a scale model matter. If you change the structure then you change what they are about. For example, if you move the buildings, then they will no longer represent the proposed buildings and the shadows they will cast. If you change the nature of the constituents you also change the meaning. It is no good building your scale model out of glass because it won't block the light in the right way. The organisation and nature of the constituents determines what the model resembles and therefore determines what it represents. In analog computation there are no rules which we can change to deal with the change to the representing vehicles. The meaning of the representing vehicles in a scale model is at a mereologically higher-level to their constituents. The same constituents in a different arrangement will have a different meaning. Even by Kim's standards, meaning is a new, causally efficacious property in the scale model.

[1] The analog-digital distinction is often understood in terms of the continuity or discreteness of the representational medium. I am using the distinction not in this way, but rather as a distinction between representing vehicles that do or do not have their content grounded in resemblance (see Copeland 2005).

[2] I have taken this example from Gerard O'Brien.

Of course we don’t have scale models in the brain. But this does not mean that an analog account of the brain is impossible as there is a more abstract ‘second-order’ form of resemblance (Shepard and Chipman 1970). First order resemblance occurs when two things share one or more physical properties. For example the paint cards you get at the hardware store represent the colour of the paint by replicating the colour of the paint. It is implausible that the ground of mental content is first-order resemblance because we are capable of representing many things with which the brain shares no physical properties. But there is a more abstract ‘second-order’ notion of resemblance. 'In second-order resemblance, the requirement that representing vehicles share physical properties with their represented objects can be relaxed in favour of one in which the *relations* among a system of representing vehicles mirror the *relations* among their objects' (O'Brien and Opie 2004, p.10). Consider for example a mercury thermometer. It represents temperature because the relation between variations in the volume of the mercury resembles the relations between variations in the temperature of the substance that the thermometer is in contact with. For example, if volume x is greater than volume y then the temperature which corresponds to x is hotter than the temperature that corresponds to y. Because of this relational similarity between the volume of the mercury and the temperature of the substance it is in contact with, it is possible to use variations in the volume of mercury to represent variations in temperature.

Importantly, the structure and nature of the constituents of the representing vehicles determines their content for second-order resemblance. As with the scale model, the meaning of a mercury thermometer depends on the structure of the parts. If you change the structure by stomping on it, you would change the meaning. Also like the scale model, the nature of the constituents determines the meaning in the thermometer. It is because mercury is a substance that expands when heated that it can be used to represent temperature; you cannot make a thermometer out of wood or custard. So meaning in a mercury thermometer is at a mereologically higher-level to its constituents.

Connectionism provides a model of how to implement analog computation in the brain. We can see second-order resemblance in the activation space of many artificial neural networks. Activation space is a multi-dimensional space that enables us to represent the relations between patterns of activation across the units of a layer of an artificial neural network. Each dimension of activation space corresponds to the possible levels of activation of one of the units. Each point in activation space represents one pattern of activation across a layer. Patterns of activation that are near each other in activation space will be similar in that they consist of similar patterns of activity across the units whereas dissimilar patterns of activation across the units will be further apart in activation space. The similarity relations between the patterns of activation can be represented in activation space via the distance between the points.

Importantly it has been shown that the activation space of many artificial neural networks resembles aspects of the task domain. Consider for example Cotrell’s face recognition network, which is able to recognize familiar faces, distinguish between faces and non-faces, and determine the gender of unfamiliar faces (Churchland 1995). When the patterns of activation across the hidden units were mapped in activation space it was discovered that all the patterns of activation that represent images of a particular individual are clustered closely together in activation space. What this means is that they are represented with patterns of activation that are quite similar. The patterns of activation that represent images of faces of the same gender are clustered together in a larger region of activation space. Finally all of the patterns of activation that represent faces are located in a large region of activation space that is separate from the patterns of activation that represent non-face images. In general, similar faces are represented with similar patterns of activation and dissimilar faces are represented with dissimilar patterns of activation. The relations between patterns of activation resemble the relations between faces.

So we can see second-order, relational similarity at work in artificial neural networks. Importantly the structure and nature of the constituents of the representing vehicles determines their content in such networks. We can see this by realising that the similarity relations between patterns of activation that are mapped in activation space are based in the intrinsic structure of the patterns of activation. Although individual patterns of activation are represented in activation space by points, the basis of their location in activation space, and the basis of their similarity relations, is their internal structure. Two patterns of activation will be located near each other in activation space because they consist of similar levels of activity across the units. Conversely, if two patterns consist of different levels of activity of the various units then they will be located in different regions of activation space. If you change the nature of a pattern of activation you will move it within activation space and therefore change what it resembles and means. And patterns of activation have the structure that they do because of the electrical nature of the units/neurons that constitute them; you cannot build a network out of beer cans or bricks. For the connectionist, meaning does depend on the structure and nature of the parts, so meaning is mereologically higher-level to the parts. It is only when you organise neurons into the right firing patterns that you have representing vehicles with meaning. Meaning is neurons arranged in a particular way in the same way that water is H_20 arranged in a particular way. So the meaning of a pattern of activation across neurons is at a higher mereological level to the neurons that constitute it. This means that the meaning of connectionist representing vehicles is not in competition for causally efficacy with their neural constituents.

We can see that on an analog connectionist account of the mind the meaning of a representing vehicle is at a mereolgocially higher-level to its constituents. So meaning can play a causal role over and above that played by its neural constituents. Yet, because the causal powers of the representing vehicles are determined by the nature of their constituents, meaning is a physical property. So, like water, meaning is higher-level, real, physical and causally efficacious. Connectionism makes the mind higher-level and therefore, by everyone’s standards, real, causally efficacious and physical.

Literature

Churchland, P. M. (1995). The engine of reason, the seat of the soul: a philosophical journey into the brain. Cambridge, Mass.: MIT Press.

Copeland, J. (2005). The Modern History of Computing. The Stanford Encyclopedia of Philosophy. E. N. Zalta. URL = <http://plato.stanford.edu/archives/fall2005/entries/computing-history/>.

Haugeland, J. (1985). Artifical Intelligence: The very idea. Cambridge, MA: Bradford Books.

Kim, J. (1998). Mind in a Physical World: An essay on the mind-body problem and mental causation. Cambridge: MIT Press.

O'Brien, G. and J. Opie (2004). Notes Towards a Structuralist Theory of Mental Representation. Representation in Mind: New Approaches to Mental Representation. H. Clapin, P. Staines and P. Slezak, Elsevier.

O'Brien, G. and J. Opie (2006). "How Do Connectionist Networks Compute?" Cognitive Processing 7(1): 30-41.

Shepard, R. N. and S. Chipman (1970). "Second-order isomorphism of internal representations: Shapes of states." Cognitive Psychology 1: 1-17.

Zwischen Humes Gesetz und „Sollen impliziert Können“ – Möglichkeiten und Grenzen empirisch-normativer Zusammenarbeit in der Bioethik (Teil II)*

Sebastian Schleidgen, Tübingen, Deutschland

1. Formallogische Grundlagen empirisch-normativer Zusammenarbeit II: Sollen impliziert Können

Während die im ersten Teil der Arbeit explizierte Annahme des Humeschen Gesetzes die Grenzen empirisch-normativer Zusammenarbeit logisch untermauert, verdeutlicht unsere zweite Grundannahme – die so genannte „Sollen impliziert Können“-Annahme – deren spezifische Möglichkeiten. Ihre erste philosophiehistorisch relevante Formulierung findet sich in I. Kants *Kritik der reinen Vernunft*. Dort schreibt Kant: „Nun muss die Handlung allerdings unter Naturbedingungen möglich sein, wenn auf sie das Sollen gerichtet ist“ (Kant 1965: 534). Die Definition und detaillierte Analyse der „Sollen impliziert Können“-Annahme in ihrem noch heute gängigen Verständnis ist jedoch H. Albert zuzurechnen. Er schreibt in *Konstruktion und Kritik*: „Wenn wir nun auf unser Problem zurückkommen, ob die Sozialwissenschaft etwas zur Beantwortung der zweiten Kantischen Frage beitragen kann, so ergibt sich eine sehr einfache Antwort, nämlich: tatsächlich kann sie, und zwar auch unter Beibehaltung des Weberschen Wertfreiheitsprinzips, außerordentlich viel dazu beitragen. Sie kann nämlich im Rahmen unseres Wissens die Frage beantworten: *Was können wir tun?* Und diese Frage hat eine fundamentale Beziehung zu der Frage: *Was sollen wir tun?* Die meisten Menschen werden nämlich vermutlich einer Behauptung zustimmen, die, auf ihre kürzeste Form gebracht, lautet: *Sollen impliziert Können.*“ (Albert 1972: 58)

Albert können wir an dieser Stelle keiner detaillierten Analyse unterziehen. Es ist jedoch schon intuitiv einleuchtend, dass die Befolgung einer moralischen Norm davon abhängig ist, inwieweit die Akteure diese Norm zu befolgen imstande sind. Diese Intuition lässt sich im Hinblick auf die Aufgaben normativer Theorie plausibilisieren: Sie soll schließlich Normen generieren, nach denen wir – qua unseres Menschseins – überhaupt handeln *können*. Es reicht nicht, dass moralische Normen *logisch* mögliche Handlungen von den Akteuren fordern, insbesondere dann nicht, wenn uns eine Erfüllung überfordern würde.

Aus der Akzeptanz der „Sollen impliziert Können“-Annahme ergibt sich jedoch ein weiteres Problem: So kann man fragen, wie Humes Gesetz und „Sollen impliziert Können“ – als Grundannahmen einer empirisch-normativen Zusammenarbeit – in einen konsistenten Zusammenhang gebracht werden können. Der Intuition zufolge scheint „Sollen impliziert Können“ doch zu bedeuten, dass sich ein – empirisch zu messendes – Können auf ein Sollen niederschlägt, mithin von einem Sein auf ein Sollen geschlossen wird: Aus den deskriptiv zu erfassenden Möglichkeiten menschlicher Akteure ergeben sich ihre moralischen Verpflichtungen. Das aber würde einen Verstoß gegen Humes Gesetz bedeuten. Nun ergibt sich dieses Problem aber nur in der intuitiven Lesart von „Sollen impliziert Können. Nach dieser Lesart wird „Sollen impliziert Können“ – formallogisch gesprochen – durch ein Bikonditional repräsentiert, d.h. Können (Ka) und Sollen (Sa) stehen im formallogischen Verhältnis Ka≡Sa. Etwas wäre demnach gesollt dann und nur dann, wenn es gekonnt wird. Das würde tatsächlich bedeuten, dass sich die moralische Richtigkeit bzw. Falschheit einer Handlung direkt aus den Handlungsmöglichkeiten eines Akteurs herleiten ließe. Sieht man sich die oben zitierten Textstellen aus der *Kritik der reinen Vernunft* bzw. *Konstruktion und Kritik* jedoch genauer an, zeigt sich, dass weder Kant noch Albert ein solch intuitives Verständnis von „Sollen impliziert Können“ im Sinn gehabt haben können. So schreibt Kant weiter: „Es mögen noch so viele Naturgründe sein, die mich zum Wollen antreiben, noch so viele sinnliche Anreize, so können sie nicht das Sollen hervorbringen, sondern nur ein noch lange nicht notwendiges, sondern jederzeit bedingtes Wollen, dem dagegen das Sollen, das die Vernunft ausspricht, Maß und Ziel, ja Verbot und Ansehen entgegen setzt.“ (Kant 1965: 534) Albert hingegen bekräftigt zwar, dass sozialwissenschaftliche Empirie zur Frage „Was sollen wir tun?“ durch ihre Erkenntnismöglichkeiten hinsichtlich der Frage „Was können wir tun?“ etwas beitragen kann, postuliert aber gleichzeitig, dass dies „unter Beibehaltung des Weberschen Wertfreiheitsprinzips“ zu geschehen habe. Damit schließen beide Autoren direkte Schlüsse vom deskriptiven Können auf normative Sollens-Aussagen und damit ein Verständnis von „Sollen impliziert Können“ als Bikonditional aus.

Das klassische Verständnis von „Sollen impliziert Können“ wird hingegen durch eine Implikation zwischen Sollen und Können, also Sa → Ka, zum Ausdruck gebracht. Auch dieses Verständnis ist unserer Meinung nach nicht adäquat: So wird Sa → Ka ja dann und nur dann falsch, wenn Sa „wahr“, Ka aber „falsch“ ist. Das würde bedeuten, das „Sollen impliziert Können“ im vorliegenden Fall verletzt ist. Nun kann man dies entweder als deskriptive Aussage verstehen oder aber versuchen, „Sollen impliziert Können“ dem vorliegenden Fall anzupassen, mithin bei ¬Ka auch ¬Sa einzufügen, um Sa → Ka wieder zu erfüllen. Das entspricht jedoch einem Schluss über Modus Tollens von ¬K auf ¬S bzw. von einem (Nicht-)Können auf ein (Nicht-)Sollen, was wiederum einen Verstoß gegen Humes Gesetz bedeutet. Möchte man das klassische Verständnis von „Sollen impliziert Können“ retten, müsste man dessen Gültigkeit auf Fälle beschränken, in denen Ka den Wahrheitswert „wahr“ besitzt. Dies würde das Prinzip jedoch ad absurdum führen, da es dann nur noch in Fällen gilt, in denen das Gesollte auch gekonnt wird.

Wie aber ist „Sollen impliziert Können“ dann zu verstehen? Unserer Meinung nach erweist sich nur ein *schwaches* Verständnis im Sinne eines Implikationsverhältnisses von Können und Sollen als plausibel. *Formallogisch* gesprochen bedeutet dies „Können impliziert Sollen“ (Ka → Sa), weshalb wir den traditionellen Ausdruck „Sollen impliziert Können“ für unglücklich gewählt halten. Dabei schränken wir die Gültigkeit von „Können impliziert Sollen“ auf solche Instanzen ein, für die

* Der erste Teil dieses Aufsatzes findet sich weiter oben; siehe den Beitrag von Michael Jungert.

Sa wahr ist. Das ist wichtig, weil eine generelle Akzeptanz von Ka → Sa für alle Instanzen in einen Normkonflikt münden würde: wären nämlich sowohl a als auch ¬a möglich, würde die Offenheit von Sa hinsichtlich seines Wahrheitswertes dazu führen, dass sowohl a als auch ¬a gesollt wären. Die Erklärung für diese Einschränkung ist einfach: Sofern kein Sollen etabliert ist, also Sa den Wert „falsch“ hat, ist eine Beschäftigung mit „Sollen impliziert Können“ oder – wie wir es nennen – „Können impliziert Sollen“ schlicht irrelevant. Es ist dann schließlich kein Sollen etabliert, demgegenüber ein Können steht.

Nun muss man sich klarmachen, was ein solches formallogisches Verständnis bedeutet: Der Satz „Können impliziert Sollen“ ist wahr unabhängig vom Wahrheitswert der Teilaussage „Können“, d.h. ob etwas tatsächlich gekonnt wird, ist irrelevant dafür, ob es gesollt wird. Das klingt zunächst contraintuitiv, ist aber deshalb plausibel, weil ein moralisches Sollen – wie wir hervorgehoben haben – ausschließlich durch Idealnormen etabliert wird, die ihre Geltung unabhängig von empirischen Erkenntnissen, mithin auch unabhängig von einem faktischen Können – beanspruchen: Sollen ist schlichtweg unabhängig von Können. „Können impliziert Sollen“ entspricht damit einem wissenschaftstheoretisch adäquaten Verhältnis von Idealnormen zur Praxis und verstößt nicht gegen Humes Gesetz.

Es bleibt die Frage, welche Rolle das Können spielt, wenn es keinerlei Einfluss auf das Sollen hat. Wie wir sagten, ist es Aufgabe normativer Theorie, Normen zu generieren, nach denen wir – qua unseres Menschseins – handeln können. Es genügt nicht, dass moralische Normen formallogische Kriterien erfüllen, um handlungsleitend zu wirken. Diese Teilaufgabe normativer Theorie entspricht aber der von uns bereits wissenschaftstheoretisch skizzierten Notwendigkeit einer Übersetzung von Ideal- in Praxisnormen. Denn das empirisch zu erfassende Können entspricht den menschlichen Handlungsmöglichkeiten innerhalb ihrer kognitiven, motivationalen sowie extern bedingten Beschränkungen. Und das Ziel der Erfassung von Beschränkungen bzw. Können besteht ja darin, die Handlungen in der Praxis soweit als möglich an den durch die Idealnormen avisierten moralischen Idealzustand anzupassen. Dabei aber verlieren – wie wir hervorhoben – die Idealnormen, als Sollens-Aussagen, keineswegs ihre Gültigkeit. Wir haben es also gewissermaßen mit zwei Sollensarten zu tun: dem durch die Idealnormen generierten „moralischen Sollen“ steht ein „praktisches Sollen“ gegenüber, das durch die Praxisnormen ausgedrückt wird und welches das moralische Sollen an die menschlichen Handlungsmöglichkeiten anpasst. Daher ist der Satz „Können impliziert Sollen“ einerseits wahr unabhängig von der Wahrheit des Gliedes „Können“ – nämlich, wenn er sich auf das moralische Sollen bezieht –, gleichzeitig aber ist eine empirische Erfassung des Könnens mit Hinblick auf die moralische Praxis zwingend notwendig – wenn es um das praktische Sollen geht: Können spielt also nur für das Sollen in der Praxis eine Rolle, die durch kognitive, motivationale und extern bedingte Beschränkungen der Akteure vorstrukturiert ist. Eine angemessene Formulierung des traditionellen Ausdrucks „Sollen impliziert Können“, die ein hinsichtlich des Sollens wirksames Können beinhaltet, ist daher „praktisches Sollen setzt Können voraus“.

An dieser Stelle sei festgehalten, dass “Sollen impliziert Können” *kein* Brückenprinzip im oben skizzierten Sinne von Anwendungsbedingungen moralischer Normen ist. Schließlich formulieren Brückenprinzipien nach dem Schema „Eine Handlung H ist moralisch geboten gemäß der Norm N genau dann wenn das empirisch zu überprüfende Kriterium K gegeben ist“ ein *bikonditionales* Verhältnis zwischen K und N: N hat in einer vorliegenden Situation dann und nur dann Geltung, wenn K gegeben ist. Wie wir gezeigt haben, liegt dem adäquaten Verständnis von „Sollen impliziert Können“ hingegen ein *Implikations*verhältnis zugrunde: Die gesollte Idealnorm N ist gültig unabhängig von einem empirischen Nachweis darüber, ob sie faktisch eingehalten werden kann oder nicht. Dieser Unterschied zwischen Brückenprinzipien im Sinne von Anwendungsbedingungen und „Sollen impliziert Können“ ergibt sich aus ihrem unterschiedlichen Status innerhalb normativer Theoriebildung: Brückenprinzipien im Sinne von Anwendungsbedingungen sind Erweiterungen moralischer Idealnormen und somit Ergebnis normativer Theoriebildung. „Sollen impliziert Können“ hingegen hat – als Grundlage der Entwicklung von Praxisnormen – lediglich eine *pragmatische* Funktion.

Bislang haben wir auf wissenschaftstheoretischem Wege Erkenntnisinteresse, -möglichkeiten und -grenzen normativer Theorie und empirischer Sozialwissenschaften sowie die sich daraus ergebenden Notwendigkeiten und Grenzen einer Zusammenarbeit im Rahmen normativ-ethischer Fragestellungen betrachtet. Diese wissenschaftstheoretischen Überlegungen haben wir anschließend durch formallogische Überlegungen untermauert. Auf dieser Basis können wir nun eine Einschätzung der von Weaver und Trevino eingeführten Ansätze vornehmen: Sowohl parallele als auch integrative Ansätze sind abzulehnen, da normative und empirische Wissenschaften – insbesondere in der Bioethik – weder vollständig voneinander getrennt noch miteinander verschmolzen werden können. Schließlich sind sie einerseits aufeinander angewiesen, andererseits unterliegt ihre Zusammenarbeit aber bestimmten wissenschaftstheoretischen und logischen Grenzen.

Demgegenüber zeichnet sich der symbiotische Ansatz gerade dadurch aus, dass die theoretischen und methodologischen Kerne empirischer und normativer Wissenschaften strikt getrennt bleiben und beide Disziplinen deshalb auf eine Zusammenarbeit nach oben dargestelltem Muster angewiesen sind (vgl. Molewijk et al. 2004: 58). Nach wissenschaftstheoretischen und logischen Kriterien kann also ausschließlich ein symbiotisches Vorgehen als adäquat bezeichnet werden.

Abschließend möchten wir nun drei Varianten einer zulässigen und konstruktiven Zusammenarbeit zwischen sozialwissenschaftlicher Empirie und normativer Theorie vorstellen und durch Beispiele verdeutlichen.

2. Konsequenzen für eine konkrete Zusammenarbeit

Wie wir gezeigt haben, gibt es einerseits offenkundige Notwendigkeiten, andererseits klare Grenzen einer Zusammenarbeit zwischen normativer Theorie und empirischer (Sozial-)Wissenschaft. Daraus ergeben sich drei konkrete Modi normativ-empirischer Kooperation, die wir abschließend darstellen möchten.

Die erste Möglichkeit adäquater empirisch-normativer Zusammenarbeit betrifft die Übersetzung normativ entwickelter Idealnormen in faktisch umsetzbare Praxisnormen. Wie wir gezeigt haben, muss eine adäquate Moraltheorie notwendig auf Idealnormen basieren, die aufgrund kognitiver und motivationaler Beschränkungen menschlicher Akteure jedoch in Praxisnormen übersetzt werden müssen. Damit wird einerseits dem Kriterium „Sollen impliziert Können“ entsprochen, andererseits aber garantiert, dem durch die Idealnormen angestrebten moralischen Idealzustand möglichst nahe zu kommen. Während Idealnormen ausschließlich mit den Mitteln normativer Theorie entwickelt werden können, ist eine Übersetzung in Praxisnormen auf die Mittel empirischer Wissenschaften angewiesen. Mit Hinblick auf die von uns dargestellten Erkenntnismöglichkeiten erweisen sich empirische *Sozial*wissenschaften als in besonderem Maße für diese Übersetzung geeignet, da sie die Erfassung kognitiver und motivationaler Beschränkungen ermöglichen. Darüber hinaus sind sie in der Lage, auch extern bedingte Beschränkungen zu erfassen. Dabei ist zu beachten, dass empirische Erkenntnisse nicht ausschließlich Beschränkungen der zu übersetzenden Idealnorm zur Folge haben müssen. Denn gerade aus der Erkenntnis motivationaler Beschränkungen hinsichtlich moralischer Normen lassen sich handlungstheoretische Steuerungsmöglichkeiten ableiten: Wenn Klarheit darüber besteht, aus welchen Gründen Akteure entgegen den Anforderungen einer moralischen Norm handeln, obwohl sie dazu kognitiv in der Lage sind, lassen sich möglicherweise Wege entwickeln, die Akteure zur Einhaltung der Norm zu motivieren. Die Ergebnisse der empirischen Sozialwissenschaften schränken die Anforderungen idealer Normen also nicht zwingend ein, sondern können im Gegenteil sogar Mittel und Wege aufzeigen, die Akteure näher an die Einhaltung der Idealnormen, mithin die Erreichung des moralischen Idealzustandes zu bringen.

Die zweite Möglichkeit adäquater empirisch-normativer Zusammenarbeit betrifft die Klärung der durch Brückenprinzipien ausgedrückten Anwendungsbedingungen moralischer Normen. Diese Bedingungen sind allerdings nicht mit jenen – ebenfalls empirisch zu überprüfenden – Beschränkungen menschlichen Handelns zu verwechseln, die wir zuvor als Grundlage der Entwicklung von Praxisnormen betrachtet haben. Vielmehr werden solche Bedingungen durch Sätze wie „Alle leidensfähigen Wesen sind gemäß der Norm S zu behandeln“ ausgedrückt. Wie wir gezeigt haben, muss es Aufgabe normativer Theorie sein, neben der Norm S auch ihre Anwendungskriterien K – hier „Leidensfähigkeit“ – zu generieren. Die Bestimmung von Fällen, in denen S in der Praxis geboten ist, ist jedoch an empirische Erkenntnisse gebunden: um festzustellen, ob gegenüber einem Wesen X tatsächlich entsprechend S gehandelt werden soll, ist empirisch zu überprüfen, ob X faktisch leidensfähig ist. Nun handelt es sich im Falle einer Anwendungsbedingung „Leidensfähigkeit“ zwar offenkundig um ein naturwissenschaftlich zu überprüfendes Kriterium. Es sind jedoch auch Brückenprinzipien entwickelt worden, die sozialwissenschaftlich zu überprüfende Kriterien einführen: Eines der bekanntesten Beispiele hierfür ist das regelutilitaristische Brückenprinzip „Wenn dies zur Stabilisierung einer Gesellschaft beiträgt, ist gemäß der Norm S zu handeln“. Bei der Klärung solcher Anwendungsbedingungen ist normative Theorie klarerweise auf sozialwissenschaftliche Empirie angewiesen.

Die dritte Möglichkeit adäquater empirisch-normativer Zusammenarbeit betrifft die Messung und Evaluation der sozialen Praxis hinsichtlich der Umsetzung moralischer Normen sowie der Genese neuer moralischer Fragestellungen: Wie wir feststellten, sind empirische Sozialwissenschaften in der Lage, Handlungs- und Diskursmuster der sozialen Praxis zu erfassen. Auf Basis dieser Erkenntnisse ist es möglich, die soziale Praxis daraufhin zu bewerten, ob und inwiefern sie den Anforderungen der zugrunde gelegten Praxisnormen entspricht. Stellt sich beispielsweise heraus, dass bestimmte Praxisnormen faktisch keine Anwendung finden, obwohl sie für moralisch wertvoll befunden wurden, können auf Grundlage dieser Erkenntnisse Maßnahmen eingeleitet werden, die eine moralkonforme Praxis im Sinne dieser Normen befördern. Darüber hinaus können empirische Sozialwissenschaften den gesellschaftlichen Diskurs auf – beispielsweise durch neuartige Technologien herausgeforderte – neue moralische Probleme und Debatten hin untersuchen und diese einer normativen Analyse zugänglich machen. So lässt sich in den letzten Jahren beispielsweise eine verstärkte „moralische Hilflosigkeit“ gegenüber neuartigen reproduktionstechnologischen Mitteln beobachten, die dringend einer normativen Analyse bedarf. Vor einer solchen Analyse muss jedoch zunächst die Feststellung eines solchen Bedürfnisses stehen, die offenkundig den empirischen Sozialwissenschaften obliegt. Damit kann auch die grundlegende Leistung empirischer Sozialwissenschaften, soziale Handlungs- und Diskursmuster zu erfassen und auszuwerten, für die empirisch-normative Zusammenarbeit fruchtbar gemacht werden.

3. Conclusio

In den beiden Teilen dieses Aufsatzes haben wir aufgezeigt, warum und an welchen Stellen eine Zusammenarbeit empirischer und normativer Wissenschaften – insbesondere in der Bioethik als *angewandter* Ethik – zwingend notwendig ist: Erkenntnisziele sowie die jeweils immanenten wissenschaftstheoretischen Fundamente bringen beide Disziplinen an ihre Erkenntnisgrenzen, wenn es um (bio-)ethische Fragestellungen geht und erfordern daher einen Rückgriff auf die Erkenntnismöglichkeiten der jeweils anderen Disziplin. Es ergibt sich also – so könnte man auch sagen – eine *Notwendigkeit* der Zusammenarbeit gerade daraus, dass Theoriekerne und Methodologie beider Disziplinen – zumindest solange man wissenschaftstheoretisch adäquat arbeitet – strikt voneinander getrennt sind und bleiben müssen.

Diese wissenschaftstheoretischen Überlegungen konnten wir durch formallogische Überlegungen zum Verhältnis von Sein und Sollen bzw. Sollen und Können weiter fundieren und symbiotische Ansätze als einzig adäquate Form empirisch-normativer Zusammenarbeit charakterisieren. Davon ausgehend wurden drei zulässige Möglichkeiten der Zusammenarbeit in der Bioethik charakterisiert: Erstens die empirische Erfassung kognitiver, motivationaler und extern bedingter Beschränkungen, welche die notwendige Übersetzung von Ideal- in Praxisnormen erst ermöglichen. Zweitens die empirische Erfassung der Anwendungsbedingungen moralischer Normen, die notwendig für die situationsspezifische Entscheidung über das Vorliegen einer moralischen Verpflichtung ist. Und drittens die empirische Evaluation der sozialen Praxis, die eine Bewertung des faktischen Umgangs mit Praxisnormen erlaubt und darüber hinaus neuartige moralische Fragestellungen ausfindig machen und der normativen Analyse übergeben kann.

Literatur

Albert, Hans 1972 Konstruktion und Kritik. Aufsätze zur Philosophie des kritischen Rationalismus, Hamburg: Hoffmann und Campe.

Kant, Immanuel 1965 Kritik der reinen Vernunft, Hamburg: Felix Meiner.

Molewijk, Bert, Stiggelbout, Anne M., Otten, Wilma, Dupuis, Heleen M., and Kievit, Job 2004 “Empirical Data and Moral Theory. A Plea for Integrated Empirical Ethics”, Medicine, Health Care, and Philosophy 7: 55-69.

Mental Causation: A Lesson from Action Theory

Markus Schlosser, Bristol, England, UK

1. What is Mental Causation?

Approaching an answer to this question, let us first assume that the kind of causation we are interested in is event-causation, where events may be construed as particulars or instantiations of properties. Given that, there is mental causation only if some mental events are causally efficacious in the sense that they stand in event-causal relations with other events. Most philosophers would add, firstly, that mental events must be causally efficacious in virtue of their mental properties (in virtue of instantiating mental properties), and secondly, that they must not overdetermine their effects. There is widespread agreement on this, and I will take it for granted here. The question I am interested in is this: What kinds of things must mental events cause for there to be genuine mental causation?

The contemporary mental causation debate, it is sometimes claimed, has its roots in Donald Davidson's seminal paper "Mental Events" (1970). One of the basic assumptions in this paper says that there is interaction between mental and physical events: some physical events cause mental events and some mental events cause physical events. Damage to muscle tissue, for instance, can cause pain, and intentions can cause behaviour. Many philosophers would agree with Davidson on this. They would agree, in particular, that mental causation requires mental-to-physical causation.

Many philosophers, however, would also acknowledge that the mental could be causally efficacious by causing other mental events. So why insist on interaction with physical events, if we could have mental causation within the domain of the mental alone? We can identify two closely related reasons for the insistence on mental-to-physical causation. Firstly, if there is mental-to-mental causation, then the mental is causally efficacious, strictly speaking. Nevertheless, this falls short of genuine mental causation. We tend to think that the mental is truly efficacious only if it causes physical events: mental events make a real difference only if they make a difference in the physical world. Secondly, genuine mental causation requires that mental events make a difference to our behaviour. They must, that is, make a difference to our bodily movements. (Some might reject the suggestion that mental-to-mental causation amounts to mental causation on the ground that we cannot rule out epiphenomenalism, if the mental has no observable effects.)

2. A Dilemma for Non-Reductive Physicalism

In the following I will sketch a version of non-reductive physicalism, and I will defend it against an influential objection. Let us assume, for the sake of argument, that psychology is irreducible and that some version of physicalism is true. As just pointed out, many philosophers think that there is genuine mental causation only if some mental events have physical effects. But mental-to-physical causation leads to a well-known problem for non-reductive physicalism: if mental events cause physical events, they merely overdetermine their effects, given the causal closure of the physical.

Given all that, non-reductive physicalism faces the following dilemma. If mental events cause physical events, they merely overdetermine their effects, and if they cause only other mental events, they are not truly efficacious. So either way, the efficacy of the mental falls short of genuine mental causation. This dilemma is based on the dichotomy between mental-to-physical and mental-to-mental causation. I will now argue that this is a false dichotomy.

3. Actions and Movements

The dichotomy between mental-to-physical and mental-to-mental causation can be avoided if one acknowledges the distinction between actions and bodily movements. This distinction is common within the philosophy of action. In the philosophy of mind, however, it is neglected, sometimes overlooked and often blurred by talk about behaviour. Given that behaviour is the most important effect of mental events, this distinction should be of interest, and I will suggest that it can help in dealing with the problem of mental causation.

There is genuine mental causation if mental events cause and causally explain actions (in virtue of their mental properties and without overdetermination). But actions, I submit, are neither mental nor physical events. If that is correct, then there can be genuine mental causation without mental-to-physical causation and without limitation to mental-to-mental causation.

Most philosophers of mind accept or presuppose, implicitly perhaps, an event-causal theory of action. According to this view, actions are events with a certain causal history. Certain events, that is, are actions in virtue of being caused by the right antecedents (in the right way). The right antecedents are mental events that rationalize the action: an event is an action only if it is caused by rationalizing mental events and in virtue of being caused in that way. The causal history, in other words, is part of an action's essence or identity.

Some actions, I take it, are mental actions. On the causal theory, mental actions are realized by and perhaps token-identical to mental events. But they are not type-identical with mental events, because being of a certain mental event-type does not determine whether or not the event is an action. The formation of an intention, for instance, may be an action or not (we form some intentions actively, others passively). Whether it is an action or not depends on its causal history. The same holds for so-called *overt* actions (roughly, actions that involve bodily movement). Take a standard example of a basic action such as raising one's arm. It is physically realized by an arm rising, and perhaps every particular arm raising is token-identical with a particular arm rising. But they are not type-identical for the reasons given. Not every arm rising is an arm raising: being a certain type of movement does not determine whether or not the event is an action.

What is the rationale behind thinking of actions in historical or etiological terms? Consider the following two widely accepted doctrines. Firstly, all actions are intentional (under some description). Secondly, an action is intentional insofar as it is done for reasons in the minimal sense that it can be rationalized in light of some of the agent's mental events, and it is an action in virtue of being done for reasons in this sense. Proponents of the causal theory have argued that an action is done for reasons only if it is caused and causally explained by the mental events that rationalize it. Hence, the causal history enters into the essence of actions.

A further reason for thinking that actions are not type-identical with the physical events that realize them is given by the fact that most of our non-basic actions are multiply realizable. Assume that I give someone a signal by raising my arm. In this case, I perform the non-basic action of giving a signal by raising my arm. Clearly, there are many different ways in which I can give someone a signal. The act-type of giving a signal is in this sense multiply realizable and therefore not identical with a certain type of movement.

So, on the view that I am suggesting, actions are caused and causally explained by mental events. Both mental events and actions are realized by physical events. Given the irreducibility of the mental, mental events are not type-identical with physical events. And given the etiological nature of action, actions are not type-identical with physical events either. In particular, overt actions are not type-identical with bodily movements (although they are realized by them). Given further that mental events cause actions in virtue of their mental properties, we obtain a view according to which the mental is causally efficacious in a way that avoids the mentioned dichotomy. Mental events cause actions in virtue of their mental properties and without overdetermination. But this is neither mental-to-physical nor merely mental-to-mental causation.

In the next section I will address a pressing objection, rather than developing the view further. Consider, though, some brief remarks concerning the interaction between mind and body. On my view, mental causation is not merely mental-to-mental, because mental events cause actions. Actions, I argued, are neither mental nor physical events. What are they? They are realized by physical events, but they belong to the domain or level of psychological and intentional explanation: we recognize them as actions only insofar as they are rationalized in the light of mental events. What about the other direction? Is pain, for instance, caused by physical events? This appears to be undeniable. But consider the following alternative. It is not implausible to suggest that damage to muscle tissue, for instance, causes certain physical events in the brain that realize the mental event of pain. On this view, no physical event *causes* pain. Rather, physical events cause other physical events, which *realize* pain. If something like this holds for all mental events, and if my view on mental causation is correct, then there is no causal interaction between the mental and the physical. Nevertheless, epiphenomalism is false as mental events cause actions.

4. Overdetermination and Downward Causation

One apparent advantage of the suggested non-reductive view is that it avoids the causal exclusion problem. Mental events do not cause physical events, hence there is no problem of causal overdetermination and exclusion. But it has been argued, most prominently by Jaegwon Kim, that non-reductive physicalism is committed to downward causation. Kim argues for this claim by considering mental-to-mental causation. Assume that the mental event *M* causes another mental event *M**. According to any version of non-reductive physicalism, *M* and *M** are realized by physical events (their supervenience base). Assume that *P* is the supervenience base of *M*, and that *P** is the supervenience base of *M**. What explains the occurrence of *M**? There are two candidates: the occurrence of *M* and the occurrence of *P**. Both explain the occurrence of *M**, and together they overdetermine *M**, albeit not causally. *M* determines *M** in virtue of being its cause, and *P** determines *M** in virtue of being its supervenience base. This creates a *prima facie* tension between the two explanations. The best way to resolve this tension, Kim argues, is to assume that *M* causes *M* by causing its supervenience base P**. This shows, according to Kim, that mental-to-mental causation presupposes mental-to-physical causation (2005, 39-40).

It would seem obvious that the same reasoning can be applied to the view that I have suggested. If actions are caused by mental events, and if they are realized by physical events, then the resulting tension and overdetermination requires us to assume downward causation: we must assume that mental events cause the physical events (bodily movements, for instance), which realize actions.

But this line of reasoning is mistaken. According to Kim, the occurrence of the supervenience base of a certain mental event realizes, determines and necessitates the occurrence of the mental event. So, in the example, the occurrence of *P** determines and necessitates the occurrence of *M**. *P** by itself necessitates *M**, as Kim says, "no matter what happened before" (39), in particular, no matter whether *M* occurred or not (unless *M* is a cause of *P**). Putting aside the question of whether non-reductive physicalism is in fact committed to this, the same does not hold for actions.

Actions, we assume, are realized by physical events. It may well be that there is a sense in which physical events determine actions (in the sense in which determinates determine determinables, perhaps). But the occurrence of a certain bodily movement, for instance, does not necessitate the occurrence of a certain type of action, for the reasons given above. A certain physical event, such as a certain bodily movement, realizes an action only if it has the right causal history. Let us replace the mental event *M** in the example by an action *A*. Given the assumptions, *A* is caused by *M* and realized by *P**. But given the introduced claims concerning the nature of action, it is not true that the occurrence of *P** necessitates the occurrence of an *A*-ing. Whether or not *P** realizes an *A*-ing depends on the causal history. Given that an action has been performed, we can assume that *M* is a rationalizing cause of *A*. Had there been no rationalizing cause, there would have been no action performed (*P** would not have realized an *A*-ing). So, in this particular case, whether an *A*-ing is performed or not depends on whether or not *M* occurred and on whether or not *M* causes the *A*-ing. Hence, *P** alone does not necessitate the occurrence of an *A*-ing.

One might reply that it is perhaps not *P** alone that necessitates *P*, but *P**'s being caused by *P*. Assuming that *P* realizes the mental event *M*, *P*'s causing *P** realizes the causal history that makes *P** an action. In that way, the occurrence of *A* is necessitated by the occurrence of physical events and physical causation, and the problem of overdetermination reappears: *M* merely overdetermines *A*, as *A* is determined by *P*'s causing *P**.

Again, this is mistaken. Non-reductive physicalism is motivated by the thought that rationality has “no echo” in the physical domain. What makes the *A*-ing an action is the fact that it is caused and *rationalized* by the right events. *P* does not rationalize *P**. *P* cannot possibly rationalize *P**, as *P* does not have intentional content. Only mental properties can possibly rationalize actions. And mental properties are, we assume, not reducible to physical properties. The reply fails. The occurrence of *P** realizes an *A*-ing because and only because it is caused by *M*. Only the occurrence of a *mental* event can rationalize behaviour.

Literature

Davidson, Donald 1970 “Mental Events”, reprinted in Donald Davidson 1980 *Essays on Actions and Events*, Oxford: Clarendon Press, 207-227.

Kim, Jaegwon 2005 *Physicalism, or Something Near Enough*, Princeton: Princeton University Press.

Supervenienz, Zeit und ontologische Abhängigkeit

Pedro Schmechtig, Dresden, Deutschland

1. Ausgangsproblem

Viele Philosophen finden die Idee attraktiv, dass zumindest die Wahrheit kontingenter Propositionen von Dingen abhängt, die in der Welt existieren. Typischerweise wird gesagt, für jede Proposition <p> und Entität α gilt, wenn α ein Wahrmacher für <p> ist, dann kann <p> unmöglich wahr sein, ohne dass α existiert. Prinzipen dieser Art charakterisieren die Verbindung zwischen Wahrheit und Welt in Form einer modalen Existenzabhängigkeit, die für jede präsentistische Zeitkonzeption eine Herausforderung darstellt. Gemäß der präsentistischen Sichtweise existieren Dinge nur in der Gegenwart. Doch natürlich ist es so, dass wir mit zahlreichen Propositionen auf vergangene oder zukünftige Ereignisse bzw. Individuen Bezug nehmen; solche Propositionen können offenkundig wahr sein, obgleich deren Wahrheit nicht von der Existenz gegenwärtiger Dinge abhängt. Entsprechend lässt sich folgendes Argument formulieren:

Wahrmacher-Argument gegen den Präsentismus:

> (P1) Wenn der Präsentismus wahr ist, dann existieren Dinge nur in der Gegenwart *[Prämisse des Präsentismus]*.
>
> (P2) Kontingente Propositionen sind wahr, weil sie durch Dinge wahrgemacht werden, die in der Welt existieren *[Prämisse des Wahrmachens]*.
>
> (P3) Propositionen in Bezug auf die Vergangenheit bzw. Zukunft sind manchmal wahr *[offensichtliche Tatsache]*.

Aufgrund von (P1) und (P2):

> (P4) Propositionen in Bezug auf die Vergangenheit bzw. Zukunft werden durch Dinge wahrgemacht, die nur in der Gegenwart existieren.

Aufgrund von (P3) und (P2):

> (P5) Wenn Propositionen mit Bezug auf die Vergangenheit bzw. Zukunft wahr sind, hängen sie (zumindest teilweise) von Dingen ab, die in der Welt existieren, aber nicht gegenwärtig sind.

Aufgrund der Unvereinbarkeit von (P4) und (P5):

> (K) Der Präsentismus kann nicht erklären, warum (P3) gilt.

Es gibt verschiedene Strategien, das vorliegende Argument zu entkräften. Man könnte beispielsweise Prämisse (P3) bestreiten, indem man sagt, dass Propositionen in Bezug auf die Vergangenheit bzw. Zukunft weder wahr noch falsch sind. Vielleicht ist es plausibel, zu behaupten, dass Propositionen mit Blick auf zukünftige Ereignisse weder wahr noch falsch sind (Tooley 1997); doch in Bezug auf vergangene Ereignisse wäre eine ähnliche Behauptung sicherlich absurd. Eine andere Möglichkeit bestünde jedenfalls darin, Prämisse (P3) zu akzeptieren und stattdessen (P2) abzulehnen. Derartige Ansätze behaupten entweder, dass Dinge neben ihrer Existenz noch eine andere ‚Seinsweise' besitzen, so dass sie – gleichwohl sie nicht existieren – trotzdem als Wahrmacher für kontingente Propositionen fungieren können (Meinong 1904). Oder aber man weist (P2) zurück, weil man denkt, dass sich die Abhängigkeit, welche im Hinblick auf die gegenwärtige Welt besteht, nicht an der ontologischen Fundierung durch entsprechende Entitäten festmacht, sondern rein semantisch über die Einführung einer primitiven Maschinerie – universelle Quantifikation plus zugehörigem *Tense*-Operator – erklären lässt (Prior 1968).[1]

Gegenüber solchen – auch als ‚Truthmaker-Denying Presentism' (Keller 2004) bezeichneten – Ansätzen wird eingewandt, dass es die wesentlich luzidere Strategie sei, an der grundlegenden Fundierungs-Intuition festzuhalten und im Austausch dafür zu überlegen, ob sich Prämisse (P2) – und damit das betreffende Wahrmacher-Prinzip – modifizieren lässt. Demnach lässt sich die bisherige Explikation der Wahrmacher-Relation durch ein sog. *Supervenienz-Prinzip* der Wahrheit ersetzen.

Ich werde kurz die Gründe benennen, die für eine solche Modifikation sprechen, und anschließend die Frage klären, ob ein Präsentist im Rahmen dieser Strategie bessere Karten hat, das angeführte Wahrmacher-Argument zu entkräften. Das Ergebnis meiner Überlegungen wird jedoch ernüchternd ausfallen.

2. Was spricht für ein modifiziertes Wahrmacher-Prinzip?

Ursprünglich wurde die Idee des Wahrmachens in direkter Anwendung auf eine konkrete (aktuale) Welt verstanden. Man fragt sich, was es heißt – bezogen auf diese einzelne Welt –, dass etwas existieren muss, aufgrund dessen eine bestimmte Proposition wahr ist. Ein Vergleich zwischen verschiedenen Welten, in denen Propositionen sowohl wahr als auch falsch sein können, spielt dabei keine Rolle. Demgegenüber wurde betont, dass sich die zentrale Intuition der Abhängigkeit von Wahrheit und Realität in anderer Weise problemloser verständlich machen lässt. Die *Essenz* der Fundierungs-Intuition beruht darauf, dass es keine Differenz zwischen Welten geben kann, in denen eine Proposition wahr oder falsch ist, ohne dass es eine Differenz bezüglich der Dinge gibt, die in diesen Welten vorkommen. Der damit verbundene Slogan ‚Wahrheit superveniert auf Sein' (Bigelow 1988) lässt sich – mit Bezug auf Notwendigkeit und mögliche Welten – durch das folgende Prinzip (Lewis 2001: 216) wiedergeben:

> *Supervenienz-Prinzip der Wahrheit (SPW)*: Für jede Proposition <p>, Welt ω und ω^* gilt: Wenn <p> in ω wahr ist, aber nicht in ω^*, dann existiert entweder etwas in ω, das nicht in ω^* existiert, oder es gibt irgendein n-tuple von Dingen, das in irgendeiner fundamentalen Relation in ω, aber nicht in ω^* steht.

Es sind vor allem drei Gründe, warum man glaubt, dass (SPW) im Vergleich zu bestehenden Wahrmacher-Prinzipien besser abschneidet. *Erstens* liefert (SPW) eine simple Lösung für das Kardinalproblem jeder Wahrmacher-Theorie. Die Idee des Wahrmachens wird als eine *cross-kategoriale* Relation der Notwendigkeit begriffen, die

[1] Eine etwas anders gelagerte Skepsis gegenüber einem ‚Truthmaker-Denying Presentism' wird von Kierland & Monton (2007) in Form eines sog. ‚brute Presentism' vertreten.

auf alle Arten von kontingenten Propositionen (Wahrmacher-Maximalismus) – also auch auf solche, die negative Prädikationen bzw. negative Existenzannahmen beinhalten – anwendbar sein sollte. Die Wahrmacher für sog. ‚negative Wahrheiten' stellen jedoch ein notorisches Problem dar. Dieses ließe sich im Rahmen von (SPW) einfach dadurch lösen, dass man sagt, es ist die Abwesenheit einer Entität in ω, die, bezogen auf eine mögliche Welt ω*, die betreffende Proposition falsch macht. *Zweitens* verhält sich (SPW) gegenüber der Frage neutral, ob es zur Fundierung strukturierter Propositionen, welche die Form <a ist F> haben, der zusätzlichen Einführung einer fundamentalen Kategorie der Sachverhalte bedarf. Oft wird behauptet, die Wahrheit strukturierter Propositionen verlangt, dass in einer Welt, in der a und F gemeinsam existieren, zusätzlich der Sachverhalt ‚a-ist-F' bestehen muss (Armstrong 2004). Dagegen besagt (SPW) lediglich Folgendes: Eine strukturierte Proposition, die in einer Welt ω wahr ist, wird auch in ω* wahr sein, sofern bezüglich der in ω und ω* vorkommenden Entitäten keine Differenz besteht. Entsprechend bedarf es für die Erklärung der Fundierung von Propositionen des Typs <a ist F> nicht der Existenz von Sachverhalten. *Drittens* scheinen herkömmliche Wahrmacher-Prinzipien das Problem zu haben, eine bestimmte Art der Veränderung nicht erklären zu können. Die Differenz zweier Welten kann – was die Veränderungen der fundierenden Entitäten angeht – manchmal existentieller Natur sein. Eine Entität α existiert in ω, die in ω* nicht vorkommt, einfach deshalb, weil α in ω* aufgehört hat zu existieren. Aufgrund des Wahrmacher-Maximalismus muss es etwas in ω* geben, das der Wahrmacher für die Proposition ‚α existiert nicht in ω*' ist. Eine derartige Entität wäre so beschaffen, dass sie nicht gemeinsam mit α in ω existieren kann. Das bedeutet jedoch, ein Vertreter des Wahrmacher-Maximalismus kann niemals sagen, dass sich zwei Welten ω und ω* allein darin unterscheiden, dass α in ω* nicht mehr existiert. Dass etwas bloß aufhört zu existieren, wäre unerklärlich, denn der Wegfall von α in ω* ist nur im Austausch mit einer positiven Entität – als Wahrmacher für die entsprechende negative Existenz-Proposition – verständlich zu machen.

3. Supervenienz-Präsentismus und das Problem der fundamentalen tensed-Eigenschaften

Trotz dieser vermeintlichen Vorteile ist nicht klar, ob eine präsentistische Zeitkonzeption im Rahmen von (SPW) besser gerechtfertigt ist. Nach der Standardstrategie (Bigelow 1996) sollte man (SPW) akzeptieren und das Wahrmacher-Argument dadurch zurückweisen, dass man (P5) negiert. Ein solcher *Supervenienz-Präsentismus* geht davon aus, dass die gegenwärtige Welt als Ganze – d.h. die Gesamtheit der Dinge, die gegenwärtig existieren – über eine Vielzahl von vergangenheitsbezogenen bzw. zukunftsorientierten Eigenschaften verfügt, auf welche die Wahrheit entsprechender Propositionen superveniert. Demnach wäre es unmöglich, dass eine Welt existiert, welche dieselben *tensed*-Eigenschaft wie die unsere besitzt, aber hinsichtlich der Vergangenheit in irgendeiner Form differiert. Während der Vertreter einer nicht-präsentistischen Zeitkonzeption sagen würde, dass unsere Welt beispielsweise die Eigenschaft hat, eine vergangenheitsbezogene Proposition wie ‚Cäsar überquerte den Rubikon' wahr zu machen, diese Eigenschaft aber fundiert ist in den aktualen Eigenschaften unserer Welt, wie sie zu anderen Zeiten war, behauptet der Präsentist, dass die Wahrheit derartiger Propositionen auf die gegenwärtigen Dinge und *wie diese Dinge sind* superveniert, d.h. auf die Eigenschaften, welche von diesen Dingen gegenwärtig instanziiert werden. Er bestreitet, dass es die Eigenschaften vergangener oder zukünftiger Ereignisse bzw. Individuen sind, welche die Wahrheit vergangenheitsbezogener oder zukunftsorientierter Propositionen begründen.

Sobald man Prämisse (P5) in Zweifel zieht und die Existenz gegenwärtiger Dinge zur grundlegenden Supervenienz-Basis erhebt, muss man einräumen, dass gegenwärtige Dinge über *fundamentale tensed*-Eigenschaften verfügen. Zumindest einige Eigenschaften müssen so sein, dass sie ohne Bezugnahme auf andere Individuen oder noch basalere Eigenschaften auskommen. Fundamentale *tensed*-Eigenschaften sind jedoch im Rahmen einer präsentistischen Ontologie in dreierlei Hinsicht äußerst prekär:

Erstens stellt sich das Problem, wie man erklären soll, dass die nicht-gegenwärtigen Instanziierungen solcher Eigenschaften – die auch ein Präsentist nicht leugnen wird – weder *in* der Vergangenheit noch *in* der Zukunft stattfinden. Verschiedene Versuche, dieses Problem zu lösen, beruhen entweder auf sehr abwegigen ontologischen Zusatzannahmen – wie z.B. nicht-instanziierten individuellen ‚Dasheiten' (Keller 2004) –, oder aber man sieht sich gezwungen, so etwas wie ‚Ersatz-B-series' einzuführen (Crisp 2007), wobei bezweifelt werden darf, dass es sich dann noch um einen strikten Präsentismus handelt (Merricks 2007).

Zweitens stellt sich die Frage, ob nicht in präsentistischer Perspektive wesentliche Vorteile von (SPW) wieder verloren gehen. Der Standardstrategie zufolge ist es hinreichend, ungebundene (‚abundantly') *tensed*-Eigenschaften zu postulieren, welche die alleinige Supervenienz-Basis bilden. Strukturierte Propositionen der Form <a ist F> machen hingegen deutlich, dass es zusätzlich zur Postulierung ungebundener Eigenschaften der Zuordnung von Individuen – als eine Art *subject matter* – bedarf, denen diese Eigenschaften zukommen. Im präsentistischen Rahmen ist es unmöglich, eine solche Zuordnung verständlich zu machen, ohne dabei die Neutralität einzubüßen, nicht gezwungen zu sein, eine fundamentale Kategorie der Sachverhalte einführen zu müssen.

Drittens stellt es generell ein Problem dar, zu glauben, die postulierten *tensed*-Eigenschaften wären in irgendeinem Sinne fundamental. Der ungebundene Charakter dieser Eigenschaften zeigt vielmehr, dass sich es dabei um *irreduzible hypothetische* Entitäten handelt. Derartige Entitäten sind jedoch als Supervervenienz-Basis untauglich, da sie nicht in der Welt verankert sind. Hypothetische Eigenschaften charakterisieren die Dinge nicht, wie sie aufgrund ihrer natürlichen Merkmale in der aktualen Welt existieren, sondern allein anhand ihres bloßen kontrafaktischen Verhaltens in Bezug auf unterschiedliche mögliche Welten. Ontologien, die von derart dubiosen hypothetischen Eigenschaften ausgehen, sind aus einer Wahrmacher-Perspektive ‚betrügerisch' (Sider 2001, Merricks 2007).

4. Asymmetrische Fundierung und globale Supervenienz

Angesichts dieser Einwände ist nicht davon auszugehen, dass die präsentistische Standardstrategie Erfolg hat. Das Scheitern der Standardstrategie bedeutet aber noch nicht, dass es allgemein verfehlt sein muss, Wahrmacher-Prinzipien in Form von (SPW) zu modulieren. Nichtsdestotrotz wirft die bestehende Diskussion ein bezeichnendes

Licht auf die Beantwortung der Frage, ob es adäquat ist, die Wahrmacher-Idee mit Hilfe von (SPW) wiederzugeben.

Der zentraler Grund, warum man glaubt, dass nur ein unbegrenztes Wahrmacher-Prinzip gerechtfertigt ist, basiert auf der Annahme einer *asymmetrischen* Fundierungs-Relation. Offenbar wollen wir sagen, dass eine Proposition <p> wahr ist, aufgrund von p, aber nicht umgekehrt, dass p vorliegt, weil <p> wahr ist. Demgegenüber scheint (SPW) eine Umkehrung zu erlauben. Angenommen, es gibt zwei mögliche Welten ω und ω^*, so dass notwendigerweise gilt: Wenn <p> in ω wahr ist und nicht in ω^*, dann existiert eine Entität α in ω, die nicht in ω^* existiert. In diesem Fall würde ebenso gelten: Sobald man die Wahrheit von <p> in ω fixiert hat, legt man gleichzeitig fest, dass α in ω und nicht in ω^* existiert.

Vertreter eines strikten Wahrmacher-Prinzips (Rodriguez-Pereyra 2005) haben aufgrund dieser Umkehrbarkeit behauptet, dass (SPW) den eigentlichen Kern der Fundierungs-Intuition nicht erfasst. Der zentrale Punkt ist folgender: Mit (SPW) wird klarerweise auf den Begriff der *globalen* Supervenienz Bezug genommen. Dieser Begriff besagt nicht, dass eine Differenz zwischen Welten – Menge der wahren Propositionen in diesen Welten – notwendigerweise in einer Differenz bezüglich des *Seins* dieser Dinge verankert sein muss; behauptet wird lediglich, dass eine Differenz besteht, *wie* die Dinge in der Welt vorkommen. Falls jedoch die Wahrheit einer Proposition nicht unmittelbar darauf superveniert, *ob* etwas existiert, sondern lediglich darin fundiert ist, *wie* bestimmte Dinge vorkommen, dann wird ein zentraler Bestandteil der Wahrmacher-Idee – nämlich die Behauptung, dass zwischen wahren Propositionen und der Beschaffenheit der Welt eine *ontologische* Abhängigkeit besteht – durch (SPW) unterminiert.

Dagegen kann eingewandt werden, dass es äußerst umstritten ist, ob zur Erklärung der Fundierungs-Intuition eine ontologische Abhängigkeitsbehauptung vonnöten ist. Man könnte beispielsweise so argumentieren (Dodd 2007), dass (SPW) zwar als globales Prinzip korrekt ist – Wahrheit superveniert darauf, *wie* die Dinge sind –, aber eben kein striktes Wahrmacher-Prinzip darstellt. Wobei Letzteres gar kein Nachteil ist, da eine adäquate Erklärung der Asymmetrie auch ohne eine ontologische Abhängigkeitsbeziehung auskommt. Demnach ist mit Abhängigkeit nicht das Bestehen einer *Relation* zwischen Propositionen und existierenden Dingen gemeint. Die besagte Asymmetrie basiert nicht – wie ursprünglich angenommen – auf einer modalen Existenzabhängigkeit; vielmehr lässt sich die Fundierungs-Intuition rein begrifflich erklären, nämlich durch das Behaupten einer speziellen Identitäts-Abhängigkeit (Lowe 2004), die innerhalb eines Erklärungsschemas der Form ‚<a ist F> ist wahr, *weil* a ist F' durch die Verwendung des Satzoperators ‚weil' angezeigt wird.

Diese Argumentation lässt zwei Optionen offen: Entweder man lehnt Prämisse (P2) des Wahrmacher-Arguments prinzipiell ab; dann erübrigt es sich, die präsentistische Konzeption mit Hilfe eines modifizierten Supervenienz-Prinzips verteidigen zu wollen. Oder aber man hält daran fest, dass (SPW) irgendwie doch im Sinne der Wahrmacher-Relation zu verstehen ist; dann aber ist man verpflichtet, dieses Prinzip vor dem Hintergrund der angesprochenen Identitäts-Abhängigkeit neu zu bestimmen.

Gegenüber der zweiten Option lässt sich allerdings Folgendes einwenden: Eine derartige Modifizierung wäre kontraproduktiv, da sie nicht auf einem globalen Verständnis von Supervenienz basiert, sondern einen strengeren Begriff erfordert. Ein strenger Supervenienz-Begriff würde jedoch zentrale Aspekte der Motivation für (SPW) zunichte machen; er wäre weder mit der bestehenden Erklärung für sog. ‚negative Wahrheiten' vereinbar, noch ließen sich damit existenzielle Veränderungen – im Sinne des bloßen Vergehens einer Entität – erklären.

Alles in allem scheint es daher keine wirksame Strategie zu sein, das angeführte Wahrmacher-Argument durch eine Modifizierung von Prämisse (P2) entkräften zu wollen. Selbst wenn sich die Annahme einer Fundierungs-*Relation* adäquat begründen lässt, ist damit nicht viel für die Rechtfertigung von (SPW) gewonnen. Wie die obige Diskussion gezeigt hat, kann ein globales Verständnis von (SPW) nur dann seine Vorteile ausspielen, sofern man eine *nicht-päsentistische* Zeitkonzeption unterstellt. Eine derartige Einschränkung ist aber nicht akzeptabel; denn ein *allgemeingültiges* Wahrmacher-Prinzip sollte nicht davon abhängen, ob eine bestimmte Auffassung über die Natur der Zeit korrekt ist oder nicht.

Literatur

Armstrong, D. M. 2004 *Truth and Truthmakers*, Cambridge.

Bigelow, J. 1996 “Presentism and Properties”, *Philosophical Perspectives 10*, 35-50.

Crisp, T. 2007 “Presentism and The Grounding Objection”, *Nous 41*, 18-137.

Dodd, J. 2007 “Negativ Truth and Truthmaker Principles”, *Synthese 156*, 383-401.

Keller, S. 2004 “Presentism und Truthmaking“, in: D. Zimmerman, *Oxford Studies in*

Metaphysics I, Oxford, 83-104.

Kierland, B./Bradley, M. 2007: “Presentism and the Objection from Being-Supervenience”,

Australasian Journal of Philosophy 85, 485-497.

Lewis, D. 2001 “Truthmaking and Difference-Making”, *Nous 35*, 602-615.

Lowe, E. J. 1994 “Ontological Dependency”, *Philosophical Papers 23*, 31-48.

Meinong, A. 1904 *Untersuchungen zur Gegenstandstheorie und Psychologie*, Leipzig.

Merricks, T. 2007 *Truth and Ontology*, Oxford.

Prior, A. 1967 *Papers on Time and Tense*, Oxford.

Rodriguez-Pereyra, G. 2005 “Why Truthmakers”, in: H. Beebee and J. Dodd (eds.),

Truthmakers, Oxford, 17-32.

Sider, T. 2001 *Four-Dimensionalism*, Oxford.

Tooley, M. 1997 *Time, Tense and Causation*, Oxford.

Reduction, Sets, and Properties

Benjamin Schnieder, Berlin, Germany

1. The Traditional Debate

Some time ago, philosophers often addressed the question about whether it is possible to reduce the category of properties to the category of sets.

The most straightforward idea was that, schematically speaking, we may identify the property of being *F* with the set of all *F*s. The straightforward problem of this proposal was that if conflicts with the combination of the generally accepted identity conditions for sets and some strong intuitions on the non-identity of certain properties: on the one hand, sets are *extensional* in that a set *x* is identical to a set *y* iff *x* has all and only the members that *y* has. On the other hand, two different properties can contingently be possessed by the same entities. Even if all and only those animals that naturally have a heart naturally have a liver, the property of naturally having a heart is not identical to the property of naturally having a liver. And even though the property of being an 80 feet long duckbill and the property of being a 5 billion dollar worth corkscrew are both exemplified by no entities at all, they are not the same property.

Because of this problem, it is generally agreed that the said straightforward identification of properties with sets is a failure. But a very similar identification is better off: David Lewis famously proposed to identify the property of being *F* with the sets of all *actual or merely possible F*s.[1] Since (presumably) no non-actual possible duckbill is a non-actual corkscrew, properties in the second example come out different on Lewis's account, and since some non-actual animals with a heart lack a liver, the properties in the first example come out different as well.

But there are still problems with Lewis's proposal: first, it may seem to correspond to a too coarse-grained individuation of properties: on his account, there are no *two* properties that are necessarily possessed by the same entities. But isn't the property of being an equilateral triangle different from the property of being an equiangular triangle, despite their being necessarily co-exemplified? However, there is a promising response available to this third challenge: in the relevant cases, our intuitions on the identity or non-identity of properties are not that stable, and they are blurred by our tendency to assimilate properties to concepts.[2] Concepts are individuated partly via the role they play in the individuation of beliefs; for this reason, we often distinguish between two concepts even if they have the same extension with respect to every possible world (someone may possess the concept of equilateral triangles but lack the concept of equiangular triangles because he lacks the concepts of angles). But properties play a more worldly role such that different properties should at least possibly correspond to differences between things; however, in no possible world there is any difference between equilateral and equiangular triangles, and hence there is only one property which is conceived of in different ways (i.e. by the employment of different concepts).

But there are other problems with Lewis's proposal. One is that it apparently cannot provide for all the properties that intuitively exist:[3] there is, it seems, the property of being a set. But this property cannot be identified with the set of all (actual and non-actual) sets, because we know from standard set-theory that there is no such entity. Lewis might at this place slightly modify his proposal and hold that *most* properties are sets, but *some* properties are proper classes rather than sets. The property of being a set, for instance, is the proper class of all sets. But firstly that response may seem a little *ad hoc*, and secondly it only pushes the problem up to a higher level: intuitively, there is the property of being a proper class. But there is no proper class containing all the proper classes. So, some properties are still missing on Lewis's account.

Finally, the proposal only gets off the ground if one accepts Lewis's form of modal realism which can provide the required ontological resources: for the proposal to work, there must be sets containing actual and non-actual duckbills as members and hence there must be non-actual duckbills (which are duckbills nonetheless).[4] Especially due to this metaphysical burden of his approach, Lewis's identification of properties with sets has not a very good reputation, to say the least.

2. A New Order: Sets as Properties

If properties cannot be reduced to sets, is there no connection between the two kinds of entity? There is! The main problem of the traditional debate was that it reversed the actual order of things: instead of identifying properties with sets, one should identify sets with properties. Sets are properties in disguise (but not *every* property is a set).

The basic idea is very simple: sets are properties of some particular sort. They are what I call *identity-properties*, such as the property of being identical to Jean-Paul Belmondo or the property of being identical to either Belmondo or Jean Seberg.

To understand the details of the proposal, recall the two standard ways of specifying a sets: they consist in either (i) providing a list of the individual members of the set—by using expressions such as '{Belmondo}', '{Belmondo, Seberg}', '{1, 2, 3, …}'—, or (ii) stating a condition such that all and only the things satisfying the condition are members of the set—by using expressions such as '{*x*: *x* is a bear}', or 'the set of all duckbills'. Focussing on these two canonical ways of specifying sets, a recipe can be given for identifying particular sets with particular identity-properties.

Re (i): If we specify a set by explicitly listing its members, the set is just the property of *being (identical to) one of those entities*. Thus, {Belmondo} is the property of being (identical to) Belmondo. {Belmondo, Godard} is the property of being (identical to) either Belmondo or Godard. More generally, {x_1, x_2, …, x_n} is the property of being (identical to) either x_1, or x_2, …, or x_n.

[1] See Lewis (1986: 50–69).

[2] For responses along this line, see, e.g., Lewis (1986: 55), Jackson (1998: 15f., 126f.), Künne (2003: 26, *et passim*), and Schnieder (2004: 59–69).

[3] For the following point cp. Egan (2004: 49n.), and Schnieder (2004: 72f.).

[4] For further criticism of Lewis's account see Egan (2004).

Re (ii): If we specify a set via some condition *C* that its members satisfy, the set is the property of being one of those entities that *actually* satisfy condition *C*. Thus, {*x*: *x* is a word on this page} is the property of being one of the words that are actually on this page. We can specify this set, and this property, in some alternative way by listing all its members. Thus, {*x*: *x* is a word on this page} = {'a', 'the', 'set', 'words', 'are', …} = the property of being either the word 'a', or the word 'the', or the word 'set', … Similarly, {*x*: *x* is a bear}, i.e. the set of all bears, is identified with the property of being identical to one of the actual bears.

3. Evaluating the Proposal

To evaluate the given proposal, it should be seen whether it can account for the generally acknowledged features of sets. On the one hand, some central metaphysical characteristics of sets should be accounted for:

M.1 Sets have their members essentially and they could not have had additional members.
M.2 Sets are extensional. For all sets *x*, *y*: *x* = *y* iff every member of *x* is a member of *y* and *vice versa*.
M.3 Sets are ontologically dependent upon their members.

These three principles are results of the indicated identification of sets with identity-properties (given some additional assumptions about properties).

Re M.1: Identity-properties are necessarily possessed by all and only those entities that possess them. So, if a set is an identity-property, and if its members are the things that posses it, then a set has its members essentially and could not have had additional members.

Re M.2: Moreover, the intensional individuation of properties (i.e. the individuation via their possible exemplifications, as described in section 1) accounts for the extensionality of sets: if set *x* and set *y* contain the same members, then *x* and y are reduced to the same identity-property (because *x* is reduced to an identity-property *I* and y to an identity-property *I** such that *I* and *I** are necessarily possessed by the same entities—hence, by the intensional individuation of properties, *I*=*I**).

Re M.3: an identity property *I* is individuated via the entities *e*, … that enter into the identities constitutive for *I*. Hence, the essence or nature of *e*, … is prior to the essence of *I*, which makes the identity property dependent upon its constitutive entities and thereby yields the ontological dependence of sets on their members.[5]

Moreover, the semantics of designators for sets comes out right: some of those expressions are rigid designators (e.g. '{Belmondo}') while others are flexible (e.g. '{*x*: *x* is a bear}'). On the current proposal, a set designator will be rigid if it rigidly specifies the members of the set, because then it rigidly specifies a particular identity-property. But if a set designator flexibly specifies the members of the designated set, then the designator will designate different identity-properties with respect to different possible worlds and therefore comes out flexible itself.

A further issue to de addressed is whether the proposal may open doors to set-theoretical paradoxes. A threat of paradox may stem from the following observation:

M.4 Sets have a limited holding capacity, i.e., there can be too many things of a sort be contained in a set.

