JAEGWON KIM

PSYCHOPHYSICAL SUPERVENIENCE

(Received 19 October, 1981)

I

Suppose we could create an exact physical replica of a living human being – exactly like him cell for cell, molecule for molecule, atom for atom. Such a replica would be indistinguishable, at least physically, from the original. For we are supposing that the replica is a perfect physical copy in every detail. The idea of such a replica, whether artificially created or naturally found, is a perfectly coherent one; in fact, it is consistent with all known laws of nature. The idea of course is a commonplace in science fiction.

Given that your replica and you are exactly alike physically, will you also share your psychological life with it? Will your replica have your psychological traits and dispositions, intellectual powers and artistic gifts, anxieties and depressions, likes and dislikes, and virtues and vices? Will it feel pain, remorse, joy and elation exactly in the way you do? That is, if two organisms have identical physical features, will they be identical in psychological characteristics as well?

According to many moral theorists, any two things sharing the same 'naturalistic' or 'descriptive' features cannot differ in respect of moral or evaluative properties. Thus, it has been said that if St. Francis is a good man, anyone who is just like him in all naturalistic respects – in this case, broadly psychological properties, such as traits of character and personality – must of necessity be an good man. This relationship between moral properties and nonmoral properties is often called 'supervenience': moral properties are said to be supervenient upon nonmoral properties just in the sense that any two things that coincide in all nonmoral properties cannot diverge with respect to any moral properties. The concept of supervenience is easily generalized so that we may speak of the supervenience relation for any two families of properties (or events, predicates, facts, etc.).[1] Briefly, a set F of properties is supervenient upon a set G of properties with respect to a domain D just in

Philosophical Studies 41 (1982) 51–70. 0031-8116/82/0411-0051$02.00

case any two things in D which are indiscernible with respect to G are necessarily indiscernible with respect to F (that is to say, any two things in D are such that necessarily if they differ with respect to F then they differ with respect to G).[2] We may call F 'the supervenient (or supervening) family' and G 'the supervenience base'.

The problem about the shared psychological life of persons and their physical duplicates can be given a perspicaceous reformulation in terms of supervenience: Are psychological properties (events, processes, etc.) supervenient upon physical properties (events, processes, etc.)? Psychological supervenience, if it obtains, would give us one important sense in which the physical determines the mental: once the physical side of our being is completely fixed, our psychological life is also completely fixed. Since the physical obviously does not supervene upon the psychological, this determination is asymmetric: the physical determines the psychological, but the psychological does not determine the physical. Thus, psychophysical supervenience is one possible way in which the psychophysical relation can be characterized; and beyond this it has implications for various problems in the philosophy of mind such as the traditional mind-body problem, psychophysical reduction, and the possibility of psychophysical laws. We will touch on some of these issues below;[3] however, our chief concern will be the question what reason there might be for accepting the thesis of psychophysical supervenience.

Lest you think that an affirmative answer to the question of psychophysical supervenience automatically yields physicalism, let me remind you that G. E. Moore, to whom the thesis of moral supervenience is often attributed, was a staunch and generally effective critic of ethical naturalism, the thesis that moral properties are definable by, or reducible to, naturalistic properties; in spite of his beleif in the supervenience of the moral upon the naturalistic, he was an advocate of the autonomy of ethics and the irreducibility of ethics to natural science. It is possible that Moore was inconsistent in holding these positions, but the inconsistency is not apparent or obvious; it would need to be demonstrated. Similar comments apply to psychophysical supervenience and physicalism. If Moore was consistent, then by symmetry of reasoning the doctrine of psychophysical supervenience ought to be compatible with the denial of physicalism.[4]

II

I believe that most of us are strongly inclined to accept the doctrine of psychophysical supervenience in some form. Your replica is not only a person, but a person who is psychologically indistinguishable from you. He will share your beliefs, memories, likes and dislikes, wants and aversions, hopes and despairs; his internal life, as well as his external life, will be just like yours. As we shall see presently, there are psychological states you and your replica will not share, but the strong intuition prevails that some form of psychophysical supervenience must hold.

In a recent paper,[5] Stephen P. Stich considers a form of psychophysical supervenience. The thesis he considers, and which he endorses, is the following, called by him "the principle of psychological autonomy": "The properties and relations to be invoked in an explanatory psychological theory must be supervenient upon the *current, internal physical* properties and relations of organisms"[6]. We can put this more simply thus: "Explanatory psychological properties and relations are supervenient upon the current internal physical properties of organisms". Stich, however, does not discuss in any detail reasons for or against this thesis of supervenience, his chief concern in the paper being the implication of this thesis for the belief-desire model for explanations of actions. We will later defend a thesis of psychophysical supervenience similar to Stich's, but let us begin by considering a broader and stronger form of the supervenience doctrine.

