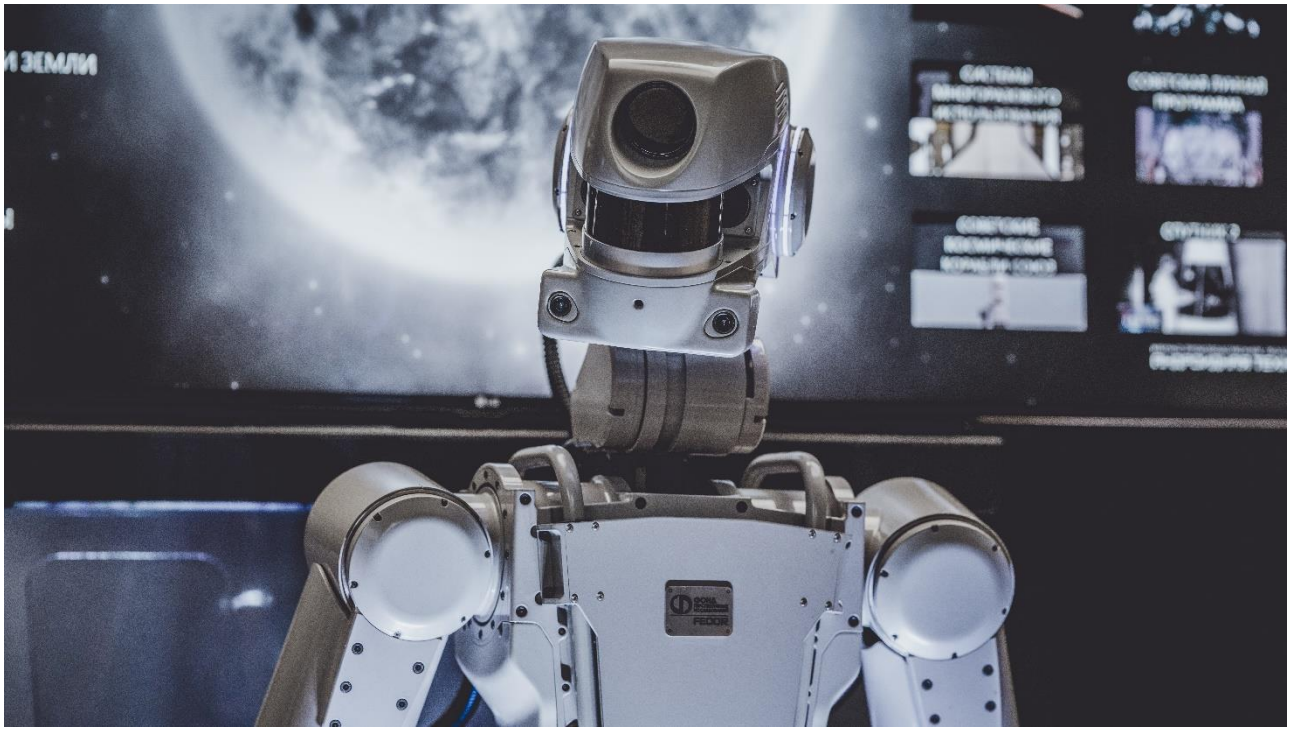

KI und Vertrauen

1 EINLEITUNG

Klimawandel, Globalisierung, Digitale Transformation – zahlreiche Veränderungen und Trends prägen unsere Welt. Die fortschreitende Technologisierung ist ein Trend, der zum Teil noch in den Kinderschuhen steckt. Technologien wie die künstliche Intelligenz (KI) haben aber ein enorm großes Potenzial, wodurch sich unsere Welt und unser Alltag sowohl beruflich wie auch privat grundlegend verändern können (Vgl. Saßmannshausen & Heupel, 2020). Neben technischen Rahmenbedingungen und Möglichkeiten sind speziell bei dieser Technologie auch ethische Aspekte zu berücksichtigen. Dem Einsatz von KI stehen einige Herausforderungen gegenüber (Vgl. Corves & Schön, 2020), die Hersteller und Nutzer von KI-Anwendungen gemeinsam bewältigen müssen, um eine positive und wertschöpfende Entwicklung schaffen zu können. (Vgl. Leopold, 2022)

„Immer wenn wir das menschliche Urteil durch eine Formel ersetzen können, sollten wir dies zumindest in Betracht ziehen“~ (Kahneman, 2012, S. 288)

Sollten wir wirklich tun was Kahneman, Nobelpreisträger der Psychologie sagte? Klar ist, dass der Einsatz von KI speziell im Management die Chance bietet Entscheidungen objektiv zu treffen (Vgl. McAfee & Brynjolfsson, 2017). Allerdings stellt sich die Frage, ob wir das überhaupt wollen. Sollte jede Entscheidung objektiv getroffen werden? Können wir uns überhaupt mit der getroffenen Entscheidung einer KI-Anwendung