Properties, on the other hand, do not seem to have a similarly limited holding capacity. For while we know that there are, for instance, too many sets to form a set, there is a property which all sets have in common, namely the property of being a set. Analogously, there is the property of being abstract which is possessed by all abstract objects and therefore by all sets. But again, there cannot be the set of all abstract objects because it would have the set of all sets as a subset.

How exactly might the limited holding capacity of sets pose a problem for the suggested reduction of sets to properties? Answer: *if* there are identity-properties belonging to more things than can form a set, then the proposal should not hold that *all* identity-properties are sets. But then, a criterion should be provided which motivates why some identity-properties are sets while others are not, and it is hard to see how such a criterion might not be *ad hoc*.

Fortunately, it can be argued that there are no identity-properties which would yield too large sets in the first place. For which properties could that be? As we have seen, the property of being a set cannot correspond to a set because the latter would be too large. But that property is not an identity-property and therefore not the kind of property that the proposal identifies with sets. What would be threatening is the identity-property of being identical to one of the actual sets. But while this property may seem innocent and thus existing at first glance, there is a reason to deny its existence. For it was said before that an identity-property is individuated by the entities that enter into its constitutive identities, such that the essence of those entities is prior to the essence of the property. Now assume that the identity-property of being one of the actual sets would exist. Then, on the present reductive theory, it would be a set, and hence one of the existing sets. But its essence should be posterior to the essence of its constituting entities, i.e. of all actual sets. Since it would be a set itself, its essence would have to be posterior to its own essence—which is impossible. Therefore, the existence of the said property can be denied. A lesson from this consideration is that—contrary to prototypical examples of properties—*identity-properties* do have a limited holding capacity, just like sets.

The last test for the proposal will be to account for the theorems of standard set theory. It would be desirable if the axioms of ZF-theory could be motivated from it. Unfortunately, the current space is too limited to go into details here. So I must conclude with the bare promise that the validity of ZF-axioms such as *intersection* can indeed be argued for from considerations about identity-properties. Once this is shown, the proposal can be seen to have numerous benefits which allow the conclusion: sets are properties.

[5] In this paragraph, I am mainly relying on Fine's views on essence and dependence; see Fine (1994 & 1995).

Literature

Egan, Andy (2004): ‘Second-Order Quantification and the Metaphysics of Properties’, *Australasian Journal of Philosophy* **82**, 48–66.

Fine, Kit (1994): ‘Essence and Modality’, in: *Philosophical Perspectives* **8**, 1–16.

Fine, Kit (1995): ‘Ontological Dependence’, *Proceedings of the Aristotelian Society* **95**, 269–90.

Jackson, Frank (1998), *From Metaphysics to Ethics*, Oxford: Clarendon Press.

Künne, Wolfgang (2003): *Conceptions of Truth*, Oxford: Clarendon Press.

Lewis, David (1986): *On the Plurality of Worlds*, Oxford: Blackwell.

Schnieder, Benjamin (2004): *Substanzen und (ihre) Eigenschaften*, Berlin: de Gruyter.

Context-Based Approaches to the Strengthened Liar Problem

Christine Schurz, Salzburg, Austria

Kripke´s paper (Kripke 1975) was the starting point for numerous formal languages that are able to express a theory of their own concept of truth without producing a liar paradox. However these theories are not regarded as modeling the concept of truth of natural language. The main reason for this is the strengthened liar paradox, or rather a formalized version of this problem. Whether it is possible to solve this problem without being committed to something (more or less) like Tarski´s hierarchical proposal (Tarski 1935), is still a highly disputed matter. The majority of those who seek for alternatives to Tarski´s proposal attempt to analyze the strengthened liar problem by uncovering certain context-dependent elements in a strengthened liar sentence.

This paper consists of two parts. First, I shall outline a formal version of the strengthened liar. Thereafter I take a look at the context-based approach to this problem.

From now on, I shall assume a formal language *L* of first order logic that is interpreted by a structure *M. M* is assumed to contain names that are assigned to the sentences of *L*, and functions that represent a certain amount of the syntax of *L*. The symbol *Tr* is a one-place predicate of *L* that is intended to represent "truth in *L*". By the fixed-point-lemma a strengthened liar sentence, i.e., a sentence φ of the form $\neg Tr(\mathrm{l})$ such that $M \models \mathrm{l} =$ '$\neg Tr(\mathrm{l})$' can be formed (the term '$\neg Tr(\mathrm{l})$' is the object-language-name of the formula $\neg Tr(\mathrm{l})$). I shall assume a theory *T* of "truth in *L*" such that neither $T \vdash \varphi$ nor $T \vdash \neg\varphi$. Such a theory results e.g. from the axioms of (Friedman and Sheard 1987) which define *Tr* together with some axioms of a theory of *M* (if, for instance, *M* is a model of arithmetic, then we can take the axioms of Robinson´s arithmetic).

The theory *T* matches with the intuition that the strengthened liar sentence is neither true nor false in as much as it gives us no information concerning the truth-value of φ. But in context of natural language we are also capable of expressing this intuition by a meaningful sentence and furthermore infer other statements from such a sentence (which together lead to the natural language strengthened liar argument). Thus, if we seek *T* to be as close as possible to the concept "is true" of natural language, we have to investigate whether it is possible for *T* to contain any formula of *L* which represents a semantical diagnosis as "φ is neither true nor false" of φ. Is it possible to add such a formula to the axioms of *T* without causing *T* to be inconsistent? This is indeed an intricate problem.

The first problem: Usually the formula $\neg Tr($'φ'$)$ is taken to be a most appropriate candidate to represent a diagnostic statement about φ. Since φ fails to be derivable from *T,* it also fails to be true according to *T* (otherwise it would be derivable from *T*), which can in *L* be represented by the formula $\neg Tr($'φ'$)$. So $\neg Tr($'φ'$)$ should be derivable from *T*. But, on the other hand, we have a "falsity-intuition" concerning $\neg Tr($'φ'$)$, taking $\neg Tr($'φ'$)$ to represent the sentence "φ is false". Since *T* gives us no information whether φ is true or false, also no formula representing "φ is false" will be derivable from *T*. Therefore $\neg Tr($'φ'$)$ should not be derivable from *T*. So in the end we have a "failure-intuition" and a "falsity-intuition" concerning the sentence $\neg Tr($'φ'$)$. Both intuitions display features of how we informally reason about the strengthened liar in natural language. So they do not have to be considered as rival intuitions of which we have to select only one that represents our "actual" reasoning or the way we should rationally reason. Maybe, if *L* is supposed to account for natural language, one has to find a way to model both of these intuitions and explain how $T \vdash \neg Tr($'φ'$)$ as well as not $T \vdash \neg Tr($'φ'$)$ can be the case. But obviously this can only be consistently realized if we assume some context-sensitive element in the sentence $\neg Tr($'φ'$)$.

Before I turn to the second problem, let me note that one could object against the analysis I just gave that it shows that $\neg Tr($'φ'$)$ is after all no appropriate choice for a formula representing a semantical diagnosis of φ. But then the only alternative is to introduce a new "untrueness"-predicate *U* to express a formula belonging to *T* that represents a statement about φ. But of course this gives rise to a new strengthened liar sentence φ' of the form *U*(l') such that $M \models \mathrm{l}' =$ '$U(\mathrm{l}')$' that leads to just the same problems. So in the end we would have just the same conflict of failure- against falsity-intuition.

The second problem: This problem results from our attempt to model the 'failure-intuition' for $\neg Tr($'φ'$)$. According to this attempt, *T* contains the sentence $\neg Tr($'φ'$)$, but, on the other hand, it should not contain φ. These conditions contradict to the rule of substitutivity of identity (I shall from now on write "(SoI)" to refer to this rule):

> (SoI) Substitutivity of identity:
> Let t_1 and t_2 be any two terms, let *P* be any one-place predicate. Then the following rule is valid:

$$\frac{t_1 = t_2,\ P(t_1)}{P(t_2)}$$

We have to make a decision whether to give up (SoI) or again put forward an explanation in terms of a context-shift. We might even again employ the context-shift we have already posited in "$\neg Tr($'φ'$)$" occurring in the argument based on the failure-intuition and the same formula occurring in the argument based on the falsity-intuition. Then we not only suppose two different contextual interpretations of "$\neg Tr($'φ'$)$", but also two different contextual interpretations of φ. Indeed some proponents of contextual approaches (e.g. (Burge 1979) and (Simmons 1993)) argue for another tension between two semantical views of the strengthened liar sentence φ. But this point is controversial. After all, φ is in the first place assumed to be an entirely pathological sentence to which no classical truth value can be assigned. $\neg Tr($'φ'$)$ is intended to reflect this assumption. So if we finally refute that φ is entirely pathological but consider φ to be also true according to another reading, then we refute the basic observation of our whole reasoning.

To conclude, in context of our formal language *L*, the strengthened liar paradox is constituted by (at least) two tensions in our metalinguistic reasoning about φ. Firstly, the tension between evaluating $\neg Tr($'φ'$)$ as true according the failure-intuition, and evaluating it as neither true nor false according to the falsity-intuition. Secondly, there is a tension between evaluating ¬*Tr*(‘φ’) and φ equally, and evaluating ¬*Tr*(‘φ’) and φ differently. The first tension indicates that ¬*Tr*(‘φ’) is context-dependent. Concerning the second tension we have two possible lines of explanation available. We can either again trust into effects of a context-shift, but we can alternatively argue against the rule (SoI).

On a first assessment one might for economical reasons decide to make an effort to solve the problem by employing an explanation in terms of context-shift only. I shall simply call this the context-based approach in order to distinguish it from the other strategies. Generally, any context-based approach has to clarify what kind of context-dependence is present in the strengthened liar reasoning. It has to determine where a context-dependent element is hidden in ¬*Tr*(‘φ’).

All context-based approaches agree that a certain change of context affects our interpretation of the truth predicate *Tr* occurring in φ. A very direct realization of this comes from (Burge 1979). He proposes to add an index i to each occurrence of *Tr*, i.e. “Tr_i”. This reminds very much of Tarski´s proposal, but Burge`s system has several advantages. Formulas, such as “Tr_i(‘Tr_i(‘P(c)’)’)” where “P(c)” represents “Snow is white”, that are meaningless or ill-formed according to Tarski´s account, become meaningful within Burge´s system. Further developments of Burge´s approach are (Koons 1992) and (Simmons 1993).

A metaphysically more sophisticated approach is posed by (Parsons 1974), (Barwise and Etchemendy 1987) and (Glanzberg 2004). According to Parsons and Glanzberg, a context-shift can lead to a shift from *M* to another *L*-structure *M** which is regarded to be an expansion of *M*. They take propositions as truth-bearers, and define an expression relation between sentences and propositions. The universe of discourse *D* of their model *M* is proposed to contain all propositions that can be expressed by L^{M} (in Glanzberg´s system, *D* actually contains all truth conditions that can constitute a proposition and that can be expressed by L^{M}). The universe can expand as context shifts. In the first place, therefore, quantifiers represent the context-dependent element in (Glanzbergs and Parsons versions of) the strengthened liar sentence. But as Glanzberg and Parsons note, it is then only a definitional matter to work with a context-dependent truth predicate, and in fact both end up with working with a context-dependent truth predicate. (Barwise and Etchemendy 1987) give a very similar account.

Context-based approaches employ an analysis that inevitably gives rise to a certain kind of hierarchy of truth predicates. This is obvious for proposals which attach an index belonging to a partially ordered set to all occurrences of “*Tr*”. Approaches that appeal to an expansion *M** of *M* with a more comprehensive domain D* lead to a hierarchical view as well. The language L^{M*} interpreted by *M** is used to express a true statement about φ' (the “ ' ” indicates that φ' is actually a translation of the original φ belonging to L^{M} into the language L^{M*}): *M** ├ *Tr**(‘φ'’), and furthermore *M** ├ ¬*Tr*(‘φ'’), where the predicate *Tr* represents “truth in L^{M}” and *Tr** represents “truth in L^{M*}”. A new strengthened liar sentence φ* of the form ¬*Tr**(l) such that l = ‘¬*Tr**(l)’ can be formed that gives rise to another context-shift that is associated with another model *M*** to reason about φ*. *M*** in turn has its own truth-predicate *Tr*** and so on and so forth.

Of course the hierarchies advanced by context-based approaches differ from Tarski´s hierarchies of truth-predicates in several respects. First, as I have already mentioned, this approach features some advantages, since several cases of formulas that correspond to ordinary and meaningful sentences in English, but that become pathological in Tarski´s system, are evaluated as our intuitions would suggest it. Furthermore, these alternative hierarchies are considered to model natural language as closely as this can possibly be done, or even to represent an intrinsic feature of natural language. The strengthened liar problem does not show that the concept of truth in natural language is inconsistent, but that the predicate “is true” is context-dependent.

Still it is prima facie undesired that the concept of truth is context-dependent. So context-based approaches have to give a motivation for this context-dependency. A very essential question is, what kind of context-shift happens to effect a shift in the extension of *Tr* and *how* it does this. Earlier proponents of context-based approaches do not employ any rigid formal concept of a context. Therefore they only give rough and informal explanations for this problem, if at all. Sometimes the presence of the strengthened liar problem is taken to provide enough evidence to prove a context-shift. In contrast, (Glanzberg 2004) gives a rigid proof of how a context-shift leads to a shift from *M* to *M**, which is based on precise notions of a context, a proposition and a truth predicate.

An additional problem which I think deserves attention apart from any profound explanation concerning the previous question is that in the languages proposed there is no way to reflect over all contexts. No formula can be used to represent an ultimate or universal perspective. Proponents of a contextual analysis have to clarify why this is so, and why our naive view that we can reflect over absolutely everything is wrong.

Apart from these two problems a context-based approach also has to motivate why it fares better than alternative approaches to explain the strengthened liar problem. Given all the difficulties related with a hierarchical account I have just outlined, it could e.g. be that contrary to our first assessment an alternative approach that relies in addition on restricting (SoI) in the end provides a better trade-off than a purely context-based approach. It has sometimes been argued against restricting (SoI) that this line of solution was ad-hoc and unmotivated. For the most part this point applies to approaches put forward so far (e.g. (Skyrms 1970)), since they are not very elaborated concerning their applicability to natural language (besides, they do not invoke any contextual considerations, and therefore do not explain the first problem from which the strengthened liar problem arises). Therefore it could be interesting to search for other kinds of contextual approaches and turn our attention to (SoI).

Literature

Barwise, Jon and Etchemendy, John 1987 *The Liar*, Oxford: Oxford University Press.

Burge, Tyler 1979 *Semantical Paradox*, Journal of Philosophy 76.

Friedman, Harvey and Sheard, Michael 1987 *An Axiomatic Approach to Self-Referential Truth*, Annals of Pure and Applied Logic 40.

Gaifman, Haim 1992 *Pointers to Truth*, Journal of Philosophy 89.

Glanzberg, Michael 2004 *A Contextual-Hierarchical Approach to Truth and the Liar Paradox*, Journal of Philosophical Logic 33.

Koons, Robert C. 1992 *Paradoxes of Belief and Strategic Rationality*, Cambridge: Cambridge University Press.

Kripke, Saul 1975 *An Outline of a Theory of Truth*, Journal of Philosophy 72.

Parsons, Terence 1974 *The Liar Paradox*, Journal of Philosophical Logic 13.

Simmons, Keith 1993 *Universality and the Liar*, Cambridge: Cambridge University Press.

Skyrms, Brian 1970 *Return of the Liar: Three-Valued Logic and the Concept of Truth*, American Philosophical Quarterly 7.

Tarski, Alfred 1935 *Der Wahrheitsbegriff in formalisierten Sprachen*, Studia Philosophica 1.

The Elimination of Meaning in Computational Theories of Mind

Paul Schweizer, Edinburgh, Scotland, UK

1. The Computational Paradigm

According to the traditional conception of the mind, semantical content is perhaps the most important feature distinguishing mental from non-mental systems. For example, in the scholastic tradition revived by Brentano (1874), the *essential* feature of mental states is their 'aboutness' or intrinsic representational aspect. And this traditional conception has been incorporated into the foundations of contemporary scientific approaches to the mind, insofar as the notion of 'mental representation' is adopted as a primary theoretical device. For example, in classical (e.g. Fodorian) cognitive science, Brentano's legacy is preserved in the view that the properly cognitive level is distinguished precisely by appeal to representational content. There are many different levels of description and explanation in the natural world, from quarks all the way to quasars, and according to Fodor (1975), it is only when the states of a system are treated as representations that we are dealing with the genuinely cognitive level.

The classical paradigm in cognitive science derives from Turing's basic model of computation as rule governed transformations on a set of syntactical elements, and it has taken perhaps its most literal form of expression in terms of Fodor's Language of Thought hypothesis (LOT), wherein mental processes are explicitly viewed as formal operations on a linguistically structured system of internal symbols. But a fundamental tension is already built into the classical picture: a central purpose of the symbolic structures is to carry content, and yet, to the extent that they are formal elements of computation, their alleged content is completely gratuitous. Computation is essentially a series of manipulations performed on *uninterpreted* syntax, and formal structure alone is sufficient for all effective procedures. The specification and operation of such procedures makes no reference whatever to the intended meaning of the symbols involved. Indeed, it is precisely this limitation to syntactic *form* that has enabled computation to emerge as a mathematically rigorous discipline. If syntax alone is not sufficient, and additional understanding or interpretation is required, then the procedure in question is, by definition, not an effective one. But then the purported content of mental 'representations' is rendered superfluous to the computations that comprise the 'cognitive' processes of cognitive science. The intended interpretation of internal syntax makes absolutely no difference to the formal mechanics of mind.

For a number of years now there has been a high profile struggle between opposing camps within the computational approach to the mind. In contrast to the classical paradigm derived from Turing, connectionist systems are based on networks of large numbers of simple but highly interconnected units that are brain-like in inspiration. But according to Fodor, the brain-like architecture of connectionist networks tells us nothing about their suitability as models of *cognitive* processing, since it still leaves open the question of whether the mind is such a network at the representational level. So a number of connectionists have taken up the challenge and seek out ways of projecting representational content onto artificial neural networks. One comparatively recent such attempt (Churchland, P.M.1998, Laakso, A. and G. Cottrell 2000, O'Brien, G. and J. Opie 2001) uses cluster analysis to locate 'vehicles' of representational content within artificial neural networks, where such clusters serve as surrogates for the classical notion of internal syntax.

However, I would contend that such attempts suffer from exactly the same built-in tension that afflicts the LOT model; namely, the purported content for which the clusters serve as vehicles does no work in the processing path leading from inputs to outputs. Just as in the classical case, the postulation of content within the connectionist framework is gratuitous, because it plays no role in the cognitive manipulation of inputs to yield the salient outputs. Indeed, if content weren't gratuitous, then computational versions of cognitive processing would be lamentably deficient in terms of their specification of the inputs. These are characterized solely in formal or syntactical terms, and content is entirely absent from the external stimuli recognized by the operations that can be defined within the model. If representational content were at all relevant, then cognitive systems would have to process content *itself*. But according to computational methods, content is not specified with the input, nor does it play any efficacious role in internal processing. So, from a perspective that takes computation as the theoretical foundation for cognition, it seems quite retrograde to posit content on top of the factors that do the actual work. Surely this is an exemplary occasion for invoking Ockham's razor.

2. Searle's Objection

Of course, John Searle's (1980) celebrated Chinese Room Argument (henceforward CRA) runs the dialectic in exactly the reverse direction: rather than taking the formal, syntactic nature of computation as a reason for eschewing content in a properly naturalistic approach to the mind, Searle instead takes it as a reason for rejecting computation as the appropriate theory of the mental. So, from the perspective of the present discussion, it is instructive to explicitly cast Searle's argument in terms of the separability of syntactical structure from its intended meaning. In what follows I will abstract away from the somewhat picturesque details of Searle's original version and express the logical core of the CRA via two premises and a conclusion:

(1) semantical content is an essential feature of the mind,

(2) syntactical manipulations cannot capture this content, therefore

(3) the mind cannot be reduced to a system of syntactical manipulations.

Preimse (1) is an expression of the traditional conception of the mind, and is accepted by both Searle and by his opponents in orthodox cognitive science and AI. Classical cognitive science and AI view the mind according to the model of rule governed symbol manipulation, and premise (1) is embraced insofar as the manipulated symbols are supposed to possess representational content. Searle's dispute with cognitive science and AI centers on his rejection of the idea that internal computation can shed any real light on mental content, which leads to his conclusion (3), and a concomitant dismissal of the research paradigm central to cognitive science and AI.

In response, a standard line for defenders of this paradigm is to try and defuse the CRA by arguing against premise (2), and claiming that the manipulated symbols really do possess some canonical meaning or privileged interpretation. However, I would urge that this is a serious strategic error for those who wish to defend the computational approach. As stated above, a distinguishing mathematical virtue of computational systems is precisely the fact that the formal calculus can be executed without any appeal to meaning. Not only is an interpretation intrinsically unnecessary to the operation of computational procedures, but furthermore, there is no unique interpretation determined by the computational syntax, and in general there are arbitrarily many distinct models for any given formal system.

Computational formalisms are syntactically closed systems, and in this regard it is fitting to view them in narrow or solipsistic terms. They are, by their very nature, independent of the ‘external world’ of their intended meaning and, as mentioned above, they are incapable of capturing a unique interpretation, since they cannot distinguish between any number of alternative models. This can be encapsulated in the observation that the relation between syntax and semantics is fundamentally *one-to-many*; any given formal system will have arbitrarily many different interpretations. And this intrinsically one-to-many character obviates the possibility of deriving or even attributing a unique semantical content merely on the basis of computational structure.

The inherent limitations of syntactical methods would seem to cast a rather deflationary light on the project of explicating *mental content* within a computational framework. Indeed, they would seem to render hopeless such goals as providing a computational account of natural language semantics or propositional attitude states. Non-standard models exit even for such rigorously defined domains as first-order arithmetic and fully axiomatized geometry. And if the precise, artificial system of first-order arithmetic cannot even impose isomorphism on its various models, how then could a *program*, designed to process a specific natural language, say Chinese, supply a basis for the claim that the units of Chinese syntax posses a *unique* meaning?

So I think that the advocates of computation make the wrong move by accepting Searle’s bait and taking on board the attendant ‘symbol grounding problem’ endemic to computational theories of mind. Instead I would accept Searle’s negative premise (2) and agree that computation is too weak to underwrite any interesting version of (1). Hence I would concur with Searle’s reasoning to the extent of accepting the salient *conditional* claim that *if* (1) is true *then* (3) is true as well. So the real crux of the issue lies in the truth-value of (1), without which the consequent of the *if-then* statement cannot be detached as a free-standing conclusion. Only by accepting the traditional, *a priori* notion of mentality assumed in premise (1), does (3) follow from the truth of (2). And it’s here that I diverge from the views of both Searle and orthodox cognitive science.

3. Representation as Heuristics

There have been a number of prominent positions advanced in negative reaction to ‘classical’ cognitive science that take anti-representationalism as one their hallmarks, including dynamical systems theory (e.g Van Gelder 1996), behaviour based robotics (e.g. Brooks 1991), approaches utilizing sensory-motor affordances (e.g. Noe 2004), who campaign on the platform of ‘intelligence without representation’. In order to locate my position on the philosophical landscape, it is salient to note that it is *not* anti-representational in this sense. On my view, there could well be internal structures that play many of the roles that people would ordinarily expect of representations, and this is especially true at the level of perception, sensory-motor control and navigation – things like spatial encodings, somatic emulators, internal mirrorings of salient aspects of the external environment. So, unlike the anti-representationalists, I do not deny that there may be internal structures and stand-ins that various people would be tempted to *call* ‘representations’.

But I would argue that this label should be construed in a weak, operational sense, and should not be conflated with the more robust traditional conception. To the extent that internal structures can encode, mirror or model external objects and states of affairs, they do so via their own causal and/or syntactic properties. And again, to the extent that they influence behaviour or the internal processing of inputs to yield outputs, they do this solely in virtue of their causal and/or syntactic properties. There is nothing about these internal structures that could support Searle’s or Brentano’s notion of original intentionality, and there is no independent or objective fact of the matter regarding their ‘real’ content or meaning.

The crucial point to notice is that these internal ‘representations’ do all their scientifically tangible *cognitive* work solely in virtue of their physical/formal/mathematical structure. There is nothing about them, qua efficacious elements of internal processing, that is ‘about’ anything else. Content is not an explicit component of the input, nor is it acted upon or transformed via cognitive computations. All that is explicitly present and causally relevant are computational structure plus supporting physical mechanisms, which is exactly what one would expect from a naturalistic account. In order for cognitive structures to do their job, there is no need to posit some additional ‘content’, ‘semantical value’, or ‘external referent’. Such representation talk may serve a useful heuristic role, but it remains a conventional, observer-relative ascription, and accordingly there’s no independent fact of the matter, and so there isn’t a sense in which it’s possible to go wrong or be mistaken about what an internal configuration is ‘really’ about. Instead, representational content is projected onto an internal structure when this plays an opportune role in characterizing the overall processing activities which govern the system’s interactions with its environment, and hence in predicting its salient input/output patterns. But it is simply a matter of convenience, convention and choice.

From the point of view of the system, these internal structures are manipulated directly, and the notion that they are ‘directed towards’ something else plays no role in the pathways leading from cognitive inputs to intelligent outputs. Hence the symbol grounding problem is a red herring – it isn’t necessary to quest after some elusive and mysterious layer of content, for which these internal structures serve as the syntactic ‘vehicle’. Syntactical and physical processes are all we have, and their efficacy is not affected by the presence or absence of meaning. I would argue that the computational paradigm is thematically inconsistent with the search for content or its supposed vehicles. Instead, computational models of cognition should be concerned only with the *processing structures* that yield the right kinds of input/output profiles, and with how such structures can be implemented in the brain. These are the factors that do the work and are sufficient to explain all of the empirical data, and they do this using the normal theoretical resources of natural science. Indeed, the postulation of content as the essential feature distinguishing mental from non-mental systems should be seen as the last remaining vestige of Cartesian dualism, and, contra Fodor, naturalized cognition has no place for a semantical 'ghost in the machine'. When it comes to computation and content, only the vehicle is required, not the excess baggage.

Literature

Brentano, F. 1874 *Psychology from an Empirical Standpoint*.

Brooks, R. 1996 "Intelligence without Representation" in *Mind Design II*, J. Haugeland (ed.), MIT Press.

Churchland, P.M. 1998 "Conceptual Similarity Across Sensory and Neural Diversity: The Fodor/Lepore Challenge Answered", *Journal of Philosophy*, 96(1): 5-32.

Fodor, J. 1975 *The Language of Thought*. Harvester Press.

Laakso, A. and G. Cottrell 2000 "Content and Cluster Analysis: Assessing Representational Similarity in Neural Systems", *Philosophical Psychology*, 13(1): 47-76.

Noë, A. 2004 *Action in Perception*, MIT Press.

O'Brien, G. and J. Opie 2001 "Connectionist Vehicles, Structural Resemblance, and the Phenomenal Mind", *Communication and Cognition*, 34: 13-38.

Searle, J. 1980 "Minds, Brains and Programs", *Behavioral and Brain Sciences*, 3: 417-424.

Van Gelder, T. 1996 "Dynamics and Cognition" in *Mind Design II*, J. Haugeland (ed.), MIT Press.

Following a Philosopher

Murilo Seabra / Marcos Pinheiro, Brasília, Brazil

1. Following rules and following philosophers

What might be to follow a philosopher? According to the latter Wittgenstein, rules are fundamentally what philosophers are concerned with, rules are what they write - that is, grammatical and not empirical propositions. This idea naturally allows us to transpose Wittgenstein's remarks on rule-following into the problem of following a philosopher. We should keep in mind that this idea applies as much to the metaphysicians and analytical philosopher's of the author's time as for himself. From his own point of view, Wittgenstein applied common rules of language to break down the bizarre representation norms - not empirical propositions about the inner workings of reality or language - proposed by philosophers. For instance, against the rule that states the privacy of meanings (sustained by Carnap and Russell as if it were an empirical proposition), Wittgenstein reminds us of an internal relation between the concepts of meaning and explanation, that is, he reminds us that meanings cannot be private as long as they can be explained. The rule that states the privacy of meanings simply takes the concept of meaning out of circulation (or else it drastically alters the concept of privacy).

But we are seldom clear about what it is 'to do the same thing' or 'to follow the same rule'. For we tend to wrongly generalize excessively simple paradigms of 'to do the same thing' as, for example:

> (1) A draws a straight line 10 cm long and asks B to do the same thing; we say that B has done the same thing only if B also draws a straight line 10 cm long.
>
> (2) A writes "2, 4, 6, 8, 10" and asks B to do the same thing; we say that B has done the same thing only if he also writes "2, 4, 6, 8, 10".

In other words, the paradigms of 'to do the same thing' which most of us have immediately present in mind belong to the most primitive kinds of repetition ever. We tend to think that B only does the same thing as A in case B acts as an impersonator. Obviously, if we keep that conception of 'to do the same thing' in mind, we will contend that following a philosopher is rewriting what he wrote with different (or perhaps even with the same) words; therefore, that to be a Russellian means something not much different from mimicking Russell.

Wittgenstein undermined the idea that B's doing the same as A must be a case of mimicry by showing that *there are different criteria for identity*. For example, letters 'C' and 'c', though graphically different, are one and the same letter. Many features are left outside the identity criteria of letters, as for instance their size - which, in turn, is taken into consideration when it comes to the identity of other sorts of object. On the other hand, if we rotate the letter 'c' at 90 degrees, we might be unable to discern it from letter 'u'. The spatial position of letters is relatively important to their identity - something that surely cannot be said about bats. Letter 'B' and 'b' are also the same, though one could contend that from the graphical point of view there are more similarities between 'b' and 'p', which are different letters. In short, the identity criteria of letters are different from the identity criteria of colors, which are different from the identity criteria of thoughts, which in turn are different from the identity criteria of bats. *Identity criteria vary according to the kind of object at hand*. To think that identity criteria always remain the same brings about absurd questions such as if two tokens of the letter 'A' can really be considered tokens of the same kind of letter, provided they do not occupy the same spatio-temporal position - an identity criterion that surely applies for bats.

2. Philosophical propositions

Let us consider for a moment identity criteria of propositions - a particularly important subject when it comes to the question of what following a philosopher might be. Wittgenstein has at least two important considerations about that in his *Investigations*. The first one is fairly straightforward: the same thought - the same proposition - can be expressed in different ways. The second consideration, though not so straightforward, is as important as the first one: there is no single set of identity criteria that holds for all kinds of proposition. Poetic propositions (and also the religious ones) have identity criteria different from philosophical propositions:

> We speak of understanding a sentence in the sense in which it can be replaced by another which says the same; but also in the sense in which it cannot be replaced by any other. (Any more than one musical theme can be replaced by another.)
> In the one case the thought in the sentence is something common to different sentences; in the other, something that is expressed only by these words in these positions. (Understanding a poem.) (Wittgenstein 2001, cf. PI 531)

The statement that a poetic proposition cannot be replaced by any other, that it cannot be expressed in different words, might be a sign of our understanding of it. However, no philosophy teacher would accept a transcription of PI 531 as a sign of her students' understanding of PI 531. This is a specially interesting case, for it shows that (1) and (2) are not particularly good paradigms for the identity criteria of philosophical propositions and therefore not particularly good paradigms for understanding what it might be to follow a philosopher. When it comes to philosophy instead of poetry or religion, one and the same thought can usually be expressed in different ways. And if B contends that A's thoughts can only be expressed in exactly the same way as A did, then we have good grounds to claim that B takes A's propositions as poetic or religious rather than as philosophical.

Wittgenstein thought that, just as there was nothing wrong with regarding 'A' and 'A' as the same letter, so there was nothing wrong with regarding "12, 14, 16, 18" as a sequel to "2, 4, 6, 8, 10" - therefore that writing "2, 4, 6, 8, 10" and writing "12, 14, 16, 18" could easily be seen as a case of doing the same thing. Transposing identity criteria which hold for a given class of objects to a class of objects where they do not hold is what brings about absurdities such as the idea that writing "2, 4, 6, 8, 10" could never be the same thing as writing "12, 14, 16, 18" (since the written characters are altogether different), or that 'A' and 'A' could never be tokens of the same letter type (simply because they are different tokens of the same letter). Extending these considerations to the problem of following a philosopher we may obtain interesting results. To follow a philosopher might be to *do* what he *did* (i.e., to do philosophy) instead of *rewriting* what he *wrote* (i.e., to interpret him). Depending on who is the philosopher, to rewrite his thoughts might be precisely *not* to follow him. Perhaps this is the case of Wittgenstein: perhaps interpreting him should be something as listening to his exhortations that we do philosophy as the cat that stares at the finger instead of the direction pointed to.

3. Philosophers and commentators

So let us raise the question: did Wittgenstein want to be followed? Did he want us to write about his thoughts or to make an effort to think for ourselves - just as he did? Perhaps the difference between the prefaces of the *Tractatus* and the *Investigations* could shed some light here. It is well known that in his first book Wittgenstein believed he had solved all philosophical problems once and for all - such as announced in the book's preface. Since the *Tractatus* had supposedly put an end to philosophy, it would only be possible to keep on interpreting its results (in the sense Wittgenstein talked about interpretation - cf. PI 201). Moreover, these interpretations would never present substantive developments, but only different ways - maybe more precise or exhaustive - of saying what the *Tractatus* had already stated. Clearly the *Tractatus* did not leave enough room for thoughts it did not contain.

The *Investigations*' preface, on the other hand, was written in a much different spirit. There Wittgenstein says he hoped "to stimulate someone to thoughts of his own" (Wittgenstein 2000, cf. PI p.x). The *Investigations* were not supposed to be the last word on philosophy, even though they still could have the last word in questions such as whether the meanings of words are the objects (either public or private) they refer to. After all philosophy had to keep moving; it is not even possible to solve all philosophical problems at once, but only to master an understanding of them that would allow us to stop doing philosophy whenever we want (Wittgenstein 2001, cf. PI 133).

But do these differences between the prefaces of both books really stand for differences in their contents? Judging from one of Wittgenstein's observations about a pre-war version of the *Investigations*, the answer seems to be affirmative:

> One could call this book a text-book. A textbook, however, not in that it provides knowledge [*Wissen*], but rather in that it stimulates thinking [*Denken*]. (Wittgenstein *apud.* Hilmy 1987, cf. p.6)

What Wittgenstein announced in the preface of his book was therefore not something alien to its content. The *Investigations* were not meant to stimulate thoughts by chance; this was rather their main goal. Wittgenstein did not want interpreters. The *Investigations* present a set of methods - of *Werkzeuge* - to the solving of philosophical problems, so that it would be pointless to just explain (instead of actually applying) them:

> Instead, we now demonstrate a method, by examples; and the series of examples can be broken off.
> – Problems are solved (difficulties eliminated), not a *single* problem.
> There is not *a* philosophical method, though there are indeed methods, like different therapies. (Wittgenstein 2001, cf. PI 133)

The remark refers to the *Investigations* themselves; it is clearly said that not only results are being presented but also the means to arrive at the results. At this point Wittgenstein brings his methods to the foreground and lets go of the actual examples by which those methods were presented. His methods are even more important than the particular philosophical problems he addresed. Wittgenstein does not direct a large portion of the *Investigations* against the idea that the meanings of words are the objects to which they refer because this idea is particularly difficult to undermine - he wants instead to present through his critique a series of methods that could and should be directed at other philosophical problems.

We hope to have left a few footprints on our pursuing the question of what it might be to follow a philosopher - and specially to have aroused some of the subtler issues involved in the very idea of following a thinker like Wittgenstein. The whole philosophy of the *Investigations* - and in many ways we are not able to hint here by lack of space - can be read as an attempt to create philosophers instead of commentators. And the unsettling possibility also remains that even a good interpretation of Wittgenstein would amount inevitably to a misinterpretation.

Literature

Hilmy, Stephen 1987 The Later Wittgenstein: The Emergence of a New Philosophical Method, Oxford: Basil Blackwell

Wittgenstein, Ludwig 2001 Philosophical Investigations, Oxford: Basil Blackwell

Davidson on Supervenience

Oron Shagrir, Jerusalem, Israel

1. Introduction

Donald Davidson introduces supervenience to the philosophy of mind in his "Mental Events". He famously writes that "mental characteristics are in some sense dependent, or supervenient, on physical characteristics" (1970:214). Curiously, however, there has been little effort to explicate what Davidson means by supervenience; philosophers typically assume Jaegwon Kim's conception of supervenience. My aim here is to explicate the passages in which Davidson discusses supervenience. I argue that Davidson's supervenience is very different from the one assumed in contemporary philosophy of mind, and is not dependence in the sense of some deeper metaphysical relation.

2. The characterization of supervenience

Davidson characterizes supervenience in several places. In "Mental Events", he writes:

> Supervenience might be taken to mean that there cannot be two events alike in all physical respects but differing in some mental respect, or that an object cannot alter in some mental respects without altering in some physical respects (1970:214).

The first part of the sentence is a characterization in terms of *indiscernibility*, namely, that "there cannot be two events alike in all physical respects but differing in some mental respect", that is, there cannot be two events that are physically indiscernible but mentally discernible. The second part is another characterization, one which construes it as a type of covariance: "an object cannot alter in some mental respects without altering in some physical respects", that is, mental changes *co-vary* with physical changes.

In later writings, Davidson provides additional covariance definitions, the gist of which is that any mental difference between objects must be accompanied by a physical difference. In "Reply to Harry Lewis", he writes:

> The notion of supervenience, as I have used it, is best thought of as a relation between a predicate and a set of predicates in a language: a predicate *p* is supervenient on a set of predicates *S* if for every pair of objects such that *p* is true of one and not of the other there is a predicate in *S* that is true of one and not of the other. (1985:242)

And in "Thinking Causes" he makes a similar claim:

> The idea I had in mind is, I think, most economically expressed as follows: a predicate *p* is supervenient on a set of predicates *S* if and only if *p* does not distinguish any entities that cannot be distinguished by *S*. (1993:4)

On a charitable reading, Davidson's characterizations are all equivalent. Supervenience is a thesis about the relations between properties or characteristics or respects, e.g., mental and physical properties, which Davidson understand as *predicates*. These properties are ascribed to particulars such as events, objects and entities. To make things more explicit, let us take two sets of properties, R and S. We can think of R as a set of mental predicates, and of S as a set of physical properties. We would say that R *supervenes* on S just in case the following condition holds:

> For every M of R and for every pair of objects (events, entities) x and y, if for every P of S, Px ↔ Py (i.e., x and y are S-indiscernible), then Mx ↔ My (i.e., x and y are M-indiscernible).

Let us compare this characterization to Kim's notions of supervenience. First, Kim (1984) famously distinguishes between a strong and a weak reading of this condition. On the strong reading, the condition applies to every pair of possible objects x and y, even if they inhabit "different worlds". On the weak reading, it applies to every pair of objects belonging to the same world (any world), but need not apply to objects across worlds. In "Thinking Causes", Davidson says that his version of supervenience is of the weak sort:

> Kim himself (correctly, I think) finds my version of supervenience very close to his 'weak' supervenience, and as not entailing connecting laws. (1993:4, n. 4)

Second, Kim demonstrates that under assumptions of closure of S, strong and weak supervenience are equivalent, respectively, to strong and weak entailment P* → M principles, where P* is a maximal S-property. Unlike Kim, Davidson never explicates supervenience in terms of entailment conditionals P* → M. Third, Kim also introduces a notion of global supervenience, which, arguably, fits better with the thesis of externalism; Davidson, we recall, is an outspoken proponent of externalism. But it turns out that global supervenience is an intricate notion that comes with very different versions. Instead of invoking global supervenience, it is sufficient to use the individual notions but not limit S to monadic, micro, local, or intrinsic properties; it could include causal relations with the physical environment, and even bits of causal history. S could even include physical properties of remote objects if such properties are indeed relevant to the ascription of mental properties.

Davidson's statement about his notion of supervenience being weak is puzzling. The problem with weak supervenience is that it does not support dependence. Weak supervenience is consistent with the scenario in which my counterpart and I have exactly the same physical properties, but different mental properties. But then the mental difference is not due to our physical properties, since nothing related to our physical makeup, including past and present causal relations with their environment, differs. It would thus seem that there *are* mental properties that do not depend on physical properties.

Davidson himself rules out such scenarios. He maintains that counterfactual scenarios like the Twin-Earth and Swampman thought-experiments, the mental differences *are* accompanied by physical differences:

> What I take Burge's and Putnam's imagined cases to show (and what I think the Swampman example shows more directly) is that people who are in all relevant physical respects similar (or 'identical' in

> the necktie sense) can differ in what they mean or think… But of course there is *something* different about them, even in the physical world; their causal histories are different. (1987:32-33)

Weak supervenience, however, lacks the modal force to support these psychophysical dependencies. Why, then, does Davidson invoke weak supervenience? One answer might be that mental properties do not strongly supervene on the intrinsic physical properties of agents, as the Twin-Earth and Swampman examples show. The mental only weakly supervenes on intrinsic physical properties: two agents of the same "world" that are physically indiscernible are also mentally indiscernible (see Davidson 1995).

However, this proposal will not do. It is true that the mental does not strongly supervenes on intrinsic physical properties, but it does strongly supervenes on intrinsic *and extrinsic* physical properties. It thus makes more sense to use strong supervenience over intrinsic and extrinsic physical properties, which reflects dependence, rather than using weak supervenience over intrinsic physical properties alone, which does not reflect dependence. One way or another, it seems that the weakness of Davidson's supervenience does not stem from the usual "across-worlds" relations.

3. Supervenience as a philosophical concept

Davidson does not say much about the philosophical import of supervenience, but it is clear that his views on this are unique. One respect in which they are unusual has to do with the philosophical role of supervenience. Supervenience is widely upheld by non-reductive monists: those who maintain that every mental event *is* a physical event, but deny psychophysical laws. Many who espouse versions of this view worry that it is insufficiently "materialistic". The concern is that non-reductive monism says too little about the relations between mental and physical *properties*. Although it denies that mental properties are physical properties, it imposes no alternative constraints on the attribution of mental properties. Monism ensures that every object with mental properties also has physical properties, but, beyond that, anything goes: monism is consistent with the possibility that my physical counterpart has no mentality whatsoever, while my cup of coffee does. Surely a monistic, to say nothing of materialistic, doctrine that allows such wild attributions is worthless. Something must be done to close this gap. And this is where supervenience kicks in. The role of supervenience is to put more significant constraints on the psychophysical relations between mental and physical properties, without reducing mental properties to physical properties. Supervenience, on this account, is a secondary thesis that makes non-reductive monism materialistically kosher.

"Mental Events" gives the impression that supervenience plays this legitimizing role in Davidson's philosophy. After presenting the tenets of anomalous monism, Davidson immediately introduces supervenience, saying that "although the position I describe denies there are psychophysical laws, it is consistent with the view that mental characteristics are in some sense dependent, or supervenient, on physical characteristics" (p. 214). From this we might conclude that Davidson, too, feels obliged to complement his monistic thesis about events with a substantive and positive thesis about the psychophysical relations between predicates. But later on, in his "Reply to Harry Lewis" (1985) and in "Thinking Causes" (1993), it turns out that this is not how Davidson sees the role of supervenience. Declaring that "supervenience in any form implies monism; but it does not imply either definitional or nomological reduction", Davidson reveals that he invoked supervenience to demonstrate that anomalous monism is consistent: "So if (non-reductive) supervenience is consistent (as the syntax-semantics example proves it is) so is *AM* [anomalous monism]" (1993:5).

Contrary to first impressions, then, supervenience is not a secondary thesis intended to correct the deficiencies of the primary doctrine of non-reductive monism. There is no evidence that Davidson deems his anomalous monism to be in need of reinforcement, whereas we do have evidence that Davidson does not take supervenience to provide such reinforcement. Davidson, of course, does resist the idea that mental properties float freely, as it were, over the physical domain, and does take supervenience as asserting that the mental *depends* on the physical realm. But this claim about dependency is not made as a substantive addition to anomalous monism. Rather, supervenience is used both to help establish monism *and* the consistency of anomalous monism. Davidson deploys supervenience once again in "Thinking Causes", this time to secure the causal relevance of mental properties. The claim made is that supervenience *entails* that an event's mental properties make a difference to its causal relations.

Davidson’s supervenience is also unique with respect to the notion of dependence. Most philosophers, following Kim, maintain that mind-body supervenience is grounded in some deeper *metaphysical* relation. The idea is that any $P^* \rightarrow M$ conditional reflects the dependence of M on P*, and this dependence is a metaphysical determination relation, e.g., identity, constitution, emergence, or realization, which underlies and explains the supervenience relations. It is thus not surprising that, in the context of supervenience, the notions of dependence and determination are often used interchangeably. The implicit assumption is that M depends on P* by virtue of M's being determined by P*, whereas determination is understood as a metaphysical determination.

Davidson's notion of dependence is different. The idea that the application of a mental predicate is grounded in some metaphysical determination of the mental by a fixed physical basis is foreign to Davidson's approach. He never hints that the mental depends on the physical by virtue of some metaphysical determination relation; and certainly does not introduce the more familiar determination relations to substantiate his supervenience thesis. In fact, in "Thinking Causes", the main argument for the causal relevance of mental properties suggests that supervenience is *not* such a determination relation. He asserts: "supervenience as I have defined it does, as we have seen, imply that if two events differ in their psychological properties, they differ in their causal properties (which we assume to be causally efficacious). If supervenience holds, psychological properties make a difference to the causal relations of an event, for they matter to the physical properties, and the physical properties matter to causal relations" (1993:14). But it is apparent that "make a difference" cannot be understood to mean "determine" in a metaphysical sense. For it refers to the mental-to-physical direction, whereas the pertinent metaphysical relation is in the physical-to-mental direction. It is most unlikely that Davidson would take supervenience to point to metaphysical determination of the mental by the physical, and still claim that supervenience implies that mental properties "matter to the physical properties".

We see that Davidson does not uphold the idea that supervenience reflects metaphysical physical-to-mental determination or dependence relation. It seems, moreover, that he also rejects the idea that dependence (of mental on the physical) grounds or accounts for supervenience. If anything, it is the other way around. Davidson says that "supervenience gives a sense to the notion of dependence here, enough sense anyway to show that mental properties make a causal difference" (1993:14). So it is not that dependence accounts for supervenience, but, if anything, dependence is explicated in terms of the supervenience of the mental on the physical.

Lastly, it is telling that Davidson invokes supervenience in *causal* contexts. In discussing the Twin-Earth and Swampman cases, Davidson insists that "of course there is *something* different about them, even in the physical world; their causal histories are different". He later describes supervenience as implying that "mental properties make a causal difference". And he links supervenience with the causal nature of the mental, stating that "Kim, as we noted, thinks my version of supervenience implies that all mental properties could be withdrawn from the world and this would make no difference to causal relations; but this supposition turned out to be incompatible with my understanding of supervenience" (1993:14); and that "[s]upervenience as I defined it is consistent with… the assumption that there are no psychophysical laws… It is not even slightly plausible that there are no important general causal connections between mental and physical properties of events. I have always held that there are such connections" (1993:14).

4. Summary

Let us sum up the distinctive features of Davidson's supervenience. Supervenience is not a secondary thesis the objective of which is to reinforce anomalous monism. It is not explicated by some deeper metaphysical determination or dependence relation, but if anything, it is supervenience that gives cogency to the notion of dependence. And it has something to do with the "causal connections between mental and physical properties of events". In addition, Davidson characterizes supervenience in terms of indiscernibility or covariance and not in terms of the entailment P* → M conditionals, and declares that his supervenience is of the weak kind. Whether we can we extract from these remarks a cohesive notion of supervenience, and whether this notion can be reconciled with anomalism is something I will discuss elsewhere.

Literature

Davidson, Donald 1970 "Mental Events", in L. Foster and J.W. Swanson (eds.), *Experience and Theory*, Amherst: University of Massachusetts Press, 79-l0l. Reprinted in his *Essays on Actions and Events*, Oxford: Clarendon Press, 207-227.

Davidson, Donald 1985 "Reply to Harry Lewis", in: B. Vermazen and M. Hintikka (eds.), *Essays on Davidson: Actions and Events*, Oxford: Clarendon, 242-244.

Davidson, Donald 1987 "Knowing One's Own Mind", *Proceedings and Addresses of the American Philosophical Association* 60:441-458. Reprinted in his *Subjective, Intersubjective, Objective*. Oxford: Clarendon Press,

Davidson, Donald 1993 "Thinking Causes", in: J. Heil and A. Mele (eds.), *Mental Causation*. Oxford: Clarendon Press, 3-17.

Davidson, Donald 1995, "Could There Be a Science of Rationality?", *International Journal of Philosophical Studies* 3:1-16.

Supervenience and ‘Should’

Arto Siitonen, Helsinki, Finland

Introduction

Concerning any entity or any fact, we may wonder how does it fit into reality as a whole. We thus raise questions of the type: *What is the place of – in the scheme of reality?* As the placeholders for '–' we may put stars, atoms, populations, societies, histories, human beings, languages, works of art, numbers etc. These are mutually related in a variety of ways. In order to account for those multifarious relations, we organize them into a system. The guiding principle of such an organization is that things and facts are not disconnected but depend on each other. Certain entities and occurrences depend for their existence on certain other entities and occurrences. We are dependent for our existence on our ancestors; we are connected to them through causal lines of heredity. There is also a more conceptual kind of dependence: for instance, to be sentient depends on being alive – non-living beings cannot be sentient beings. Stones do not perceive, let alone think.

This dependence has been expressed in philosophy by saying that the property 'sentient' *supervenes on* the property 'alive', or the fact '*a* is a sentient being' supervenes on the fact '*a* is a living being'. Supervenience means dependence, determination, and necessary condition; to these relations are added the claims of reducibility and explainability. A supervening property or fact can be reduced to the property or fact on which it supervenes; and by that, it is explained by the latter. Reducibility claim means that the former is to be "nothing but" the latter, organized in a proper way. An adequate explanation makes understandable why this is so.

We may extract from the foregoing consideration the question: *Could – be removed without removing – thereby?* If not, then the latter is a necessary precondition of the former; thus, for instance, being alive is required for being sentient. This gives a partial answer to the above question of the "place of –" in the scheme of nature, in respect to the facts and properties concerned in the example.

If supervenience is a thoroughgoing universal factor in reality – and in our account of reality –, then it sounds reasonable to extend its sphere from nature to culture. Here reductive explanation meets some challenges. Sport clubs, societies, states, etc. may be reducible to psychology – and by this, all the way down to physics. Concerning numbers, they are conceptual entities that may be reducible to logic (cf. the logicist program of Frege and Russell), but presumably not to counting, or to any other actions. As to works of art, one may claim that, e.g., a painting is reducible to the composition of its ingredients, i.e., colour spots on a canvas, or an orchestral work to the sound waves that vibrate in the air in a certain way, etc. However, the question of their aesthetic value is a harder one: would not the value that they have, exceed the evaluations given to them by various persons? Correspondingly, common morality may be reducible to social and psychological facts, but what about the claims that being moral presents to us – are not these irreducible?

Let us focus on the issue of values and norms. Above, properties and facts were considered as supervening entities. Among facts belong actions performed by human beings. Evaluating, esteeming, commanding and requiring are human actions. If the idea of supervenience is pursued consistently, these actions are traced back to facts concerning nervous systems of organisms, and subsequently all the way through to the undulations of elementary particles.

However, are the very values and norms thus reduced? Challenging this, one may appeal to the gap between 'is' and 'ought', as seen by David Hume. It may be that values and norms are not accountable by supervenience. Or, if they are, then they are reducible to facts – facts organized in a proper way, whatever that may be. Examples of such facts can be certain features in works of art, in human actions, in society. One may also try to base values and obligations on acts of esteeming and requiring. Hume did not exclude the possibility of accounting 'ought' through 'is'; he just raised the question of how the production of an 'ought' from an 'is' is achieved, and justified.

Below, the following *thesis* will be defended: it is questionable whether values and norms *sui generis* are reductively explainable – and thus, whether they supervene on facts. This worry arises due to the 'is/ought'-gap.

1. Supervenience accounting for facts

We construct systems of science with their branching subsystems; and we claim that such a system adequately represents reality. Thus, there is "the real order of things" and the order that we make in order to account for that (cf. title of Molander (1982)). This is the basis for the distinction between reality and research of reality.

The idea of supervenience is concerned with the order and with the ordering of facts. It makes ontological and epistemological commitments. It classifies facts according to certain evolutionary principles in a comprehensive way. 'Supervenience' is a philosophical concept that is applied to the methods and results of science. Francis Crick, although not employing the concept of supervenience, considers the neurobiological account of consciousness as a scientific hypothesis; cf. his work (1994). On his view of the relation between science and philosophy, cf. p. 256 ff of that work. He hopes that "philosophers will learn enough about the brain to suggest ideas about how it works" (p. 258).

Building up classificatory schemata and subsuming various occurrences under them is the first, basic task of scientific research. Carl von Linné accomplished this in the area of botany; cf. especially his work 'Species plantarum' from 1753. He also contributed to the corresponding organizing work in zoology. These studies were complemented by researchers who worked on anatomy, physiology and ecology. A new question was raised in the studies on heredity, and it was Charles Darwin who gave the explanation to this phenomenon by the theory of evolution (cf. his work 'The Origin of Species by Means of Natural Selection', 1859). Finally, the basic factor of life, the genetic code, was unravelled by Francis Crick and James D. Watson in 1953. This discovery revealed the role of nucleic acids in the generation and growth of living beings. It thus identified the physical basis of life.

The classification of living beings, and the account of their development and its core factors, contribute to answering the question "What is the place of life in the scheme of nature?" This is achieved by revealing the physical preconditions of life. Physical basis is a necessary condition for life; accordingly, if it were removed, there would not be life. However, there could be a lifeless physical universe – and there has been, before the evolution of life. Thus, the inorganic nature can be accounted for without recourse to the organic one; but accounting for the latter requires taking the former into consideration. For the organization of scientific research, this means that biology is built on the foundation of physics, but physics not on biology.

A corresponding situation ensues when mental occurrences are added to those of life. Historically speaking, consciousness and self-consciousness have developed gradually, as living beings have become sophisticated enough. Social arrangements have then evolved from mutual relations between conscious beings. Languages, i.e. signal and symbol systems of communication, have developed on this basis.

In the light of supervenience, nature is a layered system that has evolved through aeons and will presumably go on in its development – cultural evolution building itself on the basis of cosmic and biological evolution. We may trace given facts of culture back to their origin in forgone human populations, these back to the first occurrences of life, and these again to physical facts. We may also make projections concerning future: evolution will presumably continue, but how?

As was indicated above, there are occurrences, or facts, or things that supervene on something, and correspondingly there are occurrences, facts, things, on which the former supervene. This implies reduction: the former are "nothing but" the latter, organized in a proper way. A distinction is thus drawn between (i) what it is that does supervene, and (ii) what it is on which the supervening content is supervening. The relation between (i) and (ii) means that the former is reducible to the latter. One may speak of "the content factor" and of "the basis factor". For the purposes of explanation, (i) is considered as *explanandum*, (ii) as *explanans*. Thus, we may explain e.g. heat (something felt) by the acceleration of molecules (something physical).

Moreover, the original basis factor can take the role of a new content factor, etc. The result is a chain of superveniences. For instance, if we start with the fact that *a* is a sentient being, its base is the fact that *a* is a living being (cf. above). Being alive is in turn based on biochemical facts, the latter on chemical facts, and these on physical facts. This is the downward route; if we change direction, contents become bases for further contents, and we proceed upwards.

The structure of reality thus revealed and accounted for is grounded on physical facts. These yield the fixed starting position for explanations, and the final basis for reduction. The systematic order of bases and contents is mirrored by the time order of evolution and emergence: the birth of cosmos, life, consciousness, culture. The system of nature thus has a fixed start position and an open future. It has evolved from the birth of stars and planets; where its development will lead, is a moot question to which various cosmological theories try to answer. These theories have their precedent in the work by Pierre Simon de Laplace, 'Exposition du système du monde' (1796).

The order of facts that their supervenient analysis reveals, may be expected to give a wholesale answer to the question concerning the scheme of reality. In the *Introduction* above, this question was given two formulations. Its answer should give a proper classification of facts, their systematic order, and their time order. Classes of facts are strata of reality. The specific content of such a stratum is the subject matter of its attached branch of research, or branches of research. Thus, for instance, a living cell is studied by biochemistry and biology. A given stratum can be considered as basis for another stratum. The question 'what is the place of life in reality?' receives its full answer in the context of the whole system; the corresponding is true of other strata. An orderly study thus promises to give an all-encompassing account of the evolution of cosmos, life and culture. In respect to what is achieved, one may say that these are the facts, and all of them (cf. Chalmers 1996, p. 86: "*That's all*").

2. Supervenience accounting for moral facts

One may wonder how supervenience can account for moral facts. In a broad sense, these may be understood to comprise all facts that are studied in social sciences and humanities. Traditionally, the title 'moral sciences' is used as the common name for these. Moral sciences are distinguished from natural sciences. They are concerned with mental, social and cultural facts. The objects of their research are human action and its results: history, societies, states, languages, works of art etc.

In the light of supervenience, moral facts depend on social, these on mental, and these on physical facts. Moreover, moral facts can be explained on the basis of other facts, and be reduced to these.

As to morality proper, its emergence, development and character are examined in the theory of morals. One can take the fact '*a* is a moral being' as *explanandum* and look for its explanation among the more basic facts that account for it. To these belong the facts that *a* is a social, a sentient and a living being. Correspondingly, what is called 'common morality' is reducible to social and psychological facts. One can also make comparative studies of different moralities, considered either synchronically (the present cultures) or diachronically (the past and present cultures), and give explanations to their common and diverging features.

3. Accepted norms and values vs. acceptable norms and values

Morality has an outside and an inside dimension. The outside dimension is concerned with facts, such as something being done by a person or by citizens in general, or something being accepted as proper behaviour in a community. Morality in this sense comprises factual practices and factually accepted practices. These may deviate from each other; cf. the so-called "double standard morality".

The inside dimension, in contrast, gives reasons for the following questions: is that what is done *right*? Are the principles which are generally followed *right*? These questions concern the issue of morally acceptable standards.

Although the dimensions are clearly distinguishable, it is in practice difficult to keep them apart. Thus, the very word 'norm' can be thought to express something generally accepted; or it may be understood in the sense of a self-addressed unconditional duty. One may compare this to G. H. von Wright's distinction between the *descriptive* "discourse for speaking about norms" and the *prescriptive* "discourse for enunciating rules of action and other norms" ((1968), p. 11).

There is also a second, related difficulty that is concerned with attempts to defend one's behaviour by appealing to what is generally done, or generally accepted. That such a procedure is quite common, is a fact of moral psychology. But is it justifiable – and if not, why not? This question can be clarified in the light of David Hume's short remark in his work *Treatise of Human Nature* (third volume, first part, first section). He distinguishes between the expressions 'is' and 'ought'. On the basis of this, he makes two claims: (1) because the latter "expresses some new relation or affirmation…it shou'd be observ'd and explain'd", (2) "a reason should be given, for what seems altogether inconceivable, how this new relation can be a deduction from others, which are entirely different from it." (Hume (1992), p. 246).

The following conclusions can be drawn from Hume's analysis: If consent is given to claim (1), then reducing 'ought' to 'is' would mean the mistake of conflating norms with facts. If consent is given to claim (2), but the required "deduction" cannot be given, then an 'ought', in case that it is not self-evident, can be justified only by appeal to some other, more basic 'ought', but never by appeal to facts. In the light of this, acceptable norms cannot be derived from accepted norms. Moreover, there are reasons to suppose that acceptable norms form an autonomous area, for which one cannot account by any facts whatsoever.

The expression 'ought', or 'should', can be prefixed to the verbs 'be' and 'do'. Then, 'should be' may be understood to be the core of value judgements, and 'should do' the core of judgments expressing a norm. (Cf. the traditional German distinction between *Seinsollen* and *Tunsollen*). These judgments are concerned with values and norms as abstract entities (cf. the concept of number, to which values and norms are in this respect analogous). The theory of values is known as *axiology*, and the theory of norms as *deontology*.

Values as abstract entities are to be distinguished from valuations, the latter being concrete actions or mental dispositions. Correspondingly, norms differ from commands. Values and norms thus understood are nothing mystical but plain common sense: although their explication is difficult, we know quite well how and when to employ them in discourse and how to apply them to action. We then act as moral subjects *inside* the realm of morality, using normative concepts in the prescriptive sense, whether addressing them to ourselves or to others. We participate to moral discourse; we do not try to explain it or reduce it to facts.

We can step *out* and thereby switch off to descriptive, fact-stating mode of moral discourse. Speaking from this vantage point of accepted values and norms, enables us to account for them by appealing to facts of culture, society, psychology etc.

Accordingly, it is the factual side of normative utterances, i.e. valuations and commands, that can be accounted for by supervenience. Supervenience does not concern values and norms in the prescriptive sense. Acceptability is not reducible to acceptance.

Ludwig Wittgenstein made a related point in proposition 6.41 of his *Tractatus*: "In the world everything is as it is, and everything happens as it does happen: *in* it no value exists – and if it did exist, it would have no value. If there is any value that does have value, it must lie outside the whole sphere of what happens and is the case."

In 6.43 he says: "If the good or bad exercise of the will does alter the world, it can alter only the limits of the world, not the facts…". This can be interpreted in the following way: as seen from the perspective of values, the world – the totality of facts – appears in a different way than through neutral, non-committed consideration.

4. Critical remarks

It is difficult to put in words the dual dimension of value and norm expressions: firstly, valuations and obligations as accepted, empirical, factual – secondly, values and norms as acceptable, conceptual, normative. Wittgenstein even thought that the latter dimension exceeds the limits of language; cf. *Tractatus* 6.421, according to which "ethics cannot be put into words", and 6.423: "It is impossible to speak about the will in so far as it is the subject of ethical attributes."

Also, maintaining the duality in a consistent way is difficult. This is exemplified by certain occasionally encountered suspicious expressions, such as "value facts". If one wants to refer by it to valuations, one should then rather speak of facts of valuation; if it is intended in the normative sense, it marks a plain confusion. That, for instance, the life of a species in nature is intrinsically good, is not a fact but a value. The fallacy of *metabasis eis allo genos* can be committed in attribution of properties and in reasoning. As to the latter possibility, Hume speaks of an "imperceptible change" from propositions containing the copula 'is' to those "connected with an ought"; cf. Hume (1992), p. 245 f.

Let us put forward some reminders: values are not facts; valuations are facts. Norms are not facts; making normative claims is a fact. There is nothing wrong in the effort of upholding the fact/value distinction, or the factual/normative one. (There is neither anything morally wrong in this, nor anything logically wrong).

Maintaining these dichotomies implies that supervenience does not apply to values and norms themselves, but it does comprise moral facts (cf. Sec. 2 above). This is not a loss, because the idea of supervenience means explaining facts through reducing them to other, more basic facts.

Literature

Chalmers, David J. 1996 *The Conscious Mind. In Search of a Fundamental Theory*. New York & Oxford: Oxford University Press.

Crick, Francis 1994 *The Astonishing Hypothesis. The Scientific Search for the Soul*. London: Simon & Schuster.

Hume, David 1992*. A Treatise of Human Nature*. In: David Hume, Philosophical Works II, ed. T. H: Green, T. H: Grose, Aalen: Scientia Verlag (Second reprint of the new edition London 1886). (Originally 1739-40).

Molander, Bengt 1982 *The Order There Is and the Order We Make. An Investigation into the Concept of Causation*. Uppsala: Philosophical Studies published by the Philosophical Society and the Department of Philosophy No. 35.

Wittgenstein, Ludwig 1961 *Tractatus Logico-Philosophicus. Logisch-philosophische Abhandlung*. Translation by D. F. Pears & B. F. McGuinness. London: Routledge & Kegan Paul. (The first German edition 1921).

von Wright, Georg Henrik 1968 *An Essay in Deontic Logic and the General Theory of Action,* Amsterdam: North-Holland Publishing Company (Acta Philosophica Fennica, Fasc. XXI).

Rule-following as Coordination: A Game-theoretic Approach

Giacomo Sillari, Philadelphia, Pennsylvania, USA

Make the following experiment: *say* “It’s cold here”
and *mean* “It’s warm here”.
Can you do it?

Ludwig Wittgenstein, *Philosophical Investigations*, §510.

I can’t say “it’s cold here” and mean “it’s warm here”—
at least, not without a little help from my friends.

David Lewis, *Convention*.

1. Rule-following, coordination and normativity

The slogan that “meaning is normative” is best understood in the context of strategic interaction in a community of individuals. Famously, Kripke has argued in (Kripke 1982) that the central portion of the *Philosophical Investigations* describes both a skeptical paradox and its skeptical solution. Solving the paradox involves the element of the *community*, which determines conditions of assertability in the language. A battery of argument is used to show that meaning (or, in general, rule-following) cannot be explained by resorting to an individual’s mental states, or her past use, or her dispositions. By exclusion, this indicates that no descriptive fact is constitutive of meaning, and hence that “meaning is normative.” Arguably, the normativity of meaning stems from the assertability conditions holding in the society (indeed, membership in the community depends on one’s record of compliance.) But *how* exactly is the existence of such conditions sustained in the community? And is it accurate to say that there is no *fact* to the matter of rule-following?

I need an important *caveat* here: To answer these questions, I momentarily step back from analyzing *meaning* and elaborate on the more general notion of *rule-following* instead. Kripke himself uses the terms meaning and rule-following rather interchangeably in (Kripke 1982). I will conform to the ambiguous usage for ease of exposition, and mention my justification for it in the last section of this contribution.

Wittgenstein states (§§198, 199) that a rule is followed *insofar* as there exists a custom, a convention. I argue that this and similar remarks in the *Philosophical Investigations* are illuminated when looked at through the lens of David Lewis’s theory of convention. Lewis argues in (Lewis 1969) that coordination games (situations of strategic interaction in which the interest of the players roughly coincide) underlie every instance of convention, in that a convention is a regularity in the solution (equilibrium) of recurrent coordination games. The agents participating in the convention conform to the regularity because they prefer conformity over non-conformity, conditional on other agents’ conforming. They form the belief about other agents’ conformity through some coordination device: explicitly—through agreement—or tacitly—because a certain action stands out as the one that most likely (almost) everyone will pick. Such an action is *salient* to the parties. In the case of a recurrent coordination problem, a special kind of salience—*precedent*—is at play.

Conventionality in the sense of Lewis is sufficient for some degree of normativity to arise. Indeed, in a community in which a certain custom is in place—say, the custom of going by sign-posts—there is an equilibrium in the actions and beliefs of the agents involved such that the agents prefer conformity to the custom, provided that all other members in the community act according to the convention. If I do *not* go by sign-posts, or I go by them in a funny, abnormal way (for instance, going in the direction opposite to the one indicated) I act contrary to both my preferences—because I will not get where I intend to go—and the preferences of other members of the community—because, say, I will end up being late, or not showing up at all. My reputation will suffer. This indicates that, in general, parties to a convention feel, to a larger or smaller extent, the pressure to conform. As Lewis puts it, conventions are a kind of social norm. But are we entitled to cast the rule-following phenomenon in a game-theoretic account of convention?

In its most general terms, the communitarian view maintains that, while many interpretations of a given rule may arise, there is (roughly speaking) only one correct way to abide by the rule, as determined by the community. In particular, the customary action is the action that accurately corresponds to the rule. The problem with arguments of this general form is that the *same* skeptical paradox meant to show the impossibility of solipsistic rule-following applies to the community. *Which* is the customary action? And *why*? Past use is no sufficient grounds to answer such questions for the community, as it is not sufficient grounds in the solipsistic case, since the community can come up with a variety of interpretations of the rule, just as well as the individual can. However, if we introduce a *strategic* element in the behavior of community members, then the skeptical paradox disappears (or, as we shall see in the next section, gets “pushed towards bedrock.”) If we interpret rule-following as *coordination equilibrium* in a coordination problem, then there *is* a clear and compelling fact to the matter of what “going by the rule” consists of. In particular, individuals (and the population they interact with) who go by the rule net a higher payoff than do individuals who transgress the rule. Moreover, transgressing the rule comes at a price, both for the transgressor and for the agents interacting with him. Non-conformative behavior will end up being sanctioned (eventually with expulsion from the community), while conformative behavior will perpetuate itself, being based on the agreement to act according to given rules. In this sense, agreement is the agreement in preferences and beliefs that support a specific equilibrium in the recurrent coordination game.

Thus, the “little help” needed by Lewis from his friends in the answer to the challenge of §510 reported in the epigraph consists then in their *agreeing* to change their preferences and beliefs, switching in so doing from one solution of a recurrent coordination game to another. Consider §224:

> The word “agreement” and the word “rule” are *related* to one another, they are cousins. If I teach anyone the use of the one word, he learns the use of the other with it.