(A) All psychological states and processes supervene on the contemporaneous physical states of the organism.

Two points of explanation: the qualification 'contemporaneous' is intended to indicate the requirement that if a certain psychological state occurs at a time, then that state is supervenient upon a physical state (or class of physical states) of the organism occurring at the same time. The point of this qualification will become clear when we examine some apparent counterexamples to the thesis as stated. Second, by 'physical state' we have in mind what may be called 'internal physical state'; that is, we want to exclude so-called 'relational properties' of the organism, such as its distance from the moon at a given time, its being larger than this typewriter, etc. It is not a simple matter to give a precise meaning to 'internal'; but the relative looseness of this and other notions we are making use of here will not affect our discussion. Third, when we speak of 'states' and 'events', we sometimes have in mind 'generic states'

and 'generic events' – that is, *types* such as pain, itch, and belief that a stitch in time saves nine; at other times, we may be referring to concrete, dated instatiations of these generic events and states. In any particular case the context should make clear which sense is intended, although in many contexts these terms can be read in either of the two senses.

Why should anyone think (A) is true? At one point, many philosophers believed something like what Wolfgang Köhler and others called "psychophysiological isomorphism",[7] which, for our present purposes, could be stated as follows:

> *The Psychophysical Correlation Thesis*: For each psychological event M there is a physical event P such that, as a matter of law, an event of type M occurs to an organism at a time just in case an event of type P occurs to it at the same time.

The principle affirms the existence of a pervasive system of laws, of biconditional form, linking each mental event with some physical correlate, presumably some neurological state or process. Evidently, the Correlation Thesis implies the supervenience thesis (A) – assuming that the modality involved in the concept of supervenience is satisfied by the nomological modality in the statement of the Correlation Thesis. Thus, anyone who accepts the Correlation Thesis would be committed to the doctrine of psychophysical supervenience.

However, I think it would be wrong to think of the Correlation Thesis as providing *evidence* for the doctrine of psychophysical supervenience. For one thing, there are those who accept supervenience but not the Correlation Thesis, the latter being a stronger claim than the former.[8] Second, if there is an evidential relation here at all, the idea of psychophysical correlation should be seen as grounded in a belief in supervenience, and not the other way around; it seems to me that the belief that there must be laws connecting psychological events with physical events is derived from the general belief in the supervenience of the former on the latter, although a demonstration of this evidential priority would be a complex matter. In any event, the question of the possible support the Correlation Thesis might offer for psychophysical supervenience is made moot by the fact that most philosophers today would reject the Correlation Thesis.

There are various arguments in the recent literature in the philosophy of mind intended to refute the Correlation Thesis. One of the most influential

of these, advanced by the proponents of 'functionalism', runs as follows: Any mental state, such as pain, can be 'physically realized' in many diverse types of organisms and physical structures (e.g., humans, molluscs, crustaceans, perhaps Martians and robots) so that, as a matter of empirical fact, it is extremely unlikely that some *uniform* physical state exists to serve as its physical correlate. Creatures whose physicochemical structures are entirely different from our own, or from anything we know on this earth, may yet be 'psychologically isomorphic' to us in the sense that the same psychological theory is true of them. Roughly speaking, this means that their observable behavior is best explained by imputing to them certain internal states which are interconnected among themselves, and to stimuli and behavior, in the way psychological states are so interconnected for humans. And yet the biochemistry of these creatures may be so different from ours that there is no sense in which we may speak of 'the same physical state' underlying, say, pain for both humans and these creatures. We may call this 'the multiple realization argument'.

This is not the place to discuss the merits of this argument; the only point I want to make here concerns the possibility of psychophysical laws, given the multiple realizability of psychological states. First of all, notice that the multiple realization argument implies nothing about the general impossibility of psychophysical laws; at best, it shows the impossibility of psychophysical laws of a certain form (completely general biconditional laws of the form '*M* iff *P*', where *M* is a general kind of a mental state and *P* is a 'single' physical state). In fact, it is a tacit assumption of the argument that there are *species-specific* psychophysical laws, that is, laws connecting, say, pain with a certain neural correlate for each biological species. Indeed, the very notion of 'physical realization' of pain seems to presuppose the existence of nomological connections, *within each species*, between pain and some underlying neural process. If there were no such nomological link, in what sense does *this* neural state, and not some other one, 'realize' pain? And how would we know that it, and not some other state, is the physical realization of pain for this species? Similar comments apply to the talk of electronic-mechanical devices 'realizing' certain 'logical states' or psychological states. Thus, the existence of species-specific psychophysical laws of the following form is not only consistent with the multiple realization argument but in fact presupposed by it:

> Each human is such that it is in pain at a time if and only if it is in physical state P at the time.
>
> Each mullusc is such that it is in pain at a time if and only if it is in physical state Q at the time,...