abfinden? So oder so – falls wir uns durch die KI unterstützen lassen wollen müssen wir lernen ihr zu vertrauen. Doch was ist hierfür nötig und warum fällt es uns so schwer KI-Anwendungen zu vertrauen? KI agiert nicht automatisch, sondern autonom – sie agiert unerwartet und überraschend für uns. Grund hierfür ist die Andersartigkeit im Vergleich zu bisherigen Systemen, die programmiert wurden, um daraufhin klar vorgegebene Abläufe automatisch wiedergeben zu können. Ziel einer künstlichen Intelligenz dagegen ist es eher einen Rahmenprozess oder ein Entscheidungsspektrum vorzugeben, durch welches auf Basis einer Datengrundlage Entscheidungen getroffen werden. Wie genau der Prozess abläuft ist ad hoc für den Mensch oft nicht nachzuvollziehen. Für eine funktionierende Kooperation zwischen Mensch und KI und damit der Erhaltung oder Steigerung der Gesamtsystemleistung ist es notwendig, dass wir Gefühle der Unwissenheit und Überraschung durch Vertrauen ausgleichen und damit eine Akzeptanz der Technologie erreichen aus (Vgl. Saßmannshausen & Heupel, 2020). Bei KI handelt es sich im Vergleich zur Robotik „um eine Technologie, die nicht als Tool, sondern als soziale Einheit verstanden wird“ daher spricht man an dieser Stelle in erster Linie von Vertrauen und nicht von Akzeptanz. Die Herstellung von digitalem Vertrauen galt laut einigen befragten Managern bereits 2017 als wesentliche Herausforderung sowie Erfolgsfaktor für neue und bestehende (digitale) Geschäftsmodelle (Vgl. Barton, Bender & Marque, 2017; Corves & Schön, 2020). Die Frage, die sich daraus ergibt, ist was Menschen brauchen, um Vertrauen aufbauen zu können und wie sie zu KI stehen. 2020 gaben 79% der befragten Deutschen beispielsweise an, dass Leitlinien für das Vertrauen in KI sehr wichtig für sie sind, während sich lediglich 26% auf die Informationen einer KI verlassen können (Vgl. KPMG, 2021). Daraus lässt sich ableiten, dass sich die Menschen noch nicht sicher genug fühlen und aktuelle Leitlinien noch nicht ausreichen. Betrachtet man, wie Menschen grundsätzlich empfinden, wenn sie an KI denken, fällt auf, dass immer weniger Menschen negativ eingestellt sind. Der allgemeine Trend geht also eher dahin, dass Menschen einer KI positiver gegenüberstehen oder zumindest negative Vorurteile schrittweise ablegen und von einem neutralen Standpunkt aus offen für Neues sind (Vgl. TÜV, 2019). Welche Einflussfaktoren hierfür verantwortlich sind und wie das Vertrauen in KI beeinflusst und aufgebaut werden kann werden wir im Folgenden näher betrachten .

1.1 Geschichtliche Herleitung „Einführung in die kuriose Welt der Künstlichen Intelligenz“?

Bereits seit Jahrzehnten versuchen Menschen Maschinen zu erschaffen, die sich durch Gefühle wie Liebe, Hass, Freude und Angst mit einem menschlichen Bewusstsein auszeichnen. Filme wie Nummer 5 lebt! (1986), Matrix (1999) oder die Serie Black Mirror (2011-2019) spiegeln genau das wider und zeigen potenzielle Auswirkungen neuer Technologien auf die Gesellschaft. Obwohl die Verfilmungen häufig über den aktuellen Stand der Technologie hinausgehen, machen Forscher und Unternehmer rasante Fortschritte, die im Folgenden kurz dargestellt werden (Vgl. Corves & Schön, 2020).

Bereits seit den 1960-er Jahren wird rund um die Technologie KI geforscht (Vgl. Leopold, 2022). Im neuen Jahrtausend machte KI speziell in den Bereichen der statistischen Mustererkennung und lernenden Algorithmen (Machine Learning) Fortschritte (Vgl. McAfee & Brynjolfsson, 2017; Buchkremer, 2020). Mit 30-jähriger Verspätung erfüllte sich 1997 die These des Wirtschaftsnobelpreisträgers Herbert A. Simon, der meinte in 10 Jahren würden Computer in der Lage sein, menschliche Schachweltmeister zu schlagen. „Deep Blue“ von IBM schlug den damaligen Schachweltmeister Garri Kasparov in sechs Partien (Vgl. Dettmann & Kopecz, 2020). Nachdem es also nicht mehr nur darum ging, was die KI lernen kann, sondern auch wie schnell war es dem Schachprogramm Giraffe im Jahr 2015 möglich das Brettspiel in 72 Stunden auf Meisterniveau zu erlernen .2016 war ein besonderes Jahr für Google, denn dessen neuronales Netzwerk Alpha Go besiegte den damaligen Europameister und 18-fachen Weltmeistern Lee Sedol. Die KI Technologie rückte damit stärker in die Öffentlichkeit, da der Algorithmus das Go-Spiel in wenigen Wochen lernen konnte, während man bei Menschen von einer Lernzeit von zwei Jahren ausgeht (Vgl. Silver et al., 2016; Dettmann & Kopecz, 2020). Darauf folgten Siege von Algorithmen in heads-up, no-limit Texas Hold’em Poker und Dota 2 (Vgl. Dettmann & Kopecz, 2020). Aktuelle realwirtschaftliche Einsatzgebiete sind die Mustererkennung bei autonomen Fahrzeugen von Tesla oder Mobileye, Spracherkennungssysteme wie Alexa von Amazon oder Cortana von Microsoft. Im Alltag begegnen wir KI häufig in Form eines Chatbots und indirekt durch Anwendungsbereiche in der Industrie beispielsweise bei Produktionsprozessen (Vgl. Leopold, 2022). Zusammenfassend kann man sagen, dass KI Technologien mittlerweile ein fester Bestandteil unseres Alltags sind und sich künftig weiter ausbreiten werden. Ein potenzielles Hindernis für diese Ausbreitung ist das mangelnde Vertrauen, welches als Schlüssel für die notwendige Akzeptanz dient (Vgl. Siau & Wang, 2018). Das ist enorm wichtig, wenn wir uns künftig tatsächlich Arbeit von KI Anwendungen abnehmen lassen wollen. Speziell wissensbasierte Arbeiten könnten von KI übernommen werden, damit die Menschen sich stärker auf kreative Tätigkeiten fokussieren können. Hierfür ist es allerdings notwendig, sich auf die Ergebnisse der KI verlassen zu können und ihnen zu vertrauen (Vgl. McAfee & Brynjolfsson, 2017; Dettmann & Kopecz, 2020).

2 BEGRIFFSDEFINITIONEN --> GGF. KI TRUST JOURNEY AUCH DEFINIEREN

2.1 DEFINITION KI

Obwohl die künstliche Intelligenz (KI) auf eine langjährige Entwicklungsgeschichte zurückblickt (McCarthy et al., 1995), gibt es keine einheitliche Definition des Begriffes (Vgl. Allen, 1998; Wang, 2019). In der Literatur wird häufig der Begriff der intelligente Systeme verwendet. Bei dieser Interpretation wird zwischen der sogenannten schwachen KI und der starken KI unterschieden (Vgl. Lämmel & Cleve, 2020).