I believe that the view expressed in this section captures the sense in which “agreement” and “rule” are related: A custom—and hence a rule—does not hold without an agreement in preferences and beliefs—and hence in coordinative, conventional actions—on part of the members of the community.

2. Precedent and justification

Although my interpretation of §224 surely appears contentious to many, it should become clear by the end of this section that in fact it jibes with the traditional reading. I find in §241 the cue to the traditional interpretation of "agreement" in §224:

> [Human beings] agree in the *language* they use.
> That is not agreement in opinions but in form of life.

Lebensform is a rich and profound philosophical concept that does *not* reduce to the preferences and beliefs (to the *opinions*) held in a community. Still, I claim that the notion of *Lebensform* is related to the Lewisian picture of conventional behavior and that preferences and beliefs in the community in fact spring from it.

Precedent lies at bedrock, where the spade is turned (§ 217) and one acts blindly (§ 219) conforming to the convention and obeying the rule. Without reliance on precedent, no conventional strategic interaction in the sense of Lewis is possible and, as I have argued in the previous section, without strategic interaction the community is in no better position than the individual is in determining which course of action is in accord with the rule. Indeed, as Margaret Gilbert tersely points out in (Gilbert 1990), in Lewis's account of convention practical rationality does not yield any justification to act in conformity to precedent. She argues that, on the contrary, conformative action is *blind* in the Wittgensteinean sense. Consider the two person case: Given their conditional preference, one is justified in conforming if she believes that the other will conform. But the other will be justified in conforming if he believes that the first individual will. Thus, she will be justified if she believes that he believes that she will, and so on. In the endless replication of each other's reasoning, at no point anyone will come to have sufficient reason to conform.

I have argued elsewhere (Sillari 2005, 2008) that in fact precedent gives rise to the series of replications about hypothetical future conformity, which, in turn, *inductively* ground for both individuals the first-order belief that the other will conform. There is no deductive, infallible passage from past to future conformity. There rather is a causal, inductive one (cf. the interlocutor in §198: "[…] What sort of connexion is there [between the expression of a rule and my actions]?—Well perhaps this one: I have been trained to react to this sign in a particular way, and now I do so react to it.[—]But that is only to give a causal connexion […]") Wittgenstein speaks of "blind action", Gilbert speaks of an "a-rational tendency". For (McDowell 1984), understanding is "precarious and contingent", in that there is no guarantee that my grasping a concept will continue working tomorrow as well. No strong, logical, deductive nexus is to be found between precedent and future conformity. Rather, the relation between precedent and future conformity lies at bedrock, as pointed out in §481:

> If anyone said that information about the past could not convince him that something would happen in the future, I should not understand him. […] If *these* are not grounds, then what are grounds?

As flimsy as the relation might be, we all endorse it since, as the traditional interpretation of §224 indicates, we all share an agreement in *Lebensform*. Our systems of concordant beliefs about each other conformity *stem* (albeit not deductively) from such a fundamental agreement. In turn, from our concordant beliefs and conditional preferences stem our conventions and customs, and hence our capacity to obey or to go against a rule.

The game-theoretic analysis of rule-following reveals that preferences and beliefs of community members strategically determine what course of action is in accord with the rule. The formation of beliefs, however, is a bedrock notion. Can a game-theoretic analysis help us reduce the phenomenon of rule-following any further? It is well known that Wittgenstein invites us not to dig under bedrock. To ask whether it is possible, and how it may be done, I finally tackle the issue of the relation between meaning and rule-following and turn to the final section of this contribution.

3. Meaning and rule-following

In this section I focus on *meaning* by looking at a special case of coordination problems involving communication. Rather than attacking the question of meaning in *language* (a question that lies well beyond the scope of this note) I will look at the simpler case of meaning in *signaling systems* (cf. Lewis 1969). Signaling systems are a special case of coordination problems. In a signaling game certain actions (performed by the *audience*) correspond to certain states of the world (observed by the *speaker*.) The speaker sends a *signal* depending on what state of the world she observes. The audience performs a certain action depending on what signal she receives. Both speaker and audience prefer that the action corresponding to the actual state of the world be performed. For that to happen, they need to coordinate their strategies (which for the speaker are functions linking states to signals for the speaker, while they are functions linking signals to actions for the audience.) When coordination is achieved, then, a signal may assume the indicative meaning that "the state of the world is such-and-such" or the imperative meaning "perform such-and-such action!" depending on further characteristics of the situation that need not concern us here. The relevant point is that signaling problems are a special kind of coordination problems.

The builder-assistant language-game of §2 is a clear example of a signaling game that one surmises Lewis might have had it in mind when characterizing the class of signaling games: The builder is the speaker. She observes, for instance, the state of the world in which she needs a slab and she sends the signal "Slab!". The assistant is the audience. He receives, in this example, the signal "Slab!" and performs the action of bringing a slab. The *caveat* issued in the opening section of this paper can now be lifted. If rule-following is conventional action, then meaning, as the by-product of conventional action (signaling) is a special case of rule-following.

In the case of linguistic coordination, the skeptical paradox can be understood as an instance of the problem of indeterminacy of meaning. David Lewis, in *Convention* (cf. pp.199-200) as well as in later works, has tackled the problem. In particular, in (Lewis 1992) he confronts "Kripkenstein's challenge (formerly Goodman's challenge)" (p. 109) and argues that for sentences never uttered before, the rules governing the used fragment of the language determine the rules for the unused portion, too. The argument is that they do so because although extrapolation from used fragment to unused portion is "radically underdetermined", only a *minority* of extrapolations are *straight*—and acceptable—while the vast majority are *bent*: gruesome, gerrymandered—and disposable. The argument carries over also to the extrapolations we all perform daily from precedent to current use. “Straightness” of extrapolation, or of grammar, is a bedrock notion, on which we all agree. Lewis warns us that digging under bedrock—analyzing straightness of extrapolation—cannot be a linguistic enterprise, since our use of language depends on it in the first place. Digging under bedrock points, therefore, to the *ontological* distinction between properties that are natural and properties that are not.

Literature

Wittgenstein, Ludwig Philosophical Investigations, 1953

Kripke, Saul Wittgenstein on Rules and Private Language, 1982

Lewis, David Convention: A Philosophical Study, 1969

Lewis, David “Meaning without use”, Australasian Journal of Philosophy, 1992

Gilbert, Margaret “Rationality and Salience”, Philosophical Studies, 1990

McDowell, John “Wittgenstein on Following a Rule”, Synthèse, 1984

Sillari, Giacomo “A Logical Framework for Convention”, Synthèse, 2005

Sillari, Giacomo “Common Knowledge and Convention”, Topoi, 2008

Science and the Art of Language Maintenance

Deirdre C.P. Smith, Bergen, Norway

1. Classical vs. romantic understanding

In one of many reflections about John, who with his wife Sylvia join the I character and his son Chris for the first half of a motorcycle trip from the Midwest to Montana, the I character comments, “He [John] isn’t so interested in what things *mean* as in what they *are*.” (p. 59). Here Pirsig’s I character intimates a distinction between ‘classical’ and the ‘romantic’ modes of seeing the world. The classical mode is to see what things ‘mean’, their underlying form/structure. The romantic mode is to see the immediate surface/appearance of things, what they ‘are’. When the I character suggests using part of an aluminum can to ‘shim’ John’s handlebars so they stop slipping, John is doubtful. John sees an old aluminum can and is seemingly distressed by using something so base to fix his precision piece of German engineering (a BMW). He sees the surface, what it is. The I character sees beyond the surface to the properties of aluminum, how well they fit the particular demands of a shim (soft, non-rusting), and the appropriate thickness of the can’s aluminum. (p.61) The problem, Pirsig’s I character concludes, is conflicting “*visions of reality*”.

> “What you’ve got here, really, are *two* realities, one of immediate artistic appearance and one of underlying scientific explanation, and they don’t match and they don’t fit and they don’t really have much of anything to do with one another.” (p.63)

Both modes of understanding have faults. The I character notes that John romantically misunderstands what motorcycle maintenance entails. John thinks maintenance is working with hard steel *parts* in an array of shapes and sizes. The I character sees *ideas* and a working on concepts. (p. 102) In short, “That’s all a motorcycle is, a system of concepts worked out in steel. There’s no part in it, no shape in it, that is not out of someone’s mind […].” (p. 104) That said for the classical view, it has its own share of problems. The first is that understanding e.g. a motorcycle from this view presupposes already knowing how it works (the underlying system of concepts). Another difficulty is the absence of an observer, a subject, someone who rides, appreciates or tells stories about the cycle. A third limitation is that it only deals with facts, absent are value judgments of ‘good’ and/or ‘bad’. And, a final objection, perhaps the most important in relation to classical understanding’s own claims, is its cutting edge, what he calls its “intellectual scalpel: “You get the illusion that all those parts are just there and are being named as they exist. But they can be named quite differently and organized quite differently depending on how the knife moves.” (p. 80) And here Pirsig is on to something, how do we decide when and in what direction to cut?

2. Polanyi’s scientific intuition and belief

Deciding which direction to cut is a question Michael Polanyi was interested in exploring. In *Science, Faith and Society*, he uses the analogy of a burglar in the night. If in the middle of the night we hear a noise, a thumping about, in a neighbouring room we know to be unoccupied, we search for an explanation. Is the family cat going after something dangling just out of reach? Has an unlatched window been caught by the wind? Polanyi writes, “We try to guess. Was that a footfall? That means a burglar!”. (p. 23) Presented with an array of ‘facts’ we swing the blade of our intellect in one direction instead of another. Just as a motorcycle can be classified according to different schemes (making a ‘part’ difficult to order because different motorcycle manufacturers have different motorcycle mereologies), for Polanyi,

> “scientific propositions do not refer definitely to any observable facts but are like statements about the presence of a burglar next door—describing something real which may manifest itself in many indefinite ways.” (p. 29)

Although it shows a less demanding level of certainty than one might expect, the burglar scenario does show “a consistent effort at guessing”. (p. 23)

One source for this consistency Polanyi terms “scientific intuition”, a kind of ‘Gestalt’ we have for perceiving contours, arising from an underlying “urge to make contact with a reality which is felt to be there already to start with”. (p. 35) Another source he offers is found in our practices, systems of belief and their embeddedness in language. Here Polanyi draws from the work of social anthropologist E. E. Evans-Pritchard on the Zande tribe of Southern Sudan. When conflicts arise amongst the Zande they consult a poison oracle, which consists in administering a substance, Benge, to a foul. Both the way in which *Benge* is collected and the address given when it is administered are elements crucial to its proper functioning as an oracle-poison and it is to these the Zande turn for explanation when discrepancies in the oracle’s answers arise, rather than to the matter-of-fact poisonness of the *Benge* itself as a European might. For Polanyi, Zande witchcraft exemplifies the power a system of belief has in determining the outcome of the oracle-poison and further, “the power of language to embody and firmly to uphold a system of not explicitly asserted beliefs”. (Polanyi 1952) Here Polanyi concludes:

> “So long as we use a certain language, all questions that we can ask will have to be formulated in it and will thereby confirm the theory of the universe which is implied in the vocabulary and structure of the language.” (Polanyi 1952)

Thus for Polanyi, scientific intuition and the system of belief embedded in language are decisive for determining which way our intellectual scalpel cuts.

3. E.M. Forester on anonymity

Forester writes that “words have two functions to perform: they give information or they create an atmosphere.”(p.77) His arch example of information is a sign reading “Stop” on a tramline. This is an example of pure information. If the tram stops, the sign is correct, if it does not, the sign is incorrect. A sign in a marketplace reading “Beware of pickpockets, male and female.”, however, conjures up Dickensonian images of children having their sweets money stolen, old men being hustled and women unawares having patches deftly snipped from the backs of their fur coats. It produces in us a feeling of foreboding and reminds us of any number of things such as the insecurity and fragility of human life, the violent condition of the poor vs. the obliviousness of the rich, etc., i.e. an atmosphere in addition to the information it conveys. Although the beware pickpockets sign is not great literature, for Forester the atmosphere it creates is the realm of great literature. Although great literature may contain information, e.g. *Zen and the Art of Motorcycle Maintenance* about motorcycles, it is insufficient to be successfully applied by us to actually repair a motorcycle (Pirsig even says so in his author's note). So what is atmosphere and how do we gauge its usefulness? Atmosphere stems not from something conveyed through particular words, but in their arrangement, their style. In this lies their power to elicit dread, mirth and calm, possibly even simultaneously. The realm of atmosphere is one that "answers to its own laws, supports itself, internally coheres, and has a new standard of truth."(p. 81) The truth of information is its accuracy, the truth of a poem whether it "hangs together". (p. 81) "Information points to something else. A poem points to nothing but itself. Information is relative. A poem is absolute." (p.81)

Just as words have two functions, for Forster "each human mind has two personalities, one on the surface, one deeper down". (p. 82) The surface personality "has a name" such as Robert Pirsig. It is this personality that lives in the world, has idiosyncratic habits, relationships, trials and tribulations of the everyday variety. The other, is trickier to pin down, for it has no name and its depths are a ground spring running through the deep personalities of the Pirsigs and Dickenses of this world. It is something general to all humans and inspires works general and accessible to all and often across time. And in this lies the anonymity of great literature: "The poet wrote the poem, no doubt, but he forgot himself while he wrote it, and we forget him while we read." (p. 83). For Forster a signature belongs to the world of information, to the surface personality. The anonymity of great literature belongs to the realm of atmosphere, to deep personality.

4. Wittgenstein, the Life of Words and the Literariness of Language

In the end, the sanity of Pirsig's I character follows suite with the ghost of his previous self. Although Wittgenstein does not write much about insanity, some well know phrases from *Philosophical Investigations* about searching for hidden essences can be taken as a case in point, such as being on slippery ice with no friction (§107) in relation to the sublimity of logic and reaching a point when one's spade is turned (§217) in relation to the regress of rule-following. When Phaedrus continued digging even after his spade reached bedrock, he lost friction with reality and went spinning away from instead of toward it. Both Pirsig's I character, Polanyi and Forster each in their own fashion partake of this error of classical understandings 'depth' thinking, that meaning itself or its generation are something that come from inside of us: the I character for holding that the motorcycle is 'a system of concepts' that 'is primarily a mental phenomenon', the underlying gestalt urge of Polanyi's scientific intuition, and Forester's depth personality as the source of literary anonymity. However, they each offer something I think not only in line with Wittgenstein's linguistic turn on rationality but can help to illustrate it.

If we are to carry a lesson regarding language and reality from Pirsig's novel, a hands-on metaphor of 'tinkering' is where the I character successfully overcame the classical/romantic split he saw in understanding. Yet on the scale of language as a whole, tinkering has its limits. When confronted with Zande witchcraft, no slight adjustment or honing of their intellectual scalpel will lead westerners to accept the judgment of the poison oracle. It will simply not cut that way due to its mode of fabrication. We would need a different scalpel or an altogether different instrument to be at one with the Zande's conceptions of the world. But does this not imply that we can neither redirect nor expand our rationality?

This is where Forster's information – atmosphere continuum and connecting anonymity to atmosphere are illustrative. I hope the reader can agree that language conveys information and atmosphere. Wittgenstein's arguments against private language are in part a defense of it also requiring anonymity. Yet we saw above that an objection to classical understanding was the lack of a subject. Forster's solution was an internal 'ur' subject running through us all which finds its expression in atmosphere. For Wittgenstein the kind of anonymity we find in language comes neither through a depth personality, nor a special place where words live in the mind. Even though Virginia Woolf in her essay "Craftsmanship" claims the later, she also writes the following which I think approaches Wittgenstein's view:

> "Words, English words, are full of echoes, of memories, of associations–naturally. They have been out and about, on people's lips, in their houses, in the streets, in the fields, for so many centuries. And that is one of the chief difficulties in writing them today—that they are so stored with meanings, with memories, that they have contracted so many famous marriages." (p. 131)

Earlier in this essay Woolf writes regarding the 'usefulness' of words. Making a word useful is to give it a single meaning. Forcing words to be useful is a problem. Doing so causes them to mislead us since "it is their nature not to express one simple statement but a thousand possibilities." (p. 127) Put another way, language at the pure information end of Forster's continuum conveys neither accurate nor inaccurate information since it is stripped of the use generated atmosphere against which accuracy could be determined; even a tram "Stop." sign has atmosphere.

The linguistic turn of Wittgenstein's redirection of rationality is akin to Forster's atmosphere and Woolf's depiction of the life of words. Pirsig's I character makes the mistake of attributing this multifarious character of words to a mental instrument unlimited in the directions it can cut. It is rather the case that we can divide things up differently because words, our concepts, do not have single meanings. Philosophy which carves concepts intellectually or claims they can or should have such single meanings goes wrong. Yes, we must know the system, only that the system we need to know to 'tinker' in language, as Polyani recognized, is neither explicit nor explicable hierarchically, we must live it. Concepts are anonymous, but not in the logical or scientific fashion of generality/universality. Meaning is on the surface but, although it sounds strange, deeply there, i.e. over time. Although this kind of meaning is anonymous, it is not stripped of the subject like classical understanding, and therefore, not of value judgments. Subjects are vehicles for the reproduction of language and in their use of words and phrases tinker with and fine tune it. Although we can use language like the poet, forgetting ourselves, and the listener hear our words as general not subjective statements, we are not being poetic, we are simply using words conventionally. But the convention came from somewhere and this is where the subject and their idiosyncratic position in the world can make a lasting contribution.

The thoughts in this paper are born of discussions this spring with Ralph Jewell, Helle Nyvold and Christian Erbacher on Wittgenstein and literature; the notion of ‘tinkering’ comes from Jewell’s reading of Pirsig.

Literature

Forster, E.M. 1951 “Anonymity: An Enquiry”, *Two Cheers for Democracy*, London: Edward Arnold.

Pirsig, Robert M. 1974 *Zen and the Art of Motorcycle Maintenance: An Inquiry into Values*, London: Vintage.

Polanyi, Michael 1964 *Science, Faith and Society*, Chicago: University of Chicago Press.

—. 1952 “The Stability of Beliefs”, *British Journal for the Philosophy of Science* 3:11, pp 217-232. (http://www.missouriwestern.edu/rgs/polanyi/mp-stability.htm)

Wittgenstein, Ludwig 1953, 1997 *Philosophical Investigations*, Oxford: Blackwell.

Woolf, V. (1942) “Craftsmanship”, The *death of the moth and other essays*, London: Hogarth Press, reprinted from *Listener*, 5 May 1937: 868-69.

A Division in Mind. The Misconceived Distinction between Psychological and Phenomenal Properties

Matthias Stefan, Innsbruck, Austria

1. Chalmers' division of mind

According to David Chalmers (Chalmers 1996, 11-13) there are psychological and phenomenal concepts of the mind. The phenomenal concept of mind expresses that there is something it is like to be in a certain mental state. The psychological concept of mind is "the concept of mind as the causal or explanatory basis for behaviour. A state is mental in this sense if it plays the right sort of causal role in the production of behaviour [...]" (Chalmers 1996, 11). In other words, our mental life can be divided into an aspect of the way it feels like to be in that state and into an aspect of the role it plays for our behaviour. So far, this distinction is conceptual (Chalmers 1996, 12).

Soon, however, this conceptual distinction turns into an ontological one about phenomenal and psychological *properties* (e.g. Chalmers 1996, 16, 22-24). Chalmers sees this ontological distinction as exhaustive: Every mental property is either a psychological or a phenomenal property, though most mental concepts encompass both components (Chalmers 1996, 16-17). Take, for instance, pain: Pain can be analysed as phenomenal and psychological property: The phenomenal one describes the way it feels like to be in pain, the pain-experience so to say. The psychological property can be identified with the pain-behaviour and the accompanied beliefs, desires and more. Though according to Chalmers there is a real distinction between these kinds of properties, there is also a co-occurrence of them (Chalmers 1996, 22). Nevertheless, these two properties must be distinguished, since there are two *explananda*. Phenomenal properties cannot be defined in functional terms and psychological properties are not phenomenal. Therefore, we can imagine situations where a phenomenal property is instantiated without a psychological property – and vice versa (Chalmers 1996, 23).

A crucial result of the ontological division between psychological and phenomenal properties is an epistemological division in philosophy of mind: "The division of mental properties into phenomenal and psychological properties has the effect of dividing the mind-body problem into two: an easy part and a hard part." (Chalmers 1996, 24) The easy problem concerns psychological properties. As they are definable in functional terms, they do not really pose a hard problem for cognitive science. Science can explain the psychological property with the underlying physical mechanism that plays the pertinent causal role (ibid.). So the explanation of psychological properties "is a question for the sciences of physical systems. One simply needs to tell a story about the organization of the physical system that allows it to react to environmental stimulations and produce behaviour in the appropriate sorts of ways." (ibid.) Of course, the easy problem is only relatively easy, as there remain considerable difficulties. However, we know in principle how to solve it, how to explain believes, desires, wishes, memory, etc. This is not true in case of the hard problem concerning phenomenal properties. We don't know why and how psychological functions are accompanied with phenomenal states (Chalmers 1996, 25). The explanatory gap Levine famously stated some time ago is as wide open as ever (Levine 1983). We have no idea why it is something it is like to be in a mental state. Chalmers therefore writes: "As we saw above, we now have a pretty good idea of how a physical system can have psychological properties: the *psychological* mind-body problem has been dissolved. What remains is the question of why and how these psychological properties are accompanied by phenomenal properties […]" (Chalmers 1996, 25).

2. Kim's approximation to physicalism

Jaegwon Kim adopts Chalmers distinction for reaching an interpretation of reality near enough to physicalism (Kim 2005). Kim declares himself a physicalist: The world we live in is physical and so are the human person and her mind. Without going into details *why* the mental realm has to be physical, I concentrate on Kim's view *how* it is.

According to Kim mental states are identical with physical states. The relation of reduction that demonstrates this identity relation is functional reduction: As the term says, the first step in functional reduction is to define the reducible property functionally. Kim brings the well-known example of a gene that is defined as "a mechanism that encodes and transmits genetic information" (Kim 2005, 101). Once we have found a definition in terms of the causal role a property plays, we can look for the "causal realizer", that is the property in the reduction base that plays the demanded functional role. To continue the example given above, it turns out that DNA performs the mechanism that defines the concept of a gene. In the last step of functional reduction, a theory has to be given that explains "how the realizers of the property being reduced manage to perform the causal task" (ibid.). In case of the gene example, molecular biology provides the explanation demanded.

In asking whether the mental realm is reducible to the physical, Kim refers to Chalmers' distinction between psychological and phenomenal properties (Kim 2005, 162). As shown, for reducing something it first needs to be defined in functional terms. Therefore the question whether mental properties can be reduced comes up to the question which properties can be functionalized (Kim 2005, 165). Kim adopts Chalmers' proposal that psychological properties, called cognitive properties by Kim, can be explained by cognitive science, which is the same for Kim as to reduce them (see Kim 2005, 108-112 and 162). Thus, they can be identified with causal realizers in the physical basis (Kim 2005, 165). Even though we might not find full causal definitions of psychological properties, it is safe to say that they are identical with physical states (Kim 2005, 167).

Furthermore, Kim and Chalmers also agree on the irreducibility of phenomenal properties, or qualia (the classic term Kim uses). Phenomenal properties are not definable by their causal role and therefore cannot be functionally reduced: "So qualia are not functionalizable, and hence physically irreducible. Qualia, therefore, are the 'mental residue' that cannot be accommodated within the physical domain." (Kim 2005, 170) Kim takes stock: Almost all parts of the mental realm are reducible to entities in the physical realm (Kim 2005, 170-171). As there only remains a small epiphenomenal residue, according to Kim, we have got something near enough to physicalism (Kim 2005, 174).

3. The division reconsidered

I want to reconsider this distinction between psychological and phenomenal properties. I argue that this distinction circumvents appropriate understanding and explaining human behaviour and cognitive capacities. There seems to be a fatal flaw in distinguishing between properties that play a causal role and those that are just raw feelings.

To start with, I cannot find proper arguments Chalmers and Kim allege for the posed distinction. It rather seems they take the distinction between phenomenal and psychological properties for granted. Arguments such as the argument from qualia inversion (Kim 2005, 169-170) or from zombies (Chalmers 1996, e.g. 94-99) are no arguments for the distinction itself, but rather presuppose it. To make it conceivable that there could be zombies, acting like us but feeling nothing, and that there could be inverted qualia, you feeling pain and me feeling itching but both acting the same, one needs to already have accepted the presupposed distinction between phenomenal and psychological properties. In order to agree with such scenarios one has to accept what is at question here. What is there of the presupposed distinction between phenomenal and psychological properties? Why should we show pain-behaviour, if there was nothing like the feeling of pain itself? Or why should we show pain behaviour, if there was the feeling of desperation?

I want to argue against this distinction with two arguments: First, in explaining human behaviour we have to refer to the phenomenal aspect. Second, human cognitive capacities can only be understood properly if the distinction in question is given up.

First argument: It seems wrong to claim, as Chalmers and Kim do, that we can explain human behaviour appropriately if we only refer to psychological properties leaving out what it's like to be in those states. According to them, the concept of pain, for instance, refers to two kinds of properties: There is a property involved that is caused by tissue damage, for instance, and causes pain behaviour, certain believes, desires etc. This psychological property can be identified with some physical state. Every time this property is instantiated there also occurs a second property (Chalmers 1996, 17), namely the pain-feel that has no causal effects whatsoever. This view on the matter, however, is highly artificial. In case of machines this assumption might be true. We can explain the behaviour of machines in purely psychological terms, i.e. by pointing out the physical mechanisms that are activated by certain causes and having themselves a causal role. Explaining why a robot is behaving the way he does, it suffices to point out the causal mechanism that underlies its behaviour.

In case of humans however, we cannot understandably explain her behaviour by simply pointing at purely psychological properties and leaving out what it's like to be in a certain state. We act *because* of the pain-feeling, instead of it being some unnecessary adjunct. If the feeling of pain did not occur, pain-behaviour also wouldn't. The crucial point is that we would not understand *why* someone behaves the way she does, if we do not account for her phenomenal feeling. To give another example, we would not understand why a cheated husband behaves the way he does in presence of the adulterer if we leave out his hate, frustration and jealousy. His actions are only comprehensible by *knowing what its like* to be jealous and hating someone. According to Kim and Chalmers, in contrast, the poor husband's behaviour would be explained by a psychological property and therefore ultimately by a physical mechanism: He behaves the way he does because there is a causal mechanism going on in his physical system. Within such a framework, however, *rational understanding* of human behaviour is made impossible. Leaving out what it's like is nothing short of obstructing the understanding of human behaviour and actions. This is not to claim that we act as we do because of the phenomenal feeling instead of the intentional states. Rather, in case of humans, there is no distinction here.

This should become clear if we contrast human behaviour with the behaviour of machines. In case of machines we understand why they behave the way they do if we know what kind of mechanism is at work. In case of humans, however, there is a totally different way of understanding. We give up rational understanding another human person's behaviour if we overlook what it's like for someone to be in that state. There might be resemblance to human behaviour when a machine reacts in a way that is similar to pain behaviour. In fact, however, two different things are going on, namely human pain *behaviour* on the one hand and purely mechanical reactions resembling pain behaviour on the other. If we talk about “machine behaviour”, we are using the concept of behaviour in an equivocal way.

Second argument: Jonathan Lowe criticizes Chalmers because he cannot explain human cognitive capacities appropriately (Lowe 1995). According to Lowe to understand (most) human cognitive capacities one has to consider its phenomenal part. Describing the abilities to discriminate, categorize, react to the environment, etc. in purely functional terms can only explain phenomena occurring in case of machines. Explaining machine behaviour this way, however, is trivial (Lowe 1995, 270).

Lowe gives the example of experience to show that one cannot distinguish between phenomenal properties with no causal role and a psychological mechanism that properly explains representation. Experience is not just a phenomenal property occurring in case of representing the environment. Rather the representation itself is a phenomenal representation: “Not only is it ‘like something’ to enjoy such an experience, in which the phenomenal character of sensed colours impresses itself upon our awareness, but also such an experience represents – or, better, *presents* – our immediate physical environment as *being some way*” (ibid.). Representing spatial facts, for instance, such as the height of a plain, its shape, etc., is always phenomenal representation. So the intentional and conceptual content of experience is intimately tied to the phenomenal aspect of representation: “The importance of all this lies in the fact that how we *conceive* of physical objects is inextricably bound up with how they *appear* to us in perception” (Lowe 1995, 268). If Chalmers and Kim define human representation as psychological properties and therefore in purely functional terms, they oversee that human representation is always given in a certain phenomenal way. A purely functional description of environmental representation might be appropriate for information processing, storage and retrieval of machines, but it is not appropriate in the case of humans.

The example of experience shows that in human mind the phenomenal and the intentional aspect are not divisible. There are enough examples in human cognition and behaviour that show that the distinction between phenomenal and psychological properties is wrong in case of human beings. To explain machine behaviour in purely functional terms is trivial. But only because such behaviour can be explained by the underlying mechanism doesn't mean that this is also true in case of humans. Talking about functional explanation of machine behaviour is simply to change the subject. In case of humans, I therefore conclude, Chalmers' and Kim's division of the mental cannot be applied.

Literature

Chalmers, David 1996 The Conscious Mind. In Search of a Fundamental Theory, Oxford.

Kim, Jaegwon 2005 Physicalism, or Something Near Enough, Princeton.

Levine, Joseph 1983 "Materialism and Qualia: The Explanatory Gap", Pacific Philosophical Quarterly 64, 354-361.

Lowe, Jonathan 1995 "There are no Easy Problems of Consciousness", Journal of Consciousness Studies 2 (3), 266-271.

Scepticism, Wittgenstein's Hinge Propositions, and Common Ground

Erik Stei, Mainz, Germany

1. Introduction

Traditionally, systematic scepticism states that knowing a contingent proposition *p*, e. g., that I have hands, entails that I know the denial of a sceptical hypothesis (*sh*), e. g., that I am a brain in a vat. Because *sh* is usually formulated in a way that apparently makes it impossible for me to know its denial on basis of my evidence, it seems as if I do not know *p*. More formally, the sceptical argument looks like this[1]:

(1) a. K(*S, p*) → K(*S,* ¬*sh*)
b. ¬K(*S,* ¬*sh*)
c. ¬K(*S, p*)

The notorious Moorean response is to make the sceptical *modus tollens* a *modus ponens* by affirming the antecedent of (1a) in the second step of the inference, concluding K(*S,* ¬*sh*). Effectively, if we want to maintain epistemic closure, (1b) seems to be the crucial step. However, which premise we assume—K(*S, p*) or ¬K(*S,* ¬*sh*)—certainly needs some further justification. The problem is that in an epistemology classroom the sceptic's story sounds just as plausible as our ordinary knowledge-claims, which seems to leave us in an epistemically underdetermined situation.

However, I follow Williams's (2003) conjecture that the justification of either of the two propositions depends (at least in part) on the question which other propositions we (tacitly) accept in a given context. It is Wittgenstein's notion of *hinge propositions* that underpins this thought.

2. Hinge propositions

The metaphor of *hinges* (HP) goes back to a famous paragraph in *On Certainty*, in which it is employed to illustrate the necessity of accepting certain propositions in order to make sense of doubting others:

> "That is to say, the *questions* that we raise and our *doubts* depend on the fact that some propositions are exempt from doubt, are as it were like hinges on which those turn."
> (Wittgenstein 1984: 186 – § 341)[2]

Thus if "I make an experiment I do not doubt the existence of the apparatus before my eyes. I have plenty of doubts, but not *that*" (Wittgenstein 1984: 185). This, however, does not mean that there is a rigid set of self-evident propositions, i. e., Wittgenstein does not advocate a strictly foundationalist conception of justification. This is quite obvious in the following paragraph:

> "But it isn't that the situation is like this: We just *can't* investigate everything, and for that reason we are forced to rest content with assumption. If I want the door to turn, the hinges must stay put."
> (Wittgenstein 1984: 187 – § 343)

Now, how could these ideas help us with respect to the sceptical problem? A first rough answer is that *i)* which proposition we are justified in assuming as our second premise (1b) depends on certain HP and that *ii)* these HP in turn depend on the inquiry we are interested in. Thus, Wittgenstein's remarks seem to support the idea that a given knowledge claim might vary in truth value, depending on which HP we accept in a conversational context or in certain circumstances of evaluation.[3] If this is correct, then using the notion of HP might strengthen the thesis that knowledge claims are sensitive to contextual factors in a wider sense.

3. Sceptical Presuppositions

Before going on, let me briefly sketch another important feature of this view. One dubious point often made in favour of the sceptic is the Cartesian claim that scepticism operates in some kind of neutral context and does not rely on any axioms or assumptions. If this were true then the sceptic would in fact be in a privileged position. However, Williams 2007 offers many arguments for the point that this claim is in fact a platitude, arguing that scepticism does indeed depend on certain assumptions or HP. Many philosophers from quite a diverse range of positions seem to share at least the spirit of this view. From a methodological point of view, Timothy Williamson highlights that "the sceptic relies uncritically on rules of dialectical engagement (…), without questioning their appropriateness to the radical questions which scepticism raises" (Williamson 2000: 188). Another point, addressed by Barry Stroud (Stroud 2000), could be the (traditional) philosopher's wish to give a completely general account of *knowledge*. However, even if this general kind of knowledge is not available to us it does not follow that we cannot have any more specific kind of knowledge. Still another presupposition of the sceptic is that our perceptions are independent of how the world *really* is and that the meanings we assign to words and propositions are *in our heads* only. It is due to this internalist worry that scepticism seems so convincing, but that is no reason for accepting it. There is of course much more to be said on this topic, but let me adhere without further argument, relying on Wittgenstein, that there must be something the sceptic takes for granted in order to make sense of her worries in the first place.

Obviously, this analysis only transfers the problem of underdeterminacy to a meta-level. To see this, let us grant for the moment that according to the above picture the justification for assuming, say, (1b) is that it can be derived from a set of HP operative in a given context. We

[1] Where *knowledge* is symbolized by a two-place relation K, taking an epistemic subject *S* and a proposition (*p* or *sh*) as arguments. I will disregard time in this paper.

[2] Translation of Wittgenstein quotes by Denis Paul and G. E. M. Anscombe, as found on http://budni.by.ru/oncertainty.html on March, 20th, 2008

[3] I am deliberately vague on the mechanisms yielding the truth-value of knowledge claims as I do not want to enter the debate between Contextualism and Subject Sensitive Invariantism. What is important here, however, is that both views hold that a given knowledge claim can vary in truth-value.

could then deduce (1c) and conclude that *S* does not know any ordinary proposition *p*. In another context, different HP might be taken for granted that are consistent with K(*S, p*) but not with some of the sceptic's presuppositions outlined above. Then (1) would allow us to infer that *S* knows the denial of the sceptical hypothesis. Note, however, that even if the sceptic relies on certain HP this is not enough to give a direct argument against her. But it is enough to restrict the devastating impact on our ordinary knowledge claims to contexts in which sceptical HP hold. This is one way to make sense of the conjecture that the truth of a given knowledge claim can vary with context.

A point that needs more work is avoiding the account to over generate. So far it seems as if any knowledge claim could be justified, given that certain HP hold, as absurd as they may be. So we need a mechanism that *i)* restricts the admissible HP and that *ii)* can bridge the gap between internal factors like *taking something for granted* and the external factors that allow for the implication $K(S, p) \rightarrow p$, i. e., the factivity of knowledge. An additional externalist conception of justification could solve both of these problems. Consider, for the sake of illustration, a reliabilist account that incorporates the methods by which we form our beliefs about certain propositions. This would avoid not only fancyful HP as these would be analogous to unreliable methods, but also Gettier cases. It would have the further merit of allowing for *fallibility*, as a method of forming beliefs that we thought trustworthy might turn out to be false. In the spirit of the experiment-example in § 2: Future investigation might reveal that the apparatus we used is not as precise as we took it to be or even that it does not measure what we thought it did. Even though some of the beliefs we based on the data the apparatus delivered were in fact true, we would no longer consider them as *known* due to the unreliable process by which we formed our belief in them. However, nothing hinges on this specific epistemological account. The notion of HP is fairly neutral to other externalist suggestions.

4. Conversational mechanisms

The analysis sketched so far did not address the conversational application of different HP. One possibility to systematize the account seems to be the incorporation of a Stalnaker-style notion of *common ground*. The definition is as follows:

> "It is common ground that Φ in a group if all members *accept* (for the purpose of the conversation) that Φ, and all believe that all accept that Φ, and all believe that all believe that all accept that Φ, etc." (Stalnaker 2002: 716)

Acceptance, following Stalnaker, is a propositional attitude as well as a methodological stance toward a proposition. To accept some proposition Φ is to treat it as true and to "ignore, at least temporarily, and perhaps in a limited context, the possibility that it is false" (Stalnaker 2002: 716). Applying this idea to the position developed in § 3, this means, roughly, that we accept some proposition Φ as true in a context C_1 in order to undertake a reasonable investigation I_1. Given Φ is an essential presupposition in C_1, then in case we doubt Φ, this shifts the context as well as the investigation we are interested in to I_2 in C_2.[4] Thus, what depends on contextual factors is the question which propositions we are willing to accept as given and thus where to stop the regress of justification. The connection I propose is that HP determine, at least in part, which propositions we are willing to accept.

In standard cases, conversational contexts are dynamic in such a way that they are constantly extended by the incorporation of new information—this is a central feature of communication and not an instance of context shifts. It is captured by the notion of *accommodation*, which can be illustrated by the following example. Imagine a conversation between Alice and Bob, where Alice utters:

(2) I can't come to the meeting – I have to pick up my sister at the airport.

Given that Alice is a competent speaker of English and also that the basic pragmatic mechanisms, e. g., a Gricean-style *Cooperative Principle* and some conversational maxims (Grice 1989), are at work, Bob can infer from (2) that Alice believes that it is common belief that she has a sister. The latter now common belief (as Bob believes it as well after Alice's utterance of (2)) leads Bob to belief that Alice has a sister, which makes Alice's having a sister common belief in the context at issue (cf. Stalnaker 2002: 709-710). If Bob and Alice are discussing the question who is going to the meeting next Friday, Alice has merely introduced a proposition to be integrated into the common ground by decreasing the set of possible world compatible with the state of the conversation. Consider, in contrast, the following dialogue:

(3) a. [Alice]: I can't come to the meeting – I have to pick up my sister at the airport.

b. [Bob]: You can't pick up your sister at the airport. In fact there is no airport, there is no sister, and I'm not here either. You are a brain in a vat with all your impressions stimulated by a mad scientist.

In this case, it does not seem as if Bob introduced new information. Rather, his answer (3b) reveals a *defective context*. While in a *nondefective context* "the participants' beliefs about the common ground are all correct" (Stalnaker 2002: 717), this is not the case in a defective context. Before Bob's utterance of (3b), Alice believed that a proposition like *that the external world exists* was part of the common ground. Bob's answer, however, made it manifest that it was not and thus, that the context was defective. Alice now seems to have two options: either she accommodates Bob's utterance and thus accepts the shift to a sceptical context, i. e., she accepts entering a new conversation, or she rejects the context-shift by refusing to accommodate (3b). This leads to the quite natural result that in case she accepts the context-shift, we would not be inclined to ascribe her knowledge of any ordinary proposition. If, however, she refuses accommodating (3b) we are still willing to ascribe her knowledge of that proposition. I think it is the latter option that is more likely to be chosen in every-day conversation.

[4] Stalnaker does not mention context *shifts*, but only context *changes*, that decrease the set of possible worlds under consideration. I distinguish *context-shifts* as the result of a defective context in which a given conversational aim could not be achieved. It is similar to beginning an entirely new conversation.

5. Conclusion

In this paper I did *not* address the debate between contextualism and sensitive invariantism. This will be an important issue when it comes to deciding whose position is decisive for the evaluation of knowledge claims—the ascriber's or the subject's. It will have to be discussed elsewhere.

I *did* argue for the idea that Wittgenstein's *Hinge Propositions* support the thesis that the truth of a given knowledge claim can vary due to factors other than its overt variables. It could be shown how the apparent dogmatic deadlock between sceptics and Mooreans can be avoided by relativizing the truth of knowledge claims to sets of quasi-foundationalist HP. This move does not result in relativism, because it was indicated that the higher order question of which set of HP is preferable can be linked to an externalism of justification. A brief recapitulation of *Common Ground* then gave an idea of how to systematize the account in conversational contexts and explain the apparently contradicting intuitions about many of our knowledge claims.

Literature

Grice, Paul 1989 “Logic and Conversation”, in: *Studies in the Way of Words*, Harvard: Harvard University Press, 22–40.

Stalnaker, Robert 2002 “Common Ground”, *Linguistics and Philosophy* 25, 701–721.

Stroud, Barry 2000 “Understanding Human Knowledge in General”, in: *Understanding Human Knowledge*, Oxford: Oxford University Press, 99–122.

Williams, Michael 2003 “Skeptizismus und der Kontext der Philosophie”, *Deutsche Zeitschrift für Philosophie* 51, 973–991.

Williams, Michael 2007 “Why (Wittgensteinian) Contextualism is not Relativism”, *Episteme* 4, 93–114.

Williamson, Timothy 2000 *Knowledge and its Limits*, Oxford: Oxford University Press.

Wittgenstein, Ludwig 1984 *Über Gewissheit*, Werkausgabe, Band 8, Frankfurt am Main: Suhrkamp.

Neutral Monism. A Miraculous, Incoherent, and Mislabeled Doctrine?

Leopold Stubenberg, Notre Dame, Indiana, USA

Neutral monism (NM) is a general metaphysical doctrine about the nature of ultimate reality. It says that ultimate reality is, in an important sense, *one*—that is the *monism* of NM; and it says that this monistic reality is *neither mental nor physical*—that is the *neutrality* of NM. For the most part, the advocates of NM have been preoccupied with the mind-body problem. If mind and matter are composed of the same neutral reality, the gulf that separates them must be more apparent than real.

In the late 19th and early 20th centuries NM enjoyed a brief period of popularity, having been adopted by such philosophers as Ernst Mach, William James, and Bertrand Russell. But the doctrine soon dropped out of view, possibly because some of the criticisms directed against it seemed justified and serious. Prominent among those criticisms was (and is) the claim that NM is nothing but a thinly veiled versions of idealism or panpsychism. In an interesting reversal of things, Galen Strawson—one of the leading figures in the current revival of panpsychism—has subjected NM to a trenchant critique. According to Strawson, NM is either miraculous or incoherent.

After presenting Strawson's arguments against NM, I argue that Russell's version of the doctrine is immune to Strawson's powerful objections. But Russell's way out will, in turn, make it all too obvious why it has seemed to many philosophers that NM is nothing but idealism or panpsychism. I shall end the paper by explaining why the Russellian neutral monist rejects the charge that NM is a version of mentalism. The upshot of all this is twofold. First, NM (of the Russellian variety) is not open to the challenge that Strawson tried to raise against all versions of the doctrine. Second, this kind of NM must be distinguished from mentalistic doctrines like idealism and panpsychism.

The most interesting (and most controversial) premise upon which Strawson's case for panpsychism rests is this:

> *Emergence can't be brute.* It is built into the heart of the notion of emergence that emergence cannot be brute in the sense of there being absolutely no reason in the nature of things why the emerging thing is as it is (so that it is unintelligible even to God). For any feature Y of anything that is correctly considered to be emergent from X, there must be something about X and X alone in virtue of which Y emerges, and which is sufficient for Y.
> (Strawson 2006, 18)

The emergence of the liquidity of water is unproblematical; but the emergence of experience from nonexperiential matter is not. If it were to happen, it would be a miracle. Only the adoption of panpsychism makes this miracle go away. And Strawson argues that the same consideration cuts against all forms of neutral monism. If basic reality is neutral, i.e. nonexperiential, then the experiential cannot emerge out of it, on pain of a miracle. Hence neutral monism is not an option.

Strawson's second objection to NM is this:

> [NM] is incoherent, because experience—appearance, if you like—cannot itself be only appearance, i.e. not really real, because there must be experience for there to be appearance …
> (Strawson 2006, 23)

If only the neutral (i.e. the nonexperiential) is real, then the experiential cannot be fundamentally real, i.e. it must be a mere appearance. But that is absurd: there cannot be appearance without experience. x's appearing F simply consists in someone's experiencing x as F. One cannot get rid of experience be declaring it to be a mere appearance. Because appearance can only exist where there is experience.

These objections are powerful. But a closer inspection of Russell's NM seems to show that it is immune to these challenges. Guided by the supreme maxim of scientific philosophizing—"wherever possible, substitute constructions out of known entities for inferences to unknown entities." (Russell 1924, 326)—Russell adopted NM in 1919, because of the "immense simplification" (Russell 1959, 103-4) it affords. The unknown entities in need of logical construction are, on the one hand, physical objects, and, on the other, the self. The known entities that serve as construction material are events. And some of these events, the data, are immediately given.

> Everything in the world is composed of events … An 'event' … is something having a small finite duration and a small finite extension in space … When I speak of an 'event' I do not mean anything out of the way. Seeing a flash of lighting is an event, so is hearing a tire burst, or smelling a rotten egg, or feeling the coldness of a frog. These are events that are "data" … we infer that there are events which are not data and happen at a distance from our own body. Some of these are data to other people, other are data to no one … Particular colours and sounds and so on are events; their causal antecedents in the inanimate world are also events.
> (Russell 1927b, 222)

The resulting view does have an idealist ring to it. Speaking of the brain, for example, Russell has this to say:

> While its [the brain's] owner was alive, part, at least, of the contents of his brain consisted of his percepts, thoughts, and feelings. Since his brain also consisted of electrons, we are compelled to conclude that an electron is a grouping of events, and that, if the electron is in a human brain, some of the events composing it are likely to be some of the "mental states" of the man to whom the brain belongs. (Russell 1927a, 320)

"The brain consists of thoughts" (Russell 1959, 18) is Russell's most succinct expression of this thought. As for the rest of the physical world, Russell is agnostic. But he does tell us that

> If there is any intellectual difficulty in supposing that the physical world is intrinsically quite unlike that of percepts, this is reason for supposing that there is not this complete unlikeness. And there is a certain ground for such a view, in the fact that percepts are part of the physical world, and are the only part that we can know without the help of rather elaborate and difficult inferences. (Russell 1927a, 264)

This sketch of Russell's view shows that Strawson's objections to NM do not apply to Russell's version. Mental events do not emerge from nonexperiential matter. On the contrary: "mental events are part of the [stuff of the world]." (Russell 1927a, 388) And they are not mere appearance; on the contrary: they are the only part of the basic reality that is given to us. But this sketch of Russell's view also makes it quite clear why so many philosophers—Lenin, Popper, Feigl, Nagel, Chalmers—have accused (or at least suspected) Russell of endorsing some version of mentalism: idealism, panpsychism, phenomenalism, Berkleyanism, etc. Russell himself was quite aware of the mentalistic flavor of his view; in fact, he thought that "physics must be interpreted in a way which tends towards idealism". (Russell 1927a, 7) But unlike his many critics, he thought that he had avoided the slide into mentalism.

Seeing a flash of lightning—Russell's paradigm of an event—is an experience. And experiences are mental (not neutral), if anything is. So how can Russell maintain that his event monism is a *neutral* monism? Before embracing NM Russell was a dualist. The subject and its act of sensing were mental; the sensed object, the sense-datum, was physical. Once Russell discarded the subject as a basic or unconstructed entity, "the possibility of distinguishing the sensation from the sense-datum vanishes;" (Russell 1921, 142) and "the reason for distinguishing the sense-datum from the sensation disappears, and we may say that the patch of colour and our sensation in seeing it are identical." (Russell 1921, 143) The resulting entity/event—the percept—is neither mental, like the act, nor physical, like the sense-datum.

Consideration of the percept itself (e.g., seeing a flash of lightning) also yields the result that it is neutral. The percept is not material in the traditional sense. For "matter as it appears to common sense, and as it has until recently appeared in physics, must be given up." (Russell 1927b, 125) It is replaced by a logical construction out of events, among which are percepts. The term "matter" applies to a complex logical construction; the elements that go into this construction are not properly called material. Let us now turn to the controversial claim that percepts (and other events that are given to us) are not mental. Russell arrives at this conclusion by observing that neither one of the traditional criteria of the mental—intentionality and consciousness—classifies percepts as mental. Take intentionality first. The percept, as such, does not represent anything; having a percept is not, as Russell puts it, a matter of knowing anything:

> When, say, I see a person I know coming towards me in the street, it *seems* as thought the mere seeing were knowledge. It is of course undeniable that knowledge comes *through* the seeing, but I think it is a mistake to regard the mere seeing itself as knowledge. If we are to so regard it, we must distinguish the seeing from what is seen: we must say that, when we see a patch of colour of a certain shape, the patch of colour is one thing and our seeing of it is another. This view, however, demands the admission of the subject, or act … If there is a subject, it can have a relation to the patch of colour, namely, the sort of relation which we might call awareness. In that case the sensation, as a mental event, will consist of awareness of the colour, while the colour itself will remain wholly physical, and may be called the sense-datum, to distinguish it from the sensation. (Russell 1921, 141)

But once the subject (and its act) are given up, the distinction between the sensing and the sensed collapses, and we can no longer understand an episode of seeing as intentionally structured, as some sort of representation with an intentional object.

What about the second mark of the mental: consciousness, phenomenal quality, what-it's-likeness? Phenomenal quality can serve to distinguish the mental from the non-mental events only if the former do and that latter don't have this feature. But, Russell maintains, we do not know whether the events that are not given to us do or don't have these features. Russell grants that we can infer the existence of events that are not given to us. But he insists that our knowledge of these events is limited to their structural features; we cannot know anything about their intrinsic natures:

> Whenever one complex structure causes another, there must be much the same structure in the cause and in the effect, as in the case of the gramophone record and the music. This is plausible if we accept the maxim "Same cause, same effect" and its consequence, "Different effects, different causes." If this principle is regarded as valid, we can infer from a complex sensation or series of sensations the structure of its physical cause, but nothing more … (Russell 1948, 254)

We cannot infer anything about the intrinsic nature of the vast majority of events—those given to other persons and those given to nobody. For all we know, their intrinsic quality is not different from the events that are our experiences. Hence we cannot say that your experiences—the events that are given to us—must count as mental because they have phenomenal quality. They do, indeed, have phenomenal quality; but so may all other events in the universe.

We are left with the conclusion that percepts—the events Russell is left with upon identifying the act of sensing with the sensed object—are neutral. They are the stuff out of which matter is constructed; hence they are not themselves material. And because they lack intentionality or a *distinctive* phenomenal quality they cannot be called mental. Hence they are neutral. This, then, is how Russell can acknowledge that

> on the question of the stuff of the world [NM] has certain affinities with idealism—namely, that mental events are part of that stuff, and that the rest of the stuff resembles them more than it resembles traditional billiard balls (Russell 1927a, 388)

while maintaining that his NM is truly neutral and not merely a thinly veiled form of mentalism.

To sum up. Strawson has argued that NM is either committed to a miraculous form of emergence or it is incoherent. In reply I have argued that in Russell's scheme of things the mental does not emerge from the nonmental. Far from it—mental events form part of the neural material from which matter is constructed. Nor is the Russellian neutral monist obliged to make the incoherent claim that experiences are mere appearances. Far from it—experiences are among the elements of ultimate reality.

And while many passages in Russell's texts may suggest the view that his so-called NM is merely idealism or panpsychism in a new guise, I have tried to show that this impression is misleading. While acknowledging certain parallels between NM and idealism, Russell is careful to establish the neutrality of the basic elements of his metaphysics.

Literature

Chalmers, David 1996 *The Conscious Mind. In Search of a Fundamental Theory*, New York: Oxford University Press.

Feigl, Herbert 1958 "The Mental and the Physical", Minnesota Studies in the Philosophy of Science, Volume II, 370-498.

Lenin, V.I., 1909 Materialism and Empirio-Criticism, Moscow: Foreign Languages Publishing House, 1959.

Nagel, Thomas 2002 "The Psychophysical Nexus", in: Nagel, Thomas *Concealment and Exposure,* Oxford: Oxford University Press, 194-235.

Popper, Karl R. & Eccles, John (eds.) 1977 *The Self and Its Brain*, London: Routledge & Kegan Paul.

Russell, Bertrand 1921 *The Analysis of Mind,* New York: Humanities Press 1978.

Russell, Bertrand 1924 "Logical Atomism" in: Russell, Bertrand 1956 *Logic and Knowledge. Essays 1901-1955*, Edited by Robert C. Marsh, London: Allen & Unwin, 321-343.

Russell, Bertrand 1927a *The Analysis of Matter*, New York: Dover 1954.

Russell, Bertrand 1927b *An Outline of Philosophy*, London: Allen and Unwin 1979.

Russell, Bertrand 1948 *Human Knowledge. It's Scope and Limits,* New York: Simon and Schuster 1976.

Russell, Bertrand 1959 *My Philosophical Development,* London: Unwin Books 1975

Strawson, Galen 2006 "Realistic Monism" in Strawson, Galen 2006 *Consciousness and its Place in Nature*, Exeter: Imprint Academic, 3-31.

A somewhat Eliminativist Proposal about Phenomenal Consciousness

Pär Sundström, Umeå, Sweden

1. Introduction

Let eliminativism about an object, x, or property, X, be the claim that x doesn't exist or X is not instantiated. Atheists are in this sense eliminativists about all gods, Christians are eliminativists about Zeus. More universally, we tend these days to be eliminativists about impetus, caloric and phlogiston.

One may wonder how global the threat of elimination is. Among the objects and properties that we take to exist or be instantiated today, which ones may be up for "elimination" tomorrow? This paper focuses on the case of phenomenal consciousness, or "what it is like" to be in certain mental states.

It is often claimed that consciousness is secure against elimination. Flanagan, for example, urges that consciousness differs importantly from objects and properties that have been "eliminated" in the past. Phlogiston was hypothesised to exist because it could explain phenomena that are more immediately present to us, like burning and rusting. But consciousness is not, or not only, assumed to exist because it explains other, more immediately observable phenomena. Consciousness is also, and perhaps primarily, assumed to exist because it is *itself* immediately present to us. As Flanagan puts it: "consciousness as a phenomenon to be explained has a secure place at the observational periphery" (1992, 33; cf. also Chalmers 1996, 102).

It's easy to feel the force of Flanagan's suggestion. However, I shall try to cast some doubt on it in what follows.

Section 2 makes some preliminary remarks about "eliminativist" and "revisionist" outcomes of theoretical developments. Section 3 develops the somewhat eliminativist proposal about phenomenal consciousness. Section 4 elaborates on the eliminativist character of the proposal. Section 5 comments on the relation between the proposal and "monitor theories" of consciousness.

2. Eliminativist and revisionist conclusions

Every now and then, we realise that things are not quite the way we once thought they were. Nothing is quite the way Newtonians thought that mass were. The element centrally involved in burning and rusting is not the way phlogiston was taken to be. The heat of a body is not what caloric theorists thought it was. Royalties don't exercise power with a divine mandate, and solidity is not a matter of being dense all over.

In some cases where we realise that things aren't the way we thought they were, we end up saying *eliminativist* things like, "there is no x". In other cases, we end up saying *revisionist* things like, "x is not quite what we thought it was".

What determines whether, in a given case, we end up saying one thing or the other? Presumably the *magnitude* of our change of view – however exactly that should be measured – plays some role. If our views on a topic change (by some measure) to a significant extent, we are presumably more inclined to draw an eliminativist conclusion than if they revised to a lesser extent. But it's possible that more "pragmatic" factors play a role as well, for example, whether researchers judge that they will make a greater impact by using one or the other kind of formulation (c.f. Churchland 1986, 283-4; and Stich 1996, chapter 1).

However that may be, it seems to me that, whether in a given case we end up talking in an eliminativist or a revisionist way is not *as such* of any interest. What is of interest is our change of view, and I suggest that we can achieve at least an intuitive sense of how significant such a change is that is independent of how we end up speaking. For example, if Lavoisier had convinced us to say things like, "phlogiston exists but it's not what we thought it was", I suggest that we could have achieved the same appreciation of the theoretical change that he contributed to bring about.

I shall next develop a proposal concerning consciousness. Whether accepting this proposal would lead us to make eliminativist or revisionist claims, I believe that, if one comes from a certain natural and commonly occupied starting point, it would amount to a significant change of view.

3. A proposal about phenomenal consciousness

I shall develop my proposal in three steps.

Step 1: Sifting out the "Galilean qualities". Consider a visual experience of a ripe lemon in good lightning condition. Salient in this experience is a certain yellowish quality. It's somewhat tricky to make this quality a joint topic of conversation, because there are so many disagreements about it. For example, while naïve realists take it to be a property of lemons, sense-datum theorists to be a property of sense-data, some qualia theorists may take it to be a property of conscious experiences, and others take it to not be instantiated at all. There is also disagreement about how we talk, should talk, and can talk about this quality. Some find it natural to use the term "yellow" to talk about it, but others think that we don't have a public language term for it, and even that we *can't introduce* such a term (Thau 2002, section 5.13).

But despite these obstacles, I think we can make this kind of quality a joint topic of conversation. Whatever instantiates it, and whatever it can and should be called, it is the kind of quality that is most salient in our colour experiences. I will suppose that we have a shared understanding of which type of quality this is.

I will call these qualities "Galilean", since I take it to be the kind of quality that Galileo was concerned with when he discussed what qualities belong to the world and what qualities belong to the mind. (I shall later distinguish these from alleged qualities of another type.) I use "Galilean quality" broadly, for qualities that are salient in various sense perceptions like sight, smell and taste, and bodily sensations like pains and itches. However, for brevity I shall focus on the Galilean qualities that are displayed in colour experiences.

My first proposal in working towards a somewhat eliminativist view of phenomenal consciousness is that *Galilean qualities are not constitutive properties of phenomenal consciousness*. Galilean qualities may be not instantiated at all, or they may be properties of objects like lemons. In either case, they don't contribute to constitute what it is like to have a conscious experience. I call this *The Separation Thesis*, since it says that consciousness and the Galilean qualities are, in a certain sense, separate.[1]

The Separation Thesis is presumably somewhat controversial. But it is surprisingly hard to say *how* controversial it is. Even though it is quite central to our understanding of what consciousness is, few philosophers who discuss consciousness make clear whether they accept or reject it. (No doubt the difficulties in talking about the quality, noted above, contribute to this unclarity.) However, my impression – largely based on an admittedly non-scientific selection of conversations – is that few philosophers would deny The Separation Thesis out of hand. Moreover, and more importantly, I'm inclined to think there are good reasons to accept the thesis. Space does not allow an elaboration of these reasons however, and here The Separation Thesis will be just assumed.

Step 2: Noting the elusive character of whatever is left. Suppose The Separation Thesis is right. What then is phenomenal consciousness like? If what it's like to experience a ripe lemon is not in part constituted by a Galilean quality, what *is* it like? It seems that whatever is left is pretty elusive. In fact, once the Galilean qualities are assumed to not be part of consciousness, one may start to wonder whether there is much or anything left of the phenomenon at all. Perhaps consciousness doesn't have such "secure place at the observational periphery" after all?

It has been suggested to me in this context that consciousness may be constituted, wholly or in part, by a set of non-Galilean qualities. The suggestion is that an experience of a ripe lemon displays *both* a Galilean quality and a quality of another kind. Even if the former doesn't constitute what the experience is like, the latter does.

However, even if I search really hard, I fail to find in my experiences a set of qualities over and above the Galilean ones. Granted, this may be because my introspection is deficient, or because I'm blinded by some prejudice. However, it may also be because there are no extra qualities there, and that those who seem to find them are projecting on the basis of some prejudice. My present case assumes that the latter view is the right one.

Step 3: Identifying whatever is left with first-person awareness. One might think that, even if steps 1 and 2 above are taken, there remains a robust and salient phenomenon of consciousness. For compare your typical, familiar visual experience of a ripe lemon with a subliminal experience of a ripe lemon. There is a striking difference between the two. The subliminal experience goes on "in the dark" while the other one has a vivid phenomenology.

The difference is surely salient. However, it seems to me that it can plausibly be accounted for in a rather deflationary way. The proposal is that the difference is simply one of first-person awareness. In the normal case, I am aware in a peculiar first-personal way of my visual experience, or that I have it. In the subliminal case, I'm not aware – or at least, I'm not aware in that way – of my experience or that I have it.

4. The eliminativist character of the proposal

It is natural to think of phenomenally conscious states as somehow "shining" or "glowing". From that perspective, the present proposal seems to "deflate" our conception of consciousness. A tree doesn't shine any more when it is perceived than when it is not perceived; what comes and goes is only the relational property of being perceptually registered. According to the present proposal, the difference between a typical, familiar perception and a subliminal perception is of just the same kind.

The deflationary character of the proposal can also be brought out by contrasting it with certain suggestions about what "phenomenal realism" entails. According to Block, for example, you are a phenomenal realist only if you accept that consciousness resists a priori or armchairs analyses in "non-phenomenal terms" such as "representation, thought or function" (Block 2002b, 392). The present proposal would seem to qualify as non-realism about phenomenal consciousness, by Block's lights. It says that, if there is phenomenal consciousness at all, it is nothing other than a kind of representation, to wit, first-personal awareness; and, while it is presumably advisable to take into account everything you know about the world when you assess the proposal, I suspect it qualifies as an "armchair analysis", by Block's standards.

In addition to being deflationary, there is a further way in which the present proposal opens up space for an eliminativist-style development about phenomenal consciousness. Our understanding of the peculiar kind of first-person awareness we have of certain of our states is not terribly advanced, and it is certainly a live possibility that we will eventually distinguish rather different types of such awareness. For example, it is not obvious that one and the same process is in play when I am (i) aware that I experience a ripe lemon, (ii) aware that I want to marry my girlfriend, and (iii) aware that one of my tacit beliefs is that Winston Churchill had a kneecap. If consciousness is nothing other than a peculiar kind of first-personal awareness, and if there are several different kinds of such awareness, we may well end up judging that what we once thought about in terms of "consciousness" was many things.

5. Relation to monitor theories

There is a family of views according to which the phenomenal consciousness of a state is, at least in part, a matter of that state being "monitored" in some special way. On some such views, the consciousness of a state is a matter of the state being, in part, a representation of itself (e.g. Kriegel 2003). On other such views, the consciousness of a state is a matter of it being represented by another, "higher-order" state, which may be either perception-like (e.g. Lycan 1996) or belief-like (e.g. Rosenthal 2002).

[1] In Sundström (2007) I invoke The Separation Thesis to argue that the problem is often mischaracterised; I also suggest that eliminativism about consciousness should be taken seriously (section 5). The present paper traces a somewhat different route from The Separation Thesis to eliminativism about consciousness.

What is the relation between the present proposal and these views? Two remarks are in order.

First, many or all monitor theorists take monitoring to constitute only a *part* of phenomenal consciousness. For example, Kriegel distinguishes "two aspects" of what a conscious experience of blue is like: a "qualitative" aspect and a "*for-me* aspect" (2005, 23). Monitoring is supposed to account only for the second, for-me aspect. The qualitative aspect gets a separate treatment. The present proposal is different. It suggests that consciousness doesn't have a qualitative aspect at all. Whatever qualities are displayed in conscious experience are not constitutive properties of what it is like.

Second, even with respect to the "for-me aspect" of consciousness, I believe the present proposal differs from at least some monitor theories. At any rate, I think there's a difference in what the views purport to explain. The present proposal is emphatically *deflationary* about consciousness. Monitor theorists don't always or even standardly advertise their view as such, and it may not be how many or all of them regard the matter. It may be noted, though, that critics often emphasise the deflationary character of monitor theories, and complain that the theories reduce consciousness to a triviality (e.g. Block 2002a, 214; and Chalmers 1996, section 4.5). If the present proposal is right, these critics may well be right that monitor theories are seriously deflationary, but wrong to suppose that this tells against them.

Literature

Block, Ned 2002a "Concepts of consciousness", in: David Chalmers (ed.) *Philosophy of Mind: Classical and Contemporary Readings*, Oxford UP, 206-18.

Block, Ned 2002b "The harder problem of consciousness", *The Journal of Philosophy* 99, 391-425.

Chalmers, David 1996 *The Conscious Mind: In Search of a Fundamental Theory*, Oxford: Oxford UP.

Churchland, Patricia 1986 *Neurophilosophy: Toward a Unified Science of the Mind/Brain*, Cambridge, Mass.: MIT Press.

Flanagan, Owen 1992 *Consciousness Reconsidered*, Cambridge, Mass.: MIT Press.

Kriegel, Uriah 2003 "Consciousness as intransitive self-consciousness: Two views and an argument", *Canadian Journal of Philosophy* 33, 103-32.

Kriegel, Uriah 2005 "Naturalizing subjective character", *Philosophy and Phenomenological Research* 71, 23-56.

Lycan, William 1996 *Consciousness and Experience*, Cambridge, Mass: MIT Press.

Rosenthal, David 2002 "Explaining Consciousness", in: David Chalmers (ed.) *Philosophy of Mind: Classical and Contemporary Readings*, Oxford: Oxford UP, 406-21.

Stich, Stephen 1996 *Deconstructing the Mind*, Oxford: Oxford UP.

Sundström, Pär 2007 "Colour and consciousness: Untying the metaphysical knot", *Philosophical Studies* 136, 123-65.

Thau, Michael 2002 *Consciousness and Cognition*, Oxford: Oxford UP.

Impliziert der intentionale Reduktionismus einen psychologischen Eliminativismus? Fodor und das Problem psychologischer Erklärungen

Thomas Szanto, Wien & Graz, Österreich

1.

Eines der wenigen Kriterien für die Akzeptierbarkeit der Psychologie als legitimer Einzelwissenschaft unter anderen Naturwissenschaften, auf das sich die meisten nicht-reduktiven Naturalisten einigen können, ist ihre Kompatibilität mit einem Naturalismus in Bezug auf jene mentalen Zustände, die wesentlich durch Intentionalität charakterisiert sind. J. Fodor hat diesem Minimalkriterium seine kanonische Form verliehen: Was wir Fodor zufolge brauchen, ist eine Theorie, die in *nicht-semantischen* und *nicht-intentionalen* Kategorien hinreichende Bedingungen für das Vorliegen einer Repräsentationsbeziehung zwischen zwei Teilstücken („*bits*") der natürlichen Welt, nämlich einem mentalen und einem physikalischen Zustand, angibt (vgl. Fodor 1987, 98).

Fodors kognitivistisches Gesamtprojekt fußt auf zwei – scheinbar gegenläufige – metatheoretischen Prämissen: Fodor zufolge werden wir keine ernstzunehmenden psychologischen Erklärungen erzielen, wenn wir 1.) keine verallgemeinerbaren Gesetzmäßigkeiten bezüglich der intentionalen Verursachung von Verhalten beschreiben können, die mit den grundsätzlichen Erklärungsansprüchen der Volkspsychologie vereinbar sind; 2.) müssen diese alltagspsychologisch relevanten Gesetzmäßigkeiten wiederum zumindest prinzipiell in den kausalen Erklärungsrahmen nach dem Modell empirischer Wissenschaften übersetzbar sein. Wenn wir also die Realität intentionaler Zustände nicht bezweifeln und sie auch aus einer seriösen Psychologie nicht eliminieren wollen, müssen wir die intentionalen Gesetzmäßigkeiten naturalisieren (vgl. Fodor 1994, 3). Nur wenn das gelingt, haben wir Aussicht auf eine wissenschaftlich akzeptable Psychologie. Psychologie als genuine Wissenschaft wäre demnach eine Erklärung der gesetzmäßigen Struktur und der formalen Prozesse intentionaler Systeme; und genuin psychologisch wäre die Erklärung, sofern sie sich nicht auf Verallgemeinerungen anderer empirischen Einzelwissenschaften (wie etwa die Neurobiologie) reduzieren lässt. (vgl. Fodor 1990, 5)

Sollten wir eine solche Theorie nicht zustande bringen, müssten wir unsere Alltagspsychologie in die historische Senkgrube intellektueller Fehltritte verbannen – was aber nach Fodor der „greatest intellectual catastrophe in the history of our species" (Fodor 1987, xii) gleichkäme. S. Stich hat diese besorgniserregende Ansicht Fodors ironisch als „die Katastrophentheorie" diagnostiziert (Stich 1996, 92ff.). Fodors Katastrophentheorie wird durch die fragwürdige Annahme gespeist, dass ein Versagen auf dem Feld der Naturalisierung des Intentionalen *eo ipso* die „furchtbare Konsequenz" hätte, dass wir unseren festen Glauben an die *Realität des Intentionalen* aufgeben und das Feld den *intentionalen Irrealisten* räumen müssten. Fragwürdig ist diese Annahme, da gar nicht klar ist, inwiefern und ob überhaupt aus dem Versagen auf dem Feld der Naturalisierung der intentionale Irrealismus folgen soll.