It is in virtue of these lawlike connections that the state P can be said to realize pain in humans, that Q realizes pain in molluscs, and so on.

The implications of these species-specific psychophysical laws for the question of psychophysical supervenience are clear. Even if the multiple realization argument refutes the unrestricted doctrine of psychophysical correlation it has no tendency to refute a more limited correlation thesis, one that asserts the existence of a physical correlate, within each species, for every psychological event. In fact, this thesis of species-restricted psychophysical correlation appears to be an implicit commitment of the multiple realization argument, and hence of the functionalist position. To derive psychophysical supervenience from the restricted psychophysical correlation thesis, the only additional assumption needed is the self-evident proposition that if two organisms or structures are physically indistinguishable from each other, then they belong to the same species. Replicas of humans are humans; replicas of felines are felines; and replicas of Martians are also Martians.

Thus, it is clear, in general, that various forms of the psychophysical correlation thesis logically entail the supervenience thesis. But this in itself is as one would expect, and provide little enlightenment concerning the question of evidence for psychophysical supervenience. For, as observed earlier, supervenience seems more fundamental, metaphysically and methodologically, than correlation, and although evidence for the Correlation Thesis is likely to be also evidence for supervenience, it would be pointless to argue to supervenience from correlation. A more interesting question is whether supervenience itself entails the Correlation Thesis, or if not then at least the existence of some psychophysical laws.[9] Let us now consider some specific putative counterexamples to the supervenience thesis (A), postponing till later more general evidential considerations.

III

Obviously, my replica and I do not share *all* properties in common. For example, we cannot be in the same place at the same time; I was born of natural parents, but he wasn't; I have siblings but he has none; I have lived in Ann Arbor for over ten years but he hasn't; I will be alive in 1984 but perhaps he won't be. These properties that we do not share may not strike us as very significant, but they are properties nonetheless. The philosophically interesting problem is whether the properties that I do share with my replica include significant properties – significant in some clear and definite sense. That is, the interesting problem for us is what significant psychological properties are supervenient on physical properties. Let us consider some psychological properties that apparently are not supervenient.

(A) I *remember* being strafed by a jet fighter in a war over twenty years ago. My replica thinks he remembers this, too; in fact, he claims to have nightmares about this, and his mental imagery of the event is as vivid as mine. But of course he remembers no such thing, and the strafing is not part of his life experience. I also *know* and *truly believe* that I was strafed by a fighter plane, but my replica has no such knowledge, or true belief.[10]

(B) I am *thinking of* Vienna. We put my replica in the same brain state, and he has the visual imagery that I am having – say, that of an old church I was fond of visiting when I was in Vienna some years ago – and is thinking the same thoughts that I am thinking (how hot and humid that summer was in Vienna,...). And he shares my tendency to speak of Vienna (or at least to utter sentences containing the word 'Vienna' at dinner parties). Is he also thinking of Vienna? I do not think so. When I have a certain sort of visual imagery and thinking certain thoughts, that counts as 'thinking of Vienna' because of a certain historical and cognitive relationship that I have with the city Vienna, a relationship that my replica lacks. To see this more clearly, think of a person who is having the very same phenomenological visual imagery that I am having, but who has never been to Vienna and has never heard of it, and whose visual image, which is qualitatively indistinguishable from mine, can be traced to a church in his hometown in Iowa. We would hardly say of this person that he is now thinking of Vienna. These points can be made with regard to other examples: *liking* or *disliking* some particular person; *wanting* some particular object; *fearing* some particular object or event. It is important to see that while wanting to eat *this particular ham-*

burger is similar to thinking of Vienna in the respect we are presently interested in, wanting to eat *some hamburger or other* is not. The latter would be supervenient upon brain states.

(C) I *am glad* that I was invited to the Dean's party last week, but am still *embarrassed* that I could not remember the first name of the Dean's wife. Notice that if I am glad that such-and-such is so-and-so, then not only must I believe that such-and-such is so-and-so, it must be the case that such-and-such is so-and-so; and similarly if I am embarrassed that such-and-such is so-and-so, then it must be true that such-and-such is so-and-so. If Jones falsely believes that he has won a fellowship we cannot truly say Jones is glad *that* he won a fellowship. My replica, therefore, cannot be said to be glad that he was invited to the party or embarrassed that he could not recall the hostess' name. Similar comments apply to many other states of feeling and emotion.

(D) I *see a tree*. My replica has not emerged from the laboratory, but his brain is put in the same state that obtains when I see a tree. So he is having a 'treeish sense-datum', just like mine. But he is not seeing a tree. Two persons or organisms can be in the same state – the same appropriate neural state – but one may be seeing, touching, etc. a tree, and the other not.