Bei der schwachen KI wird ausschließlich ein Denkvorgang simuliert. Hierbei geht es um die Lösung von Problemen oder das Anfertigen von Analysen in einem vordefinierten Kontext unter Einsatz einer domänenspezifischen Datengrundlage (z. B. Medizin, Geoinformationen oder Sport). Im Fall der starken KI hingegen wird Maschinen ein tatsächliches künstliches Bewusstsein zugeschrieben. Die Anwendung ist sich somit ihres mentalen Zustands und ihrer Intelligenz bewusst und kann Aktionen auf der Grundlage selbst entwickelter Intention ausführen. (Vgl. Corves & Schön, 2020). Im Rahmen der vorliegenden Arbeit wird KI im Sinne der schwachen KI verstanden, da diese bereits erfolgreich angewendet wird, während die starke KI bislang ein noch eher visionäres Konzept darstellt (Vgl. Apt & Priesack, 2019). In der Praxis wird der Begriff „künstliche Intelligenz" primär auch als Sammelbegriff für vielfältige Technologien verwendet (Vgl. Mele, Spena & Peschiera, 2018).

Machine Learning (ML) gilt als Kerntechnologie der KI, da viele weitere KI-Technologien darauf basieren und KI in der Praxis vor allem mit Lernfähigkeit verbunden wird (Vgl. Mele, Spena & Peschiera, 2018; Vgl. Döbel et al., 2018). ML ermöglicht die automatisierte Erkennung von Mustern in Daten und eignet sich insbesondere für die Extraktion aus großen Datensätzen. ML-Systeme sind durch Training mit vorgegebenen Ein- und Ausgabedaten in der Lage, die Regeln für die automatisierte Lösung von Aufgaben eigenständig zu erlernen. Im Hinblick auf ML wird zwischen verschiedenen Lernstilen differenziert: Supervised, Unsupervised und Reinforced Learning (Vgl. Seitz, Willbold & Haiber, 2022).

- Das Überwachte Lernen, bei dem der Computer Daten und die Antworten zu einer Aufgabe erhält und daraus lernt.
- Das Unüberwachte Lernen, bei dem der Computer Daten ohne Antworten erhält und selbständig Muster in den Daten erkennen muss.
- Das sogenannte Verstärkende Lernen, bei dem der Computer Daten erhält, sowie das Ziel, das er beim Lernen erreichen soll.

Für das Supervised Learning müssen zusätzlich zu den Rohdaten auch entsprechenden Labels vorgegeben werden. Dadurch ist es den Systemen möglich, Objekte zu klassifizieren und Vorhersagen zu treffen. Beim Unsupervised Learning dagegen sind keine entsprechenden Markierungen notwendig. Beim Reinforcement Learning verbessern ML-Systeme ihre Algorithmen bzw. Entscheidungen, indem sie aus der Interaktion mit ihrer Umwelt Feedback erhalten (Vgl. Döbel et al., 2018; Seitz, Willbold & Haiber, 2022).

Als zusätzliches, besonders fortschrittliches Teilgebiet des ML gilt das Deep Learning. Dabei kommen sogenannte tiefe neuronale Netze zum Einsatz, die in mehrere Ebenen unterteilt sind. Je mehr Ebenen, desto komplexere Sachverhalte lassen sich erlernen. Die Nachvollziehbarkeit der Entscheidungen kann durch die Komplexität allerdings meist nichtmehr gewährleistet werden bzw. geht vollständig verloren (Vgl. Seitz, Willbold & Haiber, 2022).

2.2 Definition Vertrauen

Wie stark eine Person vertraut, beeinflusst zunächst deren Akzeptanz und in einem weiteren Schritt deren Verhalten und Interaktion mit anderen Personen, Prozessen, Institutionen oder Technologien. (Vgl. Siau & Wang, 2018; Jung & Garrel, 2021).

Die Vertrauensforschung untersucht neben anderen Aspekten auch Mechanismen des Vertrauensaufbaus. Diese Mechanismen können weiterführend auch für die Stärkung von bereits vorhandenem Vertrauen in KI-basierte Technologien eingesetzt werden (Vgl. Corves & Schön, 2020). Digitales Vertrauen (engl. Digital Trust) bezeichnet das Vertrauen, welches digitalen Technologien von Nutzerseite entgegengebracht wird (Vgl. Siau & Wang, 2018). Für die Beantwortung der Forschungsfrage ist es entscheidend welche Faktoren das (digitale) Vertrauen beeinflussen. Mit diesem Verständnis können Strategien zur Stärkung des Vertrauens in eine KI entwickelt und bewertet werden. Der Begriff des Vertrauens kommt ursprünglich aus der Psychologie, in der vordergründig das zwischenmenschliche Vertrauen betrachtet wird. Um die beeinflussenden Faktoren vollständig zu verstehen, wird im Folgenden zunächst das zwischenmenschliche Vertrauen, welches als Basis für das Verständnis von Vertrauen in KI dient, analysiert und dessen Einflussfaktoren dargestellt. Anschließend werden die Unterschiede von zwischenmenschlichem Vertrauen und digitalem Vertrauen aufgezeigt und abschließend Implikationen für das weitere Vorgehen abgeleitet.

2.2.1 Zwischenmenschliches Vertrauen

Vertrauen lässt sich auf sehr unterschiedliche Weisen definieren. Einige Forscher fokussieren sich bei der Definition auf das Element der Erwartung von Verhaltensweisen oder Ergebnissen. Ein anderer Ansatz beschreibt Vertrauen als Absicht oder Bereitschaft zu handeln. In diesem Fall geht es über die Einstellung hinaus und meint die tatsächliche Handlung, mit der man sich beispielsweise in den Zustand der Verwundbarkeit begibt. Laut Rousseau et al. (1998) und Lee & See (2004) ist die Definition von Mayer et al. (1995) die am weitesten verbreitete und akzeptierte Definition von Vertrauen. Mayer identifiziert Verwundbarkeit als kritisches Element von Vertrauen. Darüberhinausgehend kann Vertrauen als Ergebnis von Handlungen, die Menschen in bestimmte Zustände oder Situationen versetzen, definiert werden. Die Unterschiede in den