Fodor nimmt an, dass die Widerlegung des intentionalen Irrealismus nur durch eine endgültige, und das heißt vollständige Naturalisierung des Intentionalen möglich ist. Dabei setzt er jedoch die methodologische Grundprämisse der zu widerlegenden Position voraus, wonach nämlich nur das *real* ist, was prinzipiell *naturalisierbar* ist (vgl. Fodor 1994, 5).

Bevor man jedoch irgendwelche Zugeständnisse an den intentionalen Irrealismus macht, müsste man zunächst zeigen, dass dieser eine wahre ontologische These oder zumindest eine brauchbare Doktrin bezüglich des spezifischen Gegenstandsbereichs jener Wissenschaft ist, die sich mit intentionalen Zuständen beschäftigt. Die Katastrophentheorie impliziert, dass die Gültigkeit intentional-psychologischer Erklärungen notwendig von der Wahrheit des intentionalen Irrealismus abhängt. Die Wahrheit des intentionalen Irrealismus hängt wiederum von der Wahrheit ihrer ontologischen Behauptungen und deren wissenschaftstheoretischen Konsequenzen ab.

2.

In seiner ontologisch starken Version behauptet der intentionale Irrealismus, dass solche Zustände, mit denen sich die intentionale Psychologie beschäftigt, in Wirklichkeit – nämlich in jener, welche die seriösen Wissenschaften modellieren – gar nicht existieren. Der starke intentionale Irrealismus basiert auf der Ontologie des eliminativen Materialismus. Entitäten, mit denen sich Psychologen beschäftigen wenn sie intentionales Verhalten oder intentionale Zustände analysieren, haben dieser Ansicht zufolge den gleichen Realitätsgehalt wie jene Zustände, die etwa mittelalterliche Exorzisten einer Hexe zugeschrieben haben. Intentionale Psychologie ist demnach durch und durch Volkspsychologie und hat als solche ebenso wenig Platz im Kanon ernstzunehmender Wissenschaften wie religiöser oder sonstiger Aberglaube.

Demgegenüber ist die ontologisch schwächere Version des intentionalen Irrealismus rigider, was den psychologischen Erklärungsrahmen betrifft. Die schwächere Version trifft zwar keine Behauptung bezüglich der Existenz der fraglicher Entitäten und stellt lediglich fest, dass die Eigenschaften, die intentionalen Zuständen zugeschrieben werden, keinerlei Wirksamkeit innerhalb der kausalen Gesetzmäßigkeiten ausüben, welche das Untersuchungsfeld seriöser Wissenschaften konstituieren. Nun wäre dies noch insofern unproblematisch, solange man gelten lässt, dass die intentionale Psychologie es gar nicht mit Kausalerklärungen zu tun hat. Die fodorsche Version macht jedoch eine weitere Annahme, welche den Zusammenhang betrifft zwischen der kausalen Wirksamkeit, der Realität intentionaler Zustände und der Möglichkeit, psychologisch informative Aussagen über diese zu machen. Die Zusatzprämisse besagt, dass intentionale Zustände nur insofern als reale Entitäten interpretiert werden können, als sie kausal wirksam sind. Sofern intentionale Eigenschaften keine kausale Wirksamkeit ausüben, können auch Erklärungen, die auf solche Eigenschaften rekurrieren, keine Rolle bei der Erklärung der Ursachen von intentionalem Verhalten spielen und müssen folglich aus dem Untersuchungsgebiet einer seriösen Psychologie ausgeschlossen werden.

Diese ontologisch schwächere Version des intentionalen Irrealismus führt also zu einem bestimmten Typ von Eliminativismus, den man mit Blick auf den psychologischen Erklärungs*wert* der betreffenden Entitäten *psychologischen Eliminativismus* nennen könnte. Der psychologische Eliminativist lässt offen, ob intentionale Zustände nun existieren oder nicht; sofern ihre charakteristischen Eigenschaften aber keine empirisch verifizierbare, kausale Funktion ausüben, müssen wir aus methodologischen Gründen auf diese Entitäten verzichten –, zumindest sofern wir uns mit ihnen als wissenschaftlich akzeptable Kandidaten für eine seriöse Psychologie beschäftigen. Der psychologische Eliminativismus ist eine heuristische, oder besser, wissenschaftspragmatische These bezüglich der Brauchbarkeit intentional-psychologischer Erklärungen innerhalb der Logik naturwissenschaftlicher Aussagensysteme und impliziert – anders als der Eliminative Materialismus oder der Instrumentalismus – nicht notwendig eine ontologische These bezüglich der Realität intentionaler Zustände und Eigenschaften.

Obwohl Fodor qua intentionaler Realist *Anti-Reduktionist hinsichtlich der Ontologie intentionaler Zustände* ist, ist seine Theorie *hinsichtlich des Typs intentional-psychologischer Erklärungen* wesentlich reduktiv. Insofern ist sie dem intentionalen Instrumentalismus eines Dennett entgegengesetzt: Ein Instrumentalist à la Dennett schließt zwar intentionale Eigenschaften aus der Menge real-existierender Phänomene aus, garantiert einer intentionalen Einstellung („*intentional stance*“) zu kognitiven Systemen jedoch eine explanatorische Legitimität innerhalb psychologischer Aussagenssysteme. Der Kognitivist à la Fodor schlägt den umgekehrten Weg ein und räumt den Eigenschaften, auf die intentionalistische Erklärungen referieren, eine ontologische Realität ein, beraubt jedoch den Typ von Erklärung jeglicher wissenschaftlicher Legitimität.

Das Dilemma, das sich hier abzeichnet, ist folgendes: Während der Eliminativismus hinsichtlich einer psychologisch informativen Wissenschaft in der ontologisch stärkeren Lesart des intentionalen Irrealismus trivialerweise wahr ist, macht er in seiner schwächeren Lesart als psychologisches Programm gar keinen Sinn. Wenn man annimmt, dass intentionale Entitäten *nichts als* physische Entitäten sind, so sind sie trivialerweise nicht die relevanten Untersuchungsgegenstände für die (Alltags-) Psychologie, sondern für die Physik. Wenn man intentionale Eigenschaften umgekehrt auf kein ontologisch eindeutig abgezirkeltes Gebiet festlegen bzw. auf dieses reduzieren will, so wird die vermeintliche Notwendigkeit einer Naturalisierung dieser Eigenschaften hinfällig. Der intentionale Eliminativismus macht für eine Psychologie keinen Sinn, solange man nicht über ein Kriterium für eine erfolgreiche Naturalisierung intentionaler Eigenschaften verfügt, über ein Kriterium nämlich, das unabhängig von der kausalen Doktrin des intentionalen Irrealismus gültig ist –, oder aber er ist trivialerweise wahr, sofern die Entitäten einem im Vorhinein festlegten Untersuchungsgebiet zugeschrieben werden und quasi je nach Bedarf aus bestimmten Forschungsprogrammen ausgeschlossen werden können.

3.

Der Kognitivismus kann grob als der Versuch charakterisiert werden, unter alleinigem Rekurs auf die operationale Ebene mentaler Prozesse psychologisch relevante Aussagen über das Verhältnis zwischen der Realisierung mentaler Zustände und ihrer *kausalen Rolle* bei der Erklärung von Verhalten zu treffen.

Der Kognitivismus ist (zunächst) neutral hinsichtlich der Frage, ob die zu untersuchenden kognitiven Funktionen auf physikalisch gesehen niedrigere Abstraktionsniveaus reduzibel sind oder nicht. Hinsichtlich des Explikationsradius kognitivistischer Beschreibungsmodelle ist es sinnvoll, drei ontologisch verschieden gewichtete Ebenen zu unterscheiden: die Ebene der *Implementierung* kognitiver Prozesse, die Ebene ihrer *Realisierung* bzw. *Instantiierung* und schließlich die *funktionale* Ebene des Zusammenhanges zwischen den ersten beiden. Aussagen über die Ebene der Implementierung betreffen die material-ontologische Verfasstheit mentaler Systeme und Prozesse. Fragen nach der tatsächlichen Implementierung kognitiver Prozesse sind also empirische Fragen nach den Konstitutionsbedingungen menschlicher und oder künstlicher Kognition. Beschreibungen hinsichtlich der Realisierung bzw. Instantiierung auf der operationalen Ebene kognitiver Prozesse sind an und für sich genommen ontologisch ebenso neutral. Was die funktionalen Ebene betrifft, ist die Sache nicht mehr ganz so eindeutig, handelt es sich doch dabei um Behauptungen hinsichtlich der Übersetzbarkeit bzw. der kausalen Dependenzen der beiden Ebenen. Ziel des Kognitivismus ist jedenfalls, bei der Erklärung der Interaktion zwischen Implementierungs- und Realisierungsebene – also auf der funktionalen Beschreibungsebene – auf ontologische Prädikate gänzlich zu verzichten.[1] Die Frage ist freilich, inwieweit und ob dies gelingt.

Wenn für den Kognitivisten die Frage nach den Realisierungsformen und Realisierungsbedingungen intentionaler Zustände für die Naturaliserbarkeitsbehauptung irrelevant ist, so muss doch die Form der Erklärung ihrer Funktion naturalistischen Kriterien entsprechen.

Zu unterscheiden ist hierbei zwischen Typen psychologischer Erklärungen und Typen psychologisch relevanter Explananda einerseits und entsprechend zwischen der Reduktion bestimmter psychologisch relevanter Entitäten und der Reduktion bestimmter Erklärungsmodelle andererseits. In seiner eingehenden Analyse verschiedener Typen reduktiver Erklärungsmodelle innerhalb des Kognitivismus argumentiert auch J. Haugeland zu Recht, dass durchaus nicht jedes Erklärungsmodell, bei dem die Möglichkeit einer *wissenschaftlich respektablen* Erklärung abhängig gemacht wird von der Zurückführung des Explanadums auf eine niederstufigere bzw. fundamentalere, gesetzmäßig spezifizierte Ebene, die selbst unerklärt bleibt, zwangsläufig in einer vollständigen (sprich: physikalistischen) Reduktion münden muss („*complete reduction [all the way to physics]*“; Haugeland 1978, 251). Der Kognitivist hat den – metaphysischen – Traum einer einheitswissenschaftlichen Reduktion psychologischer Konzepte ausgeträumt; kognitivistische Reduktion ist eine hierarchische Korrelation verschiedener funktional spezifizierter Systeme, bei der nicht die Physik die Hierarchie der Systeme festlegt, sondern das jeweilige Erklärungsmodell (vgl. Fodor 1974).

[1] Vgl. auch Fodors (1965) frühen Versuch, verschiedene Phasen der Theoriebildung mit verschiedenen Ebenen hinsichtlich der Explanandums zu korrelieren.

Welche Systeme auf *welche* reduziert und wie diese spezifiziert werden, d.h. welche Systeme welche funktionale Rolle bei der psychologischen Erklärung spielen, ist relativ zur jeweiligen Heuristik der Erklärung. Entsprechend charakterisiert Haugeland die funktionalistische Erklärungsmodelle von Kognition als Typen systematischer Reduktion („*systematic reduction*") (Haugeland 1978, 249ff.). Systematisch-reduktive Erklärungen unterscheiden sich von nomologisch-reduktiven Erklärungen, bei denen Brückengesetze zwischen den verschiedenen Ebenen vorausgesetzt und nach quantitativen Gleichungen beschrieben werden. Systematische Reduktionen setzen nach Haugeland hingegen keine solche psychophysischen (oder allgemein gesprochen: ‚meta-funktionalen') Brückengesetze voraus und suchen auch nicht nach solchen –, wie Haugeland mit einem Seitenhieb auf Fodors Naturalisierungsstrategie vermerkt. Der wesentliche Punkt, auf den Haugeland mit seiner Konzeption systematischer Erklärungen hinauswill, ist, dass dieser spezielle Typ kognitivistischer Reduktion entgegen der Annahme vieler Kritiker weder die Explananda (die spezifisch kognitiven Zustände) noch das Explanans (die funktionale Beschreibung kognitiver Systeme) zugunsten anderer Entitäten und Erklärungen überflüssig macht bzw. eliminiert.

Trifft Haugelands Charakterisierung des Kognitivismus als einen Typ systematischer Reduktion zu, so wären – jenseits der Unterscheidung der Ontologie der Explananda und der verschiedenen Erklärungstypen – drei Thesen zu unterscheiden: 1. Die *Kompatibilitätsthese* hinsichtlich der Gültigkeit verschiedener Erklärungsmodelle (wie alltagspsychologische, physikalische etc.), 2. die *Übersetzungs*- oder *Übersetzbarkeitsthese* hinsichtlich des Vokabulars, mit dem die Explananda beschrieben werden und 3. schließlich die *Zuordnungsthese* hinsichtlich der Brückengesetze, welche die Gültigkeit der Kompatibilität und die Adäquatheit der Übersetzung regeln. Sofern der Kognitivismus die Kompatibilitätsthese bezüglich alltagspsychologischer und physikalistischer Erklärungen vertritt, sofern er nur eine Koextensionalität bezüglich der zu übersetzenden Vokabulare und keine Synonymie einfordert und solange er sich hinsichtlich der Zuordnungsthese nicht festlegt, wären kognitivistische Erklärungsmodelle nicht notwendig reduktiv (bzw. eliminativ) hinsichtlich ihrer Explananda.

Doch – *pace* Haugeland und Fodor – droht die Demarkationslinie zwischen ontologisch neutralen Beschreibungen der psychologischen Funktion mentaler Eigenschaften und dem Kognitivismus als ontologischer These bezüglich der psychophysischen Typen-Identität freilich überall dort zu verschwimmen, wo es um eine exklusive bzw. beste Beschreibung („*unique best* description") der betreffenden psychologischen Systeme geht (Block/Fodor 1972, 84f.). Das ist der Fall, wenn die Adäquatheit der psychologischen Beschreibung an die Bedingung geknüpft wird, dass die psychologischen Prädikate, die einem System attribuiert werden – zumindest prinzipiell – eins zu eins in das Vokabular übersetzbar sein müssen, mit denen nicht-psychologische/physikalische Systeme beschrieben werden. Dabei spielt es keine Rolle, ob die psychologischen Zustände als Einzelvorkommnisse (*Tokens*) selbst ontologisch spezifiziert, d.h. bestimmten physikalischen Prozessen zugeordnet werden oder nicht bzw., ob der Kognitivist ontologisch neutral ist bezüglich der Implementierung psychologischer Einzelvorkommnisse oder nicht. Denn, selbst wenn von Kognitivisten konzediert wird, dass es sich bei der Übersetzung des psychologischen in das physikalische Beschreibungsmodell nicht um eine ontologische Reduktion handelt, ist nicht zu sehen, wie diese Korrelation – unter Vorraussetzung einer exklusiv besten Beschreibung, welche den physikalisch festgelegten, kausalen Gesetzmäßigkeiten entsprechen muss –, nicht *ipso facto* zu einer solchen Reduktion führen soll. Das heißt, selbst wenn der Kognitivist annimmt, dass die verschiedenen Phasen der Theoriebildung verschiedenen irreduziblen Beschreibungsebenen entsprechen, ist nicht einsichtig – wie unter Annahme des Ideals einer physikalisch adäquaten Spezifikation des Verhältnisses zwischen mentalen und nicht-mentalen Zuständen – eine reduktive Erklärung vermieden werden soll.[2]

Das Ideal einer besten Beschreibung vereitelt mithin die Neutralität des Kognitivisten bezüglich der Zuordnungsthese. Die vermeintlich gleichberechtigte Kompatibilität von Erklärungsmodellen weicht damit einer Hierarchie von Beschreibungsebenen, bei denen die Übersetzungsregeln von Zuordnungsregeln ersetzt und diese wiederum von nicht-psychologischen Theorien vorgeschrieben werden. Wenn man, wie Fodor, einfordert, dass die Ergebnisse der funktionalen Analyse psychologischer Zustände mit den Ergebnissen des Neurologen vereinbar sein müssen, andernfalls „unakzeptabel" seien, und eine solche empirische Falsifizierbarkeit gar zu einem „Prinzip" psychologischer Erklärung erhebt, dann ist damit mehr als nur eine „flagrant empirische Annahme" in das Erklärungsmodell „eingebaut". „Wenn Übereinstimmung mit der neurologischen Wirklichkeit eine Bedingung für die Adäquatheit" von kognitivistischen Erklärungsmodellen ist (Fodor 1965, 430f.), so haben wir es hier weniger mit einer genuin psychologischen Erklärung, als vielmehr mit einer genuin reduktiven Erklärungsstrategie zu tun.

Was folgt aus all dem? Es folgt jedenfalls keine Katastrophe –, weder theoretisch, noch praktisch: Wir müssen weder unsere kognitivistischen noch unsere intentional-psychologischen Theorien in die wissenschaftshistorische Finsternis der Pseudowissenschaftlichkeit oder gar des Aberglaubens verbannen. Wir müssen aber auch nicht unsere alltagspsychologischen Praktiken bei der Interpretation unseres normalen Selbstverständnisses bzw. der typischen sozialen Interaktionsmuster rundheraus eliminieren. Worauf wir – soweit ich sehe – allerdings verzichten müssen, ist die Hoffnung auf eine genuin psychologische Wissenschaft unserer intentionalen Zustände, die nicht-reduktiv ist und gleichwohl auf die Absolution des Naturalismus zählen kann.

[2] Siehe dazu auch die Kritik N. Blocks (Block 1978, insbes. 159-163).

Literatur

Block, Ned 1978 „Schwierigkeiten mit dem Funktionalismus“, in: Dieter Münch (Hg.), *Kognitionswissenschaft. Grundlagen, Probleme, Perspektiven*. Frankfurt a.M.: Suhrkamp 1992, 159-224.

Block, Ned/Fodor, Jerry 1972 “What Psychological States Are Not”, in: Jerry A. Fodor, *Representations. Philosophical Essays on the Foundation of Cognitive Science*, Cambridge, MA: MIT Press 1981, 79-99.

Fodor, Jerry A. 1965 „Erklärungen in der Psychologie.“ In: Ansgar Beckermann (Hg.), *Analytische Handlungstheorie. Bd. 2: Handlungserklärungen*, Frankfurt a.M.: Suhrkamp 1985, 412-434.

Fodor, Jerry A. 1974 „Special Sciences – or The Disunity of Science as a Working Hypothesis“, in: *Synthese* 28, 97-115.

Fodor, Jerry A. 1987 *Psychosemantics. The Problem of Meaning in the Philosophy of Mind*, Cambridge, MA: MIT Press.

Fodor, Jerry A. 1990 *A Theory of Content and Other Essays*, Cambridge, MA: MIT Press.

Fodor, Jerry A. 1994 *The Elm and the Expert. Mentalese and Its Semantics*, Cambridge, MA: MIT Press.

Haugeland, John 1978 „The Nature and Plausibility of Cognitivism“, John Haugeland (ed.), *Mind Design. Philosophy, Psychology and Artificial Intelligence*, Cambridge, MA: MIT Press 1981, 243-281.

Stich, Stephen P. 1996 „Puritanischer Naturalismus.“ In: Geert Keil und Herbert Schnädelbach (Hg.), *Naturalismus. Philosophische Beiträge*, Frankfurt a.M.: Suhrkamp 2000, 92-112.

Structure of the Paradoxes, Structure of the Theories: A Logical Comparison of Set Theory and Semantics

Giulia Terzian, Bristol, England, UK

1. Introduction

F. Ramsey famously argued that the "logical" and the "semantical" paradoxes should be studied separately. Those of the first kind "involve only logical or mathematical terms such as class and number, and show that there must be something wrong with our logic or mathematics". On the other side are those contradictions which "cannot be stated in logical terms alone, for they all contain some reference to thought, language, or symbolism, which are not formal but empirical terms." (1925, p.353)

With the development of modern set theory and semantics over the 20th century, many have rejected this classification, arguing that there is a unique shared structure underlying most of the known paradoxes, and that therefore a joint solution is also to be expected. Arguments to this effect can be found for instance in Herzberger 1970, Feferman 1984, Priest 1994; in the next section, we will devote some attention to the second of these papers.

2. A simple sophisticated story

In 1984, S. Feferman carries out a parallel reconstruction of Russell's Paradox and the Liar Paradox, to show that both derive from the combination of three features in the background theory:

1. The language has enough syntactical resources to allow self-reference;
2. Classical logic is assumed;
3. The following unrestricted schemes are respectively assumed as basic principles (for φ a formula of the language):

   (CA) $\exists x \forall y (y \in x \leftrightarrow \varphi[y])$

   (TA) $T(\ulcorner \varphi \urcorner) \leftrightarrow \varphi$.

Restriction of each of these accordingly corresponds to a possible solution strategy for the paradoxes. Russell and Tarski pursued the first option, developing typed formalisms which "were early recognized to be excessively restrictive" (1984, p.75). To test the second strategy, Feferman constructs a common formalism for set theory and semantics, making use of three-valued logical schemes (the primary references in semantics are of course Martin and Woodruff 1975, Kripke 1975). However the resulting theory is argued to be again too restrictive, insofar as "nothing like sustained ordinary reasoning can be carried out in [three-valued] logic." (p.95) But it should also be noted that restriction of logic does not constitute a standard option for set theorists, and this on its own diminishes the prospects of obtaining a parallel solution of the paradoxes.

The rest of the paper is then devoted to the strategy of restricting basic principles: since it has been successful in ZF theory, where Russell's paradox is blocked, the aim is to prove a similar result for semantics. Although this conjecture is not supported by a direct argument in the paper, this can be made explicit as follows:

> (1) Set-theoretic and semantic paradoxes bear a structural similarity.
>
> (2) ZF set theory is *both* a faithful account of set theorists' notion of set membership, *and* it successfully deals with the set-theoretic paradoxes.
>
> Therefore:
>
> (C) Any adequate solution to the semantic paradoxes is to be expected to bear a structural similarity to ZF set theory.

A solution strategy for the paradoxes is a particular application of the theoretical framework of set theory and semantics; thus acceptance of (C) is presumably dependent on the soundness of a more general argument which should show that a structural similarity holds between set theory and semantics themselves. The upshot of the resulting account (hereafter *analogy account*) would be that set theory can inform semantics in the choice of the normative principles which would underlie a successful (axiomatic) theory of truth.

Is the analogy account sound? If the paradoxes do have a joint solution *and* thereby constitute evidence that a deeper analogy holds, then a structural analogy ought surely to hold at the level of the foundations of the theories. Hence the first condition on accepting the analogy account is that there exist semantic counterparts of the normative principles underlying the choice of the ZF axioms.

3 Why the ZF axioms?

The literature is unanimous in identifying two conceptions which enshrine the pre-theoretic intuitions concerning the notion of set, and which moreover acted both as *motivations* and *normative constraints* in the development of modern set theory.

These are limitation of size and the iterative conception:

> (LIM) The axioms ought to entail that the set-forming operation applies to a collection of objects if and only if the collection is small enough; or more formally, if and only if its objects are not in one-one correspondence with all the objects of the universe of sets.
>
> (IT) The axioms ought to entail that a collection of objects is a set if and only if it is produced in a process of the following sort: at stage 0 we have the empty set Ø; at stage 1, Ø and its singleton set {Ø}; and so on, into the transfinite; crucially, every set appears at some stage of this cumulative hierarchy. In the standard formalization:
> $V_0 = \emptyset$; $V_{\alpha+1} = \mathcal{P}(V_\alpha)$; $V_\gamma = \bigcup_{\alpha<\gamma} V_\alpha$ for limit γ.

Remarks:

> (a) Historically, LIM influenced set theorists long before IT, which was first explicitly mentioned only in a 1947 paper by Gödel (cf. Potter 2004).

> (b) A logical analysis of the two conceptions favours IT, insofar as it involves only intuitive notions, while LIM presupposes an understanding of more sophisticated concepts of set theory (cf. Boolos 1989).
>
> (c) Boolos 1989 shows that IT on its own does not entail the axioms of replacement and extensionality; these follow from LIM, but other axioms of ZF do not. The upshot is that it is possible to bridge the gap between the historical and the logical reconstructions of ZF set theory by understanding LIM and IT as jointly necessary, individually not sufficient premises to a full explication of the notion of set: in Boolos' words, “there are at least two thoughts ‘behind’ set theory” (1989, p.103).

4 Too simple?

In order to assess the soundness of the analogy account, the key question to be answered is the following: Are there any semantic principles that could be identified as counterparts to IT and LIM?

The first part of this section addresses this question; the second part raises two more issues of methodology.

A semantic limitation of size principle LIM_T should e.g. guarantee the assertability of some sentences of the form $\forall x\ (P(x) \rightarrow T(x))$ for some predicate P. To this end, LIM_T should presumably place a cardinality constraint on the extension of P, so as to ensure that problematic (Liar) sentences do not end up in the extension of T.

Suppose the extension of P is $\mathcal{L}_T$, i.e., the set of all sentences including those which contain the truth predicate. Then among these will be many (Liar-like) sentences we would not want in the extension of T: so it might seem reasonable to invoke a principle LIM_T requiring the cardinality of the extension of T to be smaller than that of $\mathcal{L}_T$.

Now suppose the extension of P is $\mathcal{L}_{arith}$ (the language of arithmetic): applying T to purely arithmetical sentences does not lead to any inconsistency, so this choice should be allowed by LIM_T. However $\mathcal{L}_T$ and $\mathcal{L}_{arith}$ are both countably infinite: hence LIM_T gives contradictory verdicts for languages with the same cardinality.

Finally suppose the extension of P is simply $\ulcorner\lambda\urcorner$. In principle LIM_T should obviously apply in this case: insofar as $\ulcorner\lambda\urcorner$ should definitely *not* be in the extension of T, the minimal language which it forms is already ‘too big’. This is an undesirable result which makes it starkly clear that the *quantitative* constraint in LIM_T is inadequate for the semantic context, where instead the key question is about *which* sentences are put inside the extension of T.

It is fairly natural to understand IT as embodying “a fundamental relation of [...] *dependence* between collections." (Potter 2004, p.36) In semantics, too, one can talk about a fundamental relation of *dependence* between some sentence φ and a set of sentences of the language. Semantic dependence subtends the notion of semantic *groundedness*[1]: a sentence containing T is grounded if its truth value ultimately depends on non-semantic states of affairs, so that working back along the dependence relation leads to a sentence which does not contain T. Liar sentences are ungrounded, because their dependence path is not linear but circular; in standard truth-gap accounts, this is equivalent to saying that they lack a truth value in the least fixed point of the (Kripkean) jump operator[2].

Dependence seems to provide a more promising case for the analogy account. To follow up this conjecture, we look at the features of a typical construction in both contexts. Take Ø as the starting point. At each level of the cumulative hierarchy, all individuals and sets appearing at all previous levels are collected into a new set: the iterated power-set operation produces a strictly increasing progression of sets, from which no element of the universe is left out.

On the other hand, the iterated jump operation produces a ‘semi’-hierarchy of extensions (anti-extensions) of the truth predicate, but in this case not all sentences of the language are guaranteed a place. Specifically, only the grounded sentences will make a regular appearance at each level (and will also be consistently part of the *same* semi-hierarchy); so the iterated collecting operation is here constrained to filter out the ungrounded sentences.

Thus on closer inspection dependence actually appears to be a fairly weak link between set theory and semantics, and moreover reveals further differences between them. Set-theoretic dependence is central to the construction of the hierarchy, in the sense that a set *exists* in virtue of being made up of lower-level elements: it is so to speak a by-product of IT. But in semantics there is no question about the ‘existence’ of a sentence: what matters instead is whether the correct ones are put inside the extension of T. The key relation here is groundedness, which determines whether a sentence can be evaluated for truth; moreover, this can only be established once we reach the least fixed point – while the existence of a set is ‘determined’ at every level at which it appears.

Finally, groundedness not only presupposes dependence, but on some accounts (e.g. Yablo 1982) also an understanding of partial predicates, non-classical logics, etc.: so there is the additional worry that the relevant set-theoretic and semantic relation are mismatching in another respect, namely that the former but not the latter underlies a *natural* conception (cf. Section 3).

In giving a formal theory of truth, the central aim is to explicate this fundamental semantic notion so as to account for its non-problematic use in everyday speech. For any such theory, a ‘sample basis’ of sentences containing the truth predicate should thus naturally be expected to be as broad as possible. Paradoxical sentences, which constitute an extremely restricted sample, could then be used to *test* the formalism, as a measure of its efficacy.

One could also choose a different approach: start from some pre-existing formalism and subject it to successive revisions, constrained primarily by the rule: “avoid contradictions”.

These diametrically opposite strategies might be called respectively ‘constructive’ and ‘regressive’; the choice of these terms is intended to parallel the distinction made by Potter 2004 between “intuitive” and “regressive” methodologies in set theory. As in the case of set theory

[1] First formally discussed in Kripke 1975; a more thorough analysis is found in Yablo 1982 and Leitgeb 2005.

[2] Let (A+;A-) be a partial interpretation of *T*; then the jump is a monotone inductive operator defined by *j* (A+;A-) = (*j* + (A+;A-); *j* - (A+;A-)), where *j* +(A+;A-) = {φ: (A+;A-) ⊨ φ} and *j* - (A+;A-) = {φ: (A+;A-) ⊨ ¬φ}.

(cf. Potter p.34), it seems desirable to adopt a constructive rather than a regressive strategy, to ensure the higher reliability of the resulting theory.

However most accounts in the literature appear to start from an analysis of the Liar paradox, and *then* proceed to develop a suitable formalism which accommodates this phenomenon as well as our intuitions about truth. This is also clearly the case in Feferman 1984; but by adopting such a plainly regressive strategy, the resulting theory is also much more exposed to the danger of misrepresenting our positive intuitions about truth.

One of the reasons for which the analogy account is attractive is that it would allow for semantics to be informed by set theory; but a genuine structural analogy should also entail that semantics can inform set theory, so that set-theoretic norms can be imported over to semantics, *and* vice versa.

A closer look at concrete theories shows that the analogy is invariably left lopsided. For instance, the axiom system proposed in Feferman 1984 is not only shown to be an unsatisfactory account of truth, but it is also clear that it is not inter-translatable with ZF, which leaves the analogy account to stand on even shakier grounds.

5 Conclusion

Feferman and others[3] have shown that the set-theoretic and semantic paradoxes can be reconstructed to follow a common principle. On the other hand, the discussion in Section 4 shows that set theory and semantics are based on structurally mismatching guiding principles, and moreover that their respective end-products – the axiom systems – do not correspond to each other, as a genuine structural similarity should guarantee.

Although the analogy account remains very attractive, this paper should hopefully have shown that it must be supported by a stronger argument if it is to resist these problems.

Literature

Boolos, George 1998 *Logic, logic, and logic*, R. Jeffrey (ed.), London: Harvard University Press.

Boolos, George 1971 “The Iterative Concept of Set", in Boolos 1998, 13-29.

Boolos, George 1989 “Iteration Again”, in Boolos 1998, 88-104.

Feferman, Solomon 1984 “Toward Useful Type-Free Theories. I”, *The Journal of Symbolic Logic* 49, 75-111.

Herzberger, Hans G. 1970 “Paradoxes of Grounding in Semantics”, *The Journal of Philosophy* 67, 145-167.

Kripke, Saul 1975 “Outline of a Theory of Truth”, *The Journal of Philosophy* 72, 690-716.

Leitgeb, Hannes 2005 “What Truth Depends On”, *Journal of Philosophical Logic* 34, 155-192.

Maddy, Penelope 1988 “Believing The Axioms. I”, *The Journal of Symbolic Logic* 53, 481-511.

Martin, Robert L. and Woodruff, Peter 1975 “On Representing `True-in-L' in L", *Philosophia* 5, 213-217.

Moore, Gregory H. 1982 *Zermelo's axiom of choice: its origins, development, and influence*, New York: Springer-Verlag.

Potter, Michael 2004 Set *theory and its philosophy: a critical introduction*, Oxford: Oxford University Press.

Priest, Graham 1994 “The Structure of the Paradoxes of Self-Reference”, *Mind* 103, 25-34.

Ramsey, Frank P. 1925 “The Foundations of Mathematics”, in: Braithwaite (ed.) 1931 *The Foundations of Mathematics and Other Logical Essays, by Frank Plumpton Ramsey*, London: Routledge Kegan Paul.

Yablo, Steve 1982 “Grounding, Dependence and Paradox”, *Journal of Philosophical Logic* 11, 117-137.

[3] E.g. Priest 1994.

The Origins of Wittgenstein's Phenomenology

James M. Thompson, Halle, Germany

While it is certainly true that the manuscripts comprising Wittgenstein's "middle" phase have enjoyed more attention since the publication of the *Nachlaß*, neither his conception of phenomenology, nor its origins have captured the interest of many within Wittgensteinian studies. The reason(s) for this situation are not fully clear and probably involve several, more or less, related factors, which I will not go into now. However, where little interest existed early on amongst Wittgenstein's interpreters, several thinkers associated with the phenomenological tradition were eager to take up the challenge of investigating these issues. This paper represents a brief overview of the possible origins of Wittgenstein's sudden and unexpected use of the term "phenomenology."[1]

While certainly not the first person to take note of Wittgenstein's use of the term "phenomenology" and "phenomenological grammar," Herbert Spiegelberg's initial article "*The Puzzle of Wittgenstein's Phänomenologie (1929-?)*" generated a great deal of attention, and marks the first serious attempt to take Wittgenstein's proclaimed phenomenology seriously. The "puzzle" began with the publication of the *Philosophical Remarks* in the original German. With this work, as Spiegelberg relates, came the "unexpectedly rich confirmation" to various allusions about a phenomenological theory and language that Wittgenstein had briefly entertained in 1929. Unfortunately, due to the lack of access to the unpublished manuscripts belonging to this period, Spiegelberg was not in a position to solve this riddle. However, his initial research and speculative efforts have significantly influenced later research regarding this topic, including my own efforts.

What Wittgenstein meant by the term "phenomenology" is certainly linked to the question of its origin. Although his use of the term is not entirely dependent upon its originary source, clearly, such information would be of great assistance in understanding what he wanted to associate himself with as well as distance himself from.

The most obvious question is whether or not Wittgenstein acquired the term from Edmund Husserl, either directly through his writings or indirectly via discussions, articles, and the like. Complicating the matter further, no comprehensive record of Wittgenstein's personal library exists. Aside from the authors Wittgenstein himself mentions, we have only second hand reports from friends and colleagues regarding books Wittgenstein had obviously been reading.

Even though we do not have any direct evidence of Wittgenstein having read Husserl, there are several anecdotes that prevent us from completely closing off this possibility or simply dismissing it out of hand. The first reference stems from notes taken during Wittgenstein's visits to the Vienna Circle between 1929 and 1930 by Waismann, which can be found in *Ludwig Wittgenstein and the Vienna Circle*.

During the course of their conversation on December 25th, 1929 the topic of *Phänomenologie* unexpectedly makes an appearance under the title *Physics and Phenomenology*. Paralleling comments in the *Philosophical Remarks*, here, Wittgenstein distinguishes his project – the logical investigation of phenomena in order to determine the structure of what is possible – from that of physics – which is only interested in establishing regularities. Toward the end of their discussion, in a section entitled *Anti-Husserl* – a title attributed to Waismann – Moritz Schlick poses the question to Wittgenstein: "What could one reply to a philosopher, who thinks the statements of phenomenology are synthetic *a priori* judgments?" (Wittgenstein 1980). Although Wittgenstein's response is rather condemning, as Spiegelberg points out, it is unclear whether or not Wittgenstein is rejecting this position with actual knowledge of Husserl or simply the position presented by Schlick. If the latter, we can hardly attribute an accurate and unbiased portrayal of Husserl's work by Schlick considering their on-going debate at that time.

Although not a member himself, Wittgenstein was certainly well acquainted with several of the Vienna Circles most influential patrons. The obvious question is: might one of members have been responsible for bringing Wittgenstein into contact with phenomenology? Felix Kaufmann would seem to be an obvious candidate, except there is no evidence that the two had anything to do with one another. And while Wittgenstein's relationship to Waismann was much closer, given that his disdain for Husserl was comparable to that of Schlick, Waismann would also seem to be an unlikely candidate.

If we are to hypothesize that Wittgenstein's sudden use of the term phenomenology is traceable to the Vienna Circle, then the most likely person to have influenced him would have been Rudolf Carnap. In his work, *The Logical Structure of the World* (1928), Carnap's conception of phenomenology reflects a certain influence of Husserl. This influence is almost certainly attributable to the contact he had with Husserl as Carnap was working on the first draft of his book. He had been staying in nearby Buchenbach between 1922 and 1925, and had attended several of Husserl's seminars in Freiburg from the summer semester of 1924 till the summer semester of 1925 (Spiegelberg 1981). While it cannot be said that Carnap was convinced of Husserl's position, his text nevertheless contains several non-critical references to the *Logical Investigations* as well as *Ideas I &II*, not to mention the adoption of Husserl's *epoché*. There are, however, two good reasons for doubting Carnap as a source for Wittgenstein's sudden use of the term phenomenology: First, their accounts of phenomenology are not very similar (although, as Spiegelberg points out, they are closer to each other's position than either is to Husserl's). This alone does not rule Carnap out, but in conjunction with Carnap's own admission that his relationship to Wittgenstein was quite strained during this time, the possibility of influence dwindles.

Another incident, which seems to lend circumstantial support for Wittgenstein's acquaintance with Husserl's work, involves a chance meeting between Wittgenstein and J. N. Findlay in 1939. Findlay mentioned to Wittgenstein

[1] This paper is a modified version of a section from my book *Wittgenstein on Phenomenology and Experience: An Investigation of Wittgenstein's 'Middle Period.'* Also, the quoted passages from Wittgenstein are my translation from the German original.

that he was working on a translation of Husserl's *Logical Investigations*, to which Wittgenstein "expressed some astonishment that he (Findlay) was still interested in this old text" (Spiegelberg 1981). While this by no means represents definitive proof, this anecdote keeps the possibility of Wittgenstein's first-hand knowledge of Husserl's phenomenology open.

Frege represents another potential source of contact between Wittgenstein and Husserl. Given that the Frege and Husserl corresponded with one another and were working on related problems, it is not unreasonable to think that Husserl's work or perhaps his ideas might have been mentioned. While I have, as of yet, not found any direct evidence for this connection in the correspondence between Wittgenstein and Frege, it nevertheless remains a promising avenue for further investigation.

Another figure who we should not leave unconsidered is Heidegger. Over the course of several years, Wittgenstein makes at least two references to his work. The first stems from a discussion with Waismann and Schlick, where Wittgenstein appears to make an unsolicited remark regarding *Being and Time* and the concept of *Angst*:

> I have a pretty good idea of what Heidegger meant by Being and angst. Man has the urge to run up against the limits of language. Think, for example, of the wonder that something exists. This wonder cannot be expressed in the form of a question, and there is not answer (Wittgenstein 1980)

In the passage, Wittgenstein continues to develop the connections between his notions of "wonder" [*Erstaunen*] and "the ethical" with those of Heidegger and Kierkegaard. This admission on the part of Wittgenstein that certain aspects of his early thought, i.e. the mystical experience of the world and the ethical, are moving in the same direction certainly indicates at least a partial familiarity with Heidegger's work.

Wittgenstein's second encounter with Heidegger is not as obvious as the first. During an early explication of language-games and the grammar of word usage, Wittgenstein is concerned with preventing "the philosopher" from "straying down hopelessly wrong paths." He then provides an example of just such a dangerous and misleading path present in language:

> If we want to deal with a sentence like 'the Nothing nothings' or the question 'what was earlier, the Nothing or the negation?' to be fair we must ask ourselves: what was the author thinking regarding this sentence? From where did he take this sentence? ... He who speaks about the opposite of Being and the Nothing as well as the Nothing as having priority over the negation, he thinks of – I believe – an island of Being surrounded by the endless sea of the Nothing (Wittgenstein 1998).

Although not named as such, the passage (and the accompanying pages) clearly points to Heidegger's lecture *What is Metaphysics*, in which the relationship of *Dasein* to "the Nothing" is treated. While Wittgenstein's attitude towards language such as "the Nothing nothings" is, indeed, critical, the passages do certainly suggest the provocative idea that Wittgenstein had first hand knowledge of Heidegger's work, even if the latter passage betrays a lack of understanding regarding Heidegger's point concerning "the Nothing" as a positive aspect of Being – and not as a mere negation of beings. When taken together, the two passages do seem to make Heidegger a promising candidate.

However, the purpose of this section is not merely to establish points of contact, but rather to investigate the origin of Wittgenstein's use of the term phenomenology. Or more precisely, was Wittgenstein's initial use of the term and corresponding project of a phenomenological language directly influenced by other phenomenologists? Keeping this distinction in mind, and given the time frame of these two references, the possibility that Wittgenstein was *influenced* by Heidegger begins to dwindle.

The first passage stems from the end of December 1929, and although that does not exclude the possibility that Wittgenstein had read *Being and Time* prior to his return to Cambridge, thus prior to his introduction of the term phenomenology, the comment alone is inconclusive. The second passage stems from the beginning of January 1932. Given that the lecture *What is Metaphysics?* was not even held until July 24th, 1929, and published later that same year, it cannot have been the impetus for Wittgenstein's phenomenology. Thus, while the possibility remains open whether or not Heidegger had any direct influence on Wittgenstein, the search for the source of his phenomenological project in all likelihood lies elsewhere.

As intriguing and provocative as these possibilities might seem, there are certainly other potential sources for Wittgenstein's use of phenomenological language, which may have little or no real connection to Husserl or Heidegger. Although now most prominently associated with the term phenomenology, Husserl by no means invented the term. Many individuals, prior to and even after the turn of the century, laid claim to the term phenomenology, among them: Hegel, Goethe, Mach, and Mauthner. And although Wittgenstein had read the work of the latter three thinkers (especially Mauthner), we do not find any real matches regarding Wittgenstein's "new" form of philosophizing.

Lastly, I would mention a theory that is neither glamorous, nor really even a theory, but more of an educated guess. On the one hand, the "theory" implies the least "causal" interaction, but, on the other, by ridding ourselves of the need for a "smoking gun" agent of change, we are probably closer to the truth of the matter. The theory contends that the term "phenomenology" was a part of the Viennese cultural landscape; that the term was simply floating freely within this uniquely charged and fertile atmosphere. Having been born and raised in Vienna to one of the wealthiest families in Europe, Wittgenstein was certainly in a position to absorb the vibrant cultural atmosphere existing at this time.

Continuing with the theme of a more general influence, it is even possible that his sister Margarete had a hand in the introduction of the term. She was the one who introduced the adolescent Ludwig to Schopenhauer's *The World and Will as Representation*, and thus to philosophy. Within the family, she was considered the most academically and culturally astute, and with her wealth she was able fully to immerse herself in the culture of that time. Margarete certainly had the opportunity to have discussed such topics with him, and even provide access to a great deal of philosophical literature. Perhaps, after his return to Cambridge (from Vienna), in order to distinguish his present phenomena-logical investigations from his earlier work, he simply adopted a familiar term without any concrete source in mind.

As I mentioned at the beginning of the section, the question regarding the origin of the term "phenomenology" in Wittgenstein's work will probably never be definitively answered. None of his known writings or notes mentions anyone specifically, and portions of his *Nachlaß* from around this time, which might have shed some light on the issue, were later destroyed per Wittgenstein’s instructions. That having been said, I would like to elaborate further my contribution to this speculative endeavor.

When one considers the kind of thinker Wittgenstein was, I would contend that the notion of any *specific* influence quickly evaporates. As Spiegelberg writes, “‘influence’ [is] a very complicated affair… [and in Wittgenstein’s case] could hardly ever amount to anything more than a stimulant and a trigger for his own thinking” (Spiegelberg 1981) With the notable exceptions of Schopenhauer, Frege, Russell and possibly Mauthner, talk of traceable influence in Wittgenstein’s writings would be, at best, an uphill fight.

Here, the problem of origin is analogous. Wittgenstein’s thought is so tightly wound around or within itself that to speak of an origin for his use of “phenomenology,” more likely than not, only misleadingly complicates the issue. By this, I am not proposing that Wittgenstein’s thought developed sealed up in some hermetic chamber; for obviously, he had been “influenced” by different thinkers and writers, even by his own admission. On several occasions, he even characterizes his own thought derogatorily as “reproductive” rather than creative or original (Wittgenstein 1977). However, the point is not whether Wittgenstein has been influenced by others, but rather how do these influences manifest themselves in his work, or concerning the question of origin, to what extent can something be regarded as being the source?

A characteristic of Wittgenstein’s work is the degree to which he has internalized the various voices presented. This is most apparent in his later works, but is actually present at every stage of his development. What this means is that Wittgenstein rarely engages in a discussion with another thinker; rather he has either so thoroughly taken over a particular viewpoint or abstracted the main tenets of a position (and continued their development) that notions of authorship begin to blur. The various positions encountered in his texts and notes are usually his own. In other words, he has personalized them to such a degree that it is not Descartes’ dualism against which Wittgenstein is arguing, but Wittgenstein himself representing this dualism – Wittgenstein contra Wittgenstein. An example of this aspect of his thought can be seen in his later critique of philosophy in the *Investigations*. While the critique is directed towards the philosophic tradition, in going about his task, Wittgenstein actually criticizes his own earlier views (mostly those contained in the *Tractatus*). Here, the faults and weakness of philosophy, he believes to be embodied in his earlier thought. Thus, by critiquing the *Tractatus*, Wittgenstein understands himself to be affecting a critique of philosophy as a whole.

In closing, I would point out that even if he acquired the term “phenomenology” in a more open and non-specific way, similar to what I have suggested above, it would be incorrect to conclude or simply insinuate that the term held no special significance for him. Quite to the contrary, had he been neutral with respect to calling his project “phenomenology,” it would never have survived the open and continuous hostility by certain members of the Vienna Circle, nor Moore’s repeated criticism of the term during Wittgenstein’s lectures.

Literature

Spiegelberg, Herbert 1981 The Puzzle of Wittgenstein's Phänomenologie (1929-?) reprinted in: The Context of the Phenomenological Movement, The Hague: Nijhoff.

Waismann, Friedrich 1980 Ludwig Wittgenstein und der Wiener Kreis (ed) B. F. McGuinness, Frankfurt am Main: Suhrkamp.

Wittgenstein, Ludwig 1998 The Bergen Electronic Edition: Wittgenstein’s Nachlaß (eds) Wittgenstein Archives at the University of Bergen, Oxford: Oxford University Press.

Wittgenstein, Ludwig 1977 Vermischte Bemerkungen (ed) G. H. von Wright, Frankfurt am Main: Suhrkamp.

Objects of Perception, Objects of Science, and Identity Statements

Pavla Toráčová, Prague, Czech Republic

I would like to present a brief investigation into the nature of theoretical identifications, in which we identify an object of our pretheoretical understanding with an object of scientific discovery. Traditional examples of such theoretical identifications are the following: „light is a stream of photons", "water is H_2O", "lightning is an electrical discharge", "gold is the element with the atomic number 79", etc.

Kripke: identity statements

I will start with the account of the nature of identity statements as provided by Saul Kripke (Kripke 1980). According to this account, identity statements like "light is a stream of photons", "heat is molecular motion", etc., are statements that involve two *rigid designators* that designate the identical natural phenomenon. A rigid designator is a term that designates the same object in every possible world. The key point is that the reference of the rigid designator is not determined by the description expressed by or associated with the term as it was suggested by the traditional theory of reference. According to that traditional theory, the referent is any object that fulfils the description expressed, possibly in a disguised form, by the term. Kripke shows that we often use terms (especially proper names, but also terms of natural species and natural phenomena) in a different way: we use them to refer to the particular object or phenomenon in its *necessary* existence, which doesn't involve the *contingent* properties the objects actually bears. Thus, the name "Kurt Gödel" designates a man who, providing his life had gone differently, might not have achieved all the things that he achieved. The name "evening star" designates the object—i.e., the planet—that, in a different possible world, is not seen in the evening sky. And similarly, term "heat" designates the natural phenomenon that, in another possible world, doesn't cause the characteristic sensations in sensitive creatures, and the term "gold" designates a natural species that might not have been yellow. The properties expressed by the term (or associated with it)—in the above examples they might be: "to have proven the incompleteness of arithmetic," "to be seen as the first bright object in the evening sky," "to produce the specific sensations in people"—are only contingently true of the objects and phenomena designated by the terms.

Now it is exactly this feature that makes the theoretical identifications possible. The structure of the identifications then may be explained as follows: Initially we identify the natural phenomenon by phenomenal qualities that are true of it in our world, and are true of it only contingently; heat, for example, we identify by the characteristic sensations it causes in us; gold by the color, brightness and solidity; water by the characteristic look, taste and feel, etc. Later in the scientific inquiry, we discover the *essential* properties of the phenomenon, i.e., the properties that are necessarily true of the phenomenon. In the statement of identity we then state the identity of the very object that is referred to by both terms—initially we refer to it without knowing its essential properties that we later scientifically discover.

Kripke says:

> „What characteristically goes on in these cases of, let's say, 'heat is molecular motion'? There is a certain referent which we have fixed, for the real world and for all possible worlds, by a contingent property of it, namely the property that it's able to produce such and such sensations in us. Let's say it's a contingent property of heat that it produces such and such sensations in people. It's after all contingent that there should ever have been people on this planet at all. So one doesn't know *a priori* what physical phenomenon, described in other terms—in basic terms of physical theory—is the phenomenon which produces these sensations. We don't know this, and we've discovered eventually that this phenomenon is in fact molecular motion. When we have discovered this, we've discovered an identification which gives us an essential property of this phenomenon. We have discovered a phenomenon which in all possible worlds will be molecular motion—which could not have failed to be molecular motion, because that's what the phenomenon *is*. On the other hand, the property by which we identify it originally, that of producing such and such sensation in us, is not a necessary property but a contingent one." (Kripke 1980, 132-133)

When we say "heat" we actually mean the phenomenon that is the same in every possible world, the phenomenon that possesses the essential property "to be molecular motion". When we use the term "heat" we *mean* this phenomenon from the very beginning although we have no knowledge of the essential property and the only way we are initially able to identify it is by the contingent property "to produce such and such sensations in us."

Kripke says:

> "In general, science attempts, by investigating basic structural traits, to find the nature, and thus the essence (in the philosophical sense) of the kind. /.../ Note that on the present view, scientific discoveries of species essence do not constitute a 'change of meaning'; the possibility of such discoveries was part of the original enterprise." (Kripke 1980, 138)

The ability of terms to refer to an object not by virtue of the object's properties which are known to us (and that are often expressed by the term)—indeed, to refer to the object, as it were, *despite* of these properties—is the crucial point of Kripke's theory of reference of proper names. It also contributes to our understanding of the theoretical identifications that have been discussed in this paper and that enjoy considerable interest among philosophers. However, there are some questions that arise: If the rigid designator designates the phenomenon independently of the description that goes with the term, be it the description of the contingent properties or the description of the essential properties, how can we know the scientifically discovered property is the essential one? Indeed, how could science have ever begun its enterprise, if the once available properties are regarded as not relevant for the necessary existence of the phenomenon that science aims to discover? To be sure, at the beginning of the scientific enterprise, the so-called "contingent" properties provide the only available knowledge of the object.

Another question that comes to one's mind is whether the scientific term really expresses the essential properties of the phenomenon—it can still be replaced by a better, and completely different, description as science proceeds forward in the analogical way the phenomenal term was replaced by the scientific one. Isn't, here, *the role* of the essential properties the same as the role of the phenomenal properties—i.e., to be regarded as the contingent property? To accept that doesn't necessarily bring any unsound relativity about the scientific truth—we can maintain that real natural phenomena posses certain essential properties, which after all makes one of our descriptions to be more adequate or true then the other, and at the same time to admit that none of our *descriptions* of the phenomena is perfectly adequate and absolutely true.

Strawson: objects of perception and objects of science

Now I would like to turn attention to ideas of P. F. Strawson, who provides an explanation of the identification of objects as scientists describe them with the objects as we know them by sense perception in his article "Perception and Its Objects" (Strawson 2002). Sense perception is, in Strawson's view, the way we initially or pretheoretically identify objects in the world—as having phenomenal qualities presented to us by our senses. Later on, when we achieve scientific descriptions of the objects in the world, we replace the phenomenal properties with the scientific ones. The traditional account of this changeover (i.e., the representative theory of perception) considers the phenomenal properties to be merely subjective sensations caused in us by the real objects that science truly describes. According to this traditional account, science teaches us that the objects *in fact* don't have the phenomenal qualities, that they, *in fact*, are not of the pale green color as we see them and of the rough surface as we touch them and of the sweet scent as we smell them and of the warm quality as we feel them. It's true that in our sense perception we perceive the objects as being green and fragrant and soft and tasty, but we are, as it were, the victims of a persistent illusion—these qualities don't belong to the objects themselves but to our experiences; they are nothing but our sensations (or sense data) which are caused by the real objects. They can be said to represent the outer objects, but the outer objects don't posses them as such, they just cause them in us.

In his article, Strawson argues against this representative approach and proposes an alternative picture of what's going on when we perceive objects and when we come to know the very same objects in a scientific investigation. First, he points out that the representative account is unconvincing. We live in a world of objects that are phenomenally propertied, where the properties belong to the objects, and not to my experiences. The book remains green even when I put it in my bag, the heat stays in the room even when I leave it. This phenomenal world is the public world, accessible to observation by others. Strawson writes:

> "Consider the character of those ordinary concepts of objects on the employment of which our lives, our transactions with each other and the world, depend: our concepts of cabbages, roads, tweed coats, horses, the lips and heir of the beloved. In using these terms we certainly intend to be talking of independent existences and we certainly intend to be talking of immediately perceptible things, bearers of phenomenal (visuo-tactile) properties. /.../ Surely we mean by a cabbage a kind of thing of which most of the specimens we have encountered have a characteristic range of colours and visual shapes and felt textures; and not something unobservable, mentally represented by a complex of sensible experiences which it causes." (Strawson 2002, 103–104)

Strawson thereby presents a common-sense realistic view with which we all pretheoretically live. In this view, phenomenal qualities belong to the objects and are considered to be the real properties existing independently of our experience of them. Its only later, when we are to explain the identity of the phenomenally propertied object and the scientific object, that we are willing to accept the illusory character of the phenomenal world: only then we can have a reason to say that the scientific description matches the real object whereas the phenomenal properties are mere subjective appearances of it, mere contingent effects of the real object. For if we don't accept the illusory character of phenomenal properties, how could we explain the question of identity?

Strawson doesn't accept this "illusionary" step and suggests his own explanation of the identity question: he uses the relativity of points of view that is present in our perception and in our ordinary ascription of the phenomenal properties to things:

> "The mountains are red-looking at this distance in this light; blue-looking at that distance at that light; and, when we are clambering up them, perhaps neither. Such-and-such a surface looks pink and smooth from a distance; mottled and grainy when closely examined; different again, perhaps, under the microscope." (Strawson 2002, 107)

We are used to shifting our point of view and so the different quality ascriptions are not seen as conflicting. Strawson suggests seeing the scientific descriptions as one of those shifts of point of view, though a more radical one. Scientists then can be seen as carrying out the original pursuit but by different means.

The pragmatic approach to identity statements

We can understand Strawson's view as the view that, in using phenomenal terms, we ordinarily *mean* objects as *essentially* phenomenally propertied, because we mean them as existing independently of us and our perception. In perceiving the object, we make the distinction between the object existing independently of us and our perceiving of it, for example between the seen object and our seeing it; this distinction is, according to Strawson, essentially embedded in our sense perception.

According to Kripke, however, terms that express phenomenal qualities designate objects that posses the phenomenal properties only *contingently*, as the effects in sensitive beings like us. The essential properties of the objects are, in Kripke's view, expressed only by the scientific descriptions.

The views of both philosophers are quite convincing, but, at the same time, both could also be challenged. Strawson's view can be challenged from the scientific point of view: We have scientifically discovered that the phenomenal properties are not the properties of the real objects, but they are the properties that the real objects cause in us. Kripke can be challenged if we are, for example, interested in the genetic question of knowledge: where does the drive for the knowledge of the essential properties come from if the phenomenal properties are irrelevant to the reference of phenomenal terms?

I believe we can reconcile these two views if we bring into play the pragmatic current present in both Kripke's and Strawson's accounts. It seems it must be the agent (or the speaker) who uses the term as referring to the properties out there, and who therefore makes the difference between the perceived thing (independent of perceiving) and the perceiving of the thing. Strawson claims there is a difference between the perceived and the perceiving present in the perception from the very beginning, and this feature points to the agent. As to Kripke's take on reference, we can ask how the term fixes the referent if it doesn't do it in virtue of its sense—and we don't have to go far to find the answer that it is the agent, who uses the term, that fixes the referent.

The question of the status of the descriptions of both the phenomenal and the scientific properties—i.e. the question if they pick out the necessary properties of the object, or if they refer to its contingent properties—then can be understood as a matter of the *manner of use* of the term. The identity statements then may be seen as consisting in the dynamic alteration of those manners of use. We are acquainted with the objects in the world as having certain properties that are essential for them; at the same time we are ready to abandon this belief and take those properties as mere appearances (or mere tentative descriptions) if another set of properties that we can take as the essential one—perhaps it allows better predictions—is available, where this shift is allowed by the other manner we use the terms—i. e., that we use them as fixing the referent independently of the properties expressed by the term or associated with it.

Literature

Kripke, Saul 1980 *Naming and Necessity*, Cambridge, Mass.: Harvard University Press.

Strawson, Peter F. 2002 “Perception and Its Objects”, in: Alva Noë and Evan T. Thompson (eds.) *Vision and Mind*, Cambridge, Mass.: MIT Press, 91-110 (originally published in: 1979, G. F. Macdonald (ed.), *Perception and Identity: Essays Presented to A. J. Ayer with His Replies*, London: Macmillan).

The Reduction of Logic to Structures

Majda Trobok, Rijeka, Croatia

Introduction

The structuralist account of logic endorsed by Koslow (Koslow 1992, 2007) is one of the most appealing contemporary formulations of structuralism in logic.

But, what does structuralism in logic amount to? Is it analogous with structuralism in mathematics or other domains? And if yes, in which sense?

In this paper I try to answer these questions by presenting the basic tenets of Koslow's theory; I then analyse his views and offer reasons for holding that some aspects of his structuralist account are flawed. Finally, I try to show that Koslow's theory of logic fails to achieve a satisfactory answer to the question of a possible reduction of logic to structure(s).

One of the fields that is paradigmatically about structures is mathematics. This can be read in two ways: mathematics is about different structures such as the vector space structure, the natural number structure, the group structure etc., while the possibility of reducing mathematical theories to set theory, gives sense to viewing mathematics as about the (common) set-theoretic structure. Philosophically speaking, the (ontological) reduction of mathematical objects to structures leads to interesting results that aim to solve some ontological, as well as epistemological, problems in the philosophy of mathematics (see (Resnik 1997), (Shapiro 1997), (Hellman 2001)), even though it also brings to the surface some difficulties such as the problem of structures that admit non-trivial automorphisms (for more details see for example (Hellman 2001, p.193)).

What about logic? Is there any analogy with mathematics, in the sense of being about (different) structures? Or is it maybe the case that logics share a universal structure? As is well known, different logics are based on different principles. Examples are legion. Let us just mention the case of relevance logic and its constraint of a necessary relevant connection between the premises and the conclusion in any argument, absolutely absent in classical logic.

Does it mean that the proposal of a common logical structure and, consequently, of a universal logic, is destined to fail? In this paper I will try to answer this question through a discussion of the tenets of the structuralist account of logic.

Koslow's structuralist account of logic

The lynch-pin of Koslow's structuralist account of logic is the notion of *implication structure* and the definition of logical (and modal) operators relative to an implication structure. Let us see what these definitions amount to and what results they imply.

An implication structure is any order pair: ((S, $\Rightarrow$); where S is a non-empty set, while "$\Rightarrow$" is an implication relation.

An implication relation is (implicitly) defined as any relation that satisfies the following conditions:

(1) *Reflexivity*: $A \Rightarrow A$, for each A in S
(2) *Projection*: $A_1, A_2, \ldots, A_n \Rightarrow A_k$, for every $k = 1, \ldots, n$, and for each A_i in S ($i = 1, \ldots, n$)
(3) *Simplification*: If $A_1, A_1, A_2, \ldots, A_n \Rightarrow B$, then $A_1, A_2, \ldots, A_n \Rightarrow B$, for all A_i ($i = 1, \ldots, n$) and B in S
(4) *Permutation*: If $A_1, A_2, \ldots, A_n \Rightarrow B$, then $A_{f(1)}, A_{f(2)}, \ldots, A_{f(n)} \Rightarrow B$, for any permutation f of $\{1, \ldots, n\}$
(5) *Dilution* (or *Thinning*): If $A_1, A_2, \ldots, A_n \Rightarrow B$, then $A_1, A_2, \ldots, A_n, C \Rightarrow B$, for any A_i ($i = 1, \ldots, n$), B and C in S
(6) *Cut*: If $A_1, A_2, \ldots, A_n \Rightarrow B$, and $B, B_1, B_2, \ldots, B_m \Rightarrow C$, then $A_1, A_2, \ldots, A_n, B_1, B_2, \ldots, B_m \Rightarrow C$, for any A_i, B_j, B and C (i, j = 1, …,n)

Someone might object that a different choice of constraints would be more fruitful and economical since clearly *Reflexivity* follows from *Projection*, and *Dilution* follows from *Projection* and *Cut*. Nevertheless, Koslow keeps the list of constraints for the sake of greater articulateness, based on Gentzen's theory.

Given the constraints, there are examples of implication relations that immediately come to mind, such as the notion of (semantic) consequence or the (syntactic) notion of deducibility for a set of sentences of some first-order logical theory. But, interestingly enough, these examples do not even remotely exhaust all the possibilities. Either in the sense of getting unusually defined logical operators, given a set of propositions of first-order logic or certain examples of logical operators defined on elements of S that are either not syntactic objects or truth bearers.

> The logical operators can act in a broad variety of settings, sentential and otherwise. In particular, the actions of the operators on structures of sets, names, and interrogatives, to cite just some non-standard examples, are mentioned because the items in these cases fail in an obvious way to be syntactical or fail to be truth-bearers.
> (Koslow 1992, p.9)

Set inclusion, for example, given any set of subsets of a non-empty set S, also fulfils all the mentioned constraints, so (S, $\subseteq$) exemplifies the implication structure. Other examples may be found in the context of the theory of individuals and erotetic logic (Koslow 1992, p. 209-229).

Such a definition might seem to be needlessly general, especially given its non-economicity, but this point is not lost on Koslow since its generality is more a virtue than a limitation. Analogously, it is possible to get some rather weird group structure examples or mathematically uninteresting equivalence relations. Of course, such examples might be more or less philosophically, mathematically or logically interesting and fruitful.

Given the definition of implication structures, the logical operators are defined *relative* to such structures, i.e. as functions defined on structures. And here again, given the possibility of non-standard implication relations, the same applies to the operators as well.

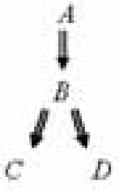
Let us take the example of the hypothetical operator $H_{\Rightarrow}$. For any elements A and B in the implication structure $(S, \Rightarrow)$, $H_{\Rightarrow}(A, B)$ is the hypothetical having A as the antecedent and B as a consequent, if and only if the following conditions are fulfilled:

(H1) $A, H_{\Rightarrow}(A, B) \Rightarrow B$

(H2) $H_{\Rightarrow}(A, B)$ is the weakest element satisfying the condition (1). It means that, for any element T of the implication structure such that $A, T \Rightarrow B$, it follows that $T \Rightarrow H_{\Rightarrow}(A, B)$

Such a definition leaves open *per se* the answer to the question as to whether the hypothetical, given an implication structure, may fail to exist or not. And the following example solves the dilemma positively.

Let us take the implication structure $(S, \Rightarrow)$, in which $S = \{A, B, C, D\}$ and the implication relation is given in the following way:



n such a structure, the hypothetical $H_{\Rightarrow}(A, B)$ does not exist (not to be confused with the fact that $A \Rightarrow B$); namely, $H_{\Rightarrow}(A, B)$ is, by definition, the weakest member T of S such that: $A, T \Rightarrow B$. Since $A \Rightarrow B$, the condition is fulfilled by any element of S, but there is no weakest element. C cannot be the weakest element since $A, D \Rightarrow B$, while $D \not\Rightarrow C$ (see the condition (H2) above). D cannot be the weakest for the same reason. (See Koslow (1992) for more examples).

As easily noted, such a definition does not put any constraints on truth conditions, syntactic features or others; as Koslow points out:

> There is no appeal to truth conditions, assertibility conditions, or any syntactical features or semantic values of the elements of the structure. (Koslow 1992, p. 78)

The fact that the elements of an implication structure are not necessarily syntactic objects having a special sign design or elements having a semantic value, is what make the explanation/definition of the logical operators free of such constraints.

The structuralist account of logic – critical remarks

I will concentrate on certain aspects of the structuralist account of logic and make a number of quite general remarks.

First, if we have a look at the six conditions (defined as (1) *Reflexivity*,…, (6) *Cut*) that any relation has to fulfil in order to be an implication relation, we may ask how is the left-hand side of the expression $A_1, \ldots, A_n \Rightarrow B$ to be construed? In the case in which S is a non-empty set of sets and the implication relation is set inclusion, the sequence $A_1, \ldots, A_n$ is the intersection of sets (Koslow 1992, p. 53). Since the intersection of sets *is* their conjunction, it turns out that in order to interpret the sequence $A_1, \ldots, A_n$, i.e. in order to determine that the implication relation is set inclusion**,** we ought to know what the intersection, i.e. the conjunction of sets is. It follows that in certain cases we ought to know how a certain logical operator is defined prior to having determined an implication relation on a non-empty set. According to the structuralist account of logic, it should be the other way round.

Secondly, one of the most interesting results of the structuralist account of logic is the definition of the operators for interrogatives:

Let Q be a set of interrogatives

S — a set of sentences inclusive of the sentential direct answers to the questions in Q

$S' = S \cup Q$.

A direct answer need not be a true one:

> We shall use the term “interrogative” to include any question that has a *direct answer*. The most important feature of the direct answers to a question is that they are statements that, whether they are true or false, tell the questioner exactly what he wants to know – neither more nor less. (Koslow 1992, p. 220)

The implication relation on S' (‘$\Rightarrow_q$’) is defined as follows (Koslow 1992, pp.218-229):

Let $M_1?, M_2?, \ldots, M_n?$ and $R?$ be questions in Q,
$F_1, F_2, \ldots, F_m$ and G be statements of S (the set of M's or the set of F's may be empty but not both), and
A_i be a direct answer to the question $M_i?$ (I = 1,…,n).

Then

(1.) $F_1, F_2, \ldots, F_m, M_1?, M_2?, \ldots, M_n? \Rightarrow_q R?$ if and only if there is some direct answer B to the question $R?$ such that
$F_1, F_2, \ldots, F_m, A_1, A_2, \ldots, A_n \Rightarrow B$

(2.) $F_1, F_2, \ldots, F_m, M_1?, M_2?, \ldots, M_n? \Rightarrow_q G$ if and only if
$F_1, F_2, \ldots, F_m, A_1, A_2, \ldots, A_n \Rightarrow G$

Such a definition is problematic. Let us see why. Let the set of F's be empty (for the sake of simplicity), and let us examine the case in which a question implies a statement (the second condition in the definition). Let the statement G be any false statement, e.g. a false answer to the question $R?$. In this case, whether $M_1? \Rightarrow_q G$ or not depends on whether $A_1 \Rightarrow G$, and the latter depends on what answer A_1 (to the question $M_1?$) we choose. If the answer we choose is a false one, then $M_1? \Rightarrow_q G$, otherwise $M_1? \not\Rightarrow_q G$. More generally, the same problem appears whenever the statement G is false. In this case, given a collection of interrogatives $M_i?$ (I = 1,…n), their respective direct answers A_i, and a set of true statements F_i (I = 1,…n), there is nothing in Koslow's definition that allows us to uniquely determine whether $F_1, F_2, \ldots, F_m, M_1?, M_2?, \ldots, M_n? \Rightarrow_q G$ or not.

Thirdly, let us consider a more general difficulty with the theory. Even though we might expect that certain results that hold given the operators classically defined hold in non-standard cases too, Koslow shows it is not the case. Let us mention the case**s** of implication structures in which $(((A \rightarrow B) \rightarrow A) \rightarrow A)$ is not a thesis or examples of structures in which the hypothetical with false antecedents can sometimes be true, and sometimes false (Koslow 1992, pp. 83-90).