(E) Finally, let us consider *actions* – especially, actions that presuppose social contexts – networks of social practices, customs, and institutions. I am signing a check to pay off my mortgage. We put my replica in the same brain state and give him a blank check. He puts his (my?) signature on it, a signature that no expert from the bank could distinguish from mine. But is he paying off his mortgage? Is he even signing a check? He does not have a bank account, not to mention a mortgage. The answer, I think, is that he is not doing any of these things; he is not signing a check, and he is not paying off any mortgage. He cannot do these things because he is not as yet a full member of the social community whose institutions and practices make these actions possible. His being in appropriate internal physical states is not enough to enable him to engage in these acts.

IV

How shall we handle these cases? Some of these cases can be handled by removing the requirement that every mental state or event supervene on

contemporaneous physical states; obvious examples include remembering. We could say that an instance of remembering occurring at a certain time does supervene on physical states but not on the synchronous ones, not on those occurring at the time the remembering takes place, but rather on a *longer temporal stretch of the physical history* of the organism that does the remembering. My Doppelgänger on Twin Earth does remember being strafed by a fighter plane, although my replica does not, and this is because the former, not the latter, has a life history similar to mine. Certain other cases discussed in the preceding section can be dealt with by enlarging the spatial scope of the supervenience base, by taking the trees seen, tables touched, and the person liked or disliked, in the supervenience base, but this is to go beyond the intended spirit of the thesis of psychophysical supervenience. I think it is important to be able to defend a form of the thesis that does not go outside the organism, a thesis that claims psychological states to be supervenient on the *internal* physical states of the organism.

I would suggest the following procedure. We first define the notion of an 'internal property' or 'internal state' of a thing, and then defend the following two theses:

> *The Supervenience Thesis*: Every internal psychological state of an organism is supervenient on its synchronous internal physical state.
>
> *The Explanatory Thesis*: Internal psychological states are the only psychological states that psychological theory needs to invoke in explaining human behavior – the only states needed for psychology.[11]

The full defense of these theses would be a major task; however, I hope to be able to say enough about them to make them plausible.

First, what is 'an internal state' or 'internal property'? In *Person and Object,*[12] R. M. Chisholm introduces the notion of a property *rooted outside the time at which it is had*:

> *G* is *rooted outside times at which it is had* $=_{\text{def}}$. Necessarily for any object *x* and for any time *t*, *x* has the property *G* at *t* only if *x* exists at some time before or after *t*.

The idea is straightforward: *G* is rooted outside the times at which it is

had just in case the possession of G by an object implies the existence of the object at a time other than the time at which it has G. Thus, consider some examples: taking the second vacation in the Rockies, taking the first of the two walks today, being twenty years old, being divorced, being a future president, and so on. A psychological example is remembering: for you now to remember a thing, you must have existed before. In analogy with this notion, we can define another:

> G is *rooted outside the objects that have it* $=_{\text{def.}}$. Necessarily any object x has G only if some contingent object wholly distinct from x exists.

The qualification 'contingent' is inserted because according to some philosophers there are 'necessary beings', beings that exist in all possible worlds. The qualification that the object other than x be *wholly distinct* from x is intended to exclude *proper parts* of x. If G is the property of being spherical, then if any object has G, then it follows necessarily that there is some object different from x, namely a spatial part of x. But this should not disqualify G from being an internal property. It will be seen that the notion we are after here corresponds, roughly, to the traditional notion of 'relational property'.

We now define 'internal':

> G is *internal* $=_{\text{def.}}$. G is neither rooted outside times at which it is had nor outside the objects that have it.

We may say that an event or state is an *internal event* or *state* of an object just in case it is the object's having an internal property at a time. So if G is an internal property, an object's having G or being G at a time is an internal event or state. An *internal process* would be a causally connected or continuous series of internal events or states involving the same objects or objects that are in some way connected or continuous with one another.

The Supervenience Thesis as stated concerns only internal psychological states, namely those psychological states whose occurrence does not imply anything about the past or future, or anything existing other than the organism or structure to which the states occur. Brief reflection should convince us that these are the states we should be concerned with.[13] Let us look at the series of counterexamples we presented earlier against the broader, unrestricted thesis of psychophysical supervenience.

Consider the group (A). Remembering is not internal. If a person now remembers anything, that entails he existed before now; so remembering does not come under the purview of the Supervenience Thesis. Nor does knowing or believing truly: if I know, or believe truly, that the moon is round, it follows that a contingent object, namely the moon, exists.