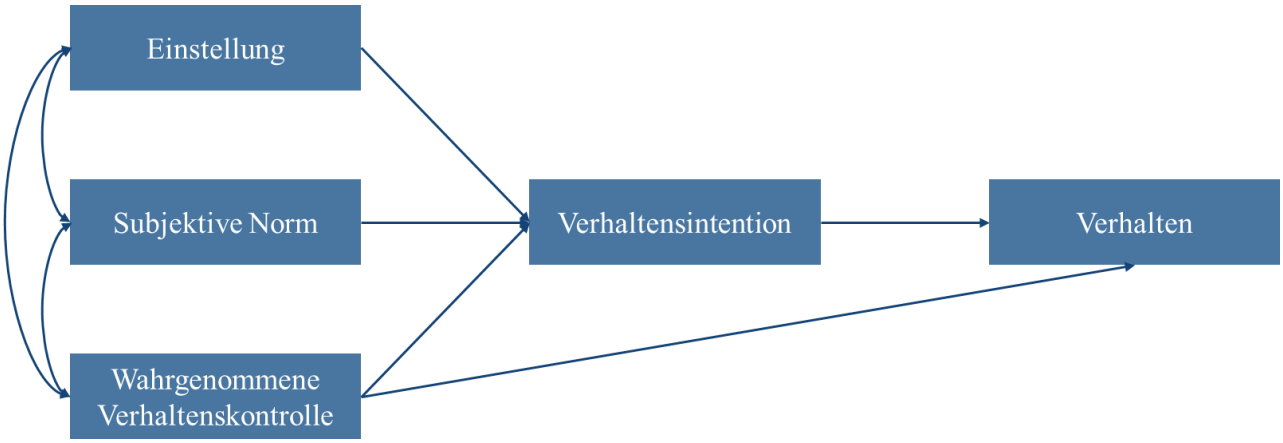

Abbildung 1: Theory of Reasoned Action nach Fishbein und Ajzen, 1975 und 1980

aufgezeigten Definitionsweisen zielen auf die Frage ab, ob Vertrauen eine Überzeugung, Einstellung, Absicht oder ein Verhalten ist. Die „Theory of Reasoned Action“ (Fishbein und Ajzen, 1975 und 1980) bietet eine schematische Verknüpfung von Verhaltensabsichten mit den elementaren Einstellungen und subjektiven Normen (Vgl. Lee & See, 2004). Diese schafft einen Rahmen, der helfen kann, die widersprüchlichen Definitionen von Vertrauen in Einklang zu bringen (Abbildung 1).

Das Modell zeigt: Einstellungen, subjektive Normen und wahrgenommene Verhaltenskontrollen bedingen sich gegenseitig und sowohl positive als auch negative Wechselwirkungen können entstehen, welche Verhaltensintentionen auslösen, die wiederum zu einer bestimmten Handlung führen können. Wichtig ist an dieser Stelle zu beachten, dass Verhaltensintentionen nicht grundsätzlich eine Aktion nach sich ziehen, da zusätzlich umweltbedingte und kognitive Zwänge darauf wirken, ob eine Verhaltensabsicht in die Tat umgesetzt wird oder nicht. Das Vertrauen, eines Menschen lässt sich also durch 2 entscheidende Faktoren beeinflussen: Wahrnehmungen und Überzeugungen, die durch die Verfügbarkeit von Informationen beeinflusst werden. (Vgl. Lee & See, 2004).

Nachdem nun festgestellt wurde, dass Vertrauen nicht eindeutig definiert werden kann, sich aber in eine Konstrukt aus Wahrnehmungen, Überzeugungen, Einstellungen und tatsächlichem Verhalten einordnen lässt muss zusätzlich berücksichtig werden, dass Vertrauen sich auf zwei Ebenen beschreiben lässt – der kognitiven und der affektiv-emotionalen Ebene. Die kognitive Vertrauensebene basiert auf der kognitiven Unwissenheit, einem Zustand, in dem sich Menschen ständig befinden, weil sie kein vollständiges Wissen erlangen können. Könnten Menschen diesem Zustand entfliehen wäre das Konstrukt des Vertrauens obsolet und man würde von reinem Kalkül sprechen. Allerdings können durch kognitive Fähigkeiten grundsätzlich nicht alle entstehenden Informations- oder Erkenntnislücken geschlossen werden. Eine Möglichkeit diesen Zustand dennoch zu überwinden ist zu vertrauen (Vgl. Corves & Schön, 2020; Petermann, 2013). Dies ist notwendig, um handlungs- und transaktionsfähig zu bleiben. Menschen versuchen grundsätzlich kognitive, gelernte rationale Maßstäbe anzulegen. Greifen diese nicht und ein Nutzer hat mangelnde Erfahrungswerte oder zu viele Alternativen (Choice Overload) werden neue Technologien häufig negativer bewertet (Vgl. Corves & Schön, 2020).

Auf der affektiv-emotionalen Ebene kann die kognitive Be- oder Überlastung dennoch überwunden werden. Eine Handlungsfähigkeit entsteht hier trotz verbleibender Ungewissheit durch unser sogenanntes Bauchgefühl. Wissenschaftlich kann man sagen, dass Bauchgefühl auf „sozialisierten, zum Teil unbewusst oder habitualisiert eingesetzten Mustern“ basiert. Auf dieser Ebene können Vertrauensvorschüsse gewährt werden, die kognitive Defizite füllen. Ein rein affektives Vertrauen, welches nicht kognitiv gesteuert wird bezeichnet man als blindes Vertrauen (Vgl. Corves & Schön, 2020)

2.2.2 Humane Vertrauensquellen und -muster

Vertrauen ist ein dynamisches Konstrukt, das sich schrittweise aufbaut (Vgl. Siau & Wang, 2018). Dieser Prozess der Vertrauensbildung und -weiterentwicklung wird in Kapitel 3 der KI Trust Journey detaillierter

erklärt, beginnt aber zunächst bei den Vertrauens- bzw. Bezugsquellen, welche als Ankerpunkte der Überprüfung einer Vertrauenswürdigkeit und Gewährung eines Vertrauensvorschusses dienen. Spricht man von Vertrauen müssen immer zwei Blickwinkel berücksichtigt werden, die des Vertrauensgebers und des Vertrauensnehmers. Vertrauensquellen können variabel zugrunde gelegt werden, demnach kann Vertrauen personenspezifisch, prozessbasiert oder institutionsbasiert sein. Die Erwartungen des Vertrauensgebers ergeben sich durch die Vertrauensquelle. Von Personen wird beispielsweise „Integrität, Kompetenz, Wohlwollen und Aufrichtigkeit“ erwartet, während bei einem Prozess als Vertrauensquelle Attribute wie Zuverlässigkeit und Einfachheit der Nutzung betrachtet werden. Von Institutionen hingegen erwarten wir beispielsweise Sicherheit und Stabilität (Vgl. Corves & Schön, 2020).