These are features of the system, not of the structure**,** so it is not odd that such results are not necessarily present in non-standard implication structures. Nevertheless, having a true hypothetical whenever the antecedent is false (or having certain expressions as thesis) is not a marginal result given the operators classically defined. Two questions arise at this point: how do we get from the semantic-and-syntactic-features-free definitions to the syntactic rules for formula formation or the (semantic) truth tables, giving results that do not follow from the structurally defined operators? And how can a system have so many basic features that are not, in some form or another, already present in the structure?

Such problems make us think that the characterization of the operators relative to the implication structure does not, in itself, obtain the semantic and syntactic results we expect to get in the standardly defined implication structures. For these reasons Koslow's structuralist account does not satisfactorily solve the task of characterizing the logical operators relative to implication structures and even though tempting in many philosophical and logical aspects, it makes us sceptical as to whether a general logical structure might, after all, be defined.

Literature

Hamilton, A. G. 1978 *Logic for Mathematicians*, Cambridge: Cambridge University Press.

Hellman, G. 2001 'Three Varieties of Mathematical Structuralism', *Philosophia Mathematica*, (3) Vol. 9, 184-211.

Hirschman, L. and Gaizauskas, R. 2001 'Natural language question answering: the view from here', *Natural Language Engineering* **7** (4), 275-300.

Koslow, A. 1992 *A Structuralist Theory of Logic*, New York: Cambridge University Press.

Koslow, A. 1999 *The Implicational Nature of Logic*: *A Structuralist Account*, in Varzi, A. C. (ed.), European Review of Philosophy, The Nature of Logic, CSLI Publications, Vol. 4, 111-155.

Koslow, A. 2007 'Structuralist Logic: Implications, Inferences and Consequences', *Logica Universalis* 1, 167-181.

Nolt, J. 1997 *Logics*, Belmont: Wadsworth Publishing Company.

Resnik, M. D. 1997 *Mathematics as a Science of Patterns*, Oxford: Clarendon Press.

Scott, D. 1971 'On Engendering an Illusion of Understanding', *Journal of Philosophy*, 68: 787-807.

Shapiro, S. 1997 *Philosophy of Mathematics: Structure and Ontology*, New York: Oxford University Press.

Szabo, M. E. 1969 *The Collected Papers of Gerhard Gentzen*, Amsterdam: North Holland.

Reducing Sets to Modalities

Rafał Urbaniak, Ghent, Belgium

The ultimate aim here is to provide a modal interpretation of the language of set theory. Let us start with the question of ontological commitment of plural quantification. First, I introduce the language of Quantified Name Logic (QNL) and provide it with a set-theoretic semantics.

The **language of QNL** is generated by the alphabet containing brackets, **name variables**:

> a_1, a_2, a_3,... (as an abbreviation, also a, b, c, d, ... possibly with numerical subscripts),

the constant ε, the existential quantifier (∃*a*) (the universal quantifier (∀) has its usual definition), and two Boolean connectives: negation ¬, and conjunction &.

The set of **well-formed formulas** of QNL is the least set satisfying the following conditions:

(i) If a_1 and a_2 are name variables, ε(a_1, a_2) is a well-formed formula,

(ii) If A_1 and A_2 are well-formed formulas and *a* is a name variable, also ¬(A_1), &(A_1, A_2) and (∃*a*)(A_1) are well-formed formulas.

I also freely move to infix notation for the language.

Quite an **imparsimonious** but a fairly **standard semantics** for QNL is given as follows.[1] Take the domain to be a set of objects and take the range of name variables to be the power set of the domain. An **S-model** of QNL is a pair <*D*, *I*> such that *D* is an arbitrary set and *I* is a total function which maps name variables into the power set of *D* (i.e. to each name variable it assigns a subset of the domain). Neither *D* nor *I*(*a*) for any *a* has to be non-empty.

Definition. Truth in an S-model is defined by the following conditions:

> <*D*, *I*> models ε(*a*, *b*) iff |*I*(*a*)| = 1 (i.e. *I*(*a*) is a singleton) and *I*(*a*) is a subset of *I*(*b*).
>
> Phrases for negation and conjunction are standard: a model models a negation iff it doesn't model the negated formula, a model models a conjunction iff it models both conjuncts.
>
> <*D*, *I*> models (∃*a*)*A* iff <*D*, I^a> models *A* for some I^a which differs from *I* at most at *a*.
>
> A sentence is **S**-valid iff it is true in any **S**-model. ∆

One of the standard objections against nominalistic acceptability of the logic of plurals is that it needs a formal semantics, the set-theoretic semantics commits the pluralist to sets, and the substitutional interpretation of plural quantification does not provide the language with the required expressive power (we "run out of tokens", if they're supposed to be finite strings over a finite alphabet). In order to provide an answer to that objection, I give a semi-substitutional semantics which avoids the objections usually raised against the substitutional interpretation of plural quantifiers.

I develop a **Kripke semantics for QNL**. It is a modal interpretation, where the plural quantifier '(∃*a*)' (suppose *A* does not contain free variables other than *a*) is intuitively read as 'it is possible to introduce a name *a,* which would make *A* substitutionally true' (the semantics is different from that of Chihara). There are good reasons to claim that QNL with Kripke semantics has the same expressive power as QNL with set-theoretic semantics (as it turns out, for any set-theoretic model there is a Kripke model which agrees with it on all QNL formulas and the other way round: for any Kripke model there is a set-theoretic model which agrees with it on all formulas).

Definition. A naming structure is a tuple <*I*, *W*> where *I* is a set (of bare individuals) and *W* is a set of possible worlds. A **possible world** is a tuple <*N*, *d*> where *I* and *N* are disjoint sets and *d* is a subset of the Cartesian product of *N* and *I*. A **bare world** is the possible world where *N* is the empty set (Φ). The following conditions all have to be satisfied:

> *B* = <Φ, Φ> belongs to *W* (i.e. the naming structure contains the bare world).
>
> For any *w* in <*N*, *d*> different from *B*, *N* is non-empty and countable.
>
> The accessibility relation on possible worlds is defined by the following condition. Let *w* = <*N*, *d*>, *w'* = <*N'*, *d'*>. *Rww'* if and only if both: (i) *N* is a proper subset of *N'*, (ii) the restriction of *d'* to *N* (i.e. the set of those *d'*-related pairs whose first elements belong to *N*) is *d*.

Let <*N, d*> = *w* belong to *W*. A naming structure ***M*** = <*I*, *W*> is ***w*-complete** if and only if: for any subset *A* of *N* there exists a *w'* = <*N'*, *d'*> in *M* such that *Rww'* and there is an *x* in *N'* such that for any *y* in *I*, *d'*(*x*, *y*) if and only if *y* belongs to *A*. ***M*** is **complete** iff for any *w* in *W*, **M** is *w*-complete. ∆

Definition. An **M**-interpretation is a triple <*M*, *w*, *v*>, where ***M*** is a naming structure, *w* = <*N*, *d*> is a possible world in **M** and *v* either assigns to every variable in QNL an element of *N*, if *N* is non-empty, or is the empty function on the set of variables of QNL otherwise. If ***M*** is a complete naming structure, then we say that this **M**-interpretation is complete. ∆

Definition. Let <*M*, *w*, *v*> be an **M**-interpretation, *w* = <*N, d*>. Also, let *a* and *b* be QNL-variables and *A* and *B* be QNL-formulas.

> <*M*, *w*, *v*> models *a*ε*b* iff *v*(*a*) and *v*(*b*) are defined and there exists a unique *x* in *I* such that <*v*(*a*),*x*> is in *d* and there is a *y* in *I* such that both <*v*(*a*),*y*> and <*v*(*b*),*y*> are in *d*.
>
> The clauses for negation and conjunction are fairly standard. A model models a negation iff *v* isn't the empty function and it doesn't model the negated formula; and it models a conjunction iff it models both conjuncts.
>
> <*M*, *w*, *v*> models (∃*a*)*A* iff for some *w'* in *M*, *Rww'* and <*M*, *w'*, *v'*> models *A,* where *v'* differs from *v* at most in what it assigns to *a*.

A sentence is **true in a naming structure M** if and only if it is satisfied in its bare world under any valuation. A sen-

[1] QNL is pretty much a variant of Boolos's logic of plurals - the expressive power of both languages (*modulo* set-theoretic semantics) is the same.

tence is valid if and only if it is true in any naming structure. A sentence is complete -valid if it is true in any complete naming structure. Δ

It turns out that this semantics is in a sense equivalent to set-theoretic semantics.

Theorem. For any QNL sentence *A*, *A* is **S**-valid if and only if *A* is complete-valid. Δ

Next, let's take a look at the modal factor involved in this semantics by comparing it to a certain two-sorted first-order modal logic of naming (MLN). The language of MLN contains two sorts of variables: individual variables

$$x, y, z, x_1, x_2, ..., y_1, y_2, ..., z_1, z_2, ...$$

and name variables

$$n, m, o, n_1, n_2, ..., m_1, m_2, ..., o_1, o_2, ...$$

Besides, it contains quantifiers ranging over objects of those two sorts, the classical propositional connectives, two modal operators (say, ◊ for possibility and □ for necessity), no predicate variables, the identity symbol and one two-place predicate constant *D*. Formation rules are standard (the only new thing is that *D* takes name variables as first arguments and individual variables as second arguments). Models of MLN are just naming structures.

Definition. An MLN-interpretation is a tuple <*M*, *w*, *i*>, where *M* = <*I*, *W*> is a naming structure, <*N*, *d*> = *w* is in *M* and *i* is (a) undefined if *N* is empty, and (b) maps all individual variables into *I* and all name variables into *N* otherwise. Satisfaction of MLN-formulas in interpretations is defined as follows:

<*M*, *w*, *i*> models *D*(*n, x*) iff *i* is defined and <*i*(*n*), *i*(*x*)> is in *d*.

<*M*, *w*, *i*> models $t_1 = t_2$ iff $i(t_1) = i(t_2)$, where each t_i is one of the variables (arguments of = don't have to be of the same sort).

The clauses for negation and conjunction are standard.

<*M*, *w*, *i*> models (∃*t*)*A* iff there is an interpretation *i'* (mapping individual variables into *I* and name variables into *N*) that differs from *i* only in what it assigns to *t* and <*M*, *w*, *i'*> models *A*.

<*M*, *w*, *i*> models ◊*(A)* iff there is a *w'* such that *Rww'*and <*M*, *w'*, *i*> models *A*.

An MLN-sentence is true in a naming structure if it is satisfied in the bare world that underlies it. Δ

Intuitively we read '◊(*A*)' as 'there is a way names could be such that *A*' and '*D*(*n*, *x*)' as '*x* is one of the objects denoted by *n*' or '*n* refers to *x*' (where it is not assumed that names do not have to refer uniquely).

Clearly, there is a translation from QNL into MLN and QNL with Kripke semantics can be embedded in the language of MLN. Since this embedding preserves models (i.e. models for QNL and MLN are the same, what changes is just the interpretation of symbols), it seems that the ontological commitment of QNL with Kripke semantics does not go beyond the ontological commitment of first-order (two-sorted) modal logic (with one relation constant). Since it is much less plausible that first-order modal logic commits one to abstract objects than that plural quantification does, this strengthen the case for the ontological innocence of QNL.

The strategy can be extended to provide an account of a cumulative hierarchy of names:

Definition. A **cumulative naming structure** is a tuple <*I, W*>, where *I* is a set of bare individuals and *W* is a set of cumulative possible worlds. A **cumulative possible world** (c.p.w., for short) is a tuple <*d*, (N_{i+})>, where (N_i) is a denumerable family of sets of names indexed with positive natural numbers, d is the subset of the Cartesian product of *U*(N_i) and *I* ∪ (N_i) (that is, of the union of all N_i's and the union of *I* with the union of all N_i's), and the following conditions are satisfied (let i ∈ N₊):

$$(\forall i)\left[\left\{y\middle|(\exists x\in N_i)\langle x,y\rangle\in d\right\}\subseteq I\cup\bigcup_{k<i}N_k\right]$$

$$(\forall i>1)\left[x\in N_i\rightarrow(\exists y\in N_{i-1})\langle x,y\rangle\in d\right]$$

$$B=\langle d,(N_i)\rangle\in W,\text{ where }(\forall i)N_i=\varnothing$$

For any i, N_i is countable.

$$(\forall w)\left[w=\langle d,(N_i)\rangle\rightarrow(\exists i)(\forall k>i)N_k=\varnothing\right]$$

If *w* = <*d,* (N_i)> and *x* is in N_i, we say that *x* is a name of level *i* in *w*. Δ

The notions of accessibility and of completeness of a naming structure are obvious generalizations of the notions that we have already introduced.

Definition. Suppose i ∈ N₊. Let <*I*, *W*> be a cumulative naming structure and let *w* = <*d,* (N_i)> and *w'* = <*d'*, (N'_i)> belong to *W*. Then, *Rww'* if and only if:

$$(\forall i)N_i\subseteq N_i'$$

$$(\exists i)N_i\subset N_i'$$

$$(\forall i)\left\{\langle x,y\rangle\middle|x\in N_i\wedge\langle x,y\rangle\in d'\right\}=\left\{\langle x,y\rangle\middle|x\in N_i\wedge\langle x,y\rangle\in d\right\}\quad\Delta$$

Definition. Let *M* = <*I*, *W*> be a cumulative naming structure and let *w* = <*d*, (N_i)> (*I* > 0) belong to *W*. For the sake of convenience we fix the notation as follows. N_i's with *i* > 1 are sets of names, and N_0 is just another name for *I*. Clearly, there exists the least natural number *k* (0 is treated as a natural number but not as a positive natural number) such that for any *I* > *k*, N_i is empty. *M* is said to be ***w*-cumulatively complete** (*w*-complete, for short) if and only if for any

$$A\subseteq\bigcup_{n\geq0,n<k}N_n$$

there is a possible world <*d*, (N'_i)> = *w'* in *W* such that *w'* is at most of level *k*+1 (that is, (∀*i* > *k*) N'_i = ∅), both *Rww'* and:

$$(\exists x\in N'_{k+1})(\forall y)\left(y\in\bigcup_{n\geq0,n\leq k}N_n\rightarrow\left(\langle x,y\rangle\in d\leftrightarrow x\in A\right)\right)$$

M is said to be ***cumulatively complete*** iff for any *w* in *W*, *M* is *w*-cumulatively complete. Δ

If *w* is of level *k* (N_k is the highest non-empty element of *w*), then the **domain of names** of *w* (denoted by $D_{N(w)}$) is the union of of all N_i for 1 ≤ *I* ≤ *k*, and the **domain of objects** of *w* (denoted by $D_{O(w)}$) is the union of $D_{N(w)}$ and *I*.

Now, I will define a language that resembles the language of set theory, and the satisfaction relation for this language. The **language of cumulative naming logic** (CNL) contains the standard (first-order) logical symbols (including identity), variables x_i that (under an interpretation) will take pure individuals as values, variables a_i that (under an interpretation) will take either names or pure individuals as values, quantifiers that can bind variables of both sorts. Besides, the language contains one primitive symbol *D* - a two-place predicate (which can take variables of both sorts as arguments in arbitrary combinations) that in the intended reading means 'denotes'. A CNL term is either an individual variable, or an a_i variable (I will use the standard simplifications regarding dropping indices and writing *a, b, c, d* instead of a_1, a_2, a_3, a_4 and *u*, *v*, *x*, *y*, *z* instead of x_1, x_2, x_3, x_4, x_5).

Definition. Complete cumulative name structures are intended **models of the language of CNL**. A CNL interpretation is a tuple <*M*, *w*, *v*> such that *M* is a complete cumulative naming structure, *w* is a c.p.w. which belongs to it, and *v* (i) maps individual variables into *I* if *I* is not empty, and is undefined on individual variables otherwise, and (ii) maps the variables a_i into $D_{O(w)}$ if this set is non-empty and is undefined on a_i variables otherwise. Let *A*, *B* be CNL formulas and let *a*, *b* be CNL terms. The **satisfaction under an interpretation** is defined by:

<*M, w, v*> models *D*(*a, b*) iff <*v*(*a*), *v*(*b*)> is in *d*.

<*M, w, v*> models *a* = *b* iff *v*(*a*) = *v*(*b*).

The clauses for Boolean connectives are standard.

<*M*, *w*, *v*> models (∃*a*)*A* iff <*M*, *w'*, *v'*> models *A*, for some *w'* such that *Rww'* and for some *v'* which differs from *v* at most at *a*. Δ

We can introduce a constant *U* (Urelement) which in any possible world under any possible valuation will refer to all and only elements of *I*. Also, instead of (∃*x*)*x* = *a* I will write *U*(*a*). Instead of *D*(*a*,*b*) I will just write *a* ∈ *b* (so '∈' here has a different meaning than it has in set theory).

Certain (translations of) principles that hold for sets in ZF (with urelements) hold also for possible names. Some of them are:

$$(\forall a)\big(U(a)\rightarrow \sim(\exists b)b\in a\big)$$

$$(\forall a)\Big[\sim U(a)\rightarrow(\exists b)\big(\sim U(b)\,\&\,(\forall c)(c\in b\leftrightarrow\varphi(c))\big)\Big]$$

$$(\exists a)\big(\square\ U(a)\,\&\,(\forall b)(b\in a\leftrightarrow U(b))\big)$$

$$(\exists a)\big(\sim U(a)\&\sim(\exists b)b\in a\big)$$

$$(\forall a,b)(\exists c)(a\in c\,\&\,b\in c)$$

If we write $a\in \mathrm{U}b$ for $(\exists c)(c\in b\,\&\,a\in c)$ the following also holds:

$$(\forall a)(\exists b)(\forall c,d)(c\in d\,\&\,d\in a\rightarrow c\in b)\ .$$

Let's abbreviate $(\forall a)(\exists b)(\forall c,d)(c\in d\,\&\,d\in a\rightarrow c\in b)$ by $\varnothing(a)$. Then the following is valid:

$$(\forall a)\Big[\sim U(a)\&\sim\varnothing(a)\rightarrow(\exists b)\big(b\in a\,\&\,(\forall c)(c\in b\rightarrow\sim c\in a)\big)\Big].$$

There are, however certain axioms of ZF whose renderings fail miserably.The CNL rendering of the **axiom of extensionality**:

$$(\forall a,b)\Big(\sim U(a)\&\sim U(b)\rightarrow\big((\forall c)(c\in a\leftrightarrow c\in b)\rightarrow a=b\big)\Big)$$

and the **axiom of power set** in its name-theoretic translation is:

$$(\forall a)(\exists b)(\forall c)\big((\forall d)(d\in c\rightarrow d\in a)\rightarrow c\in b\big).$$

The axiom of extensionality fails because it is possible that there are coextensive and yet different name tokens. The axiom of power set fails because in the case of an infinite domain it would require that a possible world contains non-denumerably many name tokens. The first problem can be fixed easily: we just *define* identity symbol in a non-standard way so that coextensive possible names are identical *ex definitione*. The second problem requires a more elaborate move that lies beyond the scope of this paper. Let me, however, just indicate what this strategy would look like.

First, we start off with a cumulative naming structure. Then we stratify the possible worlds according to how high in the semantic ascent the tokens that exist in them are. For instance, if a possible world contains only names that name individuals, it is a world of level 1. If it also contains names that name names in a world of level 1 but no names of "higher" type, it is of level 2, etc. (formal definitions are easily available). Then, the crucial move is that we allow the reference relation of a name in a possible world *w* "reach" outside of that world, that is a name *x* in *w* is now allowed to "refer" to objects that don't exist in *w*. What *x* can refer to instead are all those objects that exist in worlds of lower level than *w*. That way we still have a cumulative hierarchy and don't run into any paradoxes, but also we validate the axiom of power set because now there is no problem with a name referring to non-denumerably many name tokens, as long as those tokens don't exist in a single possible world.

Are Lamarckian Explanations Fully Reducible to Darwinian ones? The Case of “Directed Mutation” in Bacteria

Davide Vecchi, Vienna, Austria

One of the striking features of the debate on directed mutation is that it was largely based on semantical quibbles generated by the idiosyncrasies on the interpretation of the crucial terms “random”, “directed”, “Darwinian” and “Lamarckian”. I am not claiming that this is an atypical situation in science. On the contrary, it is quite ubiquitous. But, of course, only part of the controversy rotated around semantics. The rest was about substantive scientific and philosophical issues.

The official Neo-Darwinian line on bacterial mutation crystallised in the 40s and 50s thanks to Luria’s and Delbrück’s and then the Lederberg’s “crucial” experiments. For those who thought that Duhem had established once and for all that there cannot be crucial experiments, it is sufficient to take a look at D. Futuyma’s “Evolution” textbook, where such experiments are presented as the ultimate and definitive demonstration that genuine directed mutation in bacteria cannot happen. According to the “received view” bacterial mutation is not a Lamarckian response to need; mutants arise at a constant rate during growth, independently of any selection pressure and environmental influence.

The Modern-Synthesis had hardened too much by discounting possible Lamarckian phenomena and processes in one stroke, by trivialising the Lamarckian position and by simplifying too much the Darwinian picture of bacterial evolution. A classic Kuhnian paradigm was build.

The odd thing is that Delbrück himself readily admitted in 1946 that the fluctuation test had limited scope, as it could not rule out the existence of adaptive mechanisms of mutation, the reason being that the selective pressure applied to bacteria in the test was too strong. Experimental anomalies contradicting the received view accumulated in the 40 years or so that climaxed with Cairns et al. 1988 paper. But such anomalies were simply written off as mere noise by referring to the crucial experiments. The story is quite interesting and also quite typical of science. History was edited and falsity transcribed in the biology textbooks by omitting crucial details, by referring to supposedly crucial experiments of limited scope, by using obscure terminology. But eventually the bubble burst. The Lamarckian idea of directed mutation was suddenly back from limbo, and thanks to just one authoritative paper published in an authoritative journal.

The “received view” on bacterial mutation was based on a series of tenets that started to be assessed independently. Some tenets can be deemed central to the Neo-Darwinian view, while others are possibly more peripheral. The central tenet is that mutations are not more likely to be beneficial than not. Lenski and Mittler (1993) provide possibly the clearest definition:

> We define as directed a mutation that occurs at a higher rate specifically when (and even because) it is advantageous to the organism, whereas comparable increases in rate do not occur either (i) in the same environment for similar mutations that are not advantageous or (ii) for the same mutation in similar environments where it is not advantageous.

Other tenets of the Neo-Darwinian view included the auxiliary hypotheses that mutations are never environmentally induced, that mutations are solely due to replication errors, that mutation rates are constant, that mutations do not target specific parts of the genome, and that mutations are due to the breakdown of the cellular machinery of DNA repair.

Even though the partial “blindness” of bacterial mutational response to need has been substantially vindicated, all the other tenets of the Neo-Darwinian view have been partly or totally rejected. What can safely be said is that there has been a softening of the Neo-Darwinian perspective in many respects (Brisson 2003). The softening is so extensive that one can sensibly ask whether to call the emerging view on adaptive mutagenesis “Neo-Darwinian” instead of “Lamarckian” makes sense at all (Jablonka and Lamb 2005).

One of the most intriguing explanatory models of adaptive mutagenesis, generally considered to be essentially correct (cf. e.g. Foster 2004), is the hypermutable state hypothesis (HSH). The HSH was originally proposed by Hall (1988) and provides an explication of the puzzling fact that some process seems to generate purely beneficial mutations in the genes under selection that allow environmentally stressed cells in stationary phase to resume growth. The main idea of the HSH is that most bacteria under stress-induced conditions are mutationally inert, but that an unspecified but relatively small number of cells enters an hypermutable state. Those bacteria that do not mutate appropriately die (are selected against), so that their deleterious and neutral mutations are wiped out and un-retrievable, while those striking a “lucky jackpot” resume growth and reproduce, while the beneficial mutations carried become heritable by escaping the correction mechanisms (e.g. mismatch repair editing, or MMR).

Focusing on HSH, we can now ask a number of specific questions. Consider that the history of the discovery of phenomena of apparently directed generation of variation is always surrounded by a first attempt to resurrect the ghost of Lamarck, followed by a call for restraint and a process of “Darwinisation” of the phenomenon. I will now assess whether such second level reaction is justified in the case of HSH. Can we explain HSH in purely Darwinian terms? The HSH surely provides a selective explanation of the supposedly Lamarckian phenomenon of “preferential mutation” first highlighted by Cairns et al. But the HSH does not fully vindicate a Neo-Darwinian perspective at all, given that most auxiliary hypotheses surrounding the view have been rejected or substantially revised. Clearly HSH does not rely on the hypothesis of constant mutation rates (cells that enter the hypermutable state can mutate at rates 800 times higher than normal), nor on the hypothesis that mutations do not happen in stationary phase (in fact, mutations accumulate in stationary phase, that is, in that period of the bacterium life cycle, somehow equivalent to development, when it does not replicate), nor on the hypothesis that environmental factors are irrelevant to mutational responses (it has been established that there exist very specific processes of mutational response triggered by environmental stresses like starvation), nor on the hypothesis that there is no targeting of specific genomic regions, while growing evidence is accumulating in favour of seeing mutational responses as evolved adaptive responses of the organism. What can be safely said today is that random variation is not the norm. Mutational processes are highly biased and targeted affairs. Still, can a revised and softened version of Darwinism suffice to dispel the ghost of Lamarck? As Hall (1998) argues, adaptive mutation is a different process from growth-dependent or “normal” bacterial mutation, even though it clearly is a Darwinian process. The important questions on this score are: how is the hypermutation state triggered? Is it triggered by the environmental stress or by the selective pressure? Is it regulated by specific genes, or rather by some unspecified and unknown, possibly innate, capacity of the organism? We still don’t know. For the same reasons, I contend, we cannot answer the question of whether we need an explanation of HSH in terms couched via the language of intentionality and choice, or whether a Darwinian explanation couched in terms of chance and selection is all we need. Or so I will argue.

It could be argued that the phenomenon of adaptive mutagenesis is purely Darwinian. From this perspective it could be asked: what does the Lamarckian language of intentionality, purpose and choice add to explanations concerning adaptive mutations? Nothing is the answer: Lamarckian concepts are simply redundant. To say that organisms choose to mutate in particular ways, that they deliberately try to reach a particular adaptive goal, that they intend to change and adapt in specific ways, is equivalent to invite inappropriate teleological talk. Even though emphasis on the phenomenon has highlighted the need to discriminate between mutations with different characteristics concerning their basic properties (environmental conditions, time and site of occurrence), it is clear that the directional outcome is a product given by the extreme reduction of design search space, by the temporal and local increase in efficiency of the search, and by a more general strategy of mutation rate increase.

A variant of the same dismissive attitude could instead stress that instead of explanatory redundancy (based on the application of Ockham’s razor), Lamarckian explanations are reducible to Darwinian ones. This would vindicate the hypothesis that Darwinism is a more fundamental process than Lamarckism. In fact, it seems that the “intentionality” of Lamarckian systems like those supposedly governing adaptive mutational responses must arise either by means of the cumulative effects of past natural selection acting on ultimately random mutations, or more simply by a more error-prone pathological response of cells. Therefore the epistemic reducibility of Lamarckian explanations seems to be achievable. So, even though the orthodox and hardened Neo-Darwinian views on directed mutation has been revised and updated, the revision of the Darwinian view has not touched certain residual and crucial elements constitutive of some meaning of “randomness”, which make an appeal to the Lamarckian language of intentionality inappropriate. This alternative seems to illustrate the sensible view shared by most practitioners.

However, I think that this epistemological move cannot shed any light on the ontological issue pertaining to the nature of mutation, and on the basic question concerning the existence of genuine Lamarckian processes. What could a genuine Lamarckian phenomenon be like? Genuine intentionality would perhaps involve the preferential environmental induction of useful and beneficial mutations, which is not happening, or the targeting of specific genes, which is still quite controversial. But an equally promising avenue would be to consider genuine intentionality as involving the choice of starved and stresses cells to enter the hypermutation state. I think that at this juncture lies the best chance to vindicate some genuine Lamarckian credentials. If the hypermutation state is triggered by some as yet unspecified and unknown capacity of the organism, then Lamarckism could be vindicated. Or better it would be if we were unable to find a mechanistic basis for the cell response.

Intentionality does not seem to be needed to provide explanations that save the known phenomena of adaptive mutagenesis. Still, this could be simply appearance. The crucial question remains: how do we explain the fact that only a subpopulation of bacteria hypermutates? It seems that some element in the environment triggers such a state, or perhaps it is a cell that chooses to do so? Where is the evidence in favour of the first “mechanical” hypothesis? And is the first hypothesis simpler than the second? And why is it simpler to hypothesise that the bacterium *senses* an environmental signal rather than hypothesising that it *chooses* to hypermutate? In the first case we need a sensory mechanism, in the second just agency of some kind. Furthermore, even a sensory mechanism could be sufficient to phagocyte a slice of the intentional cake.

The position of the reductionist and materialist rests on the adoption of the argument from parsimony. But voiced have been raised to argue against this adoption of Ockham’s razor, and against a more general ontological conservatism (Sober 2001), especially when the intentionality hypothesis is ontologically as economical as the alternatives. The issue is that as there are no arguments in favour of the reification of intentionality as a distinct and emergent quality, so there are no arguments against.

So it could be that the more one digs into the subject the more the Lamarckian language of choice makes sense. In this sense it is not enough to point out that polymerase IV and the down regulation of MMR are involved in the process of hypermutation, that is, an error-enhancing and a less efficient error-editing process. The question remains of what triggers the state, of whether it is the bacterium that “decides” to hypermutate.

Also note that in many ways a panpsychist hypothesis, with the postulation of all its monads, is not clearly more complex and therefore amenable to be chopped off with one stroke of Ockham’s razor. Even bacterial chemotaxis is nowadays described somehow in terms of cells’ intentionality (cells are analogised to nanobrains). The panpsychist perspective according to which intentionality is everywhere remains simpler that any emergentist view. If mentality and intentionality are basic and primitive, then ascriptions of mentality are, despite the risk of anthropomorphism, always justifiable. The idea that there exists a continuum of levels of intentionality, rather than an abrupt emergence of something totally new, remains appealing.

The great appeal of the Darwinian account (e.g. based on a blind variation and selective retention formula) is that it offers a very general mode of explanation, and a clear recipe to dispense of epiphenomenal intentionality (Hull 1988). The active agent easily disappears from selectionist accounts. But can we get rid of agency so simply? Perhaps some people would be more prone to get rid of bacterial intentionality than of human cultural agency, but this could reflect an unjustifiable bias.

I conclude that there still remains the open question regarding the existence of truly Lamarckian processes, whose workings can only be accounted via intentional and mentalistic terms, even without assuming foresight on the part of cells, and even without assuming a universe of monads. My main point is merely that we have no evidence to reject the intentionality hypothesis, nor principled theoretical reasons based on simplicity considerations, and even less so on dubious metaphysical fiat. I suspect that the available evidence cannot adjudicate between the “sensible” scientific hypothesis based on mechanism, selection and chance, and the “awkward” intentional alternative couched in terms of cell’s choice and intentionality. And probably we will never know for sure.

Literature

Brisson, D. 2003 *The Directed Mutation Controversy in an Evolutionary Context*. Critical Reviews in Microbiology, 29:25-35

Cairns, J., Overbaugh, J. and Miller, S. 1988 *The origin of Mutants*. Nature, 335:142

Foster, P. (2004). *Adaptive Mutation in Escerichia coli*. Journal of Bacteriology, VOL. 186, No. 15, p. 4846-452

Hall, B.G. 1988 *Adaptive evolution that requires multiple spontaneous mutations*. Genetics 120:887-897

Hall, B.G. 1998 *Adaptive mutagenesis: a process that generates almost exclusively beneficial mutations*. Genetica 102/3: 109-125

Hull, D.L. 1988 *Science as a Process*. The University of Chicago Press

Jablonka, E. and Lamb, M. 2006 *Evolution in Four Dimensions*. MIT Press

Lenski, R. & Mittler, J.E. 1993 *The Directed Mutation Controversy and Neo-Darwinism*. Science, 259:188-194

Sober, E. 2001 *The Principle of Conservatism in Cognitive Ethology*. In Naturalism, Evolution, and Mind. Royal Institute of Philosophy Supplement 49

A Note on *Tractatus* 5.521

Nuno Venturinha, Lisbon, Portugal

Introduction

Wittgenstein scholars have long been puzzled by §5.521 of the *Tractatus*. It reads as follows:

> I separate the concept *all* from the truth-function. Frege and Russell have introduced generality in connexion with the logical product or the logical sum. Then it would be difficult to understand the propositions “$(\exists x).fx$” and “$(x).fx$” in which both ideas lie concealed.

We shall best get to the heart of the puzzle by considering the way Bertrand Russell treated the matter in his “Introduction” to the *Tractatus*. Indeed, Russell made a particular assumption which seems on the face of it to be incoherent. He alludes to “Mr Wittgenstein’s theory of the derivation of general propositions from conjunctions and disjunctions” (TLP, 15), a perspective Wittgenstein ascribes to Frege and Russell himself. In fact Wittgenstein explicitly says that both “Frege and Russell have introduced generality in connexion with the logical product or the logical sum”, thus deriving “$(x).fx$” from “*fa.fb.fc.* …” and “$(\exists x).fx$” from “$fa \vee fb \vee fc \vee \ldots$”, a procedure he rejects.

However, astonishingly as it may seem, after 1929 Wittgenstein criticized his earlier conception of generality, which took “$(x).fx$” to be a “logical product” and “$(\exists x).fx$” to be a “logical sum”. In his record of Wittgenstein’s lectures of 1930-33, G.E. Moore writes that “[h]e said that there was a temptation, to which he had yielded in the *Tractatus*, to say that $(x).fx$ is identical with the logical product ‘*fa.fb.fc.* ...’, and $(\exists x).fx$ identical with the logical sum ‘$fa \vee fb \vee fc \vee \ldots$’”, assuming that “this was in both cases a mistake” (MWL, 89). Further, Moore notes that “[h]e said that, when he wrote the *Tractatus*, he had supposed that *all* such general propositions were ‘truth-functions’”, recognizing that “in supposing this he was committing a fallacy, which is common in the case of Mathematics, *e.g.* the fallacy of supposing that 1+1+1 ... is a sum, whereas it is only a *limit*” (ibid.). And, in the same vein, G.H. von Wright reports that “[i]n one of the first conversations [he] ever had with Wittgenstein (in 1939), he said the biggest mistake he had made in the *Tractatus* was that he had identified general propositions with infinite conjunctions or disjunctions of singular propositions” (von Wright 1982, 151, n.28).

It looks, therefore, as though there is a contradiction, even if Wittgenstein’s remark at §5.521 of the *Tractatus* clearly suggests that Russell’s interpretation cannot be right. Thus H.O. Mounce, who is one of the most lucid interpreters of the generality issue in the *Tractatus*, even arguing that Russell misunderstood it, emphasizes that “we can be certain, from Wittgenstein’s own remarks on the subject, that he was confused on this matter at the time of the *Tractatus*”; and he goes on saying that “[w]hat is not at all easy to determine, however, is where precisely his confusion lies” (Mounce 1981, 67). In this paper, following the lead of Mounce, I shall try to make clear that Wittgenstein’s criticism in the 1930s is directed at his earlier view that the content of general propositions can be *enumerated*, not at the way in which he *introduced* such propositions. But, on the basis of the third of the wartime notebooks that survived and the so-called *Prototractatus*, I go deeper into Wittgenstein’s alleged ”confusion”, analysing some hitherto neglected aspects.

I

Section 73 of the “Generality” chapter of the *Big Typescript*, which bears as title “Criticism of my Earlier Understanding of Generality”, begins with a remark deriving from the first entry of 1 August 1931 in MS111. Wittgenstein says:

> My understanding of the general proposition was that $(\exists x).fx$ is a logical sum, and that although its terms weren’t enumerated *there*, they could be enumerated (from the dictionary and the grammar of language). (…) (TS213, 326: BT, 249e)

And the next remark of the section, deriving from the first entry of 1 December 1931 in MS113, runs as follows:

> Of course, the explanation of $(\exists x).\phi x$ as a logical sum and of $(x).\phi x$ as a logical product cannot be maintained. It was linked to a false view of logical analysis, with my thinking, for instance, that the logical product for a particular $(x).\phi x$ would most likely be found some day. – Of course it’s correct that $(\exists x).\phi x$ functions in some way as a logical sum, and that $(x).\phi x$ functions in some way as a product; indeed for *one* use of the words “all” and “some” my old explanation is correct, namely, in a case like “All the primary colours can be found in this picture”, or “All the notes of the C major scale occur in this theme”. But in cases like “All people die before they are 200” my explanation is not correct. (…) (TS213, 326-7: BT, 249e)

The examples given by Wittgenstein illustrate the *singular* character of our universal statements, that some refer to a simple *set*, which is thinkable as belonging to a totality that is presupposed (all colours, all musical notes), and that others refer to a whole whose *particularity* is manifest. In this latter case, there is not, therefore, any logical product, the dots in “*fa.fb.fc.* ...” being dots of innumerability, not “dots of laziness”, as Wittgenstein called them in his lectures, the sort we use to speak, for example, of the alphabet in terms of “A, B, C …” – that is, when the enumeration, though possible, is not carried out (cf. MWL, 90; also AWL, 6).

It is noteworthy, however, that the reason why the enumeration of an “infinite series”, say “1, 2, 3 ...”, is impossible is not our inability to complete it; it is rather that it belongs to the concept of *infinite* its innumerability (cf. MWL, 90; also LWL, 90). In a conversation on 22 December 1929 with Moritz Schlick and Friedrich Waismann recorded by the latter, Wittgenstein actually alludes to the whole that is made up by the four primary colours as a “finite conjunction”, constituting it a “*finite* logical product” - a *contradictio in adjecto*, since there can be no infinite logical products (cf. WVC, 45). What is projected in such cases is simply a horizon of vagueness.

Let us consider the example Wittgenstein gives in the opening of the conversation alluded to above, which is: “All men in this room are wearing trousers” (WVC, 38).

What is at stake here? That “Professor Schlick is wearing trousers, Waismann is wearing trousers, Wittgenstein is wearing trousers, and no-one else is present”, that is to say, that “Mr. Carnap is not in this room, Mr...., etc.” (ibid.). But do we really think of an infinite number of propositions about what is *not* the case? Wittgenstein now holds, contrary to his original idea, that this constitutes, rather, an “incomplete picture”, which the symbolism “must show [to be] *incomplete*” (WVC, 39-40). The question is: how to do that? How can we represent in a propositional scheme that “There is no man is this room” except by means of “~(∃*x*).*fx*”, which, being equivalent to “(*x*).~*fx*”, immediately yields a logical product “~*fa*.~*fb*.~*fc*. ...”, thus requiring an enumeration? Wittgenstein’s suggestion is that we should translate existential propositions such as “*x* is in the room” or “There is someone in the room” by means of “*fx*”, corresponding its negation solely to “~*fx*” (cf. WVC, 40, 44). What he claims is that the variable at issue is not an “apparent variable” but a “real variable”, one that does not require *individual constants* (cf. WVC, 39). Note also that in the case of the existential proposition “(∃*x*).*fx*” we have a similar case of enumeration, since it corresponds to the logical sum “There is in the room either this person or that person or that person, etc.”, an expression that, according to Wittgenstein, is nonsensical. Obviously, this does not happen with propositions like “In this square there is one of the primary colours”, to use one of the examples Moore mentions, because there the expression “*fa* ∨ *fb* ∨ *fc* ∨ …” is *conclusive*, being equivalent to “In this square there is either red or green or blue or yellow” (MWL, 89). But in all the other cases, including of course the negative ones, viz. “(∃*x*).~*fx*”, we would have infinite remissions. Now, in rejecting (Frege’s and) Russell’s notation, which he had previously adopted, Wittgenstein not only avoids the indefinite enumerability problem, but also the “twofold negation” problem, i.e. that “(∃*x*).*fx*” does not have the “right multiplicity” (WVC, 39-40). This shows, in truth, that “There is no man who is not in the room” is nonsensical and that “~(∃*x*).~*fx*” cancels the meaningfulness of “(∃*x*).*fx*”.

Wittgenstein’s conclusion is that, as Moore reports, “the cases to which the *Principia* notations (*x*).*ϕx* and (∃*x*).*ϕx* apply [...] are comparatively rare”, given that “oftener we have propositions, such as ‘I met a man’, which do not ‘presuppose any totality’”; moreover, Wittgenstein goes as far as to argue that “the cases to which the *Principia* notation apply are only those in which we could give proper names to the entities in question”, something that “is only possible in very special cases” (MWL, 91). All the others require, in effect, a concrete grammatical analysis. They cannot be seen in the light of a predefined scheme but in what they really involve.

II

We are now in a position to reconsider Wittgenstein’s “confusion”. The problems identified appear to conflict with important Tractarian themes. However, Wittgenstein’s early view actually withstands the criticisms that he himself later identifies. In fact, what he contests at §5.521 of the *Tractatus* is merely the *extralogical* way in which Frege and Russell “have introduced generality”.

Let us briefly examine Wittgenstein’s procedure for deriving general propositions, which is presented at §5.52. He writes:

> If the values of *ξ* are the total values of a function *fx* for all values of *x*, then $\mathrm{N}(\bar{\xi}) = \sim(\exists x).fx$.

His idea is that the N operator can be applied to “*fx*”, an existential proposition, which can be written in the form of “(∃*x*).*fx*”, obtaining “~(∃*x*).*fx*”, i.e., “~*fa*.~*fb*.~*fc*. …”, that is to say, *all* the propositions of *ξ* being false, which results in “(*x*).~*fx*”. The application of N to this gives us “~(*x*).~*fx*”, which in turn is equivalent to “(∃*x*).*fx*”. If we then apply N to “~*fx*”, we get “~(∃*x*).~*fx*”, that is, “(*x*).*fx*”, and by the same operation again we obtain “~(*x*).*fx*”, which is equivalent to “(∃*x*).~*fx*”. According to this proposal, the universality is, paradoxically enough, derived from existentiality, from what we do have indeed, even if it is also true that we do have an original relation to the idea of “all”. Still, to derive “(*x*).*fx*” from “*fa*.*fb*.*fc*. …” is a big step, one that can only be taken *extralogically*.

This became apparent to Wittgenstein at the time of composing the third of the surviving notebooks from the First World War. The opening entry of 13 July 1916 provides a clue:

> One keeps on feeling that even in the elementary proposition mention is made of all objects. (MS103, 23r: NB, 76e)

And in an entry from 20 July, omitted in the *Notebooks 1914-1916* along with quite a few remarks on the same subject sketched in the previous days (cf. MS103, 25r-27r), Wittgenstein observes:

> ~~The~~ My old division of all propositional forms was fundamentally correct, only another mode of generality will be required. (MS103, 27r: my translation)

This is the reason why, as we read in the *Prototractatus* manuscript, where the original version of §5.521 of the *Tractatus* was formulated, Wittgenstein “separate[s] the concept all from ~~the logical product.~~ the truth-function” (MS104, 87, §5.3201: my translation, adapted to Ogden’s). It is true all the same that he holds it as a “logical product” – and this is the core of his later criticism. But it is one thing to hold it and another to derive it.

The way Russell understood “Mr Wittgenstein’s theory of the derivation of general propositions”, while proceeding, *in fine*, as his own, “from conjunctions and disjunctions”, is thus incomprehensible, all the more since a few lines above he had written that “Wittgenstein’s method of dealing with general propositions […] differs from previous methods by the fact that the generality comes only in specifying the set of propositions concerned”, so that “when this has been done the building up of truth-functions proceeds exactly as it would in the case of a finite number of enumerated arguments *p*, *q*, *r*…” (TLP, 14). Russell will have not noticed, therefore, the real innovation of such a methodology, which lies, precisely, in the specification of “the set of propositions concerned”, not being needed an enumeration of them.

This, as a matter of fact, had already been pointed out by Wittgenstein to Russell in the postscript to a letter dated 19 August 1919. There, replying to a number of questions raised by Russell in a letter from 13 August (cf. CL, 121-3), Wittgenstein states:

> I suppose you didn’t understand the way, how I separate in the old notation of generality what is in it truth-function and what is purely generality. A general prop[osition] is A truth-function of *all* PROP[OSITION]S of a certain form. (CL, 126)

And he goes on saying, referring to the symbol "N($\bar{\xi}$)" and to Russell's feeling that "the duality of generality and existence persisted covertly in [his] system" (CL, 122):

> You are quite right in saying that "N($\bar{\xi}$)" may also be made to mean $\sim p \vee \sim q \vee \sim r \vee \sim$... But this doesn't matter! I suppose you don't understand the notation of "$\bar{\xi}$". It does not mean "for all values of ξ...". But all is said in my book about it and I feel unable to write it again. (CL, 126)

In short, Russell will have seen in Wittgenstein's method simply another way of obtaining all the quantifiers, not a truly alternative way of deriving generality, avoiding the "old" recourse to "*fa.fb.fc.* …" and "$fa \vee fb \vee fc \vee \ldots$". Yet, only by means of an *extralogical* procedure we may turn round the *singular* nature of our point of view. It is the N operator that makes it possible to realize that our relation to the universal is *constitutive*, even though the *epistemological* status of that relation is problematic, amounting propositions such as "All men are mortal" to mere "variable hypotheticals", in the phrase of F.P. Ramsey (1931, 237). Wittgenstein's refusal of an *inductive logic*, expressed at §6.31 of the *Tractatus*, turns out, in this light, to be clearer. What is not at all clear is why Wittgenstein followed §5.3201 in the *Prototractatus* notebook by a remark, which he crossed out, saying that "[e]thics is not one of the natural sciences" (MS104, 87: my translation). This, however, I cannot go into.[1]

Literature

Ambrose, Alice (ed.) ^{2}1982 *Wittgenstein's Lectures: Cambridge, 1932-1935*, Oxford: Blackwell. (AWL)

Lee, Desmond (ed.) 1980 *Wittgenstein's Lectures: Cambridge, 1930-1932*, Oxford: Blackwell. (LWL)

McGuinness, Brian (ed.) 1979 *Wittgenstein and the Vienna Circle*. Translated by Joachim Schulte and Brian McGuinness, Oxford: Blackwell. (WVC)

McGuinness, Brian and von Wright, G.H. (eds.) 1995 *Ludwig Wittgenstein: Cambridge Letters*, Oxford: Blackwell. (CL)

Moore, G.E. 1993 "Wittgenstein's Lectures in 1930-33", in: J.C. Klagge and Alfred Nordmann (eds.), *Ludwig Wittgenstein: Philosophical Occasions 1912-1951*, Indianapolis: Hackett, 45-114. (MWL)

Mounce, H.O. 1981 *Wittgenstein's Tractatus: An Introduction*, Oxford: Blackwell.

Ramsey, F.P. 1931 *The Foundations of Mathematics and other Logical Essays*. Edited by R.B. Braithwaite, London: Routledge and Kegan Paul.

von Wright, G.H. 1982 *Wittgenstein*, Oxford: Blackwell.

Wittgenstein, Ludwig ^{2}1933 *Tractatus Logico-Philosophicus*. Translated by C.K. Ogden, London: Routledge and Kegan Paul. (TLP)

Wittgenstein, Ludwig ^{2}1979 *Notebooks 1914-1916*. Edited by G.H. von Wright and G.E.M. Anscombe. Translated by G.E.M. Anscombe, Oxford: Blackwell. (NB)

Wittgenstein, Ludwig 2000 *Wittgenstein's Nachlass: The Bergen Electronic Edition*, Oxford: OUP. (MSS & TSS)

Wittgenstein, Ludwig 2005 *The Big Typescript: TS 213*. Edited and translated by C.G. Luckhardt and M.A.E. Aue, Oxford: Blackwell. (BT)

[1] Many thanks to Andrew Lugg for helpful comments and suggestions. Work on this paper was made possible by a postdoctoral fellowship from the Portuguese Foundation for Science and Technology.

The Place of Theory Reduction in the Models of Interdisciplinary Relations

Uwe Voigt, Bamberg, Germany

1. Introduction: Why Theory Reduction is Not Yet Considered in Connection with Interdisciplinary Relations – And What can Be Done About It

In the first place, this approach has to deal with the question why interdisciplinarity is not a topic for the philosophy of science. The answer to this question could be, according to Wittgenstein, that a certain picture has taken hold of the philosophers of science, or even a whole bulk of such pictures. These pictures obviously are implicit models about the way sciences do relate. The implicitness of these models prohibits their philosophical reflection. Hence, the best philosophy of science can do in this case is to make them explicit.

One way to make them explicit is demonstrating the fundamental decisions which lead to the different models. So it can be shown how they differ from one another and how they make up more or less similar "families". It can also be shown where the place of theory reduction in the according "family tree" can be found and which branches of this tree are cut off if one chooses theory reduction. The purpose of this paper is not to evaluate the different decisions in a conclusive manner but simply to name them and to list their advantages and their disadvantages.

2. A Model of the Models of Interdisciplinary Relations

By tracing the basic decisions that bring about models of interdisciplinary relations, a kind of "model of models" of these relations is constituted. The most basic decision within such a model is whether there are irreducibly many disciplines or not. Only if we answer this question positively, we face the problem of interdisciplinary relations in a strict sense, because only then there are – and forever will be – different disciplines which can relate. But do they really relate? This is the next basic decision to be made.

If we go for a "No", we reach the realm of what can be called pluralist models. According to these models, there are many disciplines, at least many types of disciplines, but there are no relations between them. This is the classic "solution" to the problem of interdisciplinarity which prevailed until the second half of the 20th century, e.g. as the separation between the "hard" sciences and the disciplines of the humanities. Such models succeed in describing the demarcation between disciplines, but they do this at the price of an equivocal concept of science. They are also, from their very foundations, unable to explain the real cooperation which is going on between disciplines of different types (Fauser 2003).

If we say, yes, there are relations between different disciplines, we choose contact models. The next question is then: What kind of contact is there between the different disciplines? How is this contact mediated? In the literature, three alternatives can be found: Contact is mediated either by common objects or by common methods or by cooperation. Accordingly, we can distinguish between object-contact models, method-contact models and cooperation-contact models.

Object-contact models are the "classical" model of interdisciplinary relations. It implies that different disciplines are linked by identical objects to which every single discipline has its own access, mediated by its own method. The contact which is supposed to be mediated this way can come about in two different forms: a hierarchical form in which one central discipline has a privileged access to the objects in question, as physics does in the model of a non-reductive naturalism (Schurz 2006, 38); or a non-hierarchical form in which the several disciplines form a cluster around their objects (Mc Cormick 2003). In both cases, object-contact models are hard to integrate into a post-Kuhnian philosophy of science which takes it for granted that science, at last in some cases, does not access but create its objects so that objects are not prior to disciplines and therefore cannot guarantee interdisciplinary contact.

Method-contact models have been popular in the second half of the 20th century when there was hope for one method to bring together all disciplines. This method was conceived of as a formal one describing dynamic structures; it was (and still is) called "cybernetics", "theory of systems" and the like. Again, there is a hierarchical (Schneider 1966) and a non-hierarchical (Meister/ Lettkemann 2004) variation of such models, depending on the decision whether there is one central discipline providing all others with its method or whether there are independent but coordinated developments of the same method in different disciplines. Again, these models run contrary to an insight of current philosophy of science: Feyerabend's remark that methods are not of huge importance for science and that it would not be desirable to give them such an importance (Feyerabend 1983).

Cooperation-contact models are a very young – and promising – brand of contact-models. They even have been developed as an alternative to models of interdisciplinarity as such (Gläser et al. 2004), but only because of the – unnecessary – assumption that these models are limited to the types discussed above. According to cooperation-contact models, interdisciplinary relations are brought about just by the cooperation between scientists from different disciplines. This cooperation is not based on common objects or common methods but precedes their discovery or creation and development. Since cooperation does not start with common criteria, it cannot be conceived of as hierarchical. Rather, it is an action which implies mutual recognition – notwithstanding the fact that, as a human action, it is also coined by political, social and other conditions (Bordieu 1988; Münch 2007). Cooperation-contact models have the advantage of working without the presuppositions found in object- and method-contact models. They also fit in with the trend to understand science as action (Gläser et al. 2004). Obviously, they have little normative power. In contrast to their "object" and "method" colleagues, they do not say how disciplines are supposed to relate, but this can turn out to be a strenghth rather than a weakness.

So far we have examined the “Yes”-branch of the model of models of interdisciplinary relations. But if we have to consider theory reduction, it obviously is to be found on the other side. The basic decision to be made, then, is that there are not irreducibly many disciplines. If we decide this way, we do not face a problem of interdisciplinary relations but rather the problem how to make the pseudo-problem of interdisciplinary go away by making all disciplines collapse into only one. So we are on the side of models which can be named as “monist”.

The advantage of monist models is that they guarantee – or at least claim to guarantee – a single, univocal concept of science, based on the promised unity of science. At the same time, such models somehow have to deal with the (in their view apparent) plurality of disciplines which even is increasing evermore (Poser 2001, 279-287). Hence, monist models are challenged by the question: If there is only one discipline, can the single members of the apparent plurality of disciplines be in some way identified with that one and only discipline? The answer “No” leads to eliminative models, because given monist presuppositions non-identity with the one and only discipline just means being no scientific discipline at all. To eliminate here means to demonstrate that the kind of objects with which a pseudo-discipline claims to deal simply do not exist and that therefore the terminology used by that pseudo-discipline is meaningless. This strategy can be – and has been – successful in single cases, as e.g. in the elimination of astrology from the realm of the sciences. The recent relevant discussion is focusing on the question whether disciplines of cognitive science can be eliminated in favor of neurobiology and in the final analysis of physics (e.g., as a classical attempt, Churchland 1986). As an overall strategy for tackling the problem of interdisciplinarity it is not very popular, though, because it flies in the face of the intuition that there are many disciplines which at least have a partial and temporal justification (Charpa 1996, 96).

Therefore the most promising answer in the monist branch seems to be “Yes”: At least some members of the apparent plurality of disciplines can be identified with the one and only discipline and, through this identification, are also justified. This is the strategy of theory reduction which, as such, but without this context, is well researched in the philosophy of science. Theory reduction can come along in various kinds, depending on which discipline one takes to be the goal of reduction. In our time, the most popular version is physicalist theory reduction (Wilson 1998); but there also is its sociological counterpart (Luhmann 1990), and the list could be continued. The final goal here, too, as in elimination always is to end up with just one scientific discipline, but before the goal is reached, the different existing disciplines at least can be tolerated since their differences from the one and only science are only apparent ones. Reductive models face similar problems as eliminative ones: They also do not seem to do justice to the given plurality of disciplines (Margolis 1987; Rosenberg 1994). Nevertheless, this plurality is just a fact and facts can change. The hard problem of theory reduction, in my view, seems to lie elsewhere, and can be found by a look at the whole model of models of interdisciplinary relations.

3. The Hard Problem of Theory Reduction

The hard problem of theory reduction can be seen in its contrast to the cooperation-contact models which are the most important plural models: Contact-models, as has been shown, imply mutual recognition between the cooperating disciplines. This recognition is withdrawn by monist models. Eliminative models do so immediately, which makes them so little attractive. Reductive models are more cautious in this respect, they even promise to give a special discipline the dignity of the one and only discipline in the way of identification. But this identification is a one-way affair. The identity of the goal-discipline of reduction is supposed to be unchanging and well-known; the identity of the discipline which is to be reduced just is an apparent one; it has been falsely taken to be something apart from the one and only science. So, in the recognition of a theory which is to be reduced, the goal-theory of reduction simply recognizes itself in a disguise which soon is to be removed. However, as Hegel has shown throughout his *Phenomenology of Spirit*, recognition from its very concept always must be mutual; it presupposes two parties recognizing one another. This problem is getting even harder as we tend to take for real only what science tells us to be real (Quine 1979). So, if there is only one scientific discipline, no one can recognize it as such, neither from the outside – for only science has the authority to do so – nor from the inside – for there can be no mutuality here. The hard problem of theory reduction, at least as a global strategy facing the problem of interdisciplinarity, therefore is: If it is successful, it leads to a situation in which the supposed one and only discipline can get no recognition at all. Hence, contact models, and especially cooperation-contact models do not only seem to be a better description of the reality of science in our days; they also seem to be a better way to deal with interdisciplinarity without endangering the whole concept of science as such.

Literature

Bourdieu, Pierre 1988: Homo academicus, Frankfurt am Main

Carrier, Martin 2006: Wissenschaftstheorie zur Einführung. Hamburg

Chalmers, Alan F. ^{5}2001: Wege der Wissenschaft. Einführung in die Wissenschaftstheorie. Berlin etc.

Charpa Ulrich 1996: Grundprobleme der Wissenschaftsphilosophie. Paderborn etc.

Churchland, Paul M. 1986: Neurophilosophy. Toward a Unified Science of the Mind-Brain. Cambridge, Mass. etc.

Hacking, Ian 1996: Einführung in die Philosophie der Naturwissenschaften. Stuttgart

Fauser, Markus 2003: Einführung in die Kulturwissenschaft. Darmstadt

Feyerabend, Paul 1983: Wider den Methodenzwang. Frankfurt am Main

Gläser, Jochen et al.: “Einleitung: Heterogene Kooperation”, in: Jörg Strübing et al. (eds.), Kooperation im Niemandsland. Neue Perspektiven auf Zusammenarbeit in Wissenschaft und Technik, Opladen 2004, pp. 7-24

Luhmann, Niklas 1990: Die Wissenschaft der Gesellschaft. Frankfurt am Main

Margolis, Joseph 1987: The Persistence of Reality. Science Without Unity. Reconciling the Human and Natural Sciences. Oxford-New York

Mc Cormick, Michael 2003: “Rats, Communication, and the Plague: Toward an Ecological History”, The Journal of Interdisciplinary History 34/1, 1-25

Meister, Martin / Lettkemann, Eric 2004: “Vom Flugabwehrgeschütz zum niedlichen Roboter. Zum Wandel des Kooperation stiftenden Universalismus der Kybernetik”, in: Jörg Strübing et al. (eds.), Kooperation im Niemandsland. Neue Perspektiven auf Zusammenarbeit in Wissenschaft und Technik, Opladen 2004, pp. 105-135

Münch, Richard 2007: Die akademische Elite. Zur sozialen Konstruktion wissenschaftlicher Exzellenz, Frankfurt am Main

Poser, Hans 2001: Wissenschaftstheorie. Eine philosophische Einführung. Stuttgart

Quine, Willard van Orman 1979: Von einem logischen Standpunkt. Frankfurt etc.

Rosenberg, Alexander 1994: Instrumental Biology or the Disunity of Science. Chicago-London

Schneider, Peter K. 1966: Die Begründung der Wissenschaften durch Philosophie und Kybernetik. Idee, Umriß und Grundprinzip einer axiomatischen Strukturtheorie Stuttgart etc.

Schurz, Gerd 2006: Einführung in die Wissenschaftstheorie. Darmstadt

Wilson, Edward O. 1998: Die Einheit des Wissens. Berlin

Ethik als irreduzibles Supervenienzphänomen

Thomas Wachtendorf, Oldenburg, Deutschland

1. Der Ursprung der Ethik

Die Frage nach der Irreduzibilität der Ethik setzt allererst die Klärung der Frage voraus, in welchem Sinne hier von *Ethik* die Rede ist. Erst danach kann anhand einer anzugebenden Definition von Supervenienz die eventuelle Nichtreduzierbarkeit von Ethik diskutiert werden.

Wittgenstein, dessen Einlassungen zur Ethik recht marginal sind, bestimmt deren Ursprung dennoch: "[Ethik] ist ein Zeugnis eines Drangs im menschlichen Bewußtsein" [VüE: 19]. Dieser Drang erscheint als ein Antrieb, der den Menschen werten lässt: "*Ich* habe die Welt zu beurteilen, die Dinge zu messen." [Tb: 2.9.1916] Zwar bleibt, was dieses wertende Ich genau ausmacht, unklar: "Das Ich, das Ich ist das tief Geheimnisvolle.“ [Tb: 5.8.16] Allerdings ist an dieser Stelle auch keine diesbezügliche Erörterung erforderlich. Genau so wenig wie eine Erklärung des Ursprungs dieses Dranges möglich ist. Es bleibt lediglich zu konstatieren, dass dieser Drang im Menschen vorhanden ist. Somit bleibt als Ursprung der Ethik ein nicht weiter erklärbarer Drang im Menschen, die Welt zu beurteilen oder zu werten. So wertet der Mensch seine Umwelt, zu der auch die Handlungen und das Verhalten der anderen gehören. Vorsprachlich wird sich dies Verhalten etwa in einer ablehnenden oder zustimmenden Reaktion ausdrücken, sprachlich als Äußerung, was selber bereits eine Handlung darstellt. Ethik ist folglich nichts, was eigenständig in der Welt existiert, ein Sachverhalt etwa (So in VüE: 13 f: "Das gleiche gilt für das *absolut Gute*; wäre es ein beschreibbarer Sachverhalt, müßte ihn jeder – unabhängig von seinen eigenen Vorlieben und Neigungen – *notwendig* herbeiführen oder sich schuldig fühlen, weil er ihn nicht herbeiführt. Ein solcher Sachverhalt, möchte ich behaupten, ist ein Hirngespinst."), sondern sie ist an wertende Handlungen gebunden (vgl.: "Die Bedeutung des Wortes »gut« ist an die Handlung, die es qualifiziert, gefesselt“ [V: 191]). Diese werden von Subjekten vollzogen und also gilt: "gut und böse [sind] Prädikate des Subjekts, nicht Eigenschaften in der Welt." [Tb: 2.8.16] Grundlage der Ethik sind also die wertenden Tätigkeiten der Sprachspielenden.

2. Ethik ist sprachvermittelt und entsteht erst durch Sprache

Der im Menschen angelegte Drang zu werten hat verschiedene Möglichkeiten der Äußerung: nichtsprachlich als bloß tätliche Reaktion, sprachlich als Äußerung.

Für die Ethik und die Verständigung gelten jedoch dieselben Muster, die Wittgenstein in der Spätphilosophie aufzeigt. Zur Verständigung bedarf es für die Verwendung von Wörtern gemeinsamer Regeln, denen auch gefolgt wird. Diese Regeln entstehen in Sprachspielen, wie am Beginn der *Philosophischen Untersuchungen* vorgeführt wird. Das besondere des Sprachspiels ist es, dass in ihm die Sprache und die "Tätigkeiten mit denen sie verwoben sind" [PU: §7] als zusammengehörig aufgefasst werden. So kann aus nichtsprachlichem Verhalten durch regelhafte Korrelation mit Lauten (sprachliches) Handeln entstehen. Durch die Tätigkeiten der Menschen und die je spezifische Art, mit der sie diese durchführen, bekommen die Wörter durch ihren mit den Tätigkeiten verbundenen Gebrauch in der Sprache ihre spezifische Bedeutung [PU: §43]. Wichtig sind in diesem Zusammenhang zwei Aspekte der Sprachspielkonzeption: Die Sprachspielenden bringen sich erstens gegenseitig durch Bestätigung oder Sanktion einzelner Tätigkeiten die geltenden Regeln bei (sie *richten sich darauf ab* [Z: Nr. 419]). Das entzieht ihnen jedoch zweitens zugleich den alleinigen Zugriff auf die geltenden Regeln. Es entsteht ein Wechselverhältnis zwischen dem Unterworfensein des einzelnen Sprachspielenden unter die Regeln auf der einen und andererseits seiner Möglichkeit zu deren Veränderung und Beeinflussung auf der anderen Seite.

An dieser Stelle ist die strukturelle Ähnlichkeit des Funktionierens von Sprache mit dem Funktionieren von Ethik nicht zu übersehen. Genau wie bei jeder anderen sprachlichen Regel betont Wittgenstein, dass es immer eine Sprachspielgemeinschaft geben müsse, damit – in diesem Falle ethische – Regeln ihre bindende Kraft erhalten: "Ein Soll hat also nur Sinn, wenn hinter dem Soll etwas steht, das ihm Nachdruck gibt – eine Macht, die straft und belohnt. Ein Soll an sich ist unsinnig" [WWK: 118]. Diese Macht ist zweigestalt: einerseits besteht sie zweifelsohne aus den Mitgliedern der Sprachspielgemeinschaft, die die Einhaltung der (sprachlichen) Regeln überwachen. Andererseits ist damit aber auch das eigene Gewissen gemeint: "Wenn mein Gewissen mich aus dem Gleichgewicht bringt, so bin ich nicht in Übereinstimmung mit Etwas. Aber was ist dies? Ist es die Welt? [...] Zum Beispiel: es macht mich unglücklich zu denken, daß ich den und den beleidigt habe. Ist das mein Gewissen?“ [Tb: 8.7.1916] Die Motivation, sich gemäß bestimmter ethischer Regeln zu verhalten, hat so einmal einen äußeren, einmal einen inneren Grund. In beiden Fällen ist gleichwohl schon vorausgesetzt, dass es bereits in einer Sprachspielgemeinschaft entstandene Regeln geben muss, deren Nichtbefolgung überhaupt als ethisch schlecht – und damit sanktionsfähig – empfunden werden kann.

Es bleibt, dass sich Ethik nur in einer Gemeinschaft von Sprachspielenden entfalten kann, wo und indem sie durch ihre sprachliche Form den Sprachspielenden erst bewusst werden kann. Denn analog zum Privatsprachenargument kann es keine privaten, sondern bloß öffentliche ethische Begriffe geben. Folglich gibt es nur eine öffentliche Ethik. Dies folgt außerdem aus den Überlegungen zum privaten Regelfolgen, wonach einer allein die Richtigkeit seiner eigenen Regelbefolgung nicht garantieren kann. Jeder hat ja bloß seine Erinnerung an ein vormaliges Regelfolgen, die er nun als Vorbild für ein erneutes Folgen verwendet. Eine Erinnerung ist jedoch immer für Täuschungen anfällig [vgl.: PU: § 265, Wachtendorf 2008: 220 f].

Genau wie die übrigen Regeln des Sprachspiels sind auch die ethischen Regeln als ein Produkt der Sprachspielgemeinschaft durch den Prozess des gemeinsamen Sprechens und Tätigseins – also beim gemeinsamen Lebensvollzug – entstanden. Genau wie auf die übrigen Regeln richten sich die Sprecher gegenseitig auf die ethischen Regeln ab. Jemand handelt und aus der billigenden oder ablehnenden Reaktion der anderen ergibt sich die notwendige Rückkoppelung, die den einzelnen jeweils in seinem Tun bestärkt oder nicht. So steht hinter dem *Soll* die erforderliche *strafende Macht* und so ist gleichzeitig die Objektivität der ethischen Regeln innerhalb der jeweiligen Sprachspielgemeinschaft sicher gestellt.

Ist also der Drang zu werten zwar im Menschen angelegt, kann er sich über das Maß eines bloß reaktiven Verhaltens hinaus erst entfalten, wenn der Mensch über eine Sprache verfügt – es entsteht die Ethik. Wenn man von Ethik in diesem Sinne spricht und damit also nicht bloß den immanenten Drang zur Wertung sondern etwas wie eine ethische Theorie meint, spricht man deshalb immer über ein Phänomen, das erst durch eine Sprache möglich wird.

Wittgensteins Spätphilosophie stellt ein sehr gutes Konzept zur Erläuterung der Funktionsweise von Ethik dar, weil sie lediglich von der Faktizität bestimmter menschlicher Eigenschaften ausgeht (sie setzt beispielsweise bloß die Sprachfähigkeit voraus) und in diesem Rahmen die Möglichkeit bietet, Ethik zu analysieren, ohne (hypostasierende) Annahmen wie etwa diejenige eines eigenständigen ethischen Reichs oder die Existenz ethischer Tatsachen machen zu müssen. Da man zudem zeigen kann [Wachtendorf: 163 ff], dass Wittgensteins frühe Vorstellung von Ethik auch noch in seiner Spätphilosophie besteht und deshalb mit dieser kompatibel ist, bietet sich die Spätphilosophie für ethische Betrachtungen auch in Wittgensteins Sinne geradezu an.

3. Ethische Sätze

Ethik findet ihren Ausdruck in Sätzen, denen die Sprachspielgemeinschaft zustimmt oder die sie ablehnt. So entsteht eine Klasse von ethischen Sätzen. Ethische Sätze sind beispielsweise: "[...] 'Du sollst das tun!' oder 'Das ist gut!' aber nicht 'Diese Menschen sagen das sei gut'. Ein ethischer Satz ist aber eine persönliche Handlung. Keine Konstatierung einer Tatsache. Wie ein Ausruf der Bewunderung. Bedenke doch daß die Begründung des 'ethischen Satzes' nur versucht den Satz auf andere zurückzuführen die Dir einen Eindruck machen. Hast Du am Schluß keinen Abscheu vor diesem keine Bewunderung für jenes so gibt es keine Begründung die diesen Namen verdiente." [DB: 43 f] Hier ist sehr schön die Eigentümlichkeit von Ethik, die sich auch in den ethischen Sätzen niederschlägt, zusammengefasst: Es geht immer um Wertungen und darum, *Abscheu* oder *Bewunderung* zu erzeugen. Da das Äußern eines ethischen Satzes eine Handlung ist, ist die Ethik selbst eine Praxis, bestehend aus Wertungen und dem Äußern von Werturteilen.