What of believing? It is now customary to distinguish between belief *de re* and belief *de dicto*, although the precise import of the distinction is still controversial.[14] Roughly speaking, *de dicto* belief is believing a certain proposition, a dictum, to be true, while *de re* belief is believing *of* some *object*, a *res*, that it is thus and so. Belief *de dicto* will in general be internal states. The belief that the tallest man is a spy does not entail the existence of a tall man or a spy; the belief that ghosts are malevolent does not entail the existence of ghosts. On the other hand, belief *de re* is plausibly viewed as noninternal when the object of belief is other than oneself. If a given belief is *de re* with respect to a certain object, then this object must exist if that *de re* belief is to exist. You cannot have a belief about Mt. Everest unless Mt. Everest exists – and unless, furthermore, you are in a certain historical-cognitive relation to it.[15] However, belief *de se*, a special case of belief *de re*, is internal; my belief that I am now sitting entails the existence of no contingent object other than myself. I think the internal-noninternal split for beliefs corresponds to the division between those beliefs which we expect to supervene on bodily states and those for which we do not have such expectations.[16]

Group (B) is analogous to belief *de re*. If, as we argued, thinking of Vienna involves as an essential ingredient some historical-cognitive contact with the city Vienna, it fails to be internal as defined: my thinking of Vienna is 'rooted outside' in both of the senses that were considered. The same goes for other *de re* psychological attitudes, such as liking and disliking, fearing, admiring, and expecting, except when these attitudes are *de se*.

Some items in group (C) will be internal and some noninternal: my being pleased that Johnson has been elected to be city council will be non-internal, but my being pleased that I am now thinking will be internal. My being pleased that I did twenty pushups this morning is of course noninternal. I believe we can expect the internal states in this group to be supervenient.

Items in group (D), involving perceptual relations to external objects, will in general be noninternal: I cannot see or touch a tree unless a tree exists, and I cannot see or touch this particular tree unless this particular tree exists.

Also, actions requiring societal contexts, the items in group (E), are noninternal; they presuppose the existence of persons in certain social relations, social institutions, and the history of these things, if these actions are to be performed.

The Supervenience Thesis concerns only those psychological states or properties that are internal in our sense, and claims that they are supervenient upon the cotemporaneous internal physical states of the organisms to which they occur. A moment's reflection should convince us that those who believe that our mental states are determined by the physical processes occurring in our bodies could not have noninternal states in mind. It is not that these noninternal states are not purely psychological, or that they have some nonpsychological mixtures;[17] my remembering that I had a severe headache yesterday is noninternal, although it presumably has no nonpsychological components. It is just that they go beyond what is *here* and *now* in the psychological space of the organism. The notion of a replica of a person as we have used here is a time-bound notion: something is a replica of me *now* but not a replica of me as I was ten years ago or as I will be ten years hence. On the other hand, some psychological states or events that occur to me now spill into other times and places, as it were. Remembering spills into the past; knowing into other places and times. In many cases, this is due to the so-called intentionality of the mental, although intentionality probably does not give us a general explanation of this phenomenon. So we cannot expect all my current psychological states to depend, or supervene, upon my current internal physical states.

There are two general ways of dealing with these apparently nonsupervenient psychological states: first, we can, as we have done, restrict the class of psychological states for which supervenience is to be claimed; second, we can broaden the supervenience base – in our case, the class of physical states – to accommodate the apparent exceptions. Thus, as previously noted, some instances of remembering could be handled by broadening the supervenience base to include a person's past physical history; in terms of the concept of a physical replica, this would amount to strengthening this concept so that a replica must match the original over a *stretch of time*, so that my replica must have qualitatively the same physical history that I have. To handle *de re* psychological states, we would need to broaden the supervenience base to include physical states of objects outside the organism; and ultimately we would need to speak of possible worlds, as Terence Horgan

does, in formulating the thesis of psychophysical supervenience.[18] We could say, following Horgan, that any two possible worlds that are indistinguishable with respect to physical details are indistinguishable from the psychological point of view, or, more briefly, that any two worlds that are physically indistinguishable are in fact one and the same world. This form of generalized supervenience thesis is of broad metaphysical interest, but its implications for specific problems concerning the mental are more difficult to guage than is the case with a thesis that is formulated in terms of individual organisms and their psychological states.

This is one reason why the philosophical interest of the Supervenience Thesis needs to be shown, and this is the task of the Explanatory Thesis. As may be recalled, the Explanatory Thesis affirms that psychological theory needs only to invoke internal psychological states in formulating explanations of human behavior. We now turn to this claim.