Durch die Hinzuziehung humaner Muster kann die Wahrscheinlichkeit ermittelt werden, mit welcher sich die Vertrauenserwartungen eines Nutzers realisieren lassen. Dieser Abgleich erfolgt sowohl auf affektiver als auch kognitiver Vertrauensebene, auf ersterer i.d.R. habitualisiert und unbewusst. Unter Zuhilfenahme dieser Muster können Vertrauenserwartungen aktiv berücksichtigt werden und Ankerpunkte gesetzt werden, die die Vertrauenswürdigkeit neuer Produkte sicherstellen oder mindestens auf ein höheres Niveau bringen. Humane Muster können in 5 Kategorien unterteilt werden: Übereinstimmung, Ähnlichkeit, Kompatibilität, Indikatoren und Transfer (Vgl. Corves & Schön, 2020). Das Zusammenspiel aus Vertrauensgeber, -nehmer, Vertrauensmustern und Vertrauensquellen ist in der untenstehenden Darstellung visualisiert worden.

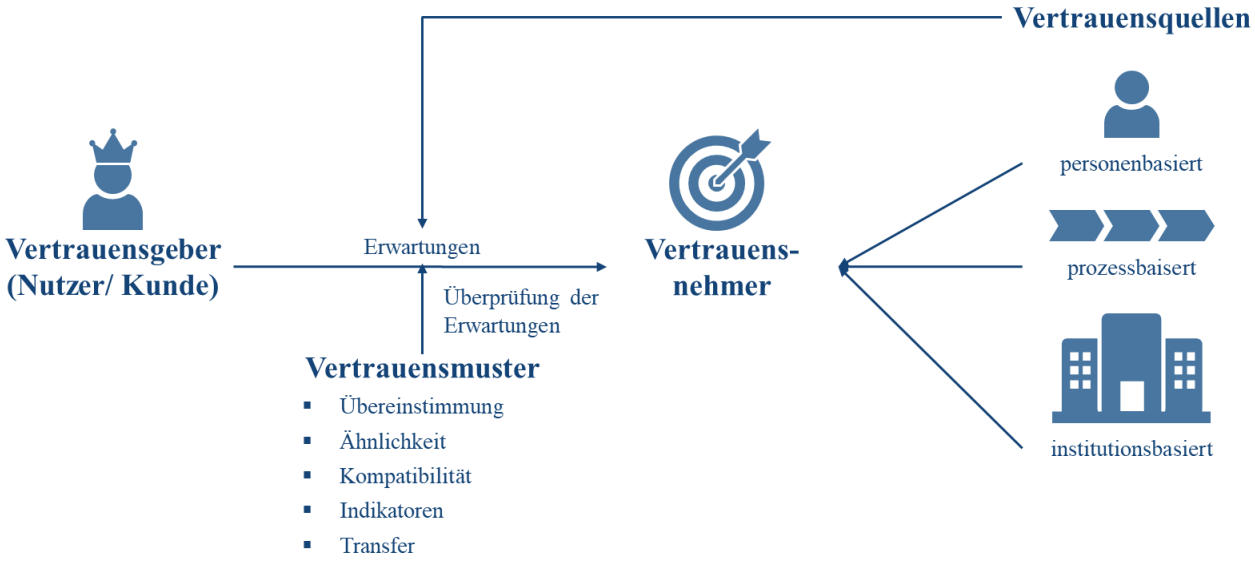

Abbildung 2: Konstrukt des Vertrauens (eigene Darstellung)

2.3 DIGITALES VERTRAUEN

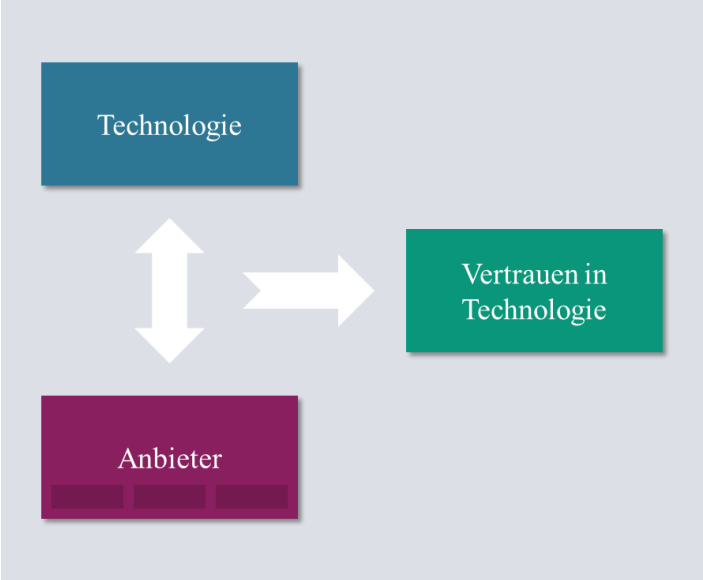

Abbildung 3: Einflussfaktoren auf digitales Vertrauen (eigene Darstellung in Anlehnung an Siau & Wang, 2018)