Den ethischen Sätzen eignet eine Besonderheit: sie gehören ähnlich wie die grammatischen Sätze zu den selbstverständlichen und unhinterfragten Grundlagen aller Tätigkeiten [Wachtendorf 2008: 184 ff]. Da man im Allgemeinen Regeln blind folgt, folgt man auch ethischen Sätzen blind. Das heißt, dass man auch solchen Regeln folgt, die in dem jeweiligen Sprachspiel in Kraft sind, obwohl man ihnen nicht explizit, sondern nur durch Teilhabe zugestimmt hat. Das eigene Verhalten ist also immer auch von ethischen Regeln geleitet, die im jeweiligen Sprachspiel Geltung haben. Es ist aus sprachlogischen Gründen gar nicht möglich, auf alle basalen Regeln zu reflektieren. Demgegenüber ist dies jedoch oftmals ein Kriterium von ethischem Verhalten: "Nur solches Verhalten unterliegt einer moralischen Billigung oder Missbilligung, das der jeweilige Akteur hätte vermeiden können und das er deshalb verantworten muss." [Birnbacher 22007: 15] Ein Kriterium für das Vermeiden-Können ist laut Birnbacher "größere Vorsicht" [Birnbacher 22007: 15]. Kann man aber überhaupt vorsichtig genug sein? Zweifelsohne gilt das für dezidiert ethische Fragen, die man in einer konkreten Situation bewusst zu lösen versucht. Hier kann man etwas anders machen, weil man die anderen Möglichkeiten kennt. Die bereits unbewusst in die ethische Überlegung eingehenden, grundlegenden Regeln sind jedoch ihrerseits einer Reflexion entzogen (oder nur sehr schwer zugänglich). So verstanden bekommt die Klasse der ethischen Sätze eine die Tätigkeiten der einzelnen unbewusst beeinflussende Funktion, wobei sie gleichzeitig dem aktiven direkten Zugriff eines einzelnen entzogen ist.

4. Supervenienz und Reduzierbarkeit

Es ist nun unschwer zu erkennen, dass die Klasse der ethischen Sätze (fortan E), wie sie bis hierhin dargestellt worden ist, in einem Supervenienzverhältis zu den Sprachspielenden steht. Klassisch wird Supervenienz definiert als: eine Eigenschaftsfamilie A superveniert über einer Eigenschaftsfamilie B, genau dann wenn man A nicht verändern kann, ohne B zu verändern. Es ist offensichtlich, dass diese Definition hier erfüllt ist, weil E keiner Änderung unterliegt, sofern nicht eine Änderung im Handeln der Sprachspielenden eintritt.

Somit ist die Ethik (selbstverständlich) abhängig von den Sprachspielenden; aber nicht von jedem einzelnen, sondern von allen in ihrer Gesamtheit. Sie ist das Ergebnis der Interaktion der Sprachspielenden untereinander und der Interdependenz zwischen diesen und den Sätzen aus E. Die einzelnen Sätze aus E haben unterschiedliche Wirkung auf die verschiedenen Sprachspielenden. Die Wechselwirkung ist *hyperkomplex*, das heißt, es gibt keine eindeutige reversible Relation zwischen Sätzen und Handlungen, sondern die Interdependenzen der einzelnen Sätze untereinander und ihre jeweilige Wirkung auf die Sprachspielenden sind derart komplex, dass überhaupt keine Zuordnung möglich ist. Denn jede Handlung kann auf unterschiedliche Weise beschrieben werden und für jeden ethischen Satz gibt es verschiedene Möglichkeiten der Befolgung. Letzteres hängt insbesondere auch von den Sitten und Gebräuchen der jeweiligen Sprachspielgemeinschaft oder von den besonderen Eigenschaften des Einzelnen ab. Außerdem kann ein ethischer Satz Konsequenzen für andere haben, die sich beispielsweise erst in der konkreten Praxis seiner Anwendung zeigen, wodurch er eine praktische Wirkung entfaltet, die wiederum sofort Konsequenzen für die Sätze der Klasse E hat. Wittgenstein deutet dieses Verhältnis in folgenden Bemerkungen an: "Was man eine Änderung in den Begriffen nennt, ist natürlich nicht nur eine Änderung im Reden, sondern auch eine im Tun." [BPP: I-910] Und "Ich will sagen: eine ganz andere Erziehung, als die unsere, könnte auch die Grundlage ganz anderer Begriffe sein." [BPP: II-707] Klarer kann man die enge Verwobenheit von Handlungen und Sprache nicht machen. In dieser Konstruktion kommt die Präreflexivität grammatischer und damit auch ethischer Sätze als Mythologie [vgl.: VüE: 38] deutlich zum Ausdruck. Erst auf der Basis einer zumindest rudimentären Klasse E kann anschließend auf einzelne ethische Sätze aus E reflektiert werden.

Gemäß dieser Darstellung ist zugleich klar, dass die Sprachspielenden nicht als kollektiver Akteur im herkömmlichen Sinne aufgefasst werden sollten. Die Frage ist hier nicht, ob die Gemeinschaft der Sprachspielenden eine eigenständige Entität ist, deren Vollzüge – sofern sie tatsächlich welche hat – zusätzlich und unabhängig von denen ihrer Mitglieder ethisch bewertet werden können, so wie etwa bei einer spontanen Demonstration sowohl das Verhalten der Demonstration insgesamt (etwa *friedlich* oder *gewaltbereit*) als auch die Handlungen der einzelnen Demonstranten jeweils eigenständig moralisch bewertet werden können. Die Sprachspielenden handeln nicht kollektiv, sondern lediglich gemeinsam. Demnach geht jeder Sprachspielende seinen eigenen Interessen nach, muss dabei jedoch zwangsläufig den gemeinsamen Regeln folgen und übt dadurch Einfluss auf diese aus. Spieltheoretisch betrachtet bekommt dadurch das Handeln der einzelnen zwar durchaus kollektive Züge. Dies ist für den hier verfolgter Zweck allerdings nicht relevant.

Aus der obigen Darstellung der Entstehung der supervenienten Klasse E folgt zugleich eine Antwort auf die Frage nach einer möglichen Reduktion der Ethik. Versteht man als Reduktionismus, dass ein System durch seine Einzelbestandteile vollständig bestimmt ist, kann davon im Zusammenhang mit Ethik keine Rede sein. Eine Reduktion ist letztlich immer entweder die Rückführung einer Theorie auf Beobachtungssätze oder von Begriffen auf Dinge oder von (mentalen) Zuständen und Zusammenhängen auf kausal-deterministische Ereignisse. In allen drei Fällen wird ein wie auch immer gearteter, (quasi-) naturwissenschaftlicher Atomismus unterstellt. Demgemäß gibt es letzte Entitäten, auf die alle von ihnen verschiedene reduziert werden können, weil sie bloß als Konfigurationen oder Beschreibungen dieser Grundentitäten verstanden werden. Im Falle der Ethik kann man entweder versuchen, die Sätze der Klasse E und die Handlungen der Sprachspielenden auf Beobachtungssätze zurückzuführen, oder man betrachtet die ethischen Sätze nur als sprachliche Darstellung mentaler Zustände. Beide Wege sind nicht gangbar. Zwar lässt sich bei der Gemeinschaft der Sprachspielenden bis zu einem gewissen Grad mit Beobachtungssätzen arbeiten. Im Falle von einzelnen Sprachspielenden ist das aufgrund der Hyperkomplexität der Ethik (die ja auch die Intentionalität einschließt) nicht möglich. Ähnlich wie beim berühmten Gavagai-Beispiel von Quine ist niemals klar, ob der im Beobachtungssatz ausgedrückte Sachverhalt wirklich dem beobachteten entspricht.

Auch der zweite Weg ist nicht zielführend: Vielleicht ist es möglich, bestimmte Werthaltungen eines konkreten Sprachspielenden mit bestimmten seiner Zustände (beispielsweise neurologisch oder psychologisch) zu identifizieren. Aber selbst wenn es gelänge, dies für alle Beteiligten zu tun, entsteht doch aus den ethischen Werthaltungen aller die supervenierende, von den einzelnen unabhängige Klasse E. Da die ethischen Werthaltungen und die Handlungen eines einzelnen von dieser Klasse interdependent abhängen, ist hier wegen der fehlenden Möglichkeit, eine eineindeutigen Zuordnung vornehmen zu können, keine Reduktion möglich. Man müsste eine vollständige Determination unterstellen, um das zu erreichen. Dagegen greifen die einschlägigen Argumente gegen einen solchen holistischen Standpunkt.

Literatur

Birnbacher, Dieter ^{2}2007 *Analytische Einführung in die Ethik*, Berlin: de Gruyter.

Wachtendorf, Thomas 2008 *Ethik als Mythologie. Sprache und Ethik bei Ludwig Wittgenstein*, Berlin: Parerga.

Wittgenstein, Ludwig ^{2}2000 *Vorlesungen 1930-1935* [V], Frankfurt/M.: Suhrkamp.

Wittgenstein, Ludwig ^{4}1999 *Vortrag über Ethik und andere kleine Schriften* [VüE], Frankfurt/M.: Suhrkamp.

Wittgenstein, Ludwig ^{12}1999 *Werkausgabe: Tractatus logico-philosophicus* [PU, TLP, Tb], Band 1, Frankfurt/M.: Suhrkamp.

Wittgenstein, Ludwig ^{9}2002 *Werkausgabe: Bemerkungen über die Farben, Über Gewißheit, Zettel Vermischte Bemerkungen* [Z], Band 8, Frankfurt/M.: Suhrkamp.

Wittgenstein, Ludwig ^{6}2001 *Werkausgabe: Wittgenstein und der Wiener Kreis* [WWK], Band 3, Frankfurt/M.: Suhrkamp.

Wittgenstein, Ludwig ^{7}1999 *Werkausgabe: Bemerkungen über die Philosophie der Psychologie* [BPP], Band 7, Frankfurt/M.: Suhrkamp.

Das ‘schwierige Problem’ des Bewusstseins – oder *wie es ist*, Person *zu sein*

Patricia M. Wallusch, Frankfurt am Main, Deutschland

1. Zwei Arten von Bewusstsein und ein altes Problem

Beckermann unterscheidet im ersten Kapitel seiner *Analytischen Einführung in die Philosophie des Geistes* fünf charakteristische Merkmale des Mentalen: *Bewusstheit*, *Unkorrigierbarkeit*, *Intentionalität*, *Nicht-Räumlichkeit*, sowie *Privatheit*. Mit Blick auf das Merkmal der *Bewusstheit* bemerkt er weiterhin, dass es sich um zumindest zwei voneinander verschiedene Merkmale handeln könne. Zum einen könne es bedeuten, dass eine Person, die sich in einem mentalen Zustand *M* befindet, auch *weiß*, dass sie in *M* ist. Er schließt mit Blick auf Freuds ´Entdeckung des Unbewussten´ aus, dass es sich – so verstanden – um ein Merkmal handelt, das sich auf alle mentalen Zustände erstreckt. Als zweites Charakteristikum mentaler Zustände führt er ihren *phänomenalen Charakter* an, d.h. die Tatsache, dass es sich für eine Person, die sich in einem mentalen Zustand *M* befindet, *irgendwie anfühlt* in *M* zu sein. Beckermann betrachtet es als ´zumindest zweifelhaft´, dass dies für alle mentalen Zustände gilt (Beckermann ^{2}2001, 9).

Gerade das letztgenannte Merkmal ist meiner Ansicht nach ein fundamentales Charakteristikum des Mentalen. Täglich befinden wir uns in unterschiedlichsten bewussten Zuständen, die diesen Charakter besitzen, haben wir Absichten und Wünsche, dringen auf uns Eindrücke verschiedenster Art ein, erleben wir die Welt als unsere Welt, – die urtümlichste Art des Bewusstseins ist die, sich des eigenen Lebens bewusst zu sein: nur ich *lebe* mein Leben, *mein Leben fühlt sich (nur) für mich auf eine ganz spezifische Weise an*. Sich seines eigenen Lebens und *Erlebens* bewusst zu sein, sehe ich als fundamentalstes Charakteristikum an, das mein – und ich unterstelle, das jedes anderen Menschen - geistiges Leben auszeichnet, – nämlich das es *irgendwie ist* es als solches zu erleben. Auch das Unbewusste im Sinne Freuds ist ein Teil dessen, was uns prinzipiell bewusst ist, weil es zu jeder Zeit (wieder) bewusst werden kann: der un- oder unterbewusste Teil unserer Selbst ist lediglich nicht aktuiert, kann aber durch einen Akt der (Wieder-)Erinnerung aktuiert werden, sei er bewusst gesetzt oder durch äußere Einflüsse ausgelöst. Was also ist das Problem? Warum scheint es so schwierig, jenen fundamentalen Bestandteil unserer Existenz zu erklären? Warum drängt es uns überhaupt dahingehend, ihn erklären zu müssen?

Seit Descartes’ Trennung von Körper und Geist verstand die Philosophie es als eine ihrer Aufgaben, diese Trennung zu bewahren oder ihre Verfehltheit nachzuweisen, indem sie Bemühungen um plausiblere Alternativen anstellte. Die Absicht, jene Trennung zu wahren, verfolgten seit jeher solche, die sich – in Anlehnung an die Cartesianische Unterscheidung – als ´Dualisten´ bezeichnen, ihre Verfehltheit nachzuweisen war das erklärte Ziel von Vertretern monistischer Positionen, die entweder in idealistischer oder materialistischer Ausprägung auftraten.

Der Dualismus hat zwar bis heute noch Bestand, doch er wird nur noch von ganz wenigen Philosophen vertreten; der Materialismus, dessen moderne Bezeichnung ´Physikalismus´ ist, hingegen erfreut sich einer großen Anhängerschaft. Dass der materialistische Ansatz zunehmend an Einfluss gewinnen konnte, lag sicherlich am zunehmenden Erfolg, den die empirischen Wissenschaften hinsichtlich der Erklärung der Phänomene der natürlichen Welt erzielten, und daran, dass eine philosophische Theorie, deren Erklärungsgrundlage also in den Gesetzmäßigkeiten einer empirisch ergründbaren Natur liegt, im Gegensatz zu anderen Ansätzen mit der Annahme einer durch Naturgesetze beschreibbaren und durch Gesetzmäßigkeiten determinierten Welt eher kompatibel ist. Dass der Physikalismus heute die einflussreichste und meistdiskutierte Position im Rahmen der Philosophie des Geistes (und nicht nur dort) ist, erklärt sich also aus dem Umstand, dass in ein solches von Wissenschaftlichkeit, Empirie und Rationalität beherrschtes Weltbild eine geistige Substanz nicht passt, – ein dualistischer Zugang setzt sie aber voraus, insofern er sich nicht nur auf Aussagen über Eigenschaften beschränkt.

Die Annahme einer ´res cogitans´ erschien dem immer aufgeklärteren Menschen zu mystisch, – zu sehr in einem religiösen Weltbild verhaftet, das im Hinblick auf seine Erklärungskraft als immer unzureichender und immer weniger adäquat erschien. Recht bald wurde das Projekt programmatisch, die mentalen Phänomene nach naturwissenschaftlichem Vorbild zu erklären, um sie so mit dem Bild einer kausal geschlossenen und determinierten Welt kompatibel zu machen, – sie in die natürlichen Vorgänge aufzulösen, aus denen sie hervorgehen. Die Position, die sich im Gefolge dieser Entwicklung herausbildete, war die des Eliminativismus. Charakteristisch für diese Position war sowohl die Leugnung einer geistigen Substanz, als auch das Bemühen darum, sämtliche geistige Eigenschaften wegzuerklären. Doch stellte sich heraus, dass es mit ihrer Hilfe nicht möglich war, alle mentalen Phänomene vollständig in physikalische Prozesse aufzulösen, sodass die Position des eliminativen Physikalismus bald von einer ihm gegenüber gemäßigteren, reduktionistisch ausgerichteten Variante abgelöst wurde. Sein Bestreben richtete sich darauf, mentale Phänomene auf die physikalischen Prozesse zu reduzieren, die ihnen zugrunde zu liegen scheinen.

Allerdings geriet auch dieser Ansatz in den vergangenen zwei Jahrzehnten zunehmend in eine Sackgasse: zwar gelang es mit seiner Hilfe tatsächlich, einen großen Teil mentaler Phänomene zu erklären, doch das phänomenale Bewusstsein gehörte zu der Art von Phänomenen, die sich einer Reduktion widersetzten. Die Lösung muss also in einem davon verschiedenen Ansatz liegen, – doch wird es durch einen nicht-reduktiven Ansatz keinesfalls einfacher; so beklagt Kim, dass mit dem Scheitern des Reduktionismus auch die Verstehbarkeit mentaler Verursachung in unerreichbare Ferne rückt (Vgl. Kim 1996, 237). Die Lage scheint aussichtslos.

2. David Chalmers Ansatz zu einer Theorie des Bewusstseins (*consciousness*)

An diesem Punkt des Scheiterns jener Theorien, knüpft Chalmers an und versucht einen nicht-reduktiven Ausweg zu weisen. Er diagnostiziert das Problem, das bisherige physikalistische Ansätze bei der Erklärung des Bewusstseins hatten, zum einen darin, dass sie reduktionistisch ausgerichtet waren und dass es sich bei den erklärten Merkmalen um die ´leichten´ Probleme handelte, – dass also entgegen dem Anspruch, den diese Theorien hinsichtlich ihrer explanativen Kraft hegten, das wirklich ´schwere´ Problem von ihnen nicht berührt wurde.

Die ´leichten´ Probleme umfassen „the ability to discriminate, categorize, and react to environmental stimuli; the integration of information by a cognitive system; the reportability of mental states; the ability of a system to access its own internal states; the focus of attention; the deliberate control of behavior; the difference between wakefulness and sleep.” (Chalmers 1995, 200) Diese bilden den Bereich des Bewusstseins im Sinn von *´awareness´*. Ihnen sei mit einem reduktionistischen Ansatz leicht beizukommen, da sie mittels computationaler oder neuronaler Mechanismen prinzipiell erklärbar seien, – auch wenn dies noch eine längere Phase schwieriger empirischer Forschung durch die Kognitions- und Neurowissenschaften bedeute. „If these phenomena were all there was to consciousness, then consciousness would not be much of a problem. Although we do not yet have anything close to a complete explanation of these phenomena, we have a clear idea of how we might go about explaining them. This is why I call these problems the easy problems.” (Chalmers 1995, 201)

Das eigentliche und schwere Problem, dem scheinbar nicht beizukommen ist, ist hingegen das der *Erfahrung*: „When we think and perceive, there is a whir of information-processing, but there is also a subjective aspect. As Nagel (1974) has put it, there is *something it is like* to be a conscious organism.” (Chalmers 1995, 201) Um diesem Problem beizukommen, konzipiert Chalmers eine Theorie, in deren Zentrum der Begriff der *Information* bzw. des *Informationsraums* steht. Diese Theorie basiert auf drei Prinzipien, zwei nicht-grundlegender und einem grundlegender Art. Das *Prinzip struktureller Kohärenz* und das *Prinzip funktionaler (´organisational´) Invarianz* sind Prinzipien nicht grundlegender Art, die es ermöglichen sollen, Entsprechungen zwischen Bewusstsein im Sinn von *awareness* und Bewusstsein im Sinn von *consciousness* sowohl auf globaler als auf struktureller Ebene herzustellen. Hinzu kommt das *Doppel-Aspekt Prinzip*, als grundlegendes Prinzip. Es besagt: „whenever we find an information space realized phenomenally, we find the same information space realized physically.“ (Chalmers 1996, 284) Dieses Prinzip gibt in Verbindung mit den erstgenannten Prinzipien Anlass zu der Hoffnung, nun endlich eine Grundlage in der Hand zu haben, mittels derer dem ´schweren´ Problem beizukommen ist. Der daraus resultierenden Theorie zufolge wird jegliche Information, die bewusst erlebt wird, zugleich physisch verkörpert. Sie basiert also auf der Annahme einer strukturellen und inhaltlichen Entsprechung zwischen diesen beiden Informationsräumen.

3. Das Problem des Bewusstseins (revisited)

Was mich bezüglich der von Chalmers vorgeschlagenen Vorgehensweise skeptisch stimmt, ist, dass in ihm dem Begriff der *Information* bzw. des *Informationsraums* eine zentrale Rolle zukommt. Mir scheint unser eigentliches Problem wiederholt nicht berührt zu sein, geht es bei dem eigentlichen Problem des Bewusstseins (*consciousness*) doch gerade um ein Phänomen, das jenseits aller bewussten Zustände mit bestimmtem (Informations-)gehalt (*awareness*) auftritt, – wie Chalmers selbst feststellte „but there is also a subjective aspect“ (Chalmers 1995, 201).

Wacome stellt einen Punkt heraus, der eine Intuition meinerseits trifft. Ihm stellt sich unser Problem folgendermaßen dar: „reducibility was conceived as a logical relation between linguistic items... but type reducibility was taken as having ontological implications.“ (Wacome 2004, 323) Ansätze des Physikalismus unterscheiden sich also lediglich hinsichtlich der Art, wie sie über die zu erklärenden Phänomene reden, – nicht hinsichtlich des Ziels, dass sie unterschiedslos auf der ontologischen Ebene verfolgen: „Physicalism, whether reductionist or nonreductionist, denies the existence of the immaterial mind.“ (Ders., 325) Doch wie könnte eine Alternative aussehen?

Kein Zweifel dürfte dahingehend bestehen, dass auch Erklärungen, die einer dualistische Intuition entstammen, nicht ignorieren können, dass auf physikalischer Ebene etwas passiert, wenn eine Person *bewusst* ist, – also ein minimaler Physikalismus als Basis vonnöten ist, um die Verstehbarkeit mentaler Verursachung prinzipiell zu ermöglichen. Dieser Aspekt ist es auch, der selbst den Neurowissenschaften noch Kopfschmerzen bereitet. Wolf Singer stellte sich in einem Vortrag, den er im März dieses Jahres hielt, der Frage „Wer regiert im Gehirn?“. Dabei betonte er, dass es zwar mehrere Zentren im Gehirn gibt (also keine Zentrale einer ´res cogitans´ wie sie Descartes noch vermutete), jedoch unabhängig von deren Aktivitäten noch eine weitere jederzeit messbare Aktivität vorhanden sei, die keinem der bisher lokalisierten Zentren zugeordnet werden kann. In dieser nicht-lokalisierbaren Aktivität könnte eben jene minimale physikalische Basis subjektiven Erlebens vorliegen, – allerdings ist die Annahme eines solchen Zusammenhangs nur eine rein spekulative Vermutung meinerseits, die plausibel erscheinen mag oder nicht, in jedem Fall aber noch der Bestätigung durch Ergebnisse interdisziplinärer Bemühungen bedürfte.

Sollte meine Vermutung jedoch nicht völlig verfehlt sein und bewusstes Erleben tatsächlich in einer nicht weiter lokalisierbaren physischen Aktivität korreliert sein, so steht dem nichts mehr entgegen, das Bewusstsein (im Sinn von *consciousness*) als eine *fundamentale Eigenschaft* anzunehmen, die sich nicht weiter erklären lässt (dies wäre nach dem Vorbild der Naturwissenschaften die klassische Gangart im Hinblick auf Phänomene, die sich über einen längeren Zeitraum der Rückführung auf einfachere Bestandteile widersetzen). Stellt sich dieser Weg als gangbar heraus, so wäre er mit neuen Perspektiven für mindestens ein weiteres Problem verbunden, nämlich der Frage danach, was Personen *wesentlich* sind. Im Zusammenhang mit dieser Frage lässt sich auch der dualistischen Intuition Rechnung tragen, die bereits angeklungen ist.

Der besondere ontologische Status des Bewusstseins, seine Nichtreduzierbarkeit, liegt laut Searle darin begründet, „because consciousness has a first-person, or subjective, ontology, and is thus not reducible to anything that has a third-person or objective ontology.“ (Searle 2007, 50) Dieser erstpersönliche Charakter zeichnet eine Äußerung wie ´Ich habe Schmerzen´ in dem Moment, in dem Schmerzen tatsächlich empfunden werden, in zweifacher Hinsicht aus. Darin wird etwas ausgedrückt, wovon nur derjenige wissen kann, der den Schmerz empfindet. Ein solcher Satz drückt Selbstwissen *de re* aus, d.h. in dem Moment der Äußerung weiß nur das äußernde Subjekt *selbst* um das in diesem Satz zum Ausdruck gebrachte Gefühl. Es ist ein Satz direkt Bezug nehmender Referenz, in dem das ´Ich´ sich als alleiniges Subjekt *dieses Schmerzes* weiß, – weiß, dass es seine Schmerzen sind (Vgl. Lowe 2006, 183f.). Diese beiden Kriterien, die *direkt Bezug nehmende Referenz* und das *Selbstwissen* müssen erfüllt sein, damit von einem *Selbst* oder einer *Person* als Subjekt der Erfahrung die Rede sein kann (Vgl. Lowe 2006, 194). Was aber ist das *Selbst* oder die *Person*?

Zwei Gründe sprechen gegen die Annahme, dass Personen mit Körpern identisch sind. Zum einen ist es vorstellbar, dass zwei Personen einen Körper haben (wie es bei siamesischen Zwillingen der Fall ist oder wie es das *Brain-Split* Argument nahelegt). Zum anderen ist es durchaus denkbar, dass ein Selbst auch dann existiert, wenn ihm die Fähigkeit sinnlicher Wahrnehmung vollständig fehlt und es somit nicht in der Lage ist, festzustellen, ob es einen Körper hat, geschweige denn, welcher seiner sein könnte. Dennoch wäre es dieser Person weiterhin möglich, um sich selbst zu wissen und auf sich selbst Bezug zu nehmen (Vgl. Lowe 2006, 184). Dass ein Selbst mit einem Körper identisch ist, darf also ausgeschlossen werden, „bodies cannot themselves *be* selves, since they could not in principle satisfy the condition of self-knowledge that selfhood entails: for even if a body could in some sense ‘know’ that a certain thought or experience ‘belonged’ to it, it could not be guaranteed (as a self is) to know that it itself was the unique subject of that thought or experience, since it could not even be guaranteed to ‘know’ that that thought or experience ‘belonged’ to *it alone*. Therefore, I know *a priori* that ‘I am this body’ is necessarily false. (The same conclusion follows if ‘bodily part’ is substituted for body throughout.)” (Lowe 2006, 196) Ebenso ausgeschlossen ist, dass eine Person identisch ist mit einem Gehirn.

Aus dem Gesagten könnte man nun durchaus zu dem Schluss gelangen, dass eine Person eine Cartesische ´res cogitans´ ist, – doch scheint er mir wenig überzeugend. Eine Person ist vielmehr eine psychologische Substanz, die psychologischen Gesetzmäßigkeiten unterliegt (Vgl. Lowe 2006, 34). Eine solche Substanz ist weit davon entfernt, immateriell zu sein. Sie ist eng verbunden mit einem Körper – durch seine Teile, derer sie sich bewusst ist, „namely, those over which it can exercise direct voluntary control and those in which it can phenomenologically localize bodily sensations“ (Lowe 2006, 13).

Literatur

Beckermann, Ansgar ^{2}2001 Analytische Einführung in die Philosophie des Geistes, Berlin: De Gruyter.

Chalmers, David 1995 Facing up to the problem of consciousness, Journal of consciousness studies 2(3), Exeter: Imprint Academic, 200-219.

Chalmers, David 1996 The conscious mind. In search of a fundamental theory, Oxford: Oxford University Press.

Kim, Jaegwon 1996 Reductive and non-reductive physicalism, in: ders. Philosophy of mind, Oxford: Westview Press, 211-240.

Lowe, E. Jonathan 2006 Subjects of experience, Cambridge: Cambridge University Press.

Searle, John R. 1984 Minds, brains and science. The Reith lectures 1984, London: British Broadcasting Corporation.

Searle, John R. 2007 Freedom and Neurobiology. Reflections on free will, language and political power, New York: Columbia University Press.

Wacome, Donald H. 2004 Reductionism’s demise: cold comfort, in: Zygon 39 (2), 321-337.

The Supervenience Argument, Levels, Orders, and Psychophysical Reductions

Sven Walter, Osnabrück, Germany

Nonreductive physicalism (NRP) dominates current discussions of the mind-body problem. According to NRP, all scientifically respectable entities which are not straightforwardly identical to physical entities are at least (asymmetrically) *dependent* upon physical entities, for instance by *supervening* upon them. Jaegwon Kim has argued for decades that NRP collapses either in epiphenomenalism, or in reductive physicalism. The punch line of his famous *Supervenience Argument* (SA) is that if mental properties indeed supervened upon physical properties without being reducible to them, then they would be causally otiose; since epiphenomenalism is absurd, mental properties must thus be reducible to physical properties.

SA is one of the most important arguments against NRP. Kim has formulated various versions since the late eighties, and in Kim (2005), he has defended it against various criticisms. The current paper assesses Kim's response to one of the most important criticisms, viz., the *Generalization Argument* according to which, if sound, SA would not only show that there is no mental causation, but also that there is no biological, no chemical, no geological causation etc.

1. The Supervenience Argument

Suppose that (an instance of; I will omit this qualification from now on) mental property *M* causes mental property *M**. Given psychophysical supervenience, there must be a physical property *P** which is (non-causally) sufficient for *M**. Why does *M** occur? Given supervenience, as long as *P** is there, *M** will be there, no matter what happened before—even if *M**'s alleged cause, *M*, had not been present (Kim 1998, 42). According to Kim, if *M* is to cause *M**, it must do so by causing *P** (Kim 2005, 40). Hence, mental-to-mental causation is possible only if mental-to-physical causation is possible; yet, it seems, the latter is possible only if mental properties are reducible to physical properties. The reason is that *P** will also have a sufficient completely physical cause *P*, since the physical world is assumed to be causally closed. But then, how can *M* cause *P**, if *P* (which is allegedly distinct from *M*) is already a sufficient cause of *P**? If *P* is a sufficient cause of *P**, then there seems nothing left for *M* to do, unless *M* is identical to *P* (barring genuine overdetermination). The alternative is thus: "reduction or causal impotence" (Kim 2005, 54). NRP is no longer a serious option.

2. The Generalization Argument

It has been argued that the argument just sketched cannot be sound since, if so, it would render all macroproperties causally impotent (Block 2003). What deprives mental properties of their causal status, according to SA, it is said, is their relationship to physical properties, viz., *supervenience without reduction*, and it seems that all macroproperties stand in this relationship to the properties below them in the micro-macro-hierarchy. Hence, if sound, SA would generalize, rendering all macroproperties causally otiose. This, Kim's critics allege, shows that it cannot be sound.

3. A *Reductio* of What?

Kim's first response is to stress that SA is intended as a *reductio*. Epiphenomenalism concerning mental properties is the absurdity that allegedly forces us to give up the irreducibility of mental properties. Hence, if this epiphenomenalism would indeed cover all macroproperties, that would only add to the force of SA because it would provide "us with one more reason to perform a *reductio* against the irreducibility premise" (Kim 2005, 69).

Yet, although one thing to dismiss as a result of the *reductio* is the irreducibility premise, another one obviously is Kim's assumption that *M* and *P* cannot *both* be causes of *M**, and from the point of view of Kim's opponents, it is *this* assumption that is reduced to absurdity.

4. Levels, Orders, and Supervenience

Kim's second response draws on a distinction between *levels* and *orders* (Kim 1998). There are, he said, two kinds of macroproperties: higher-*level* and higher-*order* properties. SA does not apply to higher-*level* properties because they do not supervene upon lower-level properties. And since most higher-*order* properties can be *reduced* to lower-order properties, SA does not apply to them either. The only macroproperties threatened are *irreducible higher-order properties*, and since phenomenal properties of conscious experience are the only properties of this kind, the *Generalization Argument* fails.

Two issues are important here: *supervenience* and *reduction*. This section tackles supervenience, section 6 reduction.

SA, Kim claimed, would apply to higher-*level* properties only if the subvenient/supervenient distinction mirrored the relation between fundamental and higher-level properties, and this is not the case. A property's *level* depends upon what object it is a property of—properties of objects with parts are higher-level properties, properties of objects with no parts are fundamental properties. Yet, since supervenience is necessarily a relation between properties of the same objects, it only generates an *intra*level hierarchy of lower- and higher-*order* properties. Higher-level properties, in contrast, are *structural* or *microbased* properties of the form $R(P_1o_1, \ldots, P_no_n)$ which do not supervene upon the properties $P_1, \ldots, P_n$, and the relation *R* that make up their microbase. Therefore, SA does not apply to them (Kim 2005, 57).

In an earlier paper, Kim himself characterized a relation between properties of objects in domains D_1 and D_2 that are coordinated by a mapping relation *R* such that for each object *x* in D_1, *R*/*x* is the image of *x* in D_2 (Kim 1988, 124). However, if *R* is the part/whole relation, his characterization amounts to an *inter*level notion of *mereological supervenience* between the properties of wholes and those of their parts. The result is that SA would apply to higher-level properties, too.

5. Determination

What prevents a microbased property *P* from being causally preempted by other properties? *P* cannot be preempted by the structural property *R*(P_1o_1, …, P_no_n), because it is *identical* to it (Kim 1998, 117–118). But what prevents *P* from being preempted by the (appropriately related) properties P_1, …, P_n? Kim's answer is that microbased properties are not *determined* by the properties in their microbase:

> We clearly cannot think of P_1, …, P_n, and *R* taken together as determining *P*. For to say that the properties 'determine' *P*, in the usual sense, is to say (at least) that necessarily any object that has them has *P*. But this condition is at best vacuous in the present case: an object that has *P* cannot be expected to have any of the P_is or *R*. The reason of course is that the P_is are the properties of the object's proper parts, and *R* is a relation, not a property.
> (Kim 1999, 117)

Hence, microbased properties fail to be determined by the properties in their microbase for the same reason they allegedly fail to supervene upon them: they are exemplified by distinct objects. And just as in the case of supervenience, the question is why a notion of determination which restricts determination to properties of the same object is the (only) correct notion to adopt. There seems to be a straightforward sense of "determines" in which microbased properties *are* determined by the properties in their microbase: a table's having a mass of ten kilograms (Kim's example) seems to be determined by its consisting of a six kilo top and a four kilo pedestal. (For further, more detailed, discussion see Walter 2008.)

6. Reduction

What remained to be addressed after section 4 was the possibility of an *intra*level causal drainage, where the higher-order properties at each level are preempted by the first-order properties of that level. Kim's response was that higher-order properties immune against SA because they are reducible, and where there is only one property, there can be no competition, and thus no preemption: "Reduction is the stopper that will plug the cosmic hole through which causal powers might drain away" (Kim 2005, 68).

But how are these reductions to be accomplished? Kim (1998) held that most higher-order properties are reducible by means of *functional reductions* (Kim 1998, 98–99), so that each level contains (except for a few non-functionalizable exceptions like phenomenal properties) strictly speaking only first-order properties. Allegedly, this dissolved the problem of intralevel causal drainage.

Kim (2005) still defends the functional account of reduction, but he seems to have abandoned the explicit distinction between orders and levels, arguing that reduction is also the key to stopping inter*level* causal drainage:

> Let us say that the property of being H_2O is the total micro-based property of water at the atomic level L (so having M_L = being H_2O). So we have:
>
> (1) Being water = having M_L.
>
> At the next level down, L-1, say the level of the Standard Model, hydrogen atoms have a certain microstructural composition as do oxygen atoms, and water has a certain microstructural composition at this level; call it M_{L-1}. Then by the same reasoning that led us to (1), we have:
>
> (2) Being water = having M_{L-1}.
>
> At the level L-2, the one below the Standard Model (if there is such a level), water is again going to have a certain microstructure at this level; this is M_{L-2}. We then have:
>
> (3) Being water = having M_{L-2}.
>
> And so on down the line, to M_{L-3} and the rest. These identities in turn imply the following series of identities:
>
> $M_L = M_{L-1} = M_{L-2} = M_{L-3}$ ...
>
> *Voilà!* These are the identities we need to stop the drainage. (Kim 2005, 68–69)

6.1 Reduction and Higher-*level* Properties

One problem with Kim's attempt to block causal drainage by appeal to reductions is that microbased properties seem to have "multiple compositions" (Block 2003, 146). Kim says the table's having a mass of ten kilograms is the microstructural property of being composed of a six kilo top and a four kilo pedestal, but it seems that the table could have the *same* property in virtue of being composed of a five kilo top and a five kilo pedestal. This raises two problems. First, if microbased properties are multiply composable or realizable, the multiple realizability of mental properties does not seem to prevent them from being microbased properties in Kim's sense. Second, the identities Kim appeals to in order to stop causal drainage would be impossible: "Kim's plugging the draining with micro-based properties depends on assuming identities (such as 'water = H_2O') and multiple composition will exclude such identities" (Block 2003, 146).

In response, Kim insists that multiple composability does not preclude identities:

> First, in spite of jade's multiple composition, each instance of jade … is either jadeite or nephrite, and I don't see anything wrong about identifying *its* being jade with *its* being nephrite (if it is nephrite) or with *its* being jadeite (if it's jadeite). … All we need is identity at the level of instances, not necessarily at the level of kinds and properties … [Second, we can; S.W.] … identify jade with a disjunctive kind, jadeite or nephrite (that is, being jade is identified with having the microstructure of jadeite or the microstructure of nephrite). … On the disjunctive approach, being jade turns out to be a causally heterogeneous property, not a causally inert one. … To disarm Block's multiple composition argument, adopting either disjunctive property/kind identities or instance (or token) identities seems sufficient. (Kim 2005, 58–59)

First, if token-identities can secure the causal efficacy of jade, despite its multiple composability, then why can they not secure the causal efficacy of irreducible mental properties, despite their multiple realizability? If all we need is identity at the level of instances, not necessarily at the level of kinds and properties, then where is the problem for NRP? Second, one wonders why Kim thinks he himself can have instance-identity without type-identity. After all, for him property-instances are events, whose identity conditions entail that the instances are identical only if the types are identical.

Concerning Kim's second option, suppose that being jade is identical to a disjunction of two microstructural properties. Given what Kim acknowledges elsewhere, the causal powers of the properties in the two microbases that form the disjunction determine the causal powers of being jade. Ascribing these properties to an object thus exhaustively fixes its causal potential, so that nothing is left for being jade to do, *even though it is identical* to a disjunction of two microstructural properties. Although being jade cannot be preempted by the disjunction of the two microstructural properties to which it is identical, it can still be preempted by the individual disjuncts.

Can multiply composable microbased properties be *functionally reduced*? No, because functional reductions are a non-starter for microbased properties, given that they are *eliminative*—as Kim has admitted in Kim (1998, 106), the property that is functionally reduced doesn't survive the reduction process.

Hence, the causal efficacy of multiply composable microbased properties can neither be vindicated by disjunctive identities, nor by token-identities, nor by functional reductions.

6.2 Reduction and Higher-*order* Properties

What about Kim's original suggestion that functional reductions can secure the causal efficacy of higher-*order* properties? As said above, functional reductions are *eliminative*. A functionally reduced property *F* has to be given up as a genuine property which can be exemplified in different species, and we retain only the predicate "*x* has *F*" and the concept *F* by which we equivocally pick out different properties in different species (Kim 1998, 106). It is thus a red herring to think that functional reductions can vindicate the causal efficacy of *the properties reduced*, because these get sundered into many different species-specific properties during the process of reduction. It is *these* that are identical to first-order properties. Hence, even if inter*level* causal drainage could somehow be stopped, *they*, i.e., the first-order properties at each level, would be the only causally efficacious properties. If this is the only kind of causally efficacious property that the proponent of SA can protect from her own argument, her position will hardly look attractive—and definitely not like "a plausible terminus for the mind-body debate" (Kim 2005, 173).

Literature

Block, Ned 2003 "Do causal powers drain away?", *Philosophy and Phenomenological Research* 67, 133–150.

Kim, Jaegwon 1988 "Supervenience for multiple domains", *Philosophical Topics* 16, 129–150.

Kim, Jaegwon 1998 *Mind in a Physical World: An Essay on the Mind-body Problem and Mental Causation*, Cambridge, MA: MIT Press.

Kim, Jaegwon 1999 "Supervenient properties and micro-based properties: A reply to Noordhof", *Proceedings of the Aristotelian Society* 99, 115–117.

Kim, Jaegwon 2005 *Physicalism, or Something Near Enough*, Princeton, NJ: Princeton University Press.

Walter, Sven 2008 "The Supervenience Argument, Overdetermination, and Causal Drainage: Assessing Kim's Master Argument", *Philosophical Psychology*.

No Bridge within Sight

Daniel Wehinger, Innsbruck, Austria

1 The explanatory gap

In his 1983 paper „Materialism and Qualia: The Explanatory Gap“, Joseph Levine states that there is – the title already says it – an „explanatory gap“ between the physical and the mental. Neuroscience has revealed many of the processes that take place in our brains when we engage in mental activities. Nevertheless, it is still kind of a mystery why, say, pain feels the way it does. Having learned which neural processes give rise to the experience of pain we are still left asking why and how they do so. Phenomenal properties, i.e. mental properties with a distinct phenomenal feel such as being in pain, cannot, so it seems, be fully explained in physical terms. There is a gap here. As a result, it seems perfectly conceivable that there is a world which is physically indiscernible from ours but lacking any phenomenal properties – the zombie world. This scenario was forcefully elaborated by David Chalmers in his "The Conscious Mind" (1996, 94-99). Chalmers draws explicitly dualist conclusions from the problem of the explanatory gap and the resulting conceivability of the zombie world. He claims that in addition to the lots and lots physical properties such as being six feet tall or weighing 100 pounds there are also some mental properties that cannot be reduced to the physical, namely the phenomenal properties introduced above. Therefore, physicalism is false and a dualism of properties must be assumed. It is this kind of dualism – property dualism – that I will be dealing with in the following.

Now, property dualism as such of course doesn't close the gap between the mental and the physical. It is rather a possible conclusion from it. Chalmers, however, claims that his theory provides him with the tools necessary for building a bridge. He aims at solving the "hard problem": the question of how phenomenal properties arise from the physical. (Chalmers 1995) This, according to Chalmers, cannot be done by appeal to physical facts. "[I]nstead, we have to look for a "Y-factor," something *additional* to the physical facts that will help explain consciousness. We find such a Y-factor in the postulation of irreducible psychophysical laws." (Chalmers 1996, 245) These laws are expected to be general and simple, i.e. they do not correlate particular types of neural processes (e.g. the firing of C-fibers) with particular types of phenomenal properties (e.g. the experience of pain). They rather have to be conceived of as the fundamental underlying laws that explain these correlations. A look at physics tells us that there are only a few fundamental physical laws. The same can be expected in the case of consciousness. (Chalmers 1996, 214-215) From the conceivability of the zombie scenario and other arguments for dualism, Chalmers concludes that the fundamental properties these laws invoke cannot be physical. They rather have to be phenomenal or protophenomenal, i.e. constituting phenomenal properties. And, just as is the case with fundamental psychophysical laws, only a few fundamental phenomenal or protophenomenal properties are to be expected. (Chalmers 1996, 126-127) By now we don't know these fundamental phenomenal or protophenomenal properties. It is the task of a future science of consciousness to discover them. A systematization of the correlations between types of neural processes and types of phenomenal properties should guide us there. And after many years of long, hard work we shall eventually come to know the fundamental psychophysical laws and the fundamental phenomenal or protophenomenal properties that underlie these correlations. The arisal of consciousness will then be explained and there will be no mystery left.

2 Why Bennett is not a dualist

In an unpublished draft of August 2006 called "Why I am Not a Dualist" Karen Bennett criticizes Chalmers view. Her aim is to establish thereby that the Chalmers-style dualist is not any better off than the physicalist: It is just as difficult for her to solve the hard problem as it is for the physicalist. Taking into account additional criteria such as ontological economy, the unification of science and so forth, physicalism wins. (Bennett unpublished, 24)

Bennett puts forward two arguments to support her view. The first argument focuses on the dualist's research strategy. The dualist believes that scientific investigation will one day reveal how phenomenal properties arise from the physical. On her way there she uses exactly the same scientific methods and tools as the physicalist: "Both will do a lot of serious neuroscience, and both will pay attention to introspective phenomenology in order to get a better understanding of 'phenomenal space'. Both will run a lab, employ postdocs, and apply for NSF funding." (Bennett unpublished, 12) Nevertheless, the dualist is sure that there will never be a fully satisfying explanation of phenomenal properties in physical terms. That is, the dualist assumes that the physicalist cannot in principle close the explanatory gap. This *a priori* prediction is rendered suspect by the fact that the dualist expects important insights from the scientific investigation of consciousness: "If the dualist thinks that scientific research can uncover hitherto unsuspected truths about the fundamental laws governing psychophysical connections, why should she not *also* think that it can uncover hitherto unsuspected truths about the physical?" (Bennett unpublished, 13) There's a real tension here, Bennett remarks: "The more you can see how research in the cognitive sciences can tell us how consciousness arises from the physical, the less secure you should be in your intuition that no purely physicalist story could ever work." (Bennett unpublished, 14)

In her second argument Bennett puts her finger on the dualist's claim that her theory allows her to close the gap between the physical and the mental. Bennett stresses the fact that "[a]ny dualist who accepts the burden to systematize the macro-correlations [i.e. the correlations between certain types of neural processes and certain types of phenomenal properties] is committed to something in the ballpark of protophenomenalism." (Bennett unpublished, 14) The generation of our everyday phenomenal properties cannot be explained by appeal to themselves. Rather, it must be assumed that the few fundamental properties out of which our conscious life is woven are quite unlike the phenomenal properties we know and had therefore better be called protophenomenal properties. (Bennett unpublished, 15) "Systematizing the relationship between the physical and the phenomenal", Bennett goes on, "is a matter of figuring out what those elements [or properties] are, and what general laws govern their relations both to the physical and to each other." (Bennett unpublished, 15) Now, the fundamental protophenomenal properties can again be either phenomenal or nonphenomenal. If they are nonphenomenal, it is hard to see how they can generate the phenomenal properties we are familiar with. "[W]e now need a story about how the phenomenal arises from the *protophenomenal*. […] The explanatory gap has not been closed; it has just been shunted into the space between the protophenomenal and the phenomenal. The hard problem rearises there." (Bennett unpublished, 16-17) If, on the contrary, protophenomenal properties are phenomenal themselves, then we need to explain how the physical can give rise to them. "More precisely, we have lost out on the attempt to systematize and unify the relationships between the physical and the phenomenal." (Bennett unpublished, 17) Bennett sums up the problem as follows: "The more similar the protophenomenal properties are to phenomenal ones, the less headway can be made on the project of systematizing the macro-correlations; we may as well take each and every phenomenal property, each and every macro-correlation, as fundamental. And the more removed the protophenomenal properties are from phenomenal ones, the less point there is to postulating them at all. We still cannot see how human experience – genuine, full blown consciousness – arises from complicated relations among such fragmentary shadows of phenomenality." (Bennett unpublished, 17) By postulating a third category of properties the dualist "only answers the letter of the hard problem". (Bennett unpublished, 18) The spirit of the hard problem – the question of how the phenomenal arises from the nonphenomenal – is as pressing as ever. (Bennett unpublished, 18)

3 The subjectivity of the mental

I believe that Bennett is right with her criticism of Chalmers's position. I believe that, just like the physicalist, the dualist cannot solve the hard problem. I disagree, however, with Bennett's claim that this amounts to an impeachment of dualism. Therefore, I will first explain why a fully satisfying explanation of how the mental arises from the physical is not to be expected and second why the prospects for dualism are nevertheless intact.

Mental properties[1] can be characterized as essentially subjective, in the sense that, while every physical property is in principle accessible to everyone, every mental property is principally accessible only from a certain subjective perspective. (Nagel 1974, 442) A tree's height can be measured by everyone around. No one is privileged here. A person's pain, however, can be experienced only by the person herself. Even though we can adopt the person's point of view, we cannot actually *feel* her pain. Therefore, the instantiation of mental properties brings with it a change of perspective. It is this change of perspective that makes the occurrence of mental properties such a mystery. In contrast, there is no big mystery involved in explaining how one neural process generates the other and how one thought follows the other. Here the entities involved are on a par, so to say: They are both objective in the first case and both subjective in the second case. Therefore, all we need to do in order to achieve a fully satisfying explanation is to keep track of every turn they take. When trying to explain, on the contrary, how neural processes give rise to mental properties and how thoughts yield neural changes we have a much harder time. Here the entities involved are tied to different perspectives. A fully satisfying explanation, however, would have to lead us all the way from the physical to the mental, or conversely from the mental to the physical without changing the perspective. In view of the subjectivity of the mental and the objectivity of the physical such an explanation is not to be expected. Adding up objective facts doesn't seem to get us any nearer to something essentially subjective. It is only by changing the perspective that we come to know there are mental, and therefore subjective, properties. Hence, a bridge between the physical and the mental is not within sight: Such a bridge would again have to be either subjective or objective and either way the hard problem, which can by now be reformulated as the question of how you get something essentially subjective out of something essentially objective, rearises.

Many physicalists concede that mental properties indeed appear to be subjective in the sense described above. If, however, they really are subjective in that sense, then they cannot be physical, for the physical is certainly objective. Therefore, the physicalist is confronted with the difficult task of having to point out how the mental, contrary to appearance, can be objective without thereby turning it into something it is not. Admittedly, it is not quite clear by now how this could be done and, thus, how the physicalist claim that everything there is, including mental properties, is physical could be made true. But this, it is argued, does not ultimately rule out the possibility of its truth. The following quote from Thomas Nagel is revealing here: "If we acknowledge that a physical theory of mind must account for the subjective character of experience, we must admit that no presently available conception gives us a clue how this could be done." But, as Nagel goes on, "[n]othing is provided by the inadequacy of physicalist hypotheses that assume a faulty objective analysis of mind. It would be truer to say that physicalism is a position we cannot understand because we do not at present have any conception of how it might be true." (Nagel 1974, 445-446) Such a line of argument, however, puts physicalism at risk of becoming something close to an article of faith.

Dualists, on the contrary, claim that mental properties are indeed the way they appear to be, i.e. subjective, and therefore nonphysical. This claim, of course, does not solve the hard problem. But this must not be seen as a lack of the dualist position. The dualist is not committed to solving the hard problem. All she is committed to is the thesis that, in addition to the lots and lots of physical properties, there are at least some mental properties that are not physical. In fact, I believe that, if the mental is essentially subjective, the insolubleness of the hard problem follows: A fully satisfying explanation of the arisal of mental properties from physical processes does not allow for a change of perspective. We are not satisfied, if the *explanans* does not lead us straightly to the *explanadum*. If, however, the mental is essentially tied to a certain subjective perspective, objective facts won't guide us there. Therefore, the insurmountability of the gap between the physical and the mental is part of the dualist claim.

[1] So far, I have restricted the discussion to phenomenal properties, following the terminology of the authors discussed. Phenomenal properties are mostly seen as one kind of mental properties, the other kind being psychological properties such as learning. Psychological properties, it is assumed, can be functionalized, whereas in the case of phenomenal properties this is questioned. (Chalmers 1996, 11-31) Since I have doubts about the phenomenal/psychological-distinction I will from now on be speaking of mental properties only. On the ontological level, of course, this doesn't have any consequences. The position discussed remains dualist all the same.

Literature

Bennett, Karen unpublished “Why I am Not a Dualist”, URL = <http://www.people.cornell.edu/pages/kb383/whyiamnotadualist.pdf>.

Chalmers, David 1995 “Facing up to the problem of consciousness”, reprinted in Jonathan Shear (ed.) 1997 *Explaining Consciousness: The Hard Problem*, Cambridge, MA: MIT Press, 9-32.

Chalmers, David 1996 *The Conscious Mind*, Oxford: Oxford University Press.

Levine, Joseph 1983 “Materialism and Qualia: The Explanatory Gap”, *Pacific Philosophical Quarterly* 64, 354-361.

Nagel, Thomas 1974 “What is it like to be a bat?”, *Philosophical Review* 83: 435-450.

On the Characterization of Objects by the Language of Science

Paul Weingartner, Salzburg, Austria

1. The Objects Described by the Laws of a Scientific Discipline

In the natural sciences we usually make use of a twofold picture for the description of the observed phenomena: the objects of experience and the laws of nature, where the behaviour of objects is governed by the laws of nature. To give some examples: Newton's laws of motion describe the spatio-temporal behaviour of mass points. Kepler 's laws determine the trajectories of planets in the solar system. The fundamental law of Quantum Mechanics, the Schrödinger equation, governs objects like atoms, electrons, neutrons … etc. Mendel's laws rule the transmission of genes into the next generation.

These objects (mass points, planets, electrons, genes … etc.) the behaviour of which is described by the laws of nature cannot be individualised. For instance mass points are not individualised objects, but stand for any object possessing mass; also the trajectories of Kepler 's laws are not individualised, they stand for any trajectory belonging to a certain category which obeys certain conditions like being periodical and having a certain type of stability[1]; further the objects like neutrons or electrons cannot be individualised by the Schrödinger equation, since this equation describes the behaviour of kinds of elementary particles for example of electrons and of neutrons. Similarly Mendel's laws do not distinguish between individual genes transmitted but are concerned with classes or categories of genes.

From this consideration it seems to follow that the objects of science which are governed by the laws are always incomplete in the sense of Meinong. Even if this is correct, we have to point out however, that the incompleteness is only relative here.

These objects are incomplete w.r.t. to the individualised object, but they are not incomplete w.r.t. what laws of nature describe. Thus for example it holds for the laws of Quantum Mechanics (for Schrödinger's equation) that they are permutational symmetric (invariant). That means that "different individual" particles of the "same kind" are treated identical by the law. Thus the laws and the respective physical reality described by these laws remain the same if we interchange any two particles of the same kind, for example two electrons. The same holds for protons, neutrons, neutinos and photons. Concerning physical systems or states permutational invariance holds only for bosons not for fermions because the latter obey Pauli's exclusion principle.

What has to be underlined here is, that invariance properties of laws of nature are just the essential characteristics of what a law is. So let us move further to the most important invariance of laws of nature: It is that the laws of nature are space time invariant. That means that the laws do not change with time, i. e. they abstract from any particular point of time (translational invariance or translational symmetry) and they do not change from one space point to another, i. e. they abstract from any particular space point. [2] From this it follows that laws of nature do not describe individualised objects. However a dynamical law describes the time development from an individual state of the system at t_1 to an individual state of the system at $t_{2.}$ But this individual state at t_2 is only derivable from the law (as a special solution of the differential equation) if the state at t_1 is known and instantiated; i.e. the later state is a definite function of the earlier. In the case of statistical laws even this is not possible. Individual microstates cannot be used as an instantiation: Statistical laws describe and predict only the states of the whole system, the macrostates, which can be realised by a huge number of different microstates; i.e. no particular individual microstate is required, anyone of the huge number will do. This means also that although the microstates cause the macrostate, no particular microstate is a necessary condition for the macrostate.

The result of this section is that the objects described by laws of nature[3] are not objects which satisfy uniqueness in the sense of Russell. Relative to individual objects satisfying uniqueness they are incomplete. But they are not incomplete w.r.t. laws, since laws have to have invariance (or symmetry) properties as their essential characteristics. The question whether real particular objects of modern physics are complete in the sense of Meinong or are individualised in the sense of satisfying uniqueness, will be treated below.

2. Russell's Ontological Presuppositions Concerning the Objects of Reference in the Sciences

2.1

Names directly designate an individual (object) which is its meaning.[4] That is the relation of denoting, designating or referring is a two-place relation and reference is identified with meaning. Russell drops the middle part of the Medieval Theory (also adopted by Meinong):

name	concept	reference
description	conceptual construction	object
	meaning, content	

2.2

In Russell's understanding, the relation of denotation (designation) or reference is the same if the objects of reference are mathematical (conceptual) entities or physical objects; i.e. this relation is independent of whether the relata are conceptual objects (which are neither spatial, nor temporal) or physical objects /(in space and time).

[1] For a detailed discussion of the conditions for dynamical (and also: statistical) laws see Mittelstaedt-Weingartner (LNt, 2005) ch. 7.

[2] If instead of space point we speak of invariance w.r.t. moving reference systems the things get more complicated. For details see Mittelstaedt-Weingartner (2005, LNt) ch. 6.

[3] For a detailed discussion see Mittelstaedt-Weingartner (2005, LNt) ch. 10.

[4] Cf. the quotation from Russell (1919, IMP), p. 174, note 3.

2.3

Mathematical entities are always rigid in the sense that they either (sharply) satisfy uniqueness or do not. Physical entities on the other hand are not always rigid. But in Russell's understanding all objects of reference are rigid.

To substantiate 2.2 and 2.3 one has to know first that according to Russell "all the objects of common-sense and developed science are logical constructions out of events"[5].

Secondly, that these logical constructions which are built from physical objects are like conceptual entities and thus rigid and impenetrable: "The events out of which we have been constructing the physical world are very different from matter as traditionally understood. The matter that we construct is impenetrable as a result of definition."[6]

Under "matter as traditionally understood" Russell understands matter as a permanent indestructible substance.

The further presuppositions or principles listed below cannot be substantiated directly by giving quotations in the literal sense from Russell's works. But they seem to be hidden by Russell's treatment of objects of reference and by consequences of such treatment.

2.4 Value-Completeness

If $(\iota x)\phi x$ satisfies uniqueness, then the object of reference is a bearer of value-definite (or value-complete) properties.

This presupposition is accepted and defined already by Kant:

Of all possible predicates (of an object as a bearer of predicates) one of each pair of opposite (contradictory) predicates must belong to it. In Kant's words: "Everything as regards its possibility is likewise subject to the principle of complete determination according to which if all possible predicates are taken together with their contradictory opposites, then one of each pair of contradictory opposites must belong to it."[7]

A physical consequence of 2.4 is that every individual (physical) object possesses always a well-defined position in space. This holds also for Russell according to whom the most elementary physical objects are his "events": "The matter in place is all the events that are there, and consequently no other event or piece of matter can be there. This is a tautology, not a physical fact"[8]. The above consequence is however typical for the domain of Classical Mechanics and does not hold generally (cf. 3 below).

2.5 Mechanical Object

If $(\iota x)\phi x$ satisfies uniqueness, then the object of reference is a bearer of such (essential) properties like mass, charge, geometrical shape, which transform covariantly under the transformation of the Galilean Group. That means that the object remains rigid under translation in space, under orientation in space, under translation in time and under inertial movement (with arbitrary velocity). In this sense "mechanical object" or "mechanical system" can be characterised by the Galilean symmetry group.[9]

From this it will be clear that the opposite implication does not hold: it does not hold that an object which satisfies the Galilean group, satisfies $(\iota x)\phi x$. Since it is a whole class of objects (the objects of Classical Mechanics) which satisfy the Galilean Group and not a single object only.

2.6 Uniqueness

If $(\iota x)\phi x$ satisfies uniqueness, then the object of reference is *unique* according to Classical Mechanics by his definite (accidental) properties: by position (*p*), momentum (*q*) and point of time (*t*).

This holds under the additional assumption of the impenetrability of the object in a space-time point (which does not follow from the dynamical laws). But also this assumption seems to be hidden in Russell's view of event and place (see the quotation in 2.4 above).

The question whether Newton has already proved the uniqueness is difficult. It is the question whether he has shown that besides the one there does not exist a different, second trajectory satisfying the same initial conditions along which the body can move obeying his laws including his law of gravitation. According to Arnold, Newton showed by checking many solutions of the laws that they depend smoothly (continuously) on the initial data. But the theoretical proof seems to have been given first by Johann Bernulli.[10]

2.7 Reidentifiability

If $(\iota x)\phi x$ satisfies uniqueness, then the object of reference is *reidentifiable* through time, i.e. has temporal identity. This reidentifiablity in turn requires two conditions to be fulfilled:

(a) There has to be a dynamical law which connects the object in state $S_1(p, q, t_1)$ with the reidentifiable object in state $S_2(p, q, t_2)$.

(b) The objects have to be impenetrable such that there can be only one object in a space-time point.

2.8 Observer-Invariance

If $(\iota x)\phi x$ satisfies uniqueness, then all observers of the object of reference (or in other words: all laboratories with rods and clocks in which the object is investigated) are equal; i.e. there is no designated observer or laboratory. In other words: all observers will arrive at the same result concerning the unique object of reference.

2.9 Trans-World-Identity

According to our understanding of "Law of Nature", the laws of nature are valid in all (physically) possible worlds which differ from our world only with respect to individual states or initial conditions.[11] Thus individual states or initial conditions are not designated by any law either in this

[5] Nagel (1944, RPS), p. 331.

[6] Russell (1927, AMt), p. 385; cf. (1925, ABC), p. 185.

[7] Kant (1787, KRV) B600. Cf. the discussion in Mittelstaedt/Weingartner (2005, LNt), p. 268, 271f and 276f.

[8] Russell (1927, AMt), p. 385.

[9] Cf. Mittelstaedt (1986, SRM), p. 219f. A more detailed and precise definition for "classical physical object" or "object of Classical Mechanics" is given in: Mittelstaedt/Weingartner (2005, LNt), p. 271f.

[10] Cf. Arnold (1990, HBN), p. 31f.

[11] For a detailed justification cf. Weingartner (1996, UWT) ch. 7 and Mittelstaedt/Weingartner (2005, LNt), p. 181ff.

world or in another (physically) possible world. Therefore it is not determined by any law whether some individual initial state of our world can be found in any of the (physically) possible other worlds. What has been said is particularly true of the dynamical laws of Classical Mechanics. Consequently, although a dynamical law connects two individual states of this (our) world and although it will connect two individual states in another world, it does not connect two individual states of two different worlds.

From this it follows immediately that trans-world-identity of individual states is not guaranteed in Classical Mechanics.

The same holds for the objects of reference of Classical Mechanics: for any such object $(\iota x)\phi x$ satisfying uniqueness, its identity cannot be guaranteed in any other possible world, independently of how the accessibility relation is defined. Although such an object is reidentifiable in *one* world, it is not from one world to another one.[12] From this it follows that an application of Kripke's semantics to Classical Mechanics will lead only to a redundant extension; since its interpretation of "possible" and "necessary" reduces to factual (true or false) in this (our) world.

3. Are the Principles about the Objects of Reference (Listed in Ch. 2) valid when applied to the Microlevel (Quantum Physics)?

3.1 Value-Completeness

Applied to stable elementary particles like electrons, protons and neutrons (or to stable composed systems) there is no general value-completeness or value-definiteness; i.e. the object of reference is, in general, not a bearer of value-definite (value-complete) properties (cf. 2.4 above). At time *t* the object, as known through measurement results, can be the bearer only of a selected or limited number of properties, i.e. those which are mutually commensurable.

The conceptual construction of the object as it is known by measurement results (Russell's logical construction) is necessarily incomplete. Therefore the description $(\iota x)\phi x$ of such an object, since it is not value definite (cf. 2.4), will not satisfy uniqueness. As a consequence of that, the conceptual construction (or the logical construction) which is incomplete cannot – in general – be identified with the reference. Thus if we interpret the conceptual construction as the meaning, it should not be identified with the reference (in contradistinction to Russell).

3.2 Permutation Invariance

The Schrödinger equation holds for *kinds* of objects, not for single individual objects. In general: the laws of Quantum Mechanics (*QM*) are permutationally invariant, i.e. they are invariant with respect to an exchange of particles of the same kind. This means that different individual (numerically different) particles of the same kind are treated identically by the laws. The laws do not distinguish between two electrons, two protons… etc.; they remain the same laws when we exchange two electrons, two protons, two neutrons or also two photons.[13]

From this it follows that one of the conditions for $(\iota x)\phi x$ – the condition that at most one x satisfies ϕx – is violated since more than one object (a whole class of objects of he same kind) satisfies the law. Thus uniqueness of the *QM*-object is not satisfied.

3.3 Uniqueness

A *QM*-object can also not be uniquely described as an individual object by accidental properties. Recall (2.6 above) that an object of Classical Mechanics can be so described, namely by the three magnitudes of position (*p*), momentum (*q*) and time (*t*). The reason that this is not possible for the *QM*-object is because the totality of accidental properties which were needed for the individualisation (uniqueness) is not available at the same time. That is the description by accidental properties is never complete and thus we cannot get uniqueness for the respective objects if they are understood to be permanent in some (reasonable) way (cf. 3.4 below).

3.4 Reidentification

The *QM*-object is not identifiable through time, there is no temporal identity. In fact there are the following two possibilities:

(a) There is a position measurement at t_1, that is we can have uniqueness of the object (or state of the system) – impenetrability presupposed – only at the time point t_1; in this case the object (or state) dissolves later at t_2 such that we do not have uniqueness anymore, i.e. no permanent object.

(b) The two states $\psi(t_1)$ and $\psi(t_2)$ are connected by a law of *QM* through time (t_1, t_2). But in this case only $\psi(t_1)$ is unique w.r.t. one object or state (or system); since $\psi(t_2)$ can then be satisfied by more than one object (state or system), it therefore does not guarantee to designate the original unique object (state, system). Although there is permanence given by the connection of the law, there is no guarantee that what is connected is the original object (system) at a later time.[14]

3.5 Trans-World-Identity

For quantum mechanical objects or systems uniqueness (the condition for using $(\iota x)\phi x$) is not satisfied. The reason is this: the characterisation by their essential and permanent properties fails because with them only classes of objects or systems (like electrons, protons, photons) can be determined. But also a characterisation by accidental properties like position and momentum at a certain time is impossible since only a part of such properties are simultaneously available. Still another possibility for a unique characterisation would be a description of a sufficiently complete historical development of the object (instead of giving only the actual properties at a certain point of time).[15] However, it is an unsolved problem how such a

[12] For a detailed justification cf. Mittelstaedt (1986, SRM), p. 241ff.

[13] For more on permutational symmetry cf. Mittelstaedt/Weingartner (2005, LNt), p. 74, 77, 82.

[14] For more details see Mittelstaedt (1986, SRM), p. 227ff.

[15] The idea to use the history of the human actions and decisions as a principle of individuation of human souls (after separation from human body –

description can be obtained and used for the individuation of quantum mechanical objects. Since trans-world-identity of objects implies reidentifiablity of *one unique* object in different worlds, it follows that trans-world-identity of quantum mechanical objects is not possible. There may be however a kind of weak analogy of trans-world-identity if the following restrictions and deviations with respect to a Kripke style semantics are made:

(1) Possible worlds are replaced by measuring processes.

(2) "There is a world *W'* (different from the actual world *W*) in which the proposition *A'* about system *S'* is true" is replaced by: "There is a measuring process *M'* (different from the actual (or earlier) measuring process *M*) with the result that proposition *A'* about system *S'* is true."

(3) The accessibility relation *R*(*S, S'*) is satisfied, when *A*(*S, M*) = *A'*(*S', M'*).

(4) There is a non-zero probability for reaching result *A'* about *S'* by measuring process *M'* from the earlier result *A* about *S* by measuring process *M*.

(5) If (4) is satisfied, there is a non-zero probability for a weak temporal identity of the system (object) *S*.

4. Are the Principles about the Objects of Reference of Ch. 2 valid when applied to SR and GR?

4.1 Value-Completeness

Special Relativity (*SR*): Value-completeness or value-definiteness of properties of an object (reference system, cf. 2.4) is not satisfied for an observer at every point of time. But in this case (we have just inertial movement and no acceleration or gravitation and Minkowski space-time) the observer may always wait until the object appears in his past light cone.

General Relativity (*GR*): As soon as acceleration or gravitation is taken into account, there are always some domains in space-time with objects that will never appear in the past light cone of the observer. This is so even if the observer moves on a geodesic, i.e. free from forces. It is plain then that the description of such objects cannot be value-complete.

4.2 Permanence

SR* and *GR: The essential properties of the object of reference (as the bearer) are no more permanent, although with one exception: charge. The other properties like mass, length and geometrical shape change in case of fast inertial motion in accordance with the Lorentz-Transformation; this holds also in local inertial reference systems of Rimanean space-time (*GR*).

4.3 Uniqueness

According to 2.6, uniqueness of objects of Classical Mechanics can be satisfied by special values of the accidental properties position (*p*), momentum (*q*) and point of time (*t*). But in 3.3 it was shown that uniqueness is no more satisfied on the microlevel (in *QM*). Concerning *SR*, uniqueness w.r.t. *p, q, t* holds only partially; namely it holds only for objects appearing in the past and future light cone of the observer (dynamical laws presupposed). With respect to *GR*, uniqueness is dependent on the space-time curvature.