V

In support of the Explanatory Thesis we shall try to make plausible the following claim: *the causal-explanatory role of any noninternal psychological state can be filled by some internal psychological state*. If this is true in general, then it will follow that no reference needs to be made in psychological theory to noninternal psychological states. Let us begin with knowing or believing truly. As we saw, these are noninternal. I know that if I turn this knob counterclockwise the burner will go on. Since I want the fire to go on, I turn the knob. My knowledge that turning the knob will cause the burner to go on plays a causal role in the explanation of my action of turning the knob. This is a simple and familiar sort of action explanation. It is clear, however, that knowledge is sufficient but not necessary to construct an action explanation: belief, or firm belief, is also sufficient. If I believe that the burner will go on if the knob is turned, then I will turn on the knob if I have the desire to have the burner go on (assuming that there is no countervailing desire). In fact it is only the element of belief in knowing that is causally productive of the action. Similar comments apply to believing truly. My truly believing that something is so is not more efficacious in producing actions than my merely believing that something is so. As Stich says, "what knowledge adds to belief is psychologically irrelevant".[19]

It is true that whether or not my action succeeds in bringing about the

intended result normally depends on whether the belief involved is true. Thus, whether my action results in the burner being turned on depends on whether my belief that it would go on if the knob is turned is correct. However, it is not part of the object of *psychological* explanation to explain why the burner went on; all it needs to explain is why I turned the knob. It might be objected that not only did I perform the action of turning the knob but I also performed that of *turning on the burner*, and that this latter action does involve – it logically entails – the burner's going on. This is correct; however, the action of turning on the burner, insofar as this is thought to involve the burner going on, is not an action that it is the proper business of psychological theory to explain or predict. The job of psychological explanation is done once it has explained the bodily action of turning the knob; whether or not this action results in my also turning on the stove, my starting cooking the dinner, my starting a house fire, and so on is dependent on events and facts quite outside the province of psychology, and are not the proper concern of psychological theory. Only *basic actions*, not 'derivative' or 'generated' actions, need to be explained by psychological theory.[20]

We now turn to remembering. Memory turns out to be noninternal for two reasons: first, it implies something about the past, and second, in most cases, like knowing, it implies the existence of something other than the rememberer. When a person firmly believes that he remembers but fails to remember in virtue of the failure of one or the other of these two conditions, then we may assume there obtains in him some internal state which is just like a genuine case of remembering except for one of these conditions failing to obtain. This internal state may be some phenomenological experience, 'memory image' or belief about the past characterized by what Russell called the 'déjà-vu' quality; but depending on how remembering or memory is construed, it need not be any sort of conscious experience. This residual element of remembering, when remembering has been stripped down to an internal psychological state, may be called 'seeming to remember'. My claim would be that this seeming remembrance can do all of the explanatory work done by remembering. Thus, when I act in a certain way in part because of my remembering a certain thing, then under the same circumstances my replica will act in the same way because of his seeming to remember the same thing. Whether or not his seeming remembrance is a genuine case of remembering will not affect his behavior. This seems plausible when we reflect that remembering affects our behavior often as a source of belief, that seeming to remember is

to remembering as believing is to knowing, and that, as we saw, insofar as behavior is concerned, belief is psychologically as efficacious as knowledge.

The foregoing exemplifies our strategy in defense of the Explanatory Thesis. The strategy is to argue that within each noninternal psychological state that enters into the explanation of some action or behavior we can locate an 'internal core state' which can assume the causal-explanatory role of the noninternal state; we would in fact argue that this internal core is the causal and explanatory core of the noninternal state. It is in virtue of this core that the noninternal state has the psychological explanatory role that it has.

But why should we believe that there is such an internal core to every explanatory psychological state that is noninternal? Causal considerations of the following general sort make such an assumption both attractive and plausible. In constructing a psychological explanation of a piece of behavior, we are attempting to ascertain a psychological causal antecedent of that behavior. Considerations of causal contiguity and continuity lead to the belief that the proximate cause of the behavior must be located within the organism emitting the behavior – that is, there must be a proximate causal explanation of that behavior in terms of an *internal* state of the organism. Why should we think that there must be an internal *psychological* state which will serve as proximate cause of behavior? This is a difficult question, but part of an answer is contained in the observation that if this internal state has all the causal powers of the corresponding noninternal psychological state in the production of behavior, then there seems to be no reason not to think of it as psychological as well. And in many cases we can identify the internal psychological core of a given noninternal psychological state, as we have done above for knowing and remembering.

With this in mind, let us turn briefly to the remaining cases. When we see a tree, there is some internal phenomenal state going on; some internal representation of the tree will be present in us. In the language of the sense-datum theory, we are sensing a tree-ish sense datum, or we are appeared to tree-ishly. The Explanatory Thesis would claim that whether there is an actual tree out there, or whether we are just having this internal presentation, makes no difference to the behavior emitted. In either case we may reach out for the real or imagined tree, answer 'Yes' when asked 'Do you see a tree?' and so forth. We finally come to actions. Let us return to the case of my replica's signing a check. The observable action he performs is the same as mine when

I sign a check and pay off the mortgage, even though his observable physical action does not issue in signing a check and mine does. But it is clear that the succes or failure of our undertaking here is not really up to us; once the appropriate physical action has been performed, it is not up to us whether that action issues in the signing of a check or the paying off of the mortgage. That depends on factors outside our immediate individual control. These nonbasic actions do not come within the purview of psychological theory; all a psychological theory of behavior needs to explain and predict is the basic actions individuals perform. Societal actions are generally nonbasic; they are produced by the basic actions that we perform, normally the basic bodily movements we can perform at will.