Der zentrale Unterschied von zwischenmenschlichem Vertrauen und digitalem Vertrauen ist, dass bei ersterem eine bipolare Beziehung besteht. Bei digitalem Vertrauen dagegen besteht mindestens eine tripolare Beziehung, wie im Bild links zusehen, aus Mensch, Technologie und Technologieanbieter (Vgl. Siau & Wang, 2018). Auch Corves & Schön zeigen unterschiedliche Zusammensetzungen in ihrem Vertrauensmodell in digitale Technologien auf. Grundsätzlich lässt sich allerdings sagen, dass die Vertrauenserwartungen darauf basieren, welche Vetrauensquelle zu Grunde liegt. Basierend auf der bipolaren Theorie kategorisieren Corves und Schön die Vertrauensquellen wie in Kapitel 2.2.2 beschrieben nach Personen (bspw. Personifizierung bei Sprachassistenten), Institutionen (bspw. Technologieanbieter), oder Prozessen (bspw. Vorhersagealgorithmen) (Vgl. Corves & Schön, 2020). Erweitert man das Konstrukt des Vertrauens allerdings auf das tripolare Zusammenspiel, in dem sowohl die Technologie selbst als auch der Anbieter eine Rolle spielen und berücksichtig die Fortschritte, bei der Entwicklung der KI in den letzten Jahren, ist es sinnvoll die Aufteilung nach Siau und Wang zu verfolgen, da diese zum einen verschiedene KI-nahe Technologien im Detail abgrenzt und zum anderen folgende Änderungen hinsichtlich des Vertrauen in Technologien generell vornimmt: die prozessbasierte Vertrauensquelle wird um weitere technologische Charakteristiken erweitert, nämlich deren Leistungsfähigkeit und geplantem Einsatzzweck. Diese beiden Aspekte ließen sich zuvor in die Betrachtung des Prozesses integrieren, da sie allerding in Bezug auf digitales Vertrauen eine wichtigere Rolle spielen, werden diese gesondert betrachtet. Des Weiteren wird die institutionsbasierte Vertrauensquelle in die Umwelt Vertrauensquellen eingegliedert, welche zusätzlich Kultur und Aufgaben umfasst. Die personenspezifische Vertrauensquelle bleibt nahezu gleich bestehen und wird lediglich in Persönlichkeit und Fähigkeiten unterteilt (Vgl. Siau & Wang, 2018). Zuletzt ist noch anzumerken, dass die beiden Herangehensweisen sich dahingehend unterscheiden, dass Siau und Wang die von Ihnen definierten Vertrauensquellen detaillierter beschreiben, während Corves und Schön bereits Erwartungshaltungen an die einzelnen Vertrauensquellen formulieren.

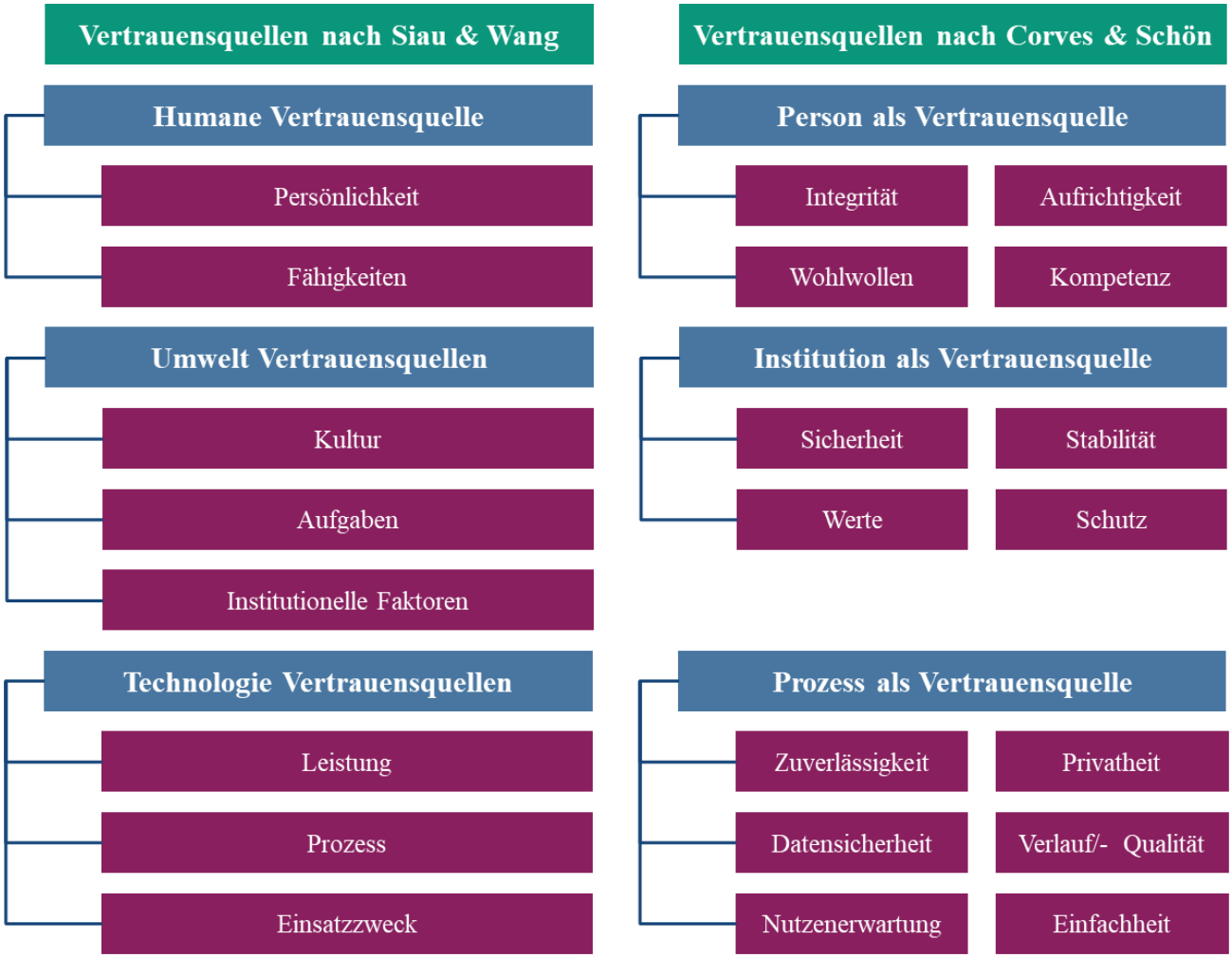

Abbildung 4: Gegenüberstellung der Vertrauensquellen nach Sau & Wang und Corves & Schön (eigene Darstellung in Anlehnung an Siau & Wang 2018 und Corves & Schön, 2020)