4.4 Reidentifiability

With respect to both, *SR* and *GR*, the object of reference is not in general reidentifiable through time; this is so because essential properties of the object like geometrical shape and mass may change depending on movement. Therefore reidentifiability holds only approximately in local reference frames of space-time.

4.5 Time and Simultaneity

With respect to both, *SR* and *GR*, there is neither a universal time, nor universal simultaneity. Each different observer (each different laboratory or reference system) has its own time and simultaneity. Therefore the object of reference is not the same for all observers.

Literature

Arnold, V.I. (1990, HBN) *Huygens and Barrow, Newton and Hooke.* Basel, Birkhäuser.

Kant, I. (1787, KRV) *Kritik der reinen Vernunft.* Riga.

Mittelstaedt, P. (1986, SRM) *Sprache und Realität in der modernen Physik.* Mannheim, Bibliographisches Institut.

Mittelstaedt, P./Weingartner, P. (2005, LNt) *Laws of Nature.* Springer, Heidelberg-Berlin.

Russell, B. (1919, IMP) *Introduction to Mathematical Philosophy.* London, Allen and Unwin.

Russell, B. (1925, ABC) *The ABC of Relativity.* London, Kegan Paul.

Russell, B. (1927, AMt) *The Analysis of Matter.* London, Kegan Paul.

Thomas Aquinas (Ver) *De Veritate.* Engl. Transl.: *The Disputed Questions on Truth.* Henry Regnery Company, Chicago 1952

Weingartner, P. (1996, UWT) "Under What Transformations are Laws Invariant?" in: Weingartner, P./Schurz, G. (eds.) *Law and Prediction in the Light of Chaos Research.* Springer, Heidelberg-Berlin, p. 47-88.

which could not serve anymore as individuating) was proposed by Thomas Aquinas (Ver), 19, as one, though not the only possibility, since it is not sufficient in all cases (like children who die immediately after birth).

The Functional Unity of Special Science Kinds

Daniel A. Weiskopf, Tampa, Florida, USA

1. Reduction vs. elimination redux

Realization is a relation between a property Φ at one level of organization and a property Ψ or family of properties Ψ_1-Ψ_n at a lower level of organization. According to the Multiple Realizability (MR) thesis, psychological properties, as well as the properties in the domain of many other special sciences, are multiply realizable. MR is true of properties that are defined by some purpose, capacity, or contribution they make to some end—generally, by their functional role. Where there are interestingly different ways of playing the role that defines the property Φ, then Φ has different realizations. For Φ to be *multiply* realizable, the Ψs must belong to distinct kinds, as defined by some independent taxonomy.

Against this consensus, Shapiro (2000) argues that the MR thesis is not even coherent. Consider the Ψs that realize Φ. Either:

(1) these realizers differ in their causally relevant properties, or

(2) they do not.

'Causally relevant properties' are those that enable something having Ψ to perform the function of a Φ. If (1) is the case, then the Ψs are different kinds. "But if they are different kinds then they are not the same kind and so we do not have a case in which a single kind has multiple realizations" (Shapiro, 2000, p. 647). That is, if the Ψs possess different causally relevant properties, then Φ itself does not constitute a kind, and hence one higher-level kind isn't being multiply realized. But if (2) is the case, then they are not different realizations and the thesis is false in this instance.

I agree that if Ψ_1 and Ψ_2 are different independently certified kinds then we have genuine MR. But I deny that their being different kinds entails that Φ is *not* a kind. Whether Φ is a kind or not depends on whether there is a sufficiently large and interesting body of empirical regularities in which Φ itself is implicated. Kindhood depends on there being a rich cluster of properties that reliably co-occur with something's being Φ, where these properties do not cluster together by chance but by the operation of some governing principle or mechanism.

It might seem that if Ψ_1 and Ψ_2 are causally different ways of bringing about Φ that this would *automatically* show that they did not participate in any (nonanalytic) common regularities. But this isn't obviously true, since distinct mechanisms can still give rise to shared properties and generalizations. We can see this by looking at an example that Shapiro himself discusses: the case of compound vs. camera eyes.

2. The eyes of others

Arthropod compound eyes and vertebrate camera eyes are all eyes in virtue of falling under the functional description 'organs for seeing'. But different mechanisms are involved in the production of sight in each kind of eye; hence, by the anti-MR argument, eyes should not be a single kind. However, both kinds of eyes can display similar *psychophysical* phenomena despite having different *optical* properties.

The main phenomenon of interest is the perception of Mach bands: regions of especially high or low brightness that occur at the high or low ends of a brightness gradient. While perception of Mach bands occurs in many organisms, including primates, cats, and horseshoe crabs (*Limulus polyphemus*), the neural circuits that underlie it differ radically across species. This can be illustrated with respect to the *Limulus* eye and the mammalian eye.

The lateral eyes of *Limulus* are composed of ~1000 cones that terminate in ommatidia (Battelle, 2006). Ommatidia contain photoreceptive cells that depolarize a central eccentric cell, the axons of which form the optic tract. Eccentric cell axons also distribute collaterals to their neighbors in adjacent ommatidia. These interwoven branching collaterals form the lateral plexus of the eye, which enables one ommatidium to be inhibited by activity in adjacent ones (Hartline and Ratliff, 1957). Lateral inhibition enhances contrast and sharpens perception of edges, and also explains the perception of Mach bands.

Mammalian eyes, while physically and optically different from compound eyes, also contain inhibitory mechanisms that produce Mach bands. In contrast to the loose organization of the lateral plexus, mammalian retinas are tightly organized into distinct layers. They also use a vastly greater range of cell types than does the *Limulus* eye. Photoreceptive cells feed into a network containing horizontal, amacrine, and bipolar cells, finally terminating at ganglion cells that project to higher regions. While there are many loci for lateral inhibition in the retina, it occurs initially in the horizontal cells linking adjacent rods and cones. These cells have highly specific connectivity patterns, as opposed to the near-random wiring of *Limulus* (Field and Chichilnisky, 2007; Sterling, 1998).

So we have two distinct mechanisms for producing lateral inhibition, and therefore two ways of constructing visual systems that can perceive Mach bands. Yet both eyes share more extensive causal properties than just enabling sight. Implementing lateral inhibition and producing Mach bands are two such interconnected properties. While these are not exhaustive of the possible kinds of eyes, they illustrate the point that distinct kinds at one level can have numerous other common causal properties, and therefore constitute a higher-level kind. Thus Shapiro's inference from different causal properties to the absence of a higher-level kind is not generally sound. Whether a functionally defined property also constitutes a kind is something to be decided on a case-by-case basis.

3. Unity through constraint?

Interestingly, Shapiro considers the case of lateral inhibition, but draws the opposite conclusion from it (Shapiro, 2004, pp. 117-120). Rather than concluding that the use of this common strategy of visual processing in different species shows its multiple realizability, he suggests that it shows that the evolution of visual systems occurs under tight constraints. These constraints mean that there will likely be only one (or at most a few) ways of building an organic system that can process visual images. And this is contrary to MR, which predicts that there should be a wide diversity of evolved mechanisms for each functional capacity.

Lateral inhibition serves a useful function in vision: it sharpens contrasts and aids in discrimination of closely spaced stimuli. Undoubtedly this accounts for its recurrence in evolution. But this recurrence does not undermine MR, since it's also clear that different species carry out lateral inhibition using physiologically distinct mechanisms (eccentric cell collaterals versus horizontal cells, for instance). Indeed, lateral inhibition occurs in multiple sensory modalities within single species (human vision, touch, and audition, for example), and it involves distinct cellular mechanisms in all of these cases.

What this suggests is that there is a common *functional* characteristic that recurs across different species as well as within individual species. But the presence of a functional characteristic does not necessarily entail the presence of any particular physical mechanism. That isn't to say that there might not be physical constraints on how nervous systems must be constructed if they're going to realize terrestrial psychological capacities. Many constraints, however, are organizational or functional, and these can't be assumed to be physically identical in all cases; at least not unless we simply adopt a taxonomy of physical mechanisms that co-classifies as physically similar anything that satisfies the relevant functional specification. But this is just to beg the question against the MR thesis.

Shapiro's other examples of constraints on the realization of cognition support this point. For instance, he argues that humanlike cognition requires sensory systems that transduce information into usable neural signals, receptors that tile sensory surfaces in varying densities, topographic maps in primary sensory areas, and broadly modular organization in the brain itself (Shapiro, 2004, pp. 105-138). But it is clear that there are at best rather abstract similarities among, say, the diverse photoreceptors in the retina and the various tactile, thermal, and chemical receptors that mediate touch. Even if they possess broadly similar receptive field organization and project to topologically organized regions of primary sensory cortex, the fine-grained detail of these neural structures will differ.

So one can agree with Shapiro that there are constraints on constructing terrestrial psychologies, but also maintain that these constraints are mainly functional. Since it is possible to build many different kinds of neural mechanisms within these constraints, this is compatible with multiple realizability.

It might seem, however, that this claim runs afoul of an argument advanced by Bechtel and Mundale (1999). They suggest that the MR thesis only appears plausible because we allow Φ to be individuated coarsely and the various Ψs to be individuated finely. "But if the grain size is kept constant, then the claim that psychological states are in fact multiply realized looks far less plausible. One can adopt either a coarse or a fine grain, but as long as one uses a comparable grain on both the brain and mind side, the mapping between them will be correspondingly systematic" (Bechtel and Mundale, 1999, p. 202).

Whether we should adopt a fine grain at the higher level or not, though, depends on whether these fine grained properties are independently certified as being theoretically 'interesting' in the relevant domain. Suppose we decide that *Limulus* eyes and cat eyes do not really both perform lateral inhibition. Rather, there is inhibition$_1$ and inhibition$_2$, two different fine-grained psychophysical capacities had by creatures with correspondingly different visual mechanisms.

Anything that performs inhibition$_{1/2}$ performs inhibition, since they are related as determinate and determinable. But inhibition and inhibition$_{1/2}$ play different explanatory roles. Explaining why a system perceives Mach bands may involve adverting to inhibition. Explaining why it perceives bands with *these* precise characteristics, though, may involve adverting to inhibition$_{1/2}$, since those constitute the particular psychophysical capacities the system has. Explaining *how* a system instantiates inhibition$_{1/2}$ may involve referring to the particular fine-grained neurobiological mechanism at work, and thus may involve matching fine-grained taxonomies. But explaining how a system instantiates inhibition might involve drawing attention *either* to the particular fine-grained mechanism at work in that organism, or to the range of possible mechanisms that can bring about that sort of capacity.

The latter involves a 'grain mismatch' of the sort that Bechtel and Mundale warn against. But it is hard to see why adopting this mixed taxonomy is a mistake. It can prove heuristically useful, for example, once one has characterized the general function of a cell type or brain region to then propose a range of possible lower level mechanisms that might realize that function, then proceed to rule them out on the basis of side effects, predicted responses to intervention, predicted anatomical consequences, and so on. Grain 'mismatches' of this sort can serve a crucial heuristic role in discovering mechanisms. So long as there is a unified field of inquiry that works with both taxonomies simultaneously, the MR advocate can do so as well.

4. Explanatory taxonomy and the special sciences

I've argued that diverse lower-level mechanisms can converge on common functional traits at various levels of organization, and that this assumption can play a heuristic role in discovering mechanisms. Now I will briefly sketch one way in which these functional groupings might be seen as kinds.

On Shapiro's view, special sciences categories have a fundamentally taxonomic function: they "collect and order the domain of a special science in a way that facilitates its investigation" (2000, p. 654). Functional concepts fix a range of 'analytic' truths about things that fall under them (e.g., eyes are for seeing). But not all functional concepts pick out categories that are equally interesting. Shapiro's view leaves us with no way of explaining why there should be such a difference. This difference shows up in theory construction because discovering the *right* functional components out of which to build an organism's control systems is non-trivial. Much of the work of building theories, models, and simulations involves finding the appropriate concepts to analyze a system.

Consider central pattern generators (CPGs). CPGs are units that produce regular oscillations endogenously or in response to input. There are many different ways to assemble such circuits (e.g., out of either multi-neuron arrays using inhibitory interneurons, or out of local dendrodendritic connections). These structures differ in their size, location, temporal characteristics, and many other physical/neural properties.

Clearly, the concept of a CPG is a functional concept. But it is still *explanatory*: positing such circuits illuminates how certain kinds of behavior and observed neural activity might take place. These abstract structures play an explanatory role independent of information about their realizers. Knowing that an organism's control systems contain a CPG in a certain location helps to explain certain of its capacities Two examples are control of swimming in the lamprey and the stomatogastric ganglion of the lobster, which regulates digestion. In virtue of producing certain sorts of effects, these functional units can be situated within a larger system of control structures.

Understanding how an organism possesses a range of capacities depends on seeing its inner organization as containing such units. The same is true of lateral inhibition, which is an abstract device for producing a range of effects in sensory processing. I suggest that, in the behavioral sciences, it is by recurrently playing this sort of role in abstract control systems that functional categories earn their status as kinds.

The *interesting* functionally defined categories, then, constitute recurrent building blocks of cognitive systems. They explain the possession of various capacities of those systems without reference to specific underlying mechanisms. They may in turn be explained by the presence of further functional units at lower levels, or by physical mechanisms. One major task in understanding cognition is finding the right decomposition of a system into abstract control units and constituents. Logic, computation theory, cybernetics and control theory, and neural network theory provide examples of how the theory of such control units might be developed. And insofar as such functional categories can usefully be applied to modeling cognition, they count as kinds.

Literature

Battelle, B.-A. 2006 "The eyes of *Limulus polyphemus* (Xiphosura, Chelicerata) and their afferent and efferent projections", *Arthropod Structure and Development* 35, 261-274.

Bechtel, W., and Mundale, J. 1999 "Multiple realizability revisited: Linking cognitive and neural states", *Philosophy of Science* 66, 175-207.

Field, G. D., and Chichilnisky, E. J. 2007 "Information processing in the primate retina: Circuitry and coding", *Annual Review of Neuroscience* 30, 1-30.

Hartline, H. K., and Ratliff, F. 1957 "Inhibitory interaction of receptor units in the eye of *Limulus*", *Journal of General Physiology* 40, 357-376.

Shapiro, L. 2000 "Multiple realizations", *Journal of Philosophy* 97, 635-654.

Shapiro, L. 2004 *The Mind Incarnate*, Cambridge, MA: MIT Press.

Sterling, P. 1998 "Retina", in: G. Shepherd (ed.), *The Synaptic Organization of the Brain* (4th ed.), Oxford: Oxford University Press, 205-254.

Transcendental Philosophy and Mind-Body Reductionism

Christian Helmut Wenzel, Puli, Taiwan

In *Wittgenstein on Language and Thought*, Thornton gives an account of naturalization that he calls "representationalism": "Representationalism attempts to explain linguistic content as resulting from mental content and then to give a reductionist account of the latter. Mental content is 'naturalised' through the provision of a causal explanation of content" (p. vii). Thus we have a two-step reduction, first from linguistic content to mental content, then from the mental to the physical. Mental content is seen in representations, "internal mental representations that stand in causal relations to things in the world" (viii). Fodor's "descriptive causal theory" and Millikan's "teleological, or natural selective" account are given as examples of such representationalism. Wittgenstein, on the other hand, Thornton shows, opposes such reductionist theories already in the first step: Representations and mental content so understood would be too isolated and internal, too detached from the outside world. Linguistic content and meaning cannot be understood this way. Instead, they should be seen as being more "outside" from the start, making sense only within language and its use in society.

How would Kant fare in such current discussions? Certainly he has much to say about representations, *Vorstellungen*, Latin *representationes*. He also has read Locke and Hume and is aware of their empiricist accounts of impressions and ideas, as well as of Descartes' *res cogitans*. Yet Kant does not take the same route they do. He is usually not mentioned in current discussions of naturalization and reductionism of the mental to the physical. Nevertheless, although he does not – returning to Thornton – talk about linguistic content, he has much to say about judgments and representations. Certainly representations must have meaning, and they often arise from perceptions. Objects appear to us, and we don't make them up. Kant was not an idealist like Berkeley. He even distanced himself from Descartes, whom he also saw as an idealist (A 226/B 274). Unlike them, he never doubted the existence of the outside world. He saw himself as an "empirical realist" instead. So how would Kant react to current physicalist-reductionist accounts of representations and meanings?

When looking at his early writings, such as his "General Natural History and Theory of the Heavens, or An Attempt to Understand the Structure and Mechanical Origin of the Whole Universe According to Newton's Principles", one might think he has a liking for naturalization. But when thinking of his *Critique of Pure Reason*, one starts to have doubts. Why is that?

Central to Kant's transcendental philosophy from his *Critique of Pure Reason* are the categories, imagination, understanding, schemata, and these can easily appear to be mental in some way. Not without reason, Kantian faculty-talk is sometimes seen as psychological. In any case, one might want to call Kant a representationalist of some kind, simply because the notion of *Vorstellung*, Latin *representatio*, holds a central place in his theoretical transcendental philosophy.

But somehow Kant cannot be a representationalist of the kind Thornton has in mind. He does not understand his theory as giving an account of "mental" content in an individual person's head, and he certainly does not try to reduce mental representations to causal stories.

Such an undertaking would undercut his transcendental project from the start, or, rather, it would not touch it, but miss it altogether. Kant is not interested in an individual person's head and in how empirical concepts arise and are acquired, as Locke was. To the contrary, he wants to establish a priori concepts that make such experience possible. These concepts, the categories, are not understood as mental in opposition to the physical. They are very special concepts. They make the physical as such possible, and the naturalists do not talk about them at all. In a sense, transcendental philosophy undercuts the mind-body naturalist's project. Kant does not start with a mind-body dualism, with categories in the head and objects out there to be schematized. He also does not go in for the Cartesian *res extensa – res cogitans* distinction.

For Kant, the categories underlie the world we experience, because objects are nothing but appearances brought under schematized categories. Not only objects, but even time and space are not out there, independently of us. The forms of time and space are subjective and make objectivity possible. If there is a "head" in the sense of transcendental philosophy, then the world has to be in it – at least the a priori aspects of it.

Kant distinguishes between an inner and an outer sense, but not between an inner world of representations in the head and an outer world next to it.

In particular, it is empirical causality that is seen, within transcendental philosophy, to depend on a priori causality, and therefore it would not make any sense to try to reduce the transcendentally mental to the empirically causal. Empirical concepts and representations might be naturalized, but not a priori ones. Transcendental philosophy and the Copernican revolution go the other way around. They are independent of any philosophical project in which causality is taken for granted, as Fodor, Millikan and others do. It is not that such projects do not make any sense. The point here is that even if they succeed, they will not answer Kant's question about the possibility of experience and objectivity. Naturalizing projects take objectivity and the physical world for granted.

It is as Barry Stroud says, in contrast to Jay Rosenberg's historicizing, evolutionary and naturalizing accounts (Rosenberg, 616-20) of the Kantian minimalist "conceptual core" (615): "The absence of any interesting necessary conditions of thought and experience must be established, and not simply asserted as likely on general historical or 'evolutionary' grounds. Even the most uncompromising 'evolutionary' attitude would not preclude us from asking what it is that makes thought or experience possible – how it is possible for thought and experience to have 'objects', or be 'of' or 'about' something. It remains to be seen that that very general question itself must be given an historical or 'evolutionary' answer, even if an historical or 'evolutionary' answer must be given to the quite different question of who and why in the development of *homo sapiens* those conditions ever in fact came to be fulfilled." (Stroud 1977, 81-82)

Doubting causality in the way Hume did is mistaken in Kant's eyes. We need an a priori concept of causality from the start to have any of the coherent experiences that we as a matter of fact do have. Particular empirical causalities can be learned about in experience, but not causality in general, universally, as such, which we need in order to have any meaningful experiences to start with.

To have an apparently simple experience such as the perception of a ship, we need to see the ship as a unit, and for that we need the category of substance. That the parts of the ship stay where they are and do not float around chaotically and dissolve, and that I distinguish the ship from my perceiving it, presupposes a priori causality. Thus just to perceive a ship, without even invoking the question of empirical causality by asking whether it is going downstream or upstream, the categories are needed.

Even deeper, it is I who perceives the ship and I am conscious of this act. It is I who gives it unity and meaning in perception and judgment. Already here I do something that requires a priori concepts. (For accounts of Kant on the I and the soul, the 'unity of thought argument' and the 'inner sense argument', and on the complexities of various kinds of immaterialism of the soul, see Ameriks 2000, especially pp. 27-47. For a defense of the view that the categories go "all the way out", see Wenzel 2005.)

We might be tempted to see even these a priori concepts and their application as being something mental again. After all, Kant thinks of the categories as subjective. But the Kantian subject is not a mind-brain that is causally affected. In the framework of his transcendental epistemology, the subject even comprises time and space, as forms of all appearances. If we think of ourselves as being affected by things "outside of us", *außer uns*, then these things are understood, transcendentally, as nothing but simply *different* and *distinct from* (logically *außer*) us and, empirically, as objects that are always *already* subjected to those subjective conditions of time and space and the categories (spatially *außer uns*).

The Kantian transcendental subject is not a mere *res cogitans*. It is more. It comprises time and space as forms of intuition. Kant holds this against Descartes. A pure science of the *res cogitans*, the "I think", would not get us anywhere. No rational knowledge of the outside world, even of ourselves, could be obtained from it. Also no limits of our empirical knowledge could be pointed out in this way, which is something important for Kant, but not for the naturalist today.

Kant distinguishes his transcendental idealism from what he calls "transcendental realism", which is the view that time and space are things in themselves, independent of us. Common sense takes this view. If one starts in this way, one can depict oneself as some kind of mind-brain-body at one location and the tree one perceives as being "outside", ten meters away. Then one can start to give a causal story of sense perception, even look into the brain and try to give a causal account of consciousness and our having representations as well, maybe with the addition of evolutionary and social aspects. Reductionism lives here, and Kripke and Putnam for instance have given accounts of what we mean by "water" and H2O in this way. Transcendental realism starts with a picture of the world that is independent of us, with water as H2O already out there. But if we then place ourselves in this world, how can we be sure that this is how it really is? How can we avoid skepticism? Thus with Putnam we run into a problem similar to Descartes' doubt. We might be a brain in a vat, nay even the whole world might not exist.

But according to transcendental idealism, neither my brain, nor the tree, nor time and space are independent of representational conditions. It is only as appearances that they are in time and space, and it is only as being subject to the categories that they are objects. This is an instance of the general view that any third-person account presupposes a first-person perspective. Cassam's criticism that "Kant's mistake was to conclude … that the unity of consciousness does not involve being presented to oneself as an object at all" might still be within this view (p. 198).

Experience requires an act of synthesis, which in turn requires unity. It must be *my* experience. For the materialist it might be the brain or the object that gives this unity. For Kant it is the *act* that must provide it. In meaningful perception and in judgment we take something as something and the "taking" itself must have unity (Allison 1996, pp. 95, 102). For Kant it is transcendental consciousness (*Reflexion* 5661, AA 18, 318-9) and the original synthetic unity of apperception (*CpR*, B 134) that provide this unity, and they do this a priori, that is, prior to experience. When the materialist points to the brain, our sense organs, and their evolutionary adaptations to their functions and the environment, and the socio-linguist points to our language and society, Kant will point out that they take time and space and empirical objects for granted, as things in themselves, and thereby beg the question. If they also want to naturalize the act of taking something as something, we may respond with Allison that "taken in an investigation of its causal conditions, any token of the act of thinking is itself something represented, an object for an I, which, considered as such, is not itself an object in the world. In short, we return in the end to the ineliminability and systematic elusiveness of this ubiquitous 'I think'" (Allison, 1996, p. 104). Furthermore, we can add that the materialists will run into the problem of skepticism, because they cannot be sure that the objects, which for them exist independently of us, are *correctly* represented by us whenever we have representations of them, that is, when they appear to us. In Kant's words, the transcendental realist then "plays the empirical idealist" (A 369).

In Kant's picture the object is nothing but its appearance, and so the correspondence problem does not arise. Truth is in judgment, not in appearance. Ironically, one may also say that in the view of transcendental philosophy appearance already gives truth, *a-letheia*, as Heidegger wanted it, insofar as appearance and its object are not two separate things (contrary to the transcendental realist's view). The object does not need to be "deduced" from its appearance (A 372). It exists only as appearance. It is its appearance. Appearance is not something extra.

Imagine the following conversation between a transcendental realist (TR) and a transcendental idealist (TI):

> TR: "I think representations are generated in the brain."
>
> TI: "You mean processes happening in the brain? Well, they happen in time and space. You imagine them as appearances."
>
> TR: "But are they not caused? Are not our representations, imaginations, perceptions all caused?"

> TI: “Well, according to their matter, *materialiter*, yes. But according to their form, *formaliter*, no. That I see a hand with five fingers, of a certain size, with a certain color, hue and shade, in a certain light, and under a certain angle, yes, there is a causal story to be told for this. But that the hand appears in time and space at all, and that it has parts, for these facts there is no causal story to be told. You can reduce material properties to their causes, but not the formal ones (for which you need the categories and time and space). Furthermore, it is these formal aspects that make your causal stories possible.

Acknowledgments

For helpful discussion and comments I wish to thank Karel Mom.

Literature

Allison, Henry E. 1996 “Kant’s Refutation of Materialism”, in *Idealism and freedom: Essays on Kant’s theoretical and practical philosophy*, Cambridge University Press, pp. 92-106; originally in *The Monist* 1989.

Ameriks, Karl. 2000 *Kant’s Theory of Mind. An Analysis of the Paralogisms of Pure Reason*, new edition (first edition 1982), Oxford University Press.

Cassam, Quassim 1997 *Self and World*, Oxford University Press.

Rosenberg, Jay 1975 “Transcendental Arguments Revisited”, *The Journal of Philosophy*, vol. 72, no. 18, Oct. 23, pp. 611-624.

Stroud, Barry 1977 “Transcendental Arguments and ‘Epistemological Naturalism’, *Philosophical Studies* 31; in Barry Stroud *Understanding Human Knowledge*, Oxford University Press 2000, pp. 71-82.

Thornton, Tim 1998 *Wittgenstein on Language and Thought. The Philosophy of Content*, Edinburgh University Press.

Wenzel, Christian Helmut 2005 “Spielen nach Kant die Kategorien schon bei der Wahrnehmung eine Rolle? Peter Rohs und John McDowell”, *Kant-Studien* 96, 4/2005, pp. 407-426.

From Topology to Logic.
The Neural Reduction of Compositional Representation

Markus Werning, Düsseldorf, Germany

When we look at the structure of thought, what we find is logic. No matter what our starting point is: the semantic analysis of linguistic expressions, the psychology of cognition, or a philosophical theory of reasoning, we usually arrive at some variant or extension of first order predicate logic that characterizes the underlying structure of thought. However, when we look at the cortex, what we find is topology. The functional role of neurons is determined by topological neighborhood relations. Given that the various kinds of neurons are by and large homogenously distributed over the cortex, the major difference in the functional role of neurons is grounded in which neurons are connected to each other, and which are not. In topological terms: Who's in the neighborhood of whom. If we presuppose the materialist assumption that the cortex is what brings about thought, any reductive explanation has to show how the logical structure of thought is necessitated by the topological organization of information in the cortex.

Unfortunately, classical textbook mathematics is of little help here, even though there are some theorems that link topology to logic. Stone's representation theorem, e.g., famously asserts the duality between the category of Boolean algebras and the category of totally disconnected compact Hausdorff spaces. We thus know how propositional logics is to be represented topologically. When our primary interest is in thought, though, first order logic rather than propositional logic ought to be our main concern. For, only first order logic (and its extensions) provides the means to represent and categorize objects. However, when it comes to first order logic, the mathematical links to topology are sparse. Building on previous work, this paper provides an explanation of how the topological organization of the cortex yields a structure expressible by (some intuitionist variant of) first order logic. The explanatory bridges are the Gestalt principles of perception and the physiological principles governing object-related neural synchronization.

The Composition of Thought

The view of thought I appeal to characterizes the triangle between language, mind and world roughly as follows: Linguistic utterances are expressions of meaning. Meanings are mental representations. More specifically, the meanings of sentences are thoughts composed of concepts by logical connectives. Concepts again have an external content and this content is responsible for an utterance having reference or denotation. The relation between concepts and their content is some relation of covariation – a causal-informational relation of sorts (Fodor, 1992). The denotation of an utterance is identical to (or otherwise determined by) the content of the concept the utterance is an expression of. This view is captured by our first hypothesis:

> **Hypothesis 1** (Covariation with Content). An expression has the denotation it has because the concept it expresses reliably co-varies with a content that is identical to the expression's denotation.

Since languages and fore and foremost first order languages have a rich constituent structure, it is rather plausible to assume that the structure of their meanings is complex, too, and that the structure of meanings in some way or another resembles the structure of their expressions. Now, the most simple way to spell out this relation of resemblance is by means of a structural match, in technical terms: a homomorphism. This homomorphism is spelled out by the principle of the compositionality of meaning:

> **Hypothesis 2** (Compositionality of Meaning). The meaning of a complex expression is a syntax-dependent function of the meanings of its syntactic constituents.

It would be surprising, furthermore, if the covariation relations between primitive concepts and their contents should not in some way or another contribute to the covariation relations between complex concepts and their contents. The quest for simplicity again leads us to the hypotheses that the contents of the primitive concepts are the sole factors to determine the content of a therefrom combined complex concept. Again, this is just what the principle of compositionality says for contents:

> **Hypothesis 3** (Compositionality of Content). The content of a complex concept is a structure-dependent function of the contents of its constituent concepts.

The aim of this paper is to make out a neuronal structure that fulfills the three hypotheses. The neuronal structure shall consist of a set of neuronal states and a set of thereon defined operations. Since the three hypotheses may serve as (minimal) identity criteria for concepts, their fulfillment by a neuronal structure will justify us in identifying the neuronal structure with a structure of concepts. The three hypotheses hence form the adequacy conditions for a neuronal reduction of concepts.

The Topology of the Cortex

For many attributes (color, orientation, size, speed, etc.) involved in perceptual processing one can anatomically identify cortical correlates. Those areas often exhibit a twofold topological structure and justify the notion of a feature map: (i) a receptor topology (e.g, retinotopy in vision, somatotopy in touch): neighboring regions of neurons code for neighboring regions of the receptive field; and (ii) a feature topology: neighboring regions of neurons code for similar attribute values. Due to physiological facts, this twofold functional topology is reflected in the topography (the physical distance relations) of the cortex.

With regard to the monkey, more than 30 cortical areas forming feature maps are experimentally known for vision alone (Felleman & van Essen, 1991). In fact, the majority view among neuroscientists now is that the cortical processing of vision below hippocampus is entirely organized in the topological way described above. The attributes involved can be very complex, though. Fig. 1 shows a number of neural maps that relate to perceptual attributes.

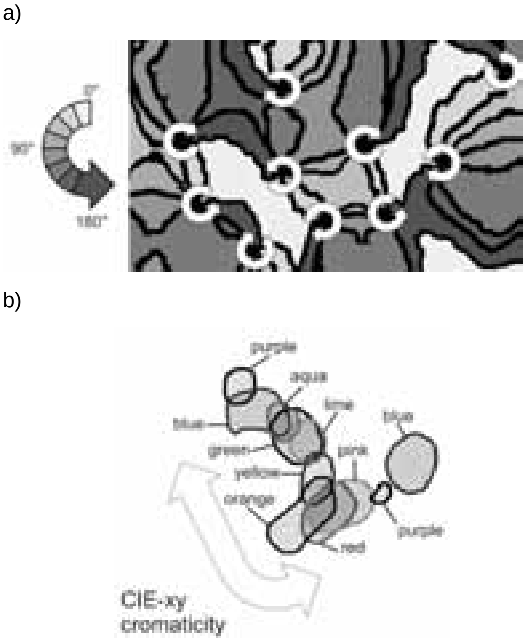


Figure 1: Cortical realizations of attributes. a) Fragment (ca. 4mm^2) of the neural feature map for the attribute orientation of cat V1 (adapted from Crair et. al., 1997). The arrows indicate the polar topology of the orientation values represented within each hypercolumn. Hypercolumns are arranged in a retinotopic topology. b) Color band (ca. 1 mm^2) from the thin stripes of macaque V2 (adapted from Xiao et. al., 2003). The values of the attribute color are arranged in a topology that follows the similarity of hue as defined by the Commission Internationale de l'Eclairages (xy-cromaticity). The topology among the various color bands of V2 is retinotopic.

The fact that values of different attributes may be instantiated by the same object, but are processed in distinct regions of cortex poses the problem of how this information is integrated in an object-specific way: the binding problem. How can it be that the color and the orientation of an object are represented in distinct regions of cortex, but still are part of the representation of one and the same object? A prominent and experimentally well supported solution postulates oscillatory neural synchronization as a mechanism of binding: Clusters of neurons that are indicative for different attribute values sometimes show synchronous oscillatory activity, but only when the values indicated are instantiated by the same object in the perceptual field; otherwise they are firing asynchronously. Synchronous oscillation, thus, might be regarded to fulfill the task of binding together various representations of attibute values to form the representation of an object with these values (Singer, 1999, for review). Using oscillatory networks as biologically motivated models, it can be demonstrated how the topological organization of information in the cortex, by mechanisms of synchronization, may yield a logically structured semantics of concepts (Maye & Werning, 2004). Oscillation functions play the role of object concepts. Clusters of feature sensitive neurons play the role of attributive concepts – or predicates.

Oscillatory Networks

From Gestalt psychology the principles governing object concepts are well known. According to some of the Gestalt principles, spatially proximal elements with similar attribute values (e.g., similar color/similar orientation) are likely to be perceived as one object or, in other words, represented by one and the same object concept. The Gestalt principles are implemented in oscillatory networks by the following mechanism: Oscillators that select input from proximal stimulus elements with like attribute values tend to synchronize, while oscillators that select input from proximal stimulus elements with unlike values (e.g., red and green for color or horizontal and vertical for orientiation) tend to desynchronize. As a consequence, oscillators selective for proximal stimulus elements with like values tend to exhibit synchronous oscillation functions when stimulated simultaneously. The oscillation in question can be regarded as one object concept. In contrast, inputs that contain proximal elements with unlike values tend to cause anti-synchronous oscillations, i.e., different object concepts.

In our model (Fig. 2) a single oscillator – marked as a cubicle – renders the statistical electrical discharge behavior of 100 to 200 biological cells and codes for an attibute value (*z*-coordinate) for a stimulus in the relevant region of the receptive field (*x*,*y*-coordinates). Differential equations describe the dynamics of the *i*-th oscillator as the temporal evolution of a variable $x_i(t)$. The oscillators for an attribute are arranged on a three-dimensional grid forming a module. Two dimensions represent the spatial domain, while the attribute values are encoded by the third dimension. Thus the twofold topology of biological feature maps is reflected in the network architecture. Spatially close oscillators that represent similar values synchronize. The desynchronizing connections establish a phase lag between different groups of synchronously oscillating clusters. Modules for different attributes can be combined by establishing synchronizing connections between oscillators of different modules in case they code for the same stimulus region

Stimulated oscillatory networks (e.g., by stimulus of Fig. 3a), characteristically, show object-specific patterns of synchronized and de-synchronized oscillators within and across modules. Oscillators that represent attributes of the same object synchronize, while oscillators that represent attributes of different objects de-synchronize. We observe that for each represented object a certain oscillation spreads through the network. The oscillation pertains only to oscillators that represent the attributes of the object in question.

Semantic Interpretation

An oscillation function $x(t)$ of an oscillator is its excitatory activity as a function of time during a time window $[0,T]$. Mathematically speaking, activity functions can be conceived of as vectors in the Hilbert space $L_2[0,T]$ of functions that are square-integrable in the interval $[0,T]$. Thus, a precise measure of synchrony can be established and a powerful algebraic framework for the semantic interpretation of the network is provided. The degree of synchrony between two oscillations lies between −1 and +1 and can be defined as the their normalized inner product

$$\Delta(x,x') = \frac{\langle x|x'\rangle}{\sqrt{\langle x|x\rangle\langle x'|x'\rangle}} \; . \qquad (1)$$

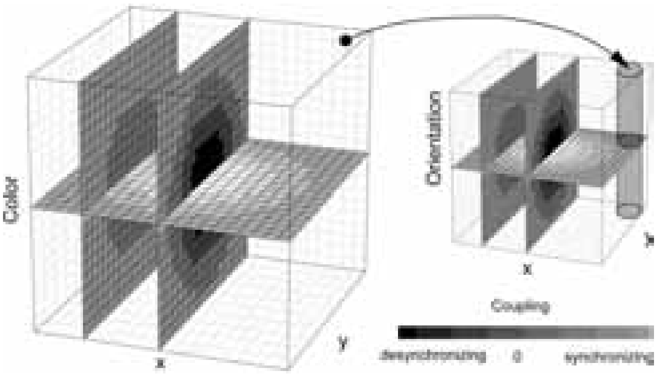


Figure 2: Oscillatory network. The network topology reflects the receptor topology (xy-plane) and the feature topology (z-axis) of the neural maps. Each module realizes one attribute. The layers in each module realize the attribute values. Oscillators activated by neighboring stimulus elements with similar attribute values synchronize (light gray). Oscillators activated by neighboring stimulus elements with unlike attribute values desynchronize (dark gray). The layers of different modules are connected in a synchronizing way that respects the common receptor topology.

The dynamics of complex systems is often governed by a few dominating states, the eigenmodes. The corresponding eigenvalues designate how much of the dynamics is accounted for by that mode. The two stable eigenmodes of a stimulated network are shown in Fig. 3b. Only the stable eigenmodes, i.e. those eigenmodes whose characteristic function does not converge to zero, will be of concern to us when it comes to semantic interpretation. The overall dynamics of the network is given by the Cartesian vector $\mathbf{x}(t) = (x_1(t), ..., x_k(t))^T$. The network state at any instant is considered as a superposition of the temporally constant, but spatially variant eigenvectors $\mathbf{v}_i$ weighted by the corresponding spatially invariant, but temporally evolving characteristic functions $c_i(t)$ of Fig. 3c:

$$\mathbf{x}(t) = \sum c_i(t)\mathbf{v}_i \ . \qquad (2)$$

The eigenmodes, for any stimulus, can be ordered along their eigenvalues λ_i so that each eigenmode can be signified by a natural number i beginning with 1 for the strongest.

The Hilbert space analysis allows us to interpret the dynamics of oscillatory networks in semantic terms. Since oscillation functions reliably co-vary with objects, they may be assigned to some of the individual terms a,b, ... ∈ Ind of a predicate language by the partial function

$$\alpha : \mathrm{Ind} \to L_2[0,T] \ . \qquad (3)$$

The sentence a = b expresses a representational state of the system (i.e., the representation of the identity of the objects denoted by the individual terms a and b) to the degree the oscillation functions α(a) and α(b) of the system are synchronous. The degree to which a sentence φ expresses a representational state of the system, for any eigenmode i, can be measured by the value $d_i(\varphi) \in [-1,+1]$. In case of identity sentences we have:

$$d_i(a = b) = \Delta(\alpha(a), \alpha(b)) . \qquad (4)$$

When we take a closer look at the eigenvector of the first eigenmode in Fig. 3b, we see that most of the vector components are exactly zero (gray shading). However, few components in the greenness and the horizontality layers are positive (light shading) and few components in the redness and the verticality layers are negative (dark shading). We may interpret this by saying that the first eigenmode represents two objects as distinct from one another. The representation of the first object is the characteristic function $+c_1(t)$ and the representation of the second object is its mirror image $-c_1(t)$ (Because of the normalization of the Δ-function, only the signs of the eigenvector components matter). These considerations justify the following evaluation of non-identity:

$$d_i(\neg a = b) = \begin{cases} +1 \text{ if } d_i(a = b) = -1 , \\ -1 \text{ if } d_i(a = b) > -1 . \end{cases} \qquad (5)$$

Feature layers function as representations of attribute values and thus can be expressed by predicates F_1, ...,F_p, i.e., to every predicate F a diagonal matrix β(F) ∈ $\{0,1\}^{k\times k}$ can be assigned such that, by multiplication with any eigenvector $\mathbf{v}_i$, the matrix renders the sub-vector of those components that belong to the feature layer expressed by F. To determine to which degree an oscillation function assigned to an individual constant a pertains to the feature layer assigned to a predicate F, we have to compute how synchronous it maximally is with one of the oscillations in the feature layer. We are, in other words, justified to evaluate the degree to which a predicative sentence Fa (read: 'a is F', e.g., 'This object is red') expresses a representational state of our system, with respect to the eigenmode i, in the following way (the f_j are the components of the vector $\mathbf{f}$):

$$d_i(Fa) = \max\{\Delta(\alpha(a), f_j) \mid \mathbf{f} = c_i(t)\beta(F)\mathbf{v}_i\} \ . \qquad (6)$$

If one, furthermore, evaluates the conjunction of two sentences by the minimum of the value of each conjunct, we may regard the first eigenvector $\mathbf{v}_1$ of the network dynamics resulting from the stimulus in Fig. 3a as a representation expressible by the sentence

> *This is a red vertical object and that is a green horizontal object.*

We only have to assign the individual terms *this* (=a) and *that* (=b) to the oscillatory functions $-c_1(t)$ and $+c_1(t)$, respectively, and the predicates *red* (=R), *green* (=G), *vertical* (=V) and *horizontal* (=H) to the redness, greenness, verticality and horizontality layers as their neuronal meanings. Simple computation then reveals:

$$d_1(Ra \wedge Va \wedge Gb \wedge Hb \wedge \neg a = b) = 1 \ . \qquad (7)$$

Using further methods of formal semantics, the semantic evaluation has been extended to sentences comprising disjunction (∨), implication (→), and existential as well as universal quantifiers (∃, ∀). Co-variation with content can always be achieved if the individual assignment α and the predicate assignment β are chosen to match the network's perceptual capabilities. Theorems that prove the compositionality of meaning and content have been provided (Werning, 2005). Werning (2003) extends this approach from an ontology of objects to an ontology of events.

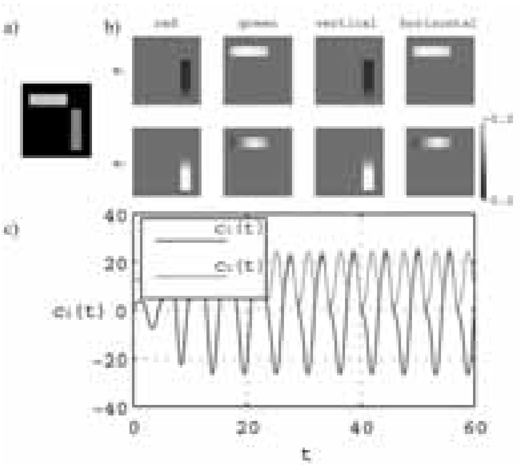


Figure 3: a) Stimulus: one vertical red bar and one horizontal green bar. It was presented to a network with 32×32×4 oscillators. b) The two stable eigenmodes. The eigenvectors $\mathbf{v}_1$ and $\mathbf{v}_2$ are shown each in one line. The four columns correspond to the four feature layers. Dark shading signifies negative, gray zero and light shading positive components. c) The characteristic functions for the two eigenmodes.

Conclusion

Oscillatory networks show how a structure of the cortex can be analyzed so that elements of this structure can be identified with mental concepts. These cortical states can be regarded as the thoughts expressed by some first order logic. They form a compositional semantics for such a logic. The cortical states can themselves be evaluated compositionally with respect to external content and thus provide denotations. The approach formulated in this paper is biologically rather well-founded. It is supported by a rich number of neurophysiological and psycho-physical data and is underpinned by various computer simulations. The eigenmode analysis of the network enables the reduction of the logical structure we encounter in thought to the the topological organization of the cortex.

Literature

Crair, M. C., Ruthazer, E. S., Gillespie, D. C., & Stryker, M. P. (1997). Ocular dominance peaks at pinwheel center singularities of the orientation map in cat visual cortex. *Journal of Neurophysiology*, 77(6), 3381–5.

Felleman, D. J., & van Essen, D. C. (1991). Distributed hierarchical processing in the primate cerebral cortex. *Cerebral Cortex*, 1, 1–47.

Fodor, J. (1992). *A theory of content and other essays*. Cambridge, MA: MIT Press.

Maye, A., & Werning, M. (2004). Temporal binding of non-uniform objects. *Neurocomputing*, 58–60, 941–8.

Singer, W. (1999). Neuronal synchrony: A versatile code for the definition of relations? *Neuron*, 24, 49–65.

Werning, M. (2003). Ventral vs. dorsal pathway: the source of the semantic object/event and the syntactic noun/verb distinction. *Behavioral and Brain Sciences*, 26(3), 299–300.

Werning, M. (2005). The temporal dimension of thought: Cortical foundations of predicative representation. *Synthese*, 146(1/2), 203–24.

Xiao, Y., Wang, Y., & Felleman, D. J. (2003). A spatially organized representation of colour in macaque cortical area V2. *Nature*, 421(6922), 535–9.

The Calculus of Inductive Constructions as a Foundation for Semantics

Piotr Wilkin, Warsaw, Poland

Since Frege started his work on formalizing natural language semantics, the framework for those semantics has been set-based. First, most theories have been based on type theory, later replaced by Zermelo-Fraenkel's set theory. Today, this paradigm is rarely questioned and even if some attempts are made to circumvent ZFC, for example to eliminate some paradoxes in quantified modal logic (Oksanen 1999), most of the development is done within that framework. This raises two major problems: set theory is structuralistic and extensional in nature. The latter needs no explanation, the former can be explained this way: in set theory, one expresses structures, not functionality. This is well seen in the definition of the ordered pair – in ZFC, one defines <a,b> as {{a},{a,b}} and only later proves that this construct satisfies the *properties* we require of the ordered pair – namely, to be able to construct an ordered pair from two elements and to retrieve, respectively, the first or second element.

What I propose here is to switch to a certain intuitionistic setting to express semantics, namely, the Calculus of Inductive Constructions. I am aware that the term "intuitionistic" will make some people, and especially philosophers, wary of the idea, since this instantly suggests a connection with Brouwer's philosophical concept of intuitionism as a philosophy of mathematics. However, I believe that while intuitionism might be questionable as a framework for mathematics, it is well-suited as a framework for natural-language semantics. This is because the subjects of both disciplines are fundamentally different. With natural-language semantics, we want to describe not an idealized area such as mathematics, but a very real one – how people use the language. While in mathematics it is often advisable to abstract away from the specific means of obtaining some result (namely, the calculation), it is hardly plausible to suggest that we can do the same with language expressions, as there are clearly some cases (as in the case of intensional contexts or, more specifically, *propositional attitudes*) where we cannot abstract from the way a specific human being reconstructs a language expression.

From Frege's times, the accepted and widespread form of a logical assertion has been $C \vdash \varphi$, where C stands for the *context* and φ represents the expression to be asserted (in the given context). This is true for provability, as well as for logical consequence in a model (obviously, the two relations differ in nature, but a common characteristic is shared – they are both binary relations between a context and an evaluated expression, whether the context be a model or a theory). However, a counterproposal to this form can be submitted: $C \vdash \varphi : t$, which reads: in context C, the expression φ is of type t. This is called a *typing judgement*. The concept of types dates back to the theory of types from *Principia Mathematica*, but this specific idea owes its roots to Church's theory of simple types (Church 1940), which was based on the lambda calculus and proposed as an alternative.

Once we enter the area of semantics, one can see the usefulness of the typing assertion over the traditional logical one. In semantics, we are often interested not only in information about true sentences, we also want to know whether a given expression is a properly constructed sentence or whether some expression is actually a definite description. Indeed, various solutions like the Montague grammar (Montague 1974) exploit this concept with the use of semantic categories, which is a concept parallel to the one by Church but designed specifically for natural language analysis in the early XX century by Polish logicians Stanislaw Lesniewski and Kazimierz Ajdukiewicz (Ajdukiewicz 1935). However, the interesting modification here lies not only in the adding of the type, but also in modifying the way we read the expression. This was first noticed by the logician Haskell Curry (Curry 1934), who recognized that in case of simply typed combinators (an analogon to Church's simple type system), if we only assign types to basic lambda terms, the proper types correspond exactly to tautologies of a minimal (consisting only of the implication, without the other connectives) intuitionistic propositional calculus. Thus, for example, the K combinator ($\lambda x \lambda y.x$) becomes a proof of the known tautology $\alpha \rightarrow (\beta \rightarrow \alpha)$. How so? According to the functional interpretation (known as the Brouwer-Heyting-Kolmogorov interpretation (Sorensen and Urzyczyn 2006)), a proof of implication is a function transforming a proof of the LHS to a proof of the RHS. So in this case, we want the function in question to take an argument being a proof of α, then return us an argument which itself is a function from β to α. The K combinator is exactly that – it takes an x (of type α), then returns us $\lambda y.x$, which is itself a function – that takes an y (of type β) and returns the previously taken x (of type α).

This approach, while generally interesting, is hardly revolutionary. However, this has to be coupled with a different idea – Church designed his simple type system not with this intuitionistic reading in mind, but to avoid a form of Russell's Paradox for his untyped lambda calculus, which is represented by the Omega term: $(\lambda x.xx)(\lambda x.xx)$ – a self-applying function applied to itself will never reduce to a normal form. It can be easily proven that any term that can be typed in the simple type system has the SN (strong normalization) property – any path of reductions of that term will terminate. The question raised by Curry's discovery was – can one construct a more advanced logic than the one represented by the simple type system, but one that is still based only on strongly normalizable terms? We now know that the answer is positive, and highly so: one can represent higher-order intuitionistic logic using this sort type system, and add inductively defined predicates to the logic to make it easier to formalize certain concepts. One such system is the Calculus of Inductive Constructions (CIC), which is the theoretical basis for an interactive proof assistant program named Coq (Herbelin et al. 2006). For a sample of the expressive power of the system, one can point to the formalized version of the four-color theorem proof – since the proof itself is a tangible mathematical object, this removes the need to "believe" a computer-based check of each of the hundreds of cases – one only has to believe the theory behind the prover, namely, the CIC itself. For a more logically interesting example, a (constructive) proof of Gödel's incompleteness theorem is also available (O'Connor 2005). Indeed, the expressive power of CIC is not far from that of ZFC.

There is not enough space here to describe in full extent the theory behind CIC. Suffice it to say, one very important distinction between CIC and the simple-type theory (STT) is that there is no longer any clear distinction between a term and a type. Therefore, we can have both $\vdash \varphi : 2+2 = 4$ and $\vdash 2+2 = 4 : \mathrm{Prop}$ - the first one states that φ is a proof of 2 + 2 = 4, the other – that 2 + 2 = 4 is a well-formed proposition. The type hierarchy in CIC is infinite and stratified, with two base types: *Set* being the type of object domains and *Prop* being the type of propositions. One example of a proposition we have already seen, an example of a *Set* typing assertion is and $\vdash$ Nat : Set, saying that the natural numbers form an object domain (I did not want to use the terms "set" or "universe" since they seem to have various connotations, but since the term "set" is used as the type designation internally in Coq, I will adopt this and use the name "set" as short for "object domain" – this is however not to be understood as a set in the set-theoretic sense). Another important difference is that apart from non-dependent products (which can be understood as functions and thus, via the Curry-Howard isomorphism, implications), we now have dependent products (which can be thought of as generalized products and via the Curry-Howard isomorphism – represented as universal quantification). For example, the assertion $\vdash \varphi : (\forall n{:}nat(n \geq 0))$ can be read as "φ is a member of a generalized product over natural numbers of domains "n $\geq$ 0", which for each n contain proofs that n $\geq$ 0".

Presenting this entire concept would have been pointless without showing why this can be useful for semantics. Indeed, I can believe that using this framework can significantly increase the clarity of describing non-extensional contexts in semantic uses. For one quick interesting demonstration, note that this approach allows us to formally explain the difference between Frege's notion of "function" and its "value-range". In CIC one cannot generally deduce $\vdash f = g$ from $\vdash \forall x.fx = gx$; extensionality is not assumed. However, this does not mean that we use some vague notion of identity – functions are constructive objects and they are the same if the constructions themselves are the same.

Now, here the inductive types come in. Inductive types are generally an extension of the intuitive notion of inductive constructions to the type system. That is: you provide ways of constructing base elements of the type, as well as inductively constructing new elements from previously constructed ones. What the type system guarantees is that the construction is unique – objects of an inductive type cannot be created in any other way than by the respective constructors and the various constructors are assumed to be provably different from each other. For example, the type Nat of natural numbers is defined as *Nat* := *O* : *Nat* | *S* : *Nat* → *Nat.* This corresponds to the traditional view from Peano Arithmetic: natural numbers are either 0 or an iterated successor of 0. They are also only that: for example, you can prove in the system (with just the above definition) that 0 is different from any successor.

We can use all of the mentioned instruments to illustrate how this framework can help us in formalizing semantics. For example, we can construct the operator *Believes* in the following way: *Believes* := *BArith* : ∀n:nat, (n = n) → *Believes*(n = n). This inductive definition means that what is "believed" is arithmetical equalities, but nothing else. Note that in this definition, both the truth value *and* the form of the expression is used – we accept only *proven* expressions, and only of the form n = n. We could widen the operator, for example, to ignore the truth value: take a definition of *Believes* := *BArith* : (∀n m:nat),*Believes*(n = m). This is a correct definition, but it no longer takes into account the logical validity of the equality. Now, we "believe" any arithmetical equality, true or not.

Note that we are not taking a model-theoretic approach here – there is no need to construct a specific intensional *model* to formalize such a semantics. We gain the benefit of a framework especially suited for analyzing intensional constructions, but one that already possesses all the "standard" logical concepts built in. One can even add the axiom of excluded middle if needed – it is consistent with CIC. There is, however, no need for the tedious task of building a specific model for formalizing an intensional fragment of the natural language – one needs only to define the proper inductive types and predicates. All this is coupled with a system for annotating the syntax much richer than standard categorial grammars – for example, we can now assign a type to the operator *very* as follows: (∀s:*Set*),(s → *Prop*) → (s → *Prop*), which allows us to uniformly type *very* for predicates of any object category (as in *very young* with *young* : *LivingBeing* → *Prop* and *very fast* with *fast* : *Car* → *Prop*, so that we can both subdivide the object category and not worry about having problems with uniformly assigning categories to higher order operators).

This proof-theoretic (or constructive, as you prefer) approach to natural-language semantics seems to me more suitable. After all, what we are after when formalizing a natural language is being able to express all the constructs from the natural language easily and not having to worry whether adding this or that construct will complicate the resulting model beyond manageability. Also, it places more emphasis on the functional way of our language use. I believe that the presented framework might be a viable and reasonable alternative to current set-theoretic based approaches.

Literature

Ajdukiewicz, Kazimierz 1935 *Die syntaktische Konnexität*, in *Studia Philosophica*, I, 1935, pp. 1-27

Church, Alonzo 1940 *A Formulation of the Simple Theory of Types*, in *The Journal of Symbolic Logic*, Vol. 5, No. 2 (Jun., 1940), pp. 56-68

Curry, Haskell 1934 *Functionality in Combinatory Logic,* in *Proceedings of the National Academy of Sciences*, vol. 20, pp. 584-590

Herbelin, Hugo et al. 2006 *The Coq Reference Manual. Chapter 4: The Calculus of Inductive Constructions*, http://coq.inria.fr/doc/Reference-Manual006.html

Montague, Richard 1974 *The Proper Treatment of Quantification in Ordinary English*, in R. Thomason (Ed.), *Formal Philosophy: Selected Papers of Richard Montague.* New Haven: Yale University Press.

O'Connor, Russell 2005 *Essential Incompleteness of Arithmetic Verified by Coq*, in *Proceedings of the 18th International Conference on Theorem Proving in Higher Order Logics (TPHOLs 2005)*

Oksanen, Mika 1999 *The Russell-Kaplan Paradox and Other Modal Paradoxes: A New Solution*, in *Nordic Journal of Philosophical Logic,* Vol. 4, No. 1, pp. 73-93

Sorensen, Martin Heine and Urzyczyn, Paweł 2006 *Lectures on the Curry-Howard Isomorphism*, Elsevier

The Four-Color Theorem, Testimony and the A Priori

Kai-Yee Wong, Hong Kong, China

1. Computer proof as experiment

The computer played an essential role in providing a key lemma, which requires certain combinatorial checks too long to do by hand, in the proof given by Kenneth Appel and Wolfgang Hanken of the Four-Color Theorem (4CT). The proof of the 4CT has generated a flurry of philosophical discussions about its significance. Some of them focused on the arguments put forth by Thomas Tymoczko 1979. According to Tymoczko, the argument for the 4CT is like 'an argument in theoretical physics where a long argument can suggest a key experiment which is carried out and used to complete the argument' (Tymoczko 1979: 78) because there is an unavoidable reliance on computers to produce the proof of the theorem. Since belief in the reliability of computers ultimately rests on empirical considerations, the proof establishes the 4CT on grounds that are in part empirical. So the 4CT, Tymoczko concludes, is a substantial piece of knowledge which can be known only a posteriori. Almost three decades on, the issues raised by Tymoczko's paper are still of great interest to many. (For a recent discussion see, see Arkoudas and Bringsjord 2007. See also Brown 1999, Burge 1998, Coady 1992: ch. 14.) In this paper I shall examine the central thread in Tymoczko's reasoning and Michael Detlefsen and Mark Luker's (1980) contention that it in fact leads, rightly, to the much more drastic conclusion that mathematical proof is *typically* empirical.

Tymoczko argues that the appeal to computers, in the case of the 4CT, involves a claim of this kind:

(1) If computer C running program P produces result R, then mathematical statement M is true.

The truth of (1) turns on the reliability of C and P, which in turn involves the claim that

(2) C does what it is supposed to do, namely, to correctly execute P.

Evidently (2) may turn out true or false, depending on a complex set of empirical factors and thus its truth is 'ultimately a matter for engineering and physics to assess' (Tymoczko 1979: 74). Mathematical knowledge obtained by computer proof is therefore grounded on a 'well-conceived computer experiment'. So, Tymoczko concludes, accepting computer-assisted proof means giving up the traditional notion of mathematical proof as something surveyable, non-empirical and a priori.

2. Proof, archives, and testimony

Expressing this argument in terms of *testimony* can help us articulate a problem for Tymoczko. What he argues for is that since testimony is empirical in character and the proof of the 4CT is based partly on the testimony of a computer, accepting the 4CT as a theorem requires changing the concept of proof.

Indeed, it seems arguable that testimony has a significant role in the transmission of mathematical knowledge and this in turn may have interesting bearing on the notion of proof. Mathematics investigation, of course, builds on past results. Now suppose Kodel, a mathematician, has produced with great care a proof with gaps filled by a set of established, 'archived' results. It is implausible to think that Kodel's proof therefore *fails* to be a proof in the proper sense, otherwise much of mathematics as we know it would go down the drain because mathematical investigation has always involved appeals to, so to speak, testimony of archived results. Reliance on such testimony is part and parcel of a mathematician's work. Thus, in a crucial respect, Kodel's proof, or for that matter, much of traditional mathematical investigation is relevantly similar to the proof of the 4CT in their reliance on testimony. If we adopt Tymoczko's view, we must say that the use of testimony injected into Kodel's proof empirical ingredients and rendered the theorem proved a piece of a posteriori knowledge, just as the 4CT is. Notice whether or not Kodel can survey the archived proofs has no bearing on this point. To see this, we can imagine that another mathematician, Podel, produced a proof with gaps filled by a vast set of archived results. Surveying the enormously complex and numerous proofs for these results is such a mammoth task that no mathematician can finish. (If this sounds far-fetched, consider the now accepted proof of the classification of all simple finite groups. The proof, carried out over many years by a large number of mathematicians, spans across about 15,000 pages on journals. This example is from Brown (1999), p. 158). The fact that these proofs have been surveyed in the past will not change the situation. For without a first-hand survey of each proof, Podel must rely on empirical evidence that the archives are a reliable source of mathematical results. It is *arbitrary* to claim that Kodel's theorem does not rely on any empirical evidence because his proof is not as *extensive* as Podel's in its appeal to the archives.

If so, one should find puzzling if not inconsistent Tymoczko's claim that the appeal to computers in the 4CT forces a change of the traditional concept of proof. For if this claim is to be true, it must be the case that mathematical proof had not come to include empirical elements *before* the proof of the 4CT. But by dint of reasoning similar to Tymoczko's own, as shown by the case of Kodel and Podel, it should be held, albeit in my view implausibly, that traditional mathematics is mostly empirical.

Yet one may think that such a conclusion is not implausible. As a matter of fact, some have argued that much of traditional mathematics is partially empirical. Detlefsen and Luker (1980) argue that if one follows the logic of Tymoczko's reasoning where it leads, one is forced to see empirical ingredients in proofs that are generally held to be paradigms of a priori mathematical arguments, such as the following one attributed to the young Gauss. Gauss proved that the sum of the first one hundred positive numbers was 5,050 in the following way: write down the following pairs: <1, 100>, <2, 99>, …, <50, 51>, observe that the numbers of each pair together make 101 and that there are 50 pairs, and then conclude that the sum of the first one hundred numbers is 5,050. According to Detlefsen and Luker, the episode of calculation by which we determinate that 101 multiplied by 50 is 5,050 is needed in the reasoning from the 'observation' to the 'conclusion'. Among the assumptions, they add, required for confidence in the result of any computation is the following:

> (3) The computing agent correctly executes the program.

Since the validity of (3) rests on evidence of an empirical sort, the reasoning embodied by Gauss's proof must be seen as *utilizing empirical premises*. Furthermore, they argue, (3) is analogous (2); so, Gauss's proof is straightly analogous to the 4CT in its appeal to experience. They then conclude that Tymoczko's reasoning, which they endorse, forces us to hold that traditional mathematics typically utilizes empirical evidence. Let's call the form of reasoning behind Tymoczko's and Detlefsen and Luker's views the *Empirical-Reliability-of-X Principle* (**ERXP**):

> If X is something we must rely on in order to see the correctness of a proof, then the empirical evidence for the reliability of X is part of the proof, or, in slightly different terms, the proof must be seem as utilizing empirical premises.

On this principle, no knowledge acquired on the basis of a proof obtained, partly or fully, through testimony of any kind can be a priori because the empirical evidence regarding the reliability of the source of testimony would have to be part of the evidence for what is claimed to be known. The problem with ERXP is that it commits us to a highly restrictive account of a priori proof, and more seriously, demonstrative reasoning *in general*. (Though someone like Chisholm would say that such an account is the correct one, see Chisholm 1989, 28-30). For in any inference involving more than a few steps, we need to rely on our memory and perception to store and retrieve premises. But the reliability of our memory and perception, and ultimately the proper functioning of our brain, is not amenable to demonstration by reason alone. One would then have to conclude that mathematical argument and demonstrative reasoning typically utilize empirical premises and are thus a posteriori. Why would Tymoczko and others take ERXP seriously, despite its implausible consequences? They seem to think that we cannot but hold that the appeal to computers must force our warrant for the 4CT to be empirical because there is no way for an agent to obtain in an a priori way any knowledge through testimony. This, however, shows that ERXP is ill-motivated because, as I shall argue, with a proper explication of the a priori we can claim that it is possible to obtain a priori knowledge through the testimony of a rational source. Space does not allow me to argue for the rational source of testimony in computer proof (but see Tyler 1998 and Arkoudas & Bringsjord 2003). Instead I shall outline an account of the a priori which shows that the empirical character of a testimony *need* not render the knowledge obtained through it empirical.

3. Apriority and two roles for experience

The key idea here is a broad construal of the *non-justificatory role* that experience is held to play in knowledge and cognition. The standard view distinguishes a non-justificatory role from a justificatory role of experience. We are familiar with the idea that sense experience is needed to enable us to acquire the concept of, say, *triangle*, before we can be said to know a priori that a triangle has three sides. In general, sense experience needed for the acquisition of a concept does not thereby become part of any warrant one may have that would make a belief expressed with the concept knowledge. I think the non-justificatory role of experience should not be limited to the acquisition of concepts. A non-justificatory role for experience in this broader sense I shall call an 'enabling role'. Burge's discussions of computer proof and content preservation illustrate very well how one can have a very broad construal of the enabling role of experience provided that one is prepared to bring into question a whole lot of widely held opinions about rationality, content preservation, and testimony. (His 'Acceptance Principle' is a case in point, see Burge 1993: 469, see also Burge 1998).

To further explain the notion of an enabling role of experience, let us consider the appeal to memory. It is often the case that memory constitutes a part of the warrant of a remembered belief, as in the case in which I remember vividly that I was convinced by someone's demonstration of a certain logical statement *p* (call this episode of remembering M). Contrast this with the case where I remember a certain logical statement *q* and utilize it to fill a gap in a proof for a certain conclusion *c* (call this episode M*). In the first case, as Steffen Borge puts it (Borge 2003: 109), my cognitive attention is not merely focused on *p* but also on the *attitude* of my earlier self towards the proposition. Here M makes a substantive contribution to the justification of a remembered belief. In the second case, my memory serves not to supply propositions about particular mental events. Rather its role is to supply for the derivation a certain step, *q*, that is part of the demonstration that entitles me to believe *c*. *c* is underwritten by the demonstration consisting of, among other steps, *q*. The memory that supplies *q*, M*, is not what the demonstration is about, neither does it make any substantive contribution to the warrant provided by the demonstration. My belief that *c* is warranted because I have proved it. My warrant needs no further *justificational forces* to be supplied by M*. The role of M* is only enabling, not justificatory. It serves to give me *access* to that warrant, but not part of my entitlement to *c*. Typically, in a complex demonstration (a proof or a deduction) memory is called upon to play an enabling role. Contingent propositions about what we happen to remember do not *thereby* become part of the premises or make a substantive contribution to the warrant provided by these premises. The use of paper and pencil extends our means of checking of longer deductions. (See Teller 1980 for a related idea.) So I think precisely in the case of hand calculation the same reasoning can be used to argue for the *mere* enabling role of sense experience and perceptual belief. In the context of computer proof we can see the appeal to computers as a yet further extension of the means of checking and executing a formally sound algorithm which serves mainly to enable our access to the a priori warrant for a mathematical result. One might say that whereas the justification provided by a first-hand proof is *not* dependent on experience *essentially*, the justification provided by the proof of the 4CT is dependent on experience essentially because it is too long for any human to check. I can agree with this if 'essentially' is used here just to mark out the complexity of the computer's proof. But the epistemic significance of 'essential dependence on experience' in this sense can be doubted. Given that it makes epistemic sense to see computers as a way to extend our means of checking proofs, it is difficult to see why the kind of dependence on experience involved in a warrant provided by a computer proof should be accorded special epistemic significance merely because of the complexity involved. Since the relevant notion of apriority pertains to the independence from experience in its justificatory, not enabling, role, the above considerations, if correct, allows us to hold that the appeal to computers does not force the warrant for relying on the testimony of a computer to be essentially dependent on experience. Of course, a detailed account of the a priori on the basis of a broad conception of the enabling role of experience would be beyond the scope of this paper.

Literature

Arkoudas, Konstantineand Selmer Bringsjord 2007. 'Computers, Justification, and Mathematical Knowledge.' *Minds and Machines,* 17, 185-202.

Borge, Steffen 2003. 'The Word of Others.' *Journal of Applied Logic,* 1, 107-118.

Brown, James R. 1999. *Philosophy of Mathematics:An Introduction to the World of Proofs and Pictures,* Routledge..

Burge, Tyler 1998. 'Computer Proof, A Priori Knowledge, and Other Minds.' *Philosophical Perspectives,* 12, 1-37.

Burge, Tyler 1993. 'Content Preservation.' *Philosophical Review,* Vol. 102, No. 4, 457-488.

Chisholm, Roderick M 1980. *Theory of Knowledge*, 3rd ed., Prentice Hall.

Coady, C. A. J. 1992. *Testimony,* Clarendon Press, Oxford

Detlefsen, Michael and Mark Luker (1980). 'The Four-Color Theorem and Mathematical Proof.' *Journal of Philosophy* 77, 803-820.

Teller, Paul 1980. 'Computer Proof.' *Journal of Philosophy* 77, 797-802.

Tymoczko, Thomas 1979. 'The Four-Color Problem and its Philosophical Significance.' *Journal of Philosophy,* 76, 57-83.