VI

In this final section we return to the question whether there are any positive reasons for thinking that psychological states – at least, the internal ones – supervene on physical states. I shall present one argument for psychophysical supervenience, and it runs as follows. First of all, I propose that we accept, in the present context, the functionalist conception of psychological states as those internal states which serve as intermediary states mediating between stimuli and behavior output.[21] This is the argument:

(1) My replica and I share all our current internal physical properties.

This premise is given *ex hypothesi*, namely, by the description of the situation to the effect that the replica is an exact physical copy of my body.

(2) But this does not mean that, at eacht and every instant, we share the same occurrent, physical properties.

This is evident. After the replica was created, we are going to have different sensory input, and engage in different activities; while I am typing, he is out playing tennis.

(3) We do share structural, dispositional properties. Our basic physical structure is identical – at least for now – and we share the same physical powers, capacities, and dispositions.

(4) One type of such dispositional properties would be the property

of responding in certain characteristic ways to different types of internal or external stimuli. Thus, my replica and I share the same system of stable lawlike relationships of the following form:

stimulus S_1 $\longrightarrow$ behavior output O_1
stimulus S_2 $\longrightarrow$ behavior output O_2
.
.
.

(5) Now the question arises how we are to *explain* these particular input-output relationships. This question arises because these particular patterns of input-output connection are not necessarily shared by other human beings (although of course we expect there will be similarities).

Typically, such explanations will proceed by positing certain *internal states* to mediate the particular input with the particular output associated with that input. Different organisms differ in the output they emit when the same input is applied because their internal states at the time are different. We now come to perhaps the most controversial assumption of this argument, the functionalist interpretation of psychological states:

(6) These internal states posited to explain sensory input-behavior output relationships are *psychological states*.

This is the functionalist conception of a psychological state: a psychological state is a 'functional state' that connects sensory inputs and behavior output in appropriate ways.

(7) If a series of psychological states, along with their mutual interconnections, are posited as the best explanation of the input-output connections in my case, then, in methodological consistency, the same psychological states must be posited in case of my replica. For he and I share the same input-output connections.

This is something like a 'generalization argument' in moral theory. I think that there clearly is a similar consistency requirement in the case of scientific methodology, and (7) is well justified. Of course, (7) is what needs to be estab-

lished, viz. that my replica and I share the same psychological properties. Thus, it follows:

(8) If two organisms or structures are physically identical, then their psychology is also identical. If two organisms coincide in the set of physical properties, they cannot diverge in the set of psychological properties. The psychological supervenes on the physical.

This completes the argument.

The leading idea of the argument is exceedingly simple: since my replica and I share the same input-output relations, and psychological states are just those states posited to explain these relations, the same psychological states must be posited for both of us. As was noted, the most obvious point of controversy in this argument is the functionalist interpretation of psychological state used at step (6). I hesitate to accept this conception of psychological state as a general characterization valid for all psychological states.[22] A fundamental question not touched by this argument is whether conscious (phenomenological) states, such as raw feels, visual images, and the like, are supervenient on bodily states, although this remark will be disputed by those who fully accept the functionalist account of psychological states. In any case, it is difficult to see what a general argument showing the supervenience of the phenomenological would look like. There may of course be broad metaphysical considerations in favor of physicalism from which the supervenience of phenomenological states could be derived. Also, the continuing discovery of lawlike connections, however rough and crude, between phenomenological experience and brain processes serves as limited but indispensable empirical evidence. The only reasonable thing to say at this point, I think, is the rather tame and unsurprising remark that the belief in psychophysical supervenience seems to be based on broad metaphysical and methodological considerations, which are yet to be spelled out, buttressed by what empirical evidence there is for specific psychophysical correlations.[23]

The University of Michigan

NOTES

[1] See Kim, 'Supervenience and nomological incommensurables', *American Philosophical Quarterly* 15 (1978), pp. 149–56; 'The concepts of supervenience' (in preparation). Also see Terence Horgan, 'Supervenience and microphysics' (forthcoming); John Haugeland, 'Weak supervenience' (forthcoming).

[2] This formulation, which closely follows the traditional wording used to explain 'supervenience', turns out to be the weaker of two distinguishable concepts of supervenience, and is in fact too weak to capture what seems to be intended by the use of this term in many contexts. These issues, however, do not fundamentally affect the present discussion. See my 'The concepts of supervenience'.

[3] See Horgan, op. cit.; Haugeland, op.cit.; Kim, 'Causality, identity, and supervenience in the mind-body problem', *Midwest Studies in Philosophy* 4 (1979), pp. 31–49; Harry A. Lewis, 'Is the mental supervenient on the physical?' (forthcoming).