Zusätzlich zu dem tripolaren Ansatz, den Siau und Wang beschreiben entsteht eine zusätzliche Komplexitätsebene, wenn eine Anwendung aus verschiedenen technologischen Elementen unterschiedlicher Anbieter besteht. Ein klassisches Beispiel dafür wären autonome Fahrzeuge. Hierbei halte es sich abstrakt gesprochen um ein Technologiebündel verschiedener Zulieferer und dem Hersteller, durch welchen das fertige Fahrzeug dem Endkunden zur Verfügung gestellt wird. Im Anwendungsfall Waymo beispielsweise handelt es sich um ein Unternehmen zur Entwicklung von Technologien für autonome Fahrzeuge. Waymo wurde im Dezember 2016 als Tochtergesellschaft von Alphabet gegründet. Die Systemsoftware des Autos heißt Google Chauffeur. Die tatsächliche physische Plattform, also das Auto wird allerdings nicht vom Unternehmen selbst produziert, sondern von unterschiedlichen anderen Automobilherstellern wie Toyota, Jaguar oder Stellantis zugekauft. Dies bedeutet, dass nicht nur das Vertrauen des Endkunden in den Hersteller, sondern auch in verschiedene verbundene Marken und Zulieferer hergestellt werden muss. Im Fall Waymo wären das mindestens Alphabet als Mutterkonzern, Google als Markenträger, Waymo als Softwarehersteller, und der jeweilige Automobilhersteller, der die physische Plattform liefert (Vgl. Waymo LLC, 2022).

Wie oben bereits beschrieben, heben Siau und Wang deutlich hervor, dass die Faktoren der Vertrauensbildung in verschiedenen technologischen und digitalen Anwendungen unterschiedlich sind. Im Folgenden werden die Faktoren im Detail beschrieben und die Besonderheiten im Bereich der angewandten KI hervorgehoben. Menschliche Charakteristika berücksichtigen im Wesentlichen die Persönlichkeit bzw. die Bereitschaft des Vertrauensgebers zu vertrauen und die Fähigkeit des Vertrauensnehmers, mit Risiken umzugehen. Die Persönlichkeit oder Vertrauensbereitschaft des Vertrauensgebers betrachten Siau und Wang sehr allgemein in Abhängigkeit von verschiedenen Erfahrungen, Persönlichkeitstypen und kulturellen Hintergründen. Die Fähigkeit des Vertrauensnehmer mit Risiken umzugehen, bezieht sich in der Regel auf die Kompetenz eines Vertrauensnehmers zur Erfüllung von Aufgaben in einem bestimmten Bereich. Wenn ein Sprachassistent beispielsweise realistische Wettervorhersagen machen kann, wird der Nutzer diese Funktion nutzen. Ansonsten wird er weiterhin selbst recherchieren (Vgl. Siau & Wang, 2018).

Die umweltbasierten Vertrauensquellen konzentrieren sich auf die Art der Aufgaben, den kulturellen Hintergrund und institutionelle Faktoren. Die Art der Aufgabe bezieht sich beispielsweise auf die Komplexität bzw. das Vorhandensein von Informationen für die Erfüllung der Aufgabe. Der kulturelle Hintergrund kann auch mit einem Land, einer bestimmten Region oder Religion in Verbindung gebracht werden. So neigen Amerikaner beispielsweise dazu, Fremden zu vertrauen, die der gleichen Gruppe angehören, und Japaner neigen dazu, denjenigen zu vertrauen, mit denen sie direkt oder indirekt verwandt sind. Institutionelle Faktoren umfassen zwei Hauptaspekte: die situative Normalität und strukturelle Sicherheiten. Das bedeutet, dass der Vertrauensgeber aufgrund der situativen Normalität erwartet, dass eine Situation normal ist und so verläuft, wie es ihm oder ihr bekannt ist. Beruft sich der Vertrauensgeber auf strukturelle Sicherheiten so erwartet er, dass kontextuelle Bedingungen, wie Versprechen, Verträge, Garantien und Vorschriften eingehalten werden. Dem Sicherheitsstreben und daraus entstehenden Ängsten vor existentieller Bedrohung steht die Bequemlichkeit gegenüber, welche diejenigen Situationen bezeichnet, in denen KI-Anwendungen uns alltägliche Aufgaben erleichtern (Vgl. Siau & Wang, 2018).

Der am laut Siau und Wang wenigsten erforschte Bereich und gleichzeitig für uns relevanteste sind die Technologie Vertrauensquellen. Hier sind die Besonderheiten einzelner Technologiezweige verordnet. Diese können aus drei Blickwinkeln analysiert werden: die Leistung der Technologie, ihre Prozesse bzw. Prozesseigenschaften und ihr Einsatzzweck. Siau und Wang heben klar hervor, dass die Technologiemerkmale, die sich auf das Vertrauen auswirken, beispielsweise bei KI bzw. ML und Robotik anders sind als bei anderen Technologien. Da KI im Vergleich zu anderen Technologien viele neue Merkmale aufweist, müssen ihre Leistung, ihr Prozess und ihr Einsatzzweck definiert, transparent sowie verständlich dokumentiert und im Vertrauensbildungsprozess berücksichtig werden (Vgl. Siau & Wang, 2018). Hierauf besonderes Augenmerk zu legen ist wichtig, da KI als „hybrides Wesen“ bezeichnet wird. Dies bedeutet, dass sich keine eindeutige Vertrauensquelle zuordnen lässt, die zentral für den Vertrauensaufbau ist, sondern alle der Technologie Vertrauensquellen Elemente zielgruppenspezifisch betrachtet muss (Vgl. Corves & Schön, 2020). Im folgenden Kapitel wird

zunächst das Konzept der Trust Journey erläutert, in welche sich ein zweistufiges Modell der Vertrauensbildung einordnen lässt. Dieses Modell (Tabelle 3) zeigt die technologischen Merkmale, die mit der Leistung, dem Prozess und dem Einsatzzweck von KI zusammenhängen, sowie deren Auswirkungen auf das Vertrauen.