The Comprehension Principle and Arithmetic in Fuzzy Logic

Shunsuke Yatabe, Toyonaka, Osaka, Japan

1 Introduction

Shapiro and Weir showed that Hume's principle has a finite model (so it can not deduce Peano arithmetic **PA**) in Aristotelian logic which does not admit the existence of the empty property [Shapiro, Weir 2000]. This shows that Hume's principle itself is not enough to develop arithmetic, and we should have a certain theory of properties to realize Frege's program [Hale 2008]. In this paper, we investigate a theory of property which satisfies what Myhill called *Frege's principle* [Myhill 1984], that "every formula with one free variable determines a property (not a set) which holds of all those and only those things which satisfy the formula", and we examine how much arithmetic we can develop by it.

It is well-known that it is impossible in classical logic: the comprehension principle, $\forall x[x \in \{y : P(y)\} \equiv P(x)]$ for any P, implies a contradiction by the Russell paradox. However, in non-classical logic, the situation is slightly different. Many first order logics, including fuzzy logic, have been known to imply no contradiction when the comprehension principle is assumed. So we can regard any singular term as an object in such logic (and this plays an important role in Fregean Platonism). Therefore, as Myhill, we can say:

> So if we want to take Frege's principle seriously, we must begin to look at some kind of non-classical logic.

Let us consider the case of the set theory **H** with the comprehension principle in Lukasiewicz infinite-valued predicate logic $\forall L$. It is a version of fuzzy logic, and is a non-classical logic weaker than classical logic which only has a weak fragment of the contraction rule. White showed that **H** is consistent [White 1979], and it is known as the strongest theory among set theories with the comprehension principle. We highlight two special features of sets (or properties) in **H** which might provide a clue of analysis: non-extensionality and full circularity.

First, the basic law **V** does not hold in **H**. Let us remember the case of classical logic: Frege's basic law **V**,

$$(\forall P)(\forall Q)[ext(P) = ext(Q) \equiv (\forall x)[P(x) \equiv Q(x)]]$$

and the definition of membership relation imply the comprehension principle and it implies a contradiction. Neo-Fregeans modified the basic law **V** and adopted *Restricted-V* (**RV**) when they developed a set theory in classical logic [Shapiro 2003]. In **H**, the basic law **V** is equivalent to the axiom of extensionality (which insists the extensional equality is equivalent to the Leibniz equality), and it implies a contradiction (it is called Grisin's paradox [Grisin 1982]). Furthermore, any version of **RV** has not been known to be consistent to **H** yet.

Second, **H** forgives circular definitions. It is because **H** proves a *general form of recursive definition*, which is a circular definition of a very strong shape, is permitted:

$$(\forall x)(\exists z)[x \in z \equiv \phi(x,z)]$$

for any formula ϕ [Cantini 2003]. Here, we define a set z by using z itself. Since the recursive definition is an essence of computation, a certain amount of arithmetic can be developed: we can define a graph of any recursive function in **H**. However, the arithmetic developed in H is not a conservative extension of **PA**: the mathematical induction scheme implies a contradiction in **H** [Yatabe 2007]. In fact, we can show, in any model of **H**, the sentence which can be interpreted as "ω contains a non-standard natural number" is truth-value 1.

On a final note we remark about an adaptation of non-classical logic. As for an antecedent, Dummett lapsed classical logic for his own philosophical purposes, anti-realism. In this sense, adherence to classical logic is not what is should be. Furthermore, this suggests that, we may argue that some rule (the law of excluded middle in his case) of classical logic corresponds a certain philosophical viewpoint (as realism). We might ask what kind of a philosophical assumption corresponds to the contraction rule as a future task[1].

2 The set theory H, extensionality and the basic law V

The comprehension principle will derive a contradiction in classical logic. However, the contraction rule is essential to derive it. Grisin proved that the comprehension principle derives no contradiction in the system *Grisin logic* which is classical logic minus the contraction rule [Grisin 1982].

So, the next question is, where the limit is: what is the strongest logic, between Girisin logic and classical logic, which does not derive a contradiction? Currently, the strongest logic is known to be Lukasiewicz infinite-valued predicate logic $\forall L$ [White 1979]. This system is known to be impossible to recursively axiomatize, so we introduce the definition of its model (because of the luck of space, here we introduce the quite informal one: we note that, this is a simplification of $([0,1], *, \Rightarrow, 0, 1)$-structure where $([0,1], *, \Rightarrow, 0, 1)$ forms a standard MV algebra [Hajek 2001]).

(1) The truth value set is [0, 1] of real numbers (it is a kind of fuzzy logic).

(2) $\|\bot\| = 0$, $\|\phi \to \varphi\| = \min(1,1 - \|\phi\| + \|\varphi\|)$

(3) $\|(\forall x)\varphi(x)\| = \inf\{\|\varphi(a)\| : a \in |M|\}$

The rest connectives are defined by using $\to,\bot$ (for example $\neg A$ is $A \to \bot$ and $A \wedge B$ is $\neg(A \to \neg(A \to B))$). We note that, we have the multiplicative conjunction $\otimes$ in $\forall L$: $\|A \otimes B\| = \max\{0, \|A\| + \|B\| - 1\}$. It is easy to see that $A \to A \otimes A$ is truth value 1 if and only if A is truth value 0 or 1. We write $A \equiv B$ instead of $(A \to B) \otimes (B \to A)$.

Let **H** be a set theory whose only axiom scheme is the comprehension principle (i.e. $\forall x[x \in \{y : P(y)\} \equiv P(x)]$ is truth value 1 for any P in any its model). We introduce two kinds of equality in **H**.

Leibniz equality: $x = y$ if and only if $(\forall z)[x \in z \equiv y \in z]$,

Extensional equality: $x =_{ext} y$ if and only if $(\forall z)[z \in x \equiv z \in y]$

[1] To answer this, we note that the contraction rule plays an essential role to imply a contradiction not only in the Russell paradox but also in the liar paradox and in the sorites paradox. Therefore we seem to need a unified framework to analyze them.

Clearly $x = y \to x =_{ext} y$ holds, but the converse does not hold in **H**.

Theorem 1 *The axiom of extensionality,* $(\forall x, y)[x = y \equiv x =_{ext} y]$ *, does not hold in* **H***.*

For the proof, see [Hajek 2005]. Here, we introduce the outline: this is by the following lemma.

Lemma 1 H *proves that Leibniz equality is a crisp relation.*

We note that formula *P*(*x*) is *crisp* (**Crisp**(*P*)) if, for any object *a*, the truth value of *P*(*a*) is either 0 or 1. Since $=_{ext}$ can have a fuzzy truth value for fuzzy sets in **H**, = and $=_{ext}$ are different.

As for the basic law **V**, it implies that = is not the Leibniz equality. For **V** defines = as the extensional equality, and if = is Leibniz equality then the axiom of extensionality holds, and this implies a contradiction in **H**. However, Frege's intension seems to define = as Leibniz equality. In this sense the basic law **V** does not hold in **H**.

In this proof, we imply a contradiction when $x =_{ext} y$ has a fuzzy truth value: this is possible when not less than one of *x* and *y* is a fuzzy set. As for another paradox, it is known that we can imply a contradiction if we assume the empty set $\phi = \{x : \perp\}$ satisfies the extensionality axiom [Cantini 2003]. So it has been unknown that the following scheme is consistent:

$$((\{x : P(x)\} \neq_{ext} \phi) \wedge Crisp(P)) \wedge ((\{x : Q(x)\} \neq_{ext} \phi) \wedge Crisp(Q))$$
$$\to [\{x : P(x)\} = \{y : Q(y)\} \equiv (\forall x)[P(x) \equiv Q(x)]]$$

for any formula *P*(*x*)*, Q*(*x*). This means that, the following law might hold (this is a version of the **RV** [Shapiro 2003]):

$$\forall P \forall Q(Good(P) \wedge Good(Q)) \to [ext(P) = ext(Q) \equiv (\forall x)[P(x) \equiv Q(x)]]$$

Therefore if this is consistent, we can think that **RV** gives an implicit definition of crisp sets, badness means fuzziness, and any fuzzy set can be regarded to represent indefinitely extensibility.

3 Circularity and arithmetic without the induction scheme

When we mention the formalization of arithmetic, we often come to axiom systems with the induction scheme as **PA**, but it is not a unique way. We also have type systems which are widely used in computer science. For example, Gödel's system **T**[2] is a simple type theory [Girard et al. 1989]. **T** has two types**, *Int*** (integers) and ***Bool*** (booleans). As for ***Int***, it has two type constructors, 0 (constant symbol) and *s* : ***Int*** → ***Int*** (successor function). And **T** does not have the induction scheme. Instead, it has a *recursion operator* **R** for recursive definition whose type is $R : U \to (U \to Int \to U) \to U \to U$ for any type U. It satisfies $Ruv0 = u$ and $Ruv(sn) = v(Ruvn)n$ (if we substitute ***U*** for ***Int***). This operator enables us to have the primitive recursion on integer numbers: for example, the addition x + y is defined by $Rx(\lambda z^{Int}.\lambda z'^{Int}.sz)y$. For, we can calculate as follows:

$$x + 0 \mapsto Rx(\lambda z^{Int}.\lambda z^{Int}.sz)0 \mapsto x$$
$$x + (sy) \mapsto Rx(\lambda z^{Int}.\lambda z^{Int}.sz)(x + y)y \mapsto s(x + y)$$

Here $t \mapsto s$ represents that s is a result of the computation whose input is *t*. So, the above computations show that $x + 0 = x$ and $x + (sy) = s(x + y)$ hold. In this way, we can represent primitive recursive functionals in **T** without using the induction scheme.

The arithmetic developed in **H** is very similar to the system **T** in that they do not have the induction scheme. **H** allows the circular definition[3] (as in section 1), and it enables us to use the general form of the recursive definition as **R**. Here we introduce three points (for more details, see [Cantini 2003][Hajek 2005]). First we can define ordinal numbers in Zermelo style: $0 = \phi$ and $sn = \{x : x = n\}$ for any finite ordinal *n*. It is easy to see, we can define Fregean term "the number of the conception *P*" (*NxPx*) by the comprehension principle if *P* is crisp and finite. Second, the set *ω* of all finite ordinals can be defined as

$$\omega =_{ext} \{x : x = 0 \vee (\exists y \in \omega)[x = sy]\}$$

(Because of the luck of extensionality, we can not require the uniqueness of *ω*). Third, any recursive function's graph can be defined. In other words, any partial recursive function is numerically representable in **H**

Let us give an example of the generalized recursive definition in **H**. For example, we can define the graph *P* of the addition as follows[4]:

$$\langle x, 0, z\rangle \in P \Leftrightarrow x = z,$$
$$\langle x, sy, sz\rangle \in P \Leftrightarrow \langle x, y, z\rangle \in P$$

Both $x + 0 = x$ and $x + (sy) = s(x + y)$ are guaranteed by very simple way. However, we do not know *P* satisfies the following conditions:

1. *P* is a crisp relation,
2. *P* defines a function *p*(*x, y*) = *z* (i.e. $(\forall x, y)[P(x, y, z) \wedge P(x, y, z') \to z = z']$),
3. *p*(*x, y*) is a total function.

If *x* and *y* are standard natural numbers, then we can calculate the unique value *z* satisfying *P*(*x, y, z*). However we have a trouble when one of *x, y* is a non-standard natural number: *P*(*x, y, z*) might be a fuzzy truth value. We neither know, whether *ω* or any graph defined by recursion becomes crisp or not. If *ω* is fuzzy, then we might think this is another expression of Dummett's "*ω* is indefinitely extensible".

H develops a fair degree of arithmetic; however it can not deduce Peano arithmetic **PA**. In fact, the following theorem holds.

Theorem 2 H proves that the induction scheme implies inconsistency.

This means that **H** is *ω*-inconsistent: in any model of **H**, the sentence which can be interpreted as "*ω* contains a non-standard natural number" is truth-value 1. For more details, see [Yatabe 2007].

Contrary, Girard showed that the weak version of mathematical induction scheme is provable in **LAST**, a set theory with the comprehension principle in light linear logic [Girard 1998] [Terui 2004]. In **LAST**, the definition of natural numbers is quite different: such definition seems to enable to prove the weak induction.

[2] We note that Gödel's original system **T** has the induction scheme.

[3] Meanwhile, the comprehension the principle can be thought as a special case of the recursive definition. Since it is the foundation placed in the calculation, so we had better to say that it is the generalized recursive definition that is essential principle in this theory.

[4] For the definition of the ordered pair in **H** , see [Cantini 2003].

Let us summarize: **H** is not a conservative extension of **PA**. The induction scheme implies a contradiction in **H** nevertheless any partial recursive function is numerically representable in **H**. This is a non-standard arithmetic, different to **PA**, but it itself is a generalization based on the concept of recursion, and that is not ad hoc.

4 Conclusion

We investigated a theory of property which satisfies what Myhill called *Frege's principle*, and we examine how much arithmetic we can develop by it. It is known that, in many logics, the comprehension principle (which represents Frege's principle) does not imply a contradiction. We concentrated the case of the set theory **H**, in Lukasiewicz infinite-valued predicate logic $\forall L$, which is known as the strongest theory among set theories with the comprehension principle.

We pointed out two features of sets in **H**: non-extensionality and full circularity. First, the basic law **V** does not hold in **H** and any versions of **RV** has not been known to be consistent to **H** yet. Second, **H** forgives circular definitions. It is because **H** proves a *general form of recursive definition*, and a certain amount of arithmetic can be developed: we can define a graph of any recursive function in **H**. However, the mathematical induction scheme leads to a contradiction in **H**, so the arithmetic developed in H is not a conservative extension of **PA**.

These results showed that we do not know about arithmetic developed by the comprehension principle enough. The problem how we can develop mathematics in **H** seems to be interesting enough from the perspective of the analysis of the broad sense of Fregean intent.

Literature

[Cantini 2003] Cantini, A. 2003 “The undecidability of Grisin's set theory”, Studia logica 74, 345-368.

[Girard 1998] Girard, J.-Y. 1998 “Light Linear Logic”, Information and Computation 143.

[Girard et al. 1989] Girard, J.-Y, Taylor, P, Lafont Y. 1989 Proofs and Types, Cambridge: Cambridge University Press

[Grisin 1982] Grisin, V. N. 1982 “Predicate and set-theoretic caliculi based on logic without contractions”, Math. USSR Izvestija 18, 41-59.

[Hajek 2001] Hajek, Petr 2001 Metamathematics of Fuzzy Logic, Dordrecht: Kluwer academic publishers

[Hajek 2005] Hajek, Petr 2005 “On arithmetic in the Cantor-Lukasiewicz fuzzy set theory”, Archive for Mathematical Logic. 44(6) 763 - 82.

[Hale 2008] Hale, Bob 2008 “The problem of Mathematical Ojects” Talks in Kyoto University, March 24, 2008.

[Myhill 1984] Myhill, John 1984 “Paradoxes”, Synthese 60, 129-43

[Shapiro 2003] Shapiro, Stewart 2003 “Prolegomenon to Any Future Neo-logicist Set Theory: Abstraction and Indefinite Extensibility” British Journal of Philosophy of Science 54, 59-91.

[Shapiro, Weir 2000] Shapiro, Stewart. Weir, Alan. 2000 ““Neo-Logicist” logic is not epistemically innocent” Philosophia Mathematica 8, 160-89.

[Terui 2004] Terui, Kazushige 2004 “Light Affine Set Theory: A Naive Set Theory of Polynomial Time”, Studia Logica, 77, 9-40.

[White 1979] White, Richard B. 1979 “The consistency of the axiom of comprehension in the infinitevalued predicate logic of Lukasiewicz” Journal of Philosophical Logic 8, 509-534.

[Yatabe 2007] Yatabe, Shunsuke 2007 “Distinguishing non-standard natural numbers in a set theory within Lukasiewicz logic” Archive for Mathematical Logic 46, 281-287.

Intentional Fundamentalism

Petri Ylikoski / Jaakko Kuorikoski, Helsinki, Finland

1. What is Intentional Fundamentalism?

In the social sciences, most debates about reductionism are related to methodological individualism. There is a wide variety of arguments for this doctrine. This paper discusses one assumption often associated with methodological individualism. We call this assumption intentional fundamentalism. According to intentional fundamentalism, the proper level of explanation in social science is the level of intentional action of individual agents. Intentional fundamentalist assumes that explanations given at the level of individual action are especially satisfactory, fundamental or even ultimate. In contrast to explanations that refer to supra-individual social structures, properties or mechanism, there is no need to provide micro-foundations for intentional explanations. French social theorist Raymond Boudon (1998, 177) expresses this idea clearly: ”When a sociological phenomenon is made the outcome of individual reasons, one does not need to ask further questions”. The idea is that in the case of supra-individual explanations there is always a black box that has to be opened before the explanation is acceptable, but in the case if intentional explanation there is no problem of black boxes: ” … the explanation is final” (Boudon 1998, 172).

We argue that the thesis of intentional fundamentalism is not acceptable. Although intentional fundamentalism can take various forms, we will discuss it only in relation to rational choice theory (RCT). We claim that the special status of (rational) intentional explanation is not compatible with the causal mechanistic account of explanation supported by many intentional fundamentalists. We will also make the case that the explanatory regress arguments presented against the possibility of non-individualistic social explanations rely on mistaken assumptions about the nature of explanation.

It is important to recognize the limited scope of our arguments. We are not arguing against the legitimacy of intentional explanations nor are we presenting a wholesale argument against the RCT. We only argue against a special explanatory status given to explanations expressed in terms of rational intentional action. We do not dispute the heuristic usefulness and practical necessity of explanations in terms of folk psychology. Adequate intentional explanations are fully legitimate causal explanations – they just do not have a privileged status. Furthermore, our argument should not be read as an argument against the idea that there exists a division of labor between the social sciences and psychology and neurosciences. This idea is unobjectionable as long as intentional fundamentalism is not the only argument supporting it.

Our argument will proceed as follows. First we will show the connection between intentional fundamentalism and the explanatory regress argument for methodological individualism. We argue that the proper understanding of the mechanistic account of explanation does not lead to an explanatory regress. If the regress does not exist, the intentional fundamentalism loses its main motivation. In latter part of the paper we will attempt to explain the appeal of intentional fundamentalism by showing how it is based on overemphasis on one particular dimension of explanatory power: cognitive salience. Intentional fundamentalists assume that different dimension of explanatory power go hand in hand when in fact there are important trade-offs between them. These trade-offs provide the basis for the argument that intentional fundamentalism is a hindrance for the search of causal explanations in the social sciences.

2. The Regress of Explanations-Argument

Methodological individualists usually argue for their position by making the case that non-individualist explanations are either explanatorily deficient or not explanatory at all. At most, the individualists allow that explanations referring to macro-social facts are placeholders for proper (individualistic) explanatory factors. The explanatory contribution of supra-individual explanations is at-best derived: they are explanatory because they are (in principle) backed up by a truly explanatory story. This is the regress of explanations argument: unless grounded at the lower level, explanations at macro level are not acceptable. The underlying general principle [P] is the following: a genuine explanation of X by Y requires that Y is itself explained or is self-explanatory. The explanatory buck has to stop somewhere.

In principle this argument is general, and it raises the possibility that the regress would continue until the level of fundamental physical particles. This would be highly unintuitive, but for the intentional fundamentalist the buck stops at the level of (self-interested) rational intentional action. This level is treated as inherently understandable, as shown in the above quotations from Boudon. The special status of intentional explanation makes the explanatory regress argument safe for methodological individualist: he can use it with its full force against non-idividualist, but it does not challenge the legitimacy of his favorite explanatory patterns. In contrast, structural explanations seem suspect as they do not have a similar privileged status. The inherent intelligibility of intentional action explains why the reductivist search for microfoundations should stop at the level of the individual.

In our view, this argument fails. First, the explanatory regress argument does not work as the methodological individualist assumes. Second, intentional explanation does not have the special properties the intentional fundamentalist presumes it to have.

The regress argument does not work because the presupposed principle [P] is not true. The explanatory relation between X and Y is independent from the question whether Y is itself explained. Belief in [P] might arise from the confusion of between justification-seeking and explanation-seeking why-questions. However, it might also arise from other (false) intuitions about explanation. One such intuition is that total understanding is not really increased if the puzzlement concerning the original *explanandum* is simply replaced with new puzzlement concerning the *explanans*. Some fifty years ago, Stephen Toulmin appealed to this intuition to argue that all sciences must presuppose an inherently understandable “ideal of natural order” (Toulmin 1961, 42). For methodological individualists the rational (self-interested) action serves in the role of ideal of natural order (Coleman & Fararo 1992, xiv). The same intuition underlines Michael Friedman’s argument that explanatory understanding cannot be based on a local dyadic relation between the *explanandum* and the *explanans*. Instead, understanding has to be a global property of the whole belief system. (Friedman 1974; Schurz and Lambert 1994)

This global conception of understanding is based on an attribution error. The fact that an explanation fits well into a systematic web of beliefs usually makes it easier to understand, but it is not what constitutes understanding. A more natural way to interpret understanding is to regard it as an ability to answer counterfactual what-if-things-had-been-different-questions. (Woodward 2003, 191; Ylikoski 2008). Explanations show what made a difference to the *explanandum*. Here understanding is a local affair: understanding is attributed based on ability make correct counterfactual inferences about the phenomenon. It does not matter for the understanding of a particular phenomenon whether the things that made a difference to it are themselves explained. That is a different, partially independent question. There is no need of a fundament.

Most intentional fundamentalists are in principle committed to the causal-mechanistic conception of explanation: they are trying to explain the properties of social wholes by laying out the causally relevant properties of their parts (individuals) and the pattern of their interaction. However, the causal-mechanistic conception of explanation and the global, fundamentalist conception of understanding do not fit together: there is no reason to expect that the putative fundament would always pick out the causally relevant factors. First, in many cases, the causally relevant factors in RCT-models are actually structural or institutional factors. Individual behavior may be institutionally constrained so as to be rational only in an “as if” –sense, or the *explanandum* may be robust with respect to the behavioral assumptions. The details of intentional level can thus be explanatorily irrelevant. (Satz and Ferejohn 1994; Lehtinen and Kuorikoski 2007) Second, reconstructing the behavior in terms of rational intentional action might misdescribe the actual causes of behavior. A growing body of social psychological research on behavior priming and post-hoc rationalization supports this claim (Wilson 2004). Hence, intentional description can be misleading. This is not the age-old argument that, on purely conceptual grounds, intentional explanation cannot be causal explanation, but an empirical claim about the relationship between folk-psychology and human behavior.

As Carl Craver (2007, Ch. 4) has pointed out, although the search for mechanistic explanations involves opening black boxes, complete mechanistic explanations are not really eliminative, because they incorporate factors in multiple different levels of mechanisms. The goal of mechanistic explanation is not to reduce different levels to single bedrock, but to understand the systematic dependencies between entities in different levels of organization. Neither does mechanistic explanation regard the lower levels as somehow sacrosanct: lower level concepts and conceptions may be altered in the face of new empirical discoveries or conceptual changes concerning upper (or even lower) level phenomena. These are bad news for the intentional fundamentalist.

3. The Dimensions of Explanatory Power

We have argued elsewhere (Ylikoski & Kuorikoski 2008) that the intuitive notion of explanatory power is related to five distinct dimensions that can be in conflict with each other. The five dimensions of explanatory goodness are non-sensitivity, precision, factual accuracy, degree of integration, and cognitive salience. All other often cited criteria for evaluating explanations, like simplicity, unification or mechanistic detail, are derivate of these basic dimensions. Could some of the appeal of intentional fundamentalism be explained in terms of these dimensions?

The appeal of intentional fundamentalism derives from the cognitive salience of our folk psychological practice. Cognitive salience refers to the ease with which the reasoning behind the explanation can be followed, how easily the implications of the explanation can be seen and how easy it is to evaluate the scope of the explanation and identify possible defeaters or caveats. To the extent that RCT belongs to the family of folk psychological theories, and once its technical concepts are internalized, people can be extremely fluent in its use. It only requires the translation of everyday folk psychological accounts to the more abstract language of beliefs and preferences. The ease of use and the broad scope of application of intentional vocabulary give rise to the impression of a strong theory. If RCT is a formalization of our everyday intentional explanatory practice, the feeling that intentional explanations are final and do not give rise to additional question is easily explained: in our everyday explanatory practice we do not usually attempt to go beyond the intentional scheme of explanation. In fact, people have difficulties in figuring out what these explanations would look like. The mistake here is to assume that cognitive salience is a reliable indicator of overall explanatory power. What the intentional fundamentalist does not recognize is the fact that in the case of RCT there are important trade-offs between the dimensions of explanatory power. To show this, we will next consider how well intentional rational explanations measure up against some other dimensions.

Let us start with the non-sensitivity. It refers to the sensitivity of the explanatory relationship to the changes in the background conditions. The less sensitive the explanation is to these changes, more robust and powerful it is. A robust explanation can answer a wider set of what-if-things-had-been-different -questions. If we focus on individual explanations, rational choice explanations can be extremely sensitive in that sometimes a small change in the beliefs or desires of the agent can bring about a drastic change in behavior. However, sometimes these explanations can also be robust, so everything depends on the kinds of changes in background assumptions we are talking about. Therefore it does not make sense to make generalizations about the sensitivity of rational choice explanations. However, one might be tempted to make such generalizations if one considers the robustness of RCT as an explanatory scheme. Like folk psychology in general, RCT provides extremely flexible vocabulary for describing behavior and for revising explanatory accounts. If one intentional description of behavior is shown to be false, it can always be replaced with a new one that incorporates the problematic observations. This flexibility gives an impression of a general and strong explanatory theory, although these are really properties of the explanatory vocabulary, not of the substantial theory. Explanatory power, understood as non-sensitivity, is in this case illusory.

The second dimension to be considered is factual accuracy. An explanation with less idealizations and abstractions is usually judged to be better than one with more. However, factual accuracy is often in conflict with the other dimensions of explanatory power: it is difficult to have an explanation that is both factually accurate, non-sensitive, integrated with other explanations and cognitively salient. The technical precision of RCT concepts can give rise to an impression of precise descriptions of psychological phenomena, but in reality it does not score highly in terms of factual accuracy. Not only does RCT significantly abstract away from the details of psychological phenomena, it also quite often factually distorts the descriptions of these phenomena. For example, the RCT reconstruction of psychological process can give a causally misleading account of it by describing as purposive behavior what was habitual or unconscious, or rationalizing decision-making processes that were everything but rational.

The final dimension we want to consider is the degree of integration with existing knowledge. It contributes to explanatory understanding by expanding the set of explanatory questions that can be answered by them separately. As RCT is just a formalization of everyday folk psychology, it is relatively well integrated with it. However, it is an open question how much this expands their joint explanatory reach. RCT fares even worse with respect to integration with other scientific theories. The difficulties of integrating RCT with the results of empirical research in psychology and social sciences are well known. When RCT is further strengthened with intentional fundamentalism, the problems become even greater. As suggested above, the ideas underlying intentional fundamentalism are not compatible with causal mechanistic explanatory ideas that motivate other sciences. Largely because of the very idea of a fundament, the fundamentalist RCT has extremely low degree of integration with other bodies of scientific knowledge.

Literature

Boudon, Raymond 1998 “Social mechanisms without black boxes”, in Hedström, Peter and Richard Swedberg (eds.) *Social mechanisms: an analytical approach to social theory*, Cambridge: Cambridge University Press, 172-203.

Coleman, James and Thomas Fararo 1992 “Introduction”, in *Rational Choice Theory. Advocacy and Critique*, New York: Sage, ix-xxii.

Craver, Carl 2007 *Explaining the Brain: Mechanisms and the Mosaic Unity of Neuroscience*, New York and Oxford: Clarendon Press.

Friedman, Michael 1974 “Explanation and Scientific Understanding”, *Journal of Philosophy* 71, 5-19.

Lehtinen, Aki and Jaakko Kuorikoski 2007 “Unrealistic Assumptions in Rational Choice Theory”, *Philosophy of the Social Sciences* 37, 115-137.

Satz, Debra and John Ferejohn 1994 “Rational Choice and Social Theory”, *Journal of Philosophy* 91, 71-87.

Schurz, Gerhard and Lambert, Karel 1994 “Outline of a Theory of Scientific Understanding”, *Synthese* 101, 65-120.

Toulmin, Stephen 1961 *Foresight and Understanding*, London: Hutchinson & Co.

Wilson, Timothy 2004 *Strangers to Ourselves*, Belknap Press.

Woodward, James 2003 *Making Things Happen. A Theory of Causal Explanation*, Oxford: Oxford University Press.

Ylikoski Petri 2008 “The illusion of depth of understanding in science”, forthcoming in De Regt, Sabinelli & Eigner (eds.) *Scientific Understanding: Philosophical Perspectives*. Pittsburgh University Press.

Ylikoski, Petri and Jaakko Kuorikoski 2008 “Dissecting Explanatory Power”, under review.

New Hope for Non-Reductive Physicalism

Julie Yoo, Easton, Pennsylvania, USA

The Problem

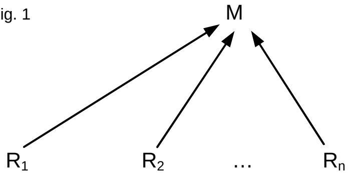
Non-reductive physicalists want to hold on to the idea that higher-level properties – mental properties, especially – have an autonomous ontological standing, and hence, their own distinctive causal powers. But their equal commitment to physicalism threatens to undermine this commitment to the autonomy of higher-level properties. This is, in effect, Kim's dilemma for non-reductive physicalism: it is inherently unstable because the physicalism denies the irreducibility thesis, while the irreducibility thesis denies physicalism (Kim 1989, 1993, 1998, 2005). In this paper, I shall to propose a novel way of solving this dilemma.

The Way Out of the Bind: Reducing Higher-Level Properties While Retaining their Powers

My argument, in a nutshell, is this. A property exists if and only it can confer causal powers upon its bearers. Properties are individuated in terms of the distinctive range of causal powers each of them are capable of conferring. Now, mental properties confer causal powers completely in virtue of the physical properties on which they logically supervene, but no mental property is identical with a physical property. The concept of *property realization* renders possible the consistent conjunction of these two strands – irreducibility and physicalism.

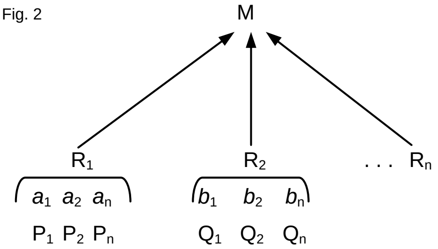
Let us get started with a picture of multiple realizability to get us oriented with respect to its sister notion of property realization. On standard accounts of multiple realizability, it is believed that different physical properties – different physical types of kinds – P_1, P_2, … P_n, can each necessitate one and the same type of mental M (or higher-level) property:

Fig. 1



While this picture isn't wrong, the schematic nature of the diagram engenders an oversimplified way of thinking about realization that is misleading. It is misleading in that it gives the impression that a given realization R is a *single* physical property. But this convenient simplification obscures an important detail, much attended to by the emergentists, to their credit, and it is that a mental property is a property of a system made up of concrete aggregates and whose micro-physical properties and relations make the operations of the whole system or mechanism possible. To say that a property is multiply realizable is to say that a number of different kinds of systems or mechanisms – ones that are not only made up of different kinds of materials, but that also have different kinds of architectural configurations – can execute one and the same function. Think of the indispensible corkscrew: there is the rack and pinion model, the waiter's corkscrew, and the rigid coiled wire with a convenient handle. Each of these different kinds of mechanisms counts as a distinct type of realization; that is, the rack and pinion model is R_1, the waiter's corkscrew is R_2, and so on. A more careful look at what's going on suggests that a concrete instance of a given realization is best represented as being made up of inter-related micro-physical aggregates a_1, a_2, …, a_n, having micro-physical properties P_1, P_2, …, P_n, whose causal powers "constitute" the causal powers of M:

Fig. 2



The realizing base of M is the micro-physical properties P_1, P_2, …, P_n of the constituent aggregates a_1, a_2, …, a_n, which they instantiate in virtue of being the type of material or substance they are.[1] Now, if physicalism is true, then each of these properties have causal powers that fully determine the causal powers of M. M, to use an emergentist term, is a *resultant* of each of the micro-physical properties P_1, P_2, … P_n. The question, then, is how M is capable of having its own *distinctive* causal powers, given that M's powers are fully derived from its realizing base. The answer lies in drawing a connection between two things: a causal power conception of the nature of properties (Shoemaker 1984), now often called the "Eleatic theory of properties,"[2] and the phenomenon of *universalizability*, as described by Robert Batterman (2000, 2001). I begin with universality.

To the extent that each mental property, by all appearances, has a unique array of causal powers (the total facts about which a complete empirical functionalist analysis of mental properties would deliver) it satisfies a necessary condition for being a distinctive genuine property. Now, one of the remarkable things about nature is that there are higher-level regularities that are realized by very heterogeneous mechanisms. That is, wildly different kinds of systems manage to accomplish the same kinds of tasks. This phenomenon is what Jerry Fodor 1997 has called a "metaphysical mystery" – how, in essence, things that are so different can still give rise to things that are the same:

[1] Actually, Fig. 2 may also be too schematic. Depending on how we want to individuate the *a*'s, it is possible for a given *a* to have many properties P, not just a single property in the way Fig. 2 represents. Take a particular corkscrew *x* and say that *x* is the rack and pinion kind. The individual *x* has the macro-property M of being a corkscrew in virtue of *x*'s being composed of two handle bars, a flange, spindle, metal spiral, etc.. These are the *a*'s that have the P's whose collective instantiation makes it possible for the *a*'s to constitute *x* and enable it to do its thing. Now, the handle bars alone would have the properties of being handle-bar shaped, being rigid, being made of metal, and so on, so a given *a* is most likely to have properties P. I won't be too fussy about this detail, as Fig. 2 is sufficient as a working model.)

[2] The Eleatic theory of properties is so called because the conception goes back to Plato's Eleatic stranger in the *Sophist*, who suggests that the mark of being is power. This idea is also codified in what Kim calls "Alexander's principle," named after the emergentist, Samuel Alexander, who argued that a property must have causal powers to exist.

> The very existence of the special sciences testifies to reliable macro-level regularities that are realized by mechanisms whose physical substance is quite typically heterogeneous. - - Damn near everything we know about the world suggests that unimaginably complicated to-ings and fro-ings of bits and pieces at the extreme *micro*-level manage somehow to converge on stable *macro*-level properties.
>
> On the other hand, the 'somehow' really is entirely mysterious, … [for we don't] see why there should be (how there could be) [macro-level regularities] unless, at a minimum, macro-level kinds are homogeneous in respect of their micro-level constitution. Which, however, functionalists in psychology, biology, geology, and elsewhere, keep claiming that they typically aren't. (Fodor 1997, pp. 160 – 61)

We see the mystery across the special sciences: an animal's fitness, for example, can be realized by its reproductive success, lack of predators, abundance of food, and the good fortune of not suffering odd accidents; pain, to use everyone's favorite example, can be realized by C-fiber stimulation, silicon chips, or hydraulic mechanisms – all at least in principle. With all the heterogeneity at this lower level, it is indeed mysterious how they manage to give rise to stable, recurring, patterns and regularities that appear to hold with nomological force. (The converse is equally mysterious, namely, how such few types of fundamental particles, laws, and forces, give rise to the complexity and heterogeneity of our world, but this is a slightly different matter that will not be addressed here.)

In his "Multiple Realizability and Universality," Batterman explains how the mysterious phenomenon, called "universality," is explained in physics. Multiple realizability is an instance of universal behavior. Universal behavior is where vastly different systems – systems with different microstructures, different architectures, different properties – exhibit identical behavior when characterized at some level of description:

> To begin to get an idea about the concept of universality as well as its ubiquity, consider the following homely example. One wants to explain the observed common behavior of pendulums – one wants, for example, to understand why pendulums with bobs of different colors, rods of different lengths, different masses composed of diverse materials, etc., all have periods (for small oscillations) that are directly proportional to the square root of the length of the rod from which the bob is hanging. At one level the explanation is quite straightforward: one solves the very simple equation of motion for such a system. - - But there is another why-question which is far from simple. Why, one might ask, are factors such as the color and (to a large degree) the constitution or micro-structural makeup of the bobs irrelevant for answering our why-question about the period of the pendulums? Why is this equation, rather than one in which, say, a color parameter plays a prominent role, explanatory? In other words, what allows us to bracket, or set aside as 'noise' these other features of the individual pendulums as inessential or irrelevant for the explanation of the behavior of interest? These latter questions concern the explanation of universal behavior.
> (Batterman 2000, pp. 120 – 121)

As Batterman explains, physics has managed to identify universal behavior in all kinds of systems – systems of thermodynamics near their critical points, certain limit theorems of probability – and construct very detailed explanations of those variables that are relevant for the behavior of a system's macro-level behavior (the length of the bob of the pendulum) and those that are mere negligible to the salient macro-level properties of the system (the color of a pendulum). The lesson to take away from this, according to Batterman, is that there are "*physical reasons why the details of the makeup of the individual realizers may be largely irrelevant for the upper level behavior of the system*" (Batterman's emphasis, 2000, p. 124).

Now, while it is interesting that some of explanatory methods are sophisticated enough to handle universal behavior, *our* question is how our world is *constructed* so that it can display universal behavior while being composed of such a diverse heterogeneity of more basic physical (micro-physical) constituents. The question is decidedly about the metaphysics of the phenomenon, the "truth-makers" in the world that make our explanations of universality true. On my view, we need to look to the Eleatic theory of properties to lay out that metaphysics.

On the Eleatic theory, properties are those things in virtue of which the objects having them can enter into causal relations. Thus, a property X is not the thing that *has* causal power K. Many people speak this way, but what they really mean is that a property *confers* causal powers. The things that *have* causal powers are individuals, like physical objects or events or other kinds of concrete particulars. A property, then, is individuated in terms of the unique array of causal powers it confers upon the individuals that have it. Its unique array is what I call its *causal profile* (see also Gillett 2002). The notion is drawn from two observations. First, that a property in isolation from other properties is not enough to confer its bearer with a causal power, as many other contributing properties must also be instantiated. A property, then, is one among many others that contribute to the causal power of an object:

(α) A property X contributes to a causal power K in a given circumstance just in case
  i. X is necessary for K, and
  ii. X, together with a set of properties Γ, is minimally sufficient to confer K.

Thus, the relationship between a property and a causal power is not one-to-one. And this is the second observation: it is, in fact, many-many. For any property X, its instantiation in different circumstances can confer different causal powers, and for each type of causal power, different individuals can have that type of causal power though it instantiates different properties.

(β) Possibly, for any property X, X_i contributes to K_i and X_j contributes to K_j and $X_i = X_j$, but $K_i \neq K_j$.

(χ) Possibly, for any causal power K, X_i contributes to K_i and X_j contributes to K_i and $K_i = K_j$ but $X_i \neq X_j$.

The individuation conditions for a property can be stated thus:

(δ) X and Y are the same property just in case they have exactly the same *causal profile*: for all actual and possible K_i and K_j, X contributes to K_i and Y contributes to K_j and $K_i = K_j$.

The basic idea of (δ) is that X and Y are the same property if under all possible circumstances – all possible sets of properties Γ with which they can be conjoined – they contribute to all and only the same causal powers. For instance, if the property of having heat contributes to various causal powers in various circumstances – melt wax, boil water, bake brownies, … – and the property of having a mean kinetic energy contributes to exactly the same causal powers in those same circumstances, then heat and mean kinetic energy are the same property. But given that mean kinetic energy is only one realization of heat among other realizers, this means that there is a circumstance where heat and mean kinetic energy contribute to different causal powers and thus count as different properties.

This theory about the nature of properties and their conditions for individuation gives us a very intuitive way of explaining the phenomenon of universality. On my view, the capacity to behave in universal ways is built into the profile of each physical property. That is, the things that make up the profile of a higher-level property are the profiles of many lower-level properties acting in concert. This is not emergentism. Let us return to the theological story about our cosmology to see why. A physicalist says that to create our world, God created physical particles, physical properties, and their governing laws, and nothing more. I would add that God also created ways for those physical properties to combine with each other in *universal-behaving ways*, ways that the special sciences are so adept at describing. These higher-level properties are genuine because higher-level predicates refer to entities that have unique causal profiles. The comparison with fractals may be instructive here: larger patterns, which are made up of smaller patterns, have properties that are unique to them. But those properties are entirely derived from or are “resultants” out of the properties of the smaller constituent patterns; it just so happens that the smaller patterns are constructed in such a way that they generate the larger patterns with their distinctive properties. So we get the higher-level properties and regularities because of the complex ways that physical systems can behave, and solely due to their physical nature.

Perhaps another example will be helpful here. A wall will have certain properties that its individual component bricks do not have. An obvious property will be its mass *m*. For the purposes of illustration, let’s say that all bricks have somewhat different masses. Now supposed that the wall were to engage in universal behavior with respect to its mass. Then bricks of different masses would regularly combine to create a wall with mass *m*, in a variety of different contexts (inside, outside, during the day, during the night, …), and in a number of different ways (the left side gets done first, then the right, or the layers get added one level at a time, or diagonally, …). They all lead to instances of *wall-of-mass-m*. This is quite remarkable, but there is nothing but physical ingredients and physical laws at work here. There need be no irreducible *wall-of-mass-m* property that imposes its causal powers from above. All the causal powers for the wall with its mass come strictly from the causal powers of the individual bricks and their mode of combination. Emergence doesn’t come into the picture. It is a part of the nature of these bricks to form an object with the property of being a *wall-of-mass-m*. But this is no more mysterious than the fact that it is a part of the nature of an electron to attract protons. Insofar as the property of being *wall-of-mass-m* makes a causal difference, and makes it in its uniquely distinctive ways, then by the theory of properties I have laid out, this is a genuine property, as genuine as the properties at the fundamental level.

The Eleatic theory is purely democratic when it comes to determining which properties exist. The level doesn’t matter. What matters is its profile. If my approach to the metaphysics of universality is right, then a whirl of many properties P_1, P_2, … P_n and their corresponding profiles can give rise to a “larger” stable property M and with its corresponding profile, all thanks to nothing but the nature of the properties P_1, P_2, … P_n.

Non-Reductive Physicalism and Downward Causation

One would be right to wonder whether my view preserves the original non-reductive physicalist conception of the world as having different “layers” or “levels” that are hierarchically arranged. On one way of looking at it, my view places everything in one grand level, so that there is no hierarchy of different levels that exist in a metaphysically robust way. Instead, all the entities and properties postulated by the sciences, special and micro-physical, are equally fundamental and mutually irreducible, living side by side, and running in and out of each other’s lives. This picture is consistent with my view. As long as it does not violate the supervenience relations between properties that belong to traditionally different levels – the biological supervening upon the chemical, and mental supervening upon the biological, and so on – it certainly does not force us to alter our ways of how the sciences are related to each other.

The important part of this proposal is that the supervening properties do indeed have their own causal powers, and hence, ultimately brings about physical changes when the supervening properties are instantiated. But it does so, not by exercising downward causation as on the classical emergentists view, but by constraining how the physical changes come about. The view, then, that I present, may better be described as an account of *downward determination*.

Literature

Alexander, S. (1920). *Space, Time, and Deity*, 2 vols. London: Macmillan.

Batterman, R. (2000), “Multiple Realizability and Universality,” *British Journal of the Philosophy of Science*, 51: 115 – 145.

Batterman, R. (2001), *The Devil In the Details: Asymptotic Reasoning in Explanation, Reduction, and Emergence*. Oxford: Oxford University Press.

Beckermann, A., Flohr, H., Kim, J., eds., (1992), *Emergence or Reduction?* Berlin: Walter de Gruyter.

Bedau, M. (1997), “Weak Emergence,” *Philosophical Perspectives*, 11: 375 – 399.

Bedau, M. (2002), “Downward Causation and the Autonomy of Weak Emergence,” *Principia*, 6(1): 5 – 50.

Broad, C.D. (1925), *The Mind and Its Place In Nature*, London: Routledge and & Paul.

Chalmers, D. (1996), *The Conscious Mind: In Search of a Fundamental Theory*, Oxford: Oxford University Press.

Crane, T. (2001), “The Significance of Emergence,” in Loewer and Gillett, eds., (2001), *Physicalism and Its Discontents*. Cambridge: Cambridge University Press.

Davidson, D. (1992), “Thinking Causes,” in Heil, J. and Mele, A., eds., (1993), *Mental Causation*, Oxford: Clarendon Press.

Gillett, C. (2002), “The Dimensions of Realization: A Critique of the Standard View,” *Analysis*, 62: 316 – 323.

Gillett, C. (2003), “The Varieties of Emergence: Their Purposes, Obligations, and Importance,” *Grazer Philosophische Studien*, 65: 89 – 115.

Fodor, J. (1997), “Special Sciences: Still Autonomous After All these Years,” *Philosophical Perspectives*, 11: 149 – 63.

Heil, J. (1999), “Multiple Realizability,” American Philosophical Quarterly, 36: 189 – 208.

Heil, J. and Mele, A., eds., (1993), *Mental Causation*, Oxford: Clarendon Press.

Horgan, T. (1993), “From Supervenience to Superdupervenience: Meeting the Demands of a Material World,” *Mind* 102: 555 – 586.

Humphreys, P. (1996), “Aspects of Emergence,” *Philosophical Topics*, 24 (1), pp. 53 – 70.

Humphreys, P. (1997), “How Properties Emerge,” *Philosophy of Science*, 64, pp. S337 – S345.

Kim, J. (1984), “Concepts of Supervenience,” *Philosophical and Phenomenological Research*, 45: 153 – 76.

Kim, J. (1989), "The Myth of Nonreductive Materialism," *Proceedings of the American Philosophical Association*, 63: 31-47, reprinted in Kim (1993).

Kim, J. (1992), “’Downward Causation’; in Emergentism and Reductionism,” in Beckermann, et al, eds. (1992), *Emergence or Reduction?* Berlin: Walter de Gruyter.

Kim, J. (1993), "The Non-Reductivist's Trouble With Mental Causation," in Heil and Mele, eds., (1993): 189-210.

Kim, J. (1998), *Mind in a Physical World*, Cambridge: Cambridge University Press.

Kim, J. (1999), “Making Sense of Emergence,” *Philosophical Studies*, 95, pp. 3 – 36.

Kim, J. (2005), *Physicalism, or Something Near Enough*, Princeton: Princeton University Press.

Klee, R.L. (1984), “Micro-Determinism and Concepts of Emergence,” Philosophy of Science, 51: 44 – 63.

Loewer, B. and Gillett, C., eds., (2001), *Physicalism and Its Discontents*. Cambridge: Cambridge University Press.

Lowe, E.J. (2000), “Causal Closure Principles and Emergentism,” *Philosophy*, 75, pp. 571 – 585.

McLaughlin, B. (1992), “The Rise and Fall of British Emergentism,” in Beckermann, et al, eds. (1992), *Emergence or Reduction?* Berlin: Walter de Gruyter.

McLaughlin, B. (1995), “Varieties of Supervenience,” in Savellos, E. and Yalcin, U., eds., *Supervenience: New Essays*, (1995), Cambridge: Cambridge University Press.

Melnyk, A. (2003), *A Physicalist Manifesto: Thoroughly Modern Materialism*. Cambridge: Cambridge University Press.

Meyering, T. C. (2000), “Physicalism and Downward Causation In Psychology and the Special Sciences,” *Inquiry*, 43: 181 – 202.

Mill, J.S. (1843), *The System of Logic*, London: Longmans, Green, Reader, and Dyer. (8th edition, 1872).

Morgan, (1923), *Emergent Evolution*, London: Williams and Norgate.

O’Connor, T. (1994), “Emergent Properties,” *American Philosophical Quarterly*, 31, pp. 91 – 104.

O’Connor, T. (2000), *Persons and Causes*, Oxford: Oxford University Press.

Rueger, A. (2000), “Robust Supervenience and Emergence,” *Philosophy of Science*, 67: 466 – 489.

Savellos, E. and Yalcin, U., eds., (1995), *Supervenience: New Essays*, Cambridge: Cambridge University Press.

Shaffer, J. (2003), “Is There a Fundamental Level?”, *Nous*, 37: 498-517.

Shoemaker, S. (1980), “Causality and Properties,” in P. van Inwagen (ed.), *Time and Cause*, Dordrecht, Netherlands: R. Reidel Publishing Co., pp. 109 – 135.

Shoemaker, S. (2002), “Kim on Emergence,” *Philosophical Studies*, 108: 53 – 63.

Sperry, R. (1969), “A Modified Concept of Consciousness,” *Psychological Review* 76: 532 – 536.

Van Gulick, R. (2001), “Reduction, Emergence and Other Recent Options on the Mind/Body Problem: A Philosophical Overview,” *Journal of Consciousness Studies*, vol. 8, no. 9 – 10: 1 – 34.

Yablo, S. (1992), "Mental Causation," *The Philosophical Review*, 101: 245-280.

Are Tractarian Objects Whitehead's Pure Potentials?

Piotr Żuchowski, Łódź, Poland

Without doubt Whitehead and Wittgenstein construct their philosophies from different perspectives. It's enough to consider the way they treat language: for Whitehead language is not transparent as it is for Wittgenstein – there's no any isomorphism or homomorphism between the world and language as one can find it in Tractatus. One cannot reveal ultimate constituents of reality by process of logical analysis of language. Basically such an analysis can disclose no more than fundamental presuppositions, prejudices trenched in scientific and common-sense minds of a given epoch. However, despite all the differences there's striking parallel between both systems. What may be astonishing is that one can find just a few works exploring these parallels (if any concerning ontological aspects of both systems).

I assume that Wittgenstein's Tractatus contains a kind of ontological system. To put it other words: Wittgenstein wanted to say something about the very structure of reality. Hence Fact Ontology which is stated in the thesis that the world divides into facts is not mere counterpart of logical analysis of language but it is an explicit ontological position, though analysis of language is the only way we can apply to get to that position. One can find, however, various interpretations of the thesis 1.2. There are two main: the first - let me name it "extensional" - is based on the assumption that facts merely determine the scope (extension) of the world but facts are not the final constituents of reality – these are objects which are usually supposed to be kind of substances. Thus 1.2 says no more than to know all facts is to know all objects that are included in the world – all facts determine all objects but not conversely [1.11, 1.12]. The second interpretation - "contentual" - holds that facts are the final constituents of reality, the reality is **made of** facts. Then the problem what Tractarian objects are immediately arises to its fullest extent. In this case one can also find opposite views. Nominalistic interpretation (which is supplied by the extensional interpretation mentioned above) holds that objects are individual things, while realistic interpretation holds that there are not only individual things among objects but also properties and relations (ie. traditional universals). Neither of above interpretations presupposes that we can point out examples of objects or that we have direct acquaintance with them.

I would like to propose the third probable interpretation of Tractarian objects according to which there are no individual things among objects (substances that bear qualities) but just entities that philosophical tradition generally considered to be universals (i.e. properties and relations). This is what ontology of facts holds – the world consist of facts, not of things. Objects are derivative beings, they exist only as constituents of facts, though they are not parts of facts. Facts do not have any material, concrete parts they could be spilt into. This is why one can find two modes of existence in Tractatus: the first (primary) that belongs to facts, which are existing state of affairs [2], the second (derivative) is proper to the substance of the world and is indicated as subsistence [2.024]. Objects can enter into the world only as elements of facts. They are thus abstract aspects of facts. It follows that there are no substances or other entities, which endure their existence on the successive moments of time. Momentary facts are thus ultimate bricks of reality, and objects are their necessary elements but they are not primal, they do not build the reality. However strange it may seem at first sight, this kind of antisubstanialistic position is developed in Whitehead's process metaphysics, it also corresponds better to the view of reality derived from modern physics.

In Process and Reality Whitehead introduces eight categories of existence, of which so called actual occasions (referred here as facts) are fundamental. They are the final realities (the only reasons). Entities that belong to remaining seven categories exist only as elements of facts. This applies primarily to eternal objects (referred further as EO), which are also described as forms of definiteness or pure potentials and which seem to me correspond directly to Tractarian objects.

Now, if we admit interpretation introduced above all striking parallels between Whitehead's and Wittgenstein ontologies shall emerge. Due to the lack of space I shall confine myself in this paper basically to discussing those parallels which are related to characteristics of Tractarian objects and Whiteheadian EOs, though some other relevant similarities will also be mentioned.

In the first place let's consider the problem of change. As it was mentioned above since facts should be momentary they cannot be subject of change. Momentary does not mean that they are not time-extended, geometrical point-like, since it is impossible to construct extension from unextended parts. Facts are out of time, they cannot remain identical in the successive moments. Time is derivative from the succession of time moments constituted by facts coming into being. They become and perish. Thus the slightest "change" means we are dealing with another fact. For both philosophers some kind of (eternal) objects are necessary to provide the stability of world-structures and to avoid heraklitean consequences of a total flux of all things [2.026]. What undergo changes then is the structure, configuration of objects, ie. state of affairs, or to use Whiteheadian terms, the pattern made of EO which differs in successive facts by some of its elements. The change consists in adding or loosing an object to some state of affairs leading by necessity to a new situation. Hence if we talk about subject of change, we have to refer to complex which is identical in the successive facts. This complex, however, does not endure, since it is not in time, it can be said to "haunt" time - or to use more proper Whiteheadian term - it just ingresses into successive facts (it is thus connected to time in a different manner than facts). It seems an inevitable consequence for Fact Ontology to conceive time as atomic, quantified, consisting of discrete epochs constituted by facts (in opposition to substance ontologies for which continuous time seems to be more natural) and hence it is bluntly absurd on the ground of Fact Ontology to ask what is in time. According to Whitehead internal relations between facts provide that common characteristics can be inherited by effect from the cause. This element of mutual relations of facts is absent in Tractatus and exposed in process metaphysics, nevertheless in both systems the totality of objects with their internal relations constitutes the substance of the world conceived as enunciated above.

It follows also that no new objects become, the substance is given once for ever. In this respect it is independent of what is the fact, facts cannot affect the web of objects (though even here one could find some points for discussion). More, since the substance comprises all possibilities, every possible world, no matter how different from ours, should have the same substance.

The congruence of both ontologies seems even deeper when we consider characteristic of (eternal) objects.

Whitehead gives twofold characteristics of EO: *Individual essence* of EO determines its particular, unique individuality; for example red colour (particular shade of red) is what it is without any reference or relation to other beings whatsoever (facts, objects). The same colour may determine entities in many different ways staying identical self. To put it other words, no matter how given EO ingresses into actuality, how it is realised, each time it provides identical unique contribution to reality. EO is thus transcendent in respect to every actual entity, its relations to facts are external to it. In this sense EO is abstract – but to be abstract does not mean to be disconnected from a fact.

Transcendence described above is a reason why we should think of EO as relations rather than properties (beside the fact that Whiteheadian facts are not substances that may gain or loose properties). Relations, contrary to properties, cannot be ascribed to one object, thus EOs are relations with many arguments that transcend each of them. Now, one of the main points of Whiteheads philosophical objection is so called theory of simple location, according to which all entities are related to space in a simple manner, that is one and only one location can be ascribed to them (no matter whether we take absolute or relative space). For Whitehead ingression of EO into actuality yields complicated web of their relations to space (different for example in a mirror image or in a simple perception of a red object). Whitehead stresses that relations are primal to properties. EO's ingression into actuality is not embodiment of a property in a thing (since there are no things) but it relates facts. It does not mean that EOs are relations as such (at least) not all of them, but that they are by their nature relational, contrary to properties which are private (excluding) in character. To point location of EO one has to point the whole web of relations (ex. I see *here* a green leaf being *there*).

EOs as pure potentials for determination have to be in mutual relations among themselves. It stems from their pure potentiality that all possible relations to other EOs should be included in it - every relationship, which is possible, is thereby in the realm of possibility. It belongs to their essence that they can jointly determine facts. There are two types of EOs: simple and complex. Ingression of simple EOs does not necessarily imply ingression of other EOs, though "in fact" EOs always determine facts as patterns. There have to be simple EOs that are the fundaments of each hierarchy of patterns. Similarly in Tractatus there have to be objects which are simple, they do not comprise other objects. The web of mutual relations of EO is their *relational essence*. Its function is twofold: it determines all relations of given EO to all other EO and it also includes general possibility of determining facts. General means here that no EO simple or complex could be a cause of how matters really happen. So it belongs to EO's nature to be generally realizable, otherwise they would be nonentities.

The above characteristic corresponds almost entirely to Wittgensteinian characteristic of objects. On the one hand they have form which consists of relations to other objects. It is essential for objects that they should be constituents of possible states of affairs, and the form consists of all possible configurations, no other possible configurations are left to be found or added later, there is nothing accidental in the form. Relations to other objects are then internal to the object. Object have relational essence, it is their internal structure. This structure is independent of what is the case, the actual flow of things cannot affect the structure, otherwise there would be nothing stable in the reality. On the other hand the substance of the world is not only form but also content [2.025]. The content should consist of some internal qualities other than the logical structure: for example redness as a first-order internal quality and being a color which could be conceived as a second-order quality i.e. internal structure determining what composition may the red color come into (Stenius 1981: 79-81). These internal qualities could correspond to individual essence of EO.

But there is a problem with the above characteristic. As 2.0231 states, if two objects have the same logical form, the only distinction between them, apart from their external properties, is that they are different (external properties mean here being a component of existing sate of affairs). It follows that the form excludes other individual characteristic of an object whatsoever. All properties of an object are relational, object possesses them only with respect to other objects, being a component of all possible states of affairs. An object as such is undifferentiated from other objects. How then objects obtain their individuality? Are they to be conceived as points of geometrical plane – each point is individual only due to infinite boundle of relations to other points (ie. due to its relational essence). This consequence is accepted by nominalistic interpretation mentioned above (Wolniewicz 1968: 85). Similarly Whitehead holds that in isolation EOs are undifferentiated nonentities. More, according to Whitehead relational essence is not unique to a given EO, each EO stands in a uniform web of relations. No matter which system we consider one can ask then, what determines which relations are possible and which are not? Could laws of reality, of logic be different? How "rich" is the structure: are there only logical relations of exclusion, implication and are objects only kind of logical variables in these relations?

Another difficulty I should point out is connected with the interpretation of Fact Ontology according to which primary mode of existence belong exclusively to facts, other entities are derivative beings. If it is so, then one can protest that there could not be any unalterable substance of the world, which after all seems to be necessary element. That's the reason why Whitehead postulates that there should be a primordial fact that valuates whole multiplicity of EOs, establishing all possible connections among mere multiplicity of EOs. Whitehead describes it as a kind of primordial God's vision and regards such an entity indispensable element of any sound metaphysics. It seems to me however that this primordial valuation can be conceived as an act of establishing Laws of Nature for a given reality, setting up boundary conditions by some quantum fluctuations from which our reality begins as it is described in some multiverse cosmological scenarios. It follows that this primordial fact introduces a matrix that would serve as a substance by providing a stable structure determining facts coming into being. However, if substance of the world is to be something common to all possible worlds, it requires that objects, which are to be connected, should have some content - internal qualities other than the logical structure, as I tried to argue.

Both systems may be called Fact Ontologies, though while Whitehead’s philosophy is procesualism, Wittgenstein’s is not. Wittgenstein does not have much, if anything, to say about the process facts become and how they can be bounded together; for Whitehead – contrary – these are main questions. Undoubtedly Wittgenstein would treat Whitehead’s description of process of facts’ becoming as “improper” metaphysics, nonetheless structurally both ontologies are deeply similar.

Literature

Stenius, Erik, 1981, *Wittgenstein’s Tractatus*, Greenwood Press, Publishers, Westport, Connecticut

Wolniewicz, Bogusław, 1968, *Rzeczy i Fakty*, Panstwowe Wydawnictwo Naukowe

ontos verlag

Frankfurt • Paris • Lancaster • New Brunswick
P.O. Box 1541 • D-63133 Heusenstamm bei Frankfurt
www.ontosverlag.com • info@ontosverlag.com
Tel. ++49-6104-66 57 33 • Fax ++49-6104-66 57 34

Publications of the Austrian Ludwig Wittgenstein Society. New Series

Volume 1

Friedrich Stadler, Michael Stöltzner (Eds.)

Time and History

Proceedings of the 28. International Ludwig Wittgenstein Symposium in Kirchberg am Wechsel, Austria 2005
ISBN 3-938793-17-1
621pp., Hardcover € 79,00

Time and History presents the invited papers of the 28th International Wittgenstein Symposium 2005 in Kirchberg/W. (Austria). Renowned scientists and scholars address the issue of time from a variety of disciplinary and cross-disciplinary perspectives in four sections: philosophy of time, time in the physical sciences, time in the social and cultural sciences, temporal logic, time in history/history of time, and Wittgenstein on time. Questions discussed include general relativity and cosmology, the physical basis of the arrow of time, the linguistics of temporal expressions, temporal logic, time in the social sciences, time in culture and the arts. Outside the natural sciences, time typically appears as history and in historiography in different forms, like a history of our conceptions of time. The first chapter of the book is dedicated to the major positions in contemporary philosophy of time. Is there a real sense of past, present, and future, or is time just a special coordinate among others? What does it mean that identity persists over time? The importance of Wittgenstein for present-day philosophy notwithstanding, his ideas about time have hitherto received only little attention. The final chapter, for the first time, provides an extensive discussion of his respective views.

Volume 2

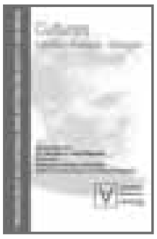
Alois Pichler, Simo Säätelä (Eds.)

Wittgenstein: The Philosopher and his Works

ISBN 3-938793-28-7
461pp., Hardcover € 98,00

This wide-ranging collection of essays contains eighteen original articles by authors representing some of the most important recent work on Wittgenstein. It deals with questions pertaining to both the interpretation and application of Wittgenstein's thought and the editing of his works. Regarding the latter, it also addresses issues concerning scholarly electronic publishing. The collection is accompanied by a comprehensive introduction which lays out the content and arguments of each contribution.
Contributors: Knut Erik Tranøy, Lars Hertzberg, Georg Henrik von Wright, Marie McGinn, Cora Diamond, James Conant, David G. Stern, Eike von Savigny, P.M.S. Hacker, Hans-Johann Glock, Allan Janik, Kristóf Nyíri, Antonia Soulez, Brian McGuinness, Anthony Kenny, Joachim Schulte, Herbert Hrachovec, Cameron McEwen.

Volume 3



Christian Kanzian, Edmund Runggaldier (Eds.)

Cultures. Conflict - Analysis - Dialogue

Proceedings of the 29th International Ludwig Wittgenstein-Symposium in Kirchberg, Austria 2006.
ISBN 978-3-938793-66-4
431pp., Hardcover, EUR 59,00

What can systematic philosophy contribute to come from conflict between cultures to a substantial dialogue? – This question was the general theme of the 29th international symposium of the Austrian Ludwig Wittgenstein Society in Kirchberg. Worldwide leading philosophers accepted the invitation to come to the conference, whose results are published in this volume, edited by Christian Kanzian & Edmund Runggaldier. The sections are dedicated to the philosophy of Wittgenstein, Logics and Philosophy of Language, Decision- and Action Theory, Ethical Aspects of the Intercultural Dialogue, Intercultural Dialogue, and last not least to Social Ontology. Our edition include (among others) contributions authored by Peter Hacker, Jennifer Hornsby, John Hyman, Michael Kober, Richard Rorty, Hans Rott, Gerhard Schurz, Barry Smith, Pirmin Stekeler-Weithofer, Franz Wimmer, and Kwasi Wiredu

Volume 4

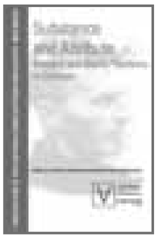
Georg Gasser (Ed.)

How Successful is Naturalism?

ISBN 13: 978-938793-67-1
ca. 300pp., Hardcover, EUR 69,00

Naturalism is the reigning creed in analytic philosophy. Naturalists claim that natural science provides a complete account of all forms of existence. According to the naturalistic credo there are no aspects of human existence which transcend methods and explanations of science. Our concepts of the self, the mind, subjectivity, human freedom or responsibility is to be defined in terms of established sciences. The aim of the present volume is to draw the balance of naturalism's success so far. Unlike other volumes it does not contain a collection of papers which unanimously reject naturalism. Naturalists and anti-naturalists alike unfold their positions discussing the success or failure of naturalistic approaches. "How successful is naturalism?" shows where the lines of agreement and disagreement between naturalists and their critics are to be located in contemporary philosophical discussion.

Volume 5



Christian Kanzian, Muhammad Legenhausen (Eds.)

Substance and Attribute

Western and Islamic Traditions in Dialogue
ISBN 13: 978-3-938793-68-8
ca. 250pp., Hardcover, EUR 69,00

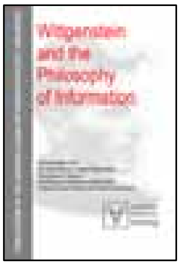
The aim of this volume is to investigate the topic of Substance and Attribute. The way leading to this aim is a dialogue between Islamic and Western Philosophy. Our project is motivated by the observation that the historical roots of Islamic and of Western Philosophy are very similar. Thus some of the articles in this volume are dedicated to the history of philosophy, in Islamic thinking as well as in Western traditions. But the dialogue between Islamic and Western Philosophy is not only an historical issue, it has also systematic relevance for actual philosophical questions. The topic Substance and Attribute particularly has an important history in both traditions; and it has systematic relevance for the actual ontological debate.
The volume includes contributions (among others) by Hans Burkhardt, Hans Kraml, Muhammad Legenhausen, Michal Loux, Pedro Schmechtig, Muhammad Shomali, Erwin Tegtmeier, and Daniel von Wachter.

ontos
verlag

Publications of the
Austrian Ludwig Wittgenstein Society. New Series

Frankfurt • Paris • Lancaster • New Brunswick
P.O. Box 1541 • D-63133 Heusenstamm bei Frankfurt
www.ontosverlag.com • info@ontosverlag.com
Tel. ++49-6104-66 57 33 • Fax ++49-6104-66 57 34

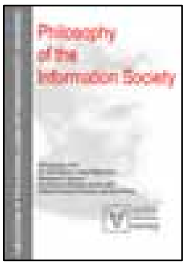
Volume 6



Alois Pichler, Herbert Hrachovec (Eds.)
Wittgenstein and the Philosophy of Information
Proceedings of the 30th International Ludwig Wittgenstein-Symposium in Kirchberg, Volume 1
ISBN 978-3-86838-001-9
356pp., Hardcover, EUR 79,00

This is the first of two volumes of the proceedings from the 30th International Wittgenstein Symposium in Kirchberg, August 2007. In addition to new contributions to Wittgenstein research (by N. Garver, M. Kross, St. Majetschak, K. Neumer, V. Rodych, L. M. Valdés-Villanueva), the book contains articles with a special focus on digital Wittgenstein research and Wittgenstein's role for the understanding of the digital turn (by L. Bazzocchi, A. Biletzki, J. de Mul, P. Keicher, D. Köhler, K. Mayr, D. G. Stern), as also discussions - not necessarily from a Wittgensteinian perspective - of issues in the philosophy of information, incl. computational ontologies (by D. Apollon, G. Chaitin, F. Dretske, L. Floridi, Y. Okamoto, M. Pasin and E. Motta).

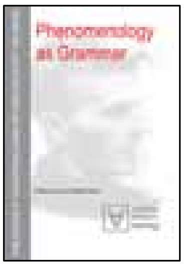
Volume 7



Herbert Hrachovec, Alois Pichler (Eds.)
Philosophy of the Information Society
Proceedings of the 30th International Ludwig Wittgenstein-Symposium in Kirchberg, Volume 2
ISBN 978-3-86838-002-6
326pp., Hardcover, EUR 79,00

This is the second of two volumes of the proceedings from the 30th International Wittgenstein Symposium in Kirchberg, August 2007. It contains selected contributions on the Philosophy of media, Philosophy of the Internet, on Ethics and the political economy of information society. Also included are papers presented in a workshop on electronic philosophy resources and open source/open access.

Volume 8



Jesús Padilla Gálvez (Ed.)
Phenomenology as Grammar
ISBN 978-3-938793-91-6
224pp., Hardcover, EUR 59,00

This volume gathers papers, which were read at the congress held at the University of Castilla-La Mancha in Toledo (Spain), in September 2007, under the general subject of phenomenology. The book is devoted to Wittgenstein's thoughts on phenomenology. One of its aims is to consider and examine the lasting importance of phenomenology for philosophic discussion. For E. Husserl phenomenology was a discipline that endeavoured to describe how the world is constituted and experienced through a series of conscious acts. His fundamental concept was that of intentional consciousness. What did drag Wittgenstein into working on phenomenology? In his "middle period" work, Wittgenstein used the headline "Phenomenology is Grammar". These cornerstones can be signalled by notions like language, grammar, rule, visual space *versus* Euclidean space, *minima visibilia* and colours. L. Wittgenstein's main interest takes the form of a research on language.