[4] These questions of consistency are subtle and not easy to answer definitively. For further discussion, see Simon Blackburn, 'Moral realism', J. Casey (ed.), *Morality and Moral Reasoning* (Methuen, London, 1971); Lewis, op. cit.; Kim, op. cit.

[5] 'Autonomous psychology and the belief-desire thesis', The Monist 61 (1978), pp. 573–91.

[6] Stich, op. cit., p. 575.

[7] See the selections under 'Psychophysiological isomorphism' by Köhler, Max Wertheimer, and others, in: A Source Book in the History of Psychology, ed. by Richard J. Herrnstein and Edwin G. Boring (Harvard, Cambridge, 1966). The isomorphism thesis defended by these psychologists is in fact a good deal stronger than the thesis that merely affirms the existence of a neural correlate for each mental state.

[8] For example, Donald Davidson, 'Mental events' in Lawrence Foster and J. W. Swanson (eds.), Experience and Theory (University of Massachusetts Press, Amherst, 1970).

[9] On this question, see Davidson, op. cit.; Kim, 'Supervenience and nomological incommensurables'; Lewis, op. cit.; Ted Honderich, 'Psychophysical lawlike connections and their problem' (forthcoming).

[10] There are some subtle questions about how the content of my replica's belief is to be specified, e.g., whether it refers to him or me. See Stich, op. cit., for some of the complications.

[11] So my strategy in this paper can be thought of as splitting Stich's 'principle of psychological autonomy' into two independent theses and considering the second thesis (the Explanatory Thesis) as providing a philosophical rationale for the first (the Supervenience Thesis).

[12] Open Court, La Salle Ill., 1976, p. 127.

[13] And not just when psychophysical supervenience is at issue; I think internal psychological states are just those states for which we should look for physical (neurophysiological) correlates, and which we would expect to enter into psychophysical laws. At least, we could say this: it would be absurd to look for neural correlates for noninternal psychological states, such as knowing (as distinguished from believing) and thinking of Vienna.

[14] See, e.g., Tyler Burge, 'Belief de re', Journal of Philosophy 74 (1977), pp. 338–62; John Pollock, 'De re belief' (ms.).

[15] I expand somewhat on this theme in 'Perception and reference without causality', Journal of Philosophy 74 (1977), pp. 606–20.

[16] This will be disputed by Stich (ibid.) who takes the sameness of truth value as a necessary condition for two beliefs ('belief tokens') to be the same belief ('belief type'). On this criterion, my replica's *de se* belief which he expresses by the sentence 'I have two brothers' is not the same belief as my belief which I would express by the use of the same sentence; for the former is false while the latter is true. From this it follows, for

Stich, that my replica does not share my belief that I have two brothers, from which it further follows that this *de se* belief is nonsupervenient. I reject Stich's criterion of belief identity as an appropriate one in the present context, and would hold that my replica and I share the same belief, in this case, in virtue of each of us exemplifying the property expressed by 'x believes that x has two brothers' or 'x believes himself to have two brothers'. As an analogous case consider: it would be inappropriate to say that my replica and I have *different wants* when each of us wants to eat, on the ground that my replica wants it to be the case that *he eats*, whereas I want it to be the case that *I eat*, and that these two states of affairs desired by us respectively are different. My replica and I have the same want because each of us exemplifies the property expressed by 'x wants it to be the case the x eats'. There are many complex and subtle issues here, and we obviously cannot adequately deal with them here. See Stich, op. cit.; Tyler Burge, 'Individualism and the mental', Midwest Studies in Philosophy 4 (1979), pp. 73–121; Stephen L. White, 'Partial character and the language of thought' (forthcoming).

[17] As perhaps hinted at by Stich, op. cit., p. 574.

[18] Horgan, op. cit.; also see Haugeland, op. cit.

[19] Op. cit., p. 574.

[20] I have heard Richard Brandt defend this claim, although he has not done so in print. See for a discussion of some relevant related issues William P. Alston, 'Conceptual prolegomena to a psychological theory of itentional action', in S. C. Brown (ed.), Philosophy of Psychology (Harper & Row, New York, 1974).

[21] This is somewhat simplified. See, e.g., Hilary Putnam, 'The nature of mental states' in his: Mind, Language and Reality: Philosophical Papers II (Cambridge University Press, London, 1975); N. J. Block, 'Troubles with functionalism', Minnesota Studies in the Philosophy of Science IX (University of Minnesota Press, Minneapolis, 1978).

[22] See, e.g., Block, op. cit.; N. J. Block and J. A. Fodor, 'What psychological states are not', Philosophical Review 81 (1972), pp. 159–81.

[23] I have benefited from discussing with Terence Horgan many of the issues touched on in this paper.