3 KI TRUST JOURNEY

Corves und Schön definieren eine Trust Journey (Abbildung 5) analog zu einer Customer Journey „als eine Reise von dem initialen Vertrauensvorschuss bis zum berechtigten Vertrauen des Kunden.“ (Corves & Schön, 2020). Die bereits erläuterten Vertrauensquellen, Vertrauenserwartungen und Vertrauensmuster bilden die Vertrauensanker, welche sowohl auf affektiver als auch kognitiver Ebene verortet sein können. Die so genannten Vertrauensanker sind entscheidende Momente für die gesamte Trust Journey, sie können Kernpunkte für den Vertrauensaufbau sein und über das digitale Wohlbefinden entscheiden bzw. dieses steigern und das digitale Kundenerlebnis verbessern. Vertrauensanker sind also analog zu den Touchpoints zu betrachten, welche in einer klassischen Customer Journey die Berührpunkte des Kunden mit der Marke oder dem Unternehmen darstellen. Der Trust Journey Ansatz ist bisher noch nicht so populär bietet aber die Grundlage für die Identifikation wichtiger Ansatzmöglichkeiten für den Vertrauensaufbau mittels phasenspezifischen Vertrauensanker. Aktuell werden häufig nur einzelne Phasen wie die kognitive Kontrolle fokussiert wodurch die Technologie selbst dafür verantwortlich ist ob potenzielle Nutzer ihr vertrauen oder nicht. Durch den Trust Journey Ansatz wird ein umfänglicheres Bild über alle Phasen des Vertrauens geschaffen, welches weitere Eingriffsmöglichkeiten aufzeigt (Vgl. Corves & Schön, 2020).

Eine Trust Journey wird in verschiedene Phasen des Vertrauens unterteilt, die jeweils analysiert und strategische Maßnahmen geplant werden können. Die menschliche Wahrnehmung ist geprägt von individuell entstandenen Einstellungen und Emotionen, diese bilden sich in der Regel durch bisherige Erfahrungen und Einflüsse des persönlichen Umfeldes, sowohl im sozialen als auch im institutionellen Sinne (Vgl. Siau & Wang, 2018). In Bezug auf das Vertrauen bezeichnet man diesen Zustand, in dem sich Menschen befinden, bevor der Prozess des aktiven Vertrauensaufbaus beginnt, als Vertrauensvorschuss. Der Vertrauensvorschuss ist also der erste Schritt hin zu einem kontinuierlichen Vertrauensaufbau und dem daraus entstehenden berechtigten Vertrauen (Vgl. Corves & Schön, 2020). Der erste Eindruck, den wir von einer Person oder auch einer Technologie haben bezeichnet man als initiales Vertrauen, welches sich allerdings bereits auf das künftige Vertrauen auswirkt . Auf die Phase der initialen Vertrauensbildung folgt die kontinuierliche Vertrauensentwicklung. Diese beschreibt hauptsächlich die Pflege und Erhaltung des Vertrauens. Im Folgenden werden beide Phasen erläutert und dargestellt was auf kognitiver und affektiver Ebene in den jeweiligen Phasen beachtet werden kann.

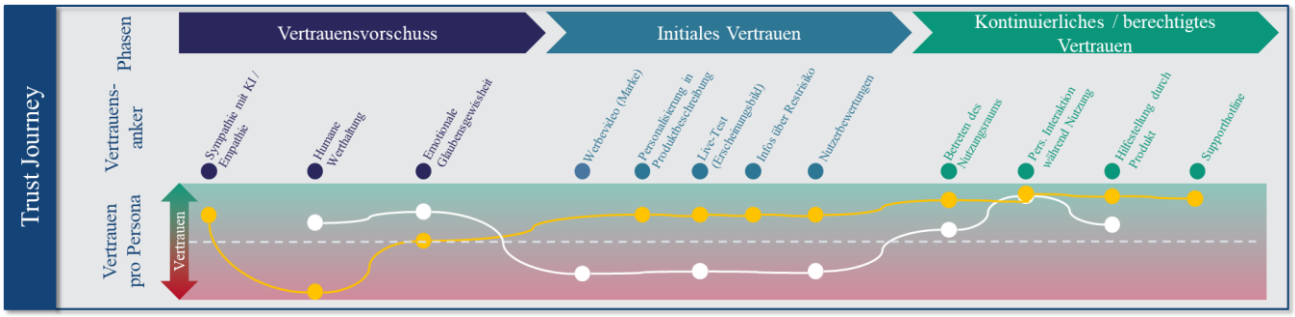

Abbildung 5: Trust Journey mit Vertrauensankern (eigene Darstellung)

3.1 VERTRAUENSVORSCHUSS

Die Gewährung eines Vertrauensvorschusses ist notwendig, um die initiale Vertrauensbildung zu ermöglichen. Durch Vertrauensvorschüsse kann die Komplexität des Vertrauensbildungsprozesses reduziert werden (Vgl. Widuckel et al., 2015), indem kognitive Defizite gefüllt werden (Vgl. Corves & Schön, 2020). Gewährt der Vertrauensgeber einen Vorschuss vertraut er sozusagen blind und ermöglicht dem Vertrauensnehmer zu beweisen, dass er sein Vertrauen verdient hat. Die Phase des Vertrauensvorschusses ist für den weiteren Verlauf kritisch zu betrachten, denn im Falle, dass dieser nicht gewährt wird, erfährt der Nutzer in der Regel eine negative Zuspitzung, welche die initiale Vertrauensbildung blockieren würde. Daher ist die positive Beurteilung und Entwicklung des Vertrauensvorschusses unumgänglich für den weiteren Prozess des Vertrauensaufbaus. Diese positive Beurteilung wird meist durch affektive Vertrauensanker getriggert, wie beispielsweise den Aufbau von Sympathien und das Eingehen auf die individuelle Wertehaltung eines Nutzers ohne Bezug zum Produkt bzw. der Marke, die in den späteren Phasen im Vordergrund steht (Vgl. Corves & Schön, 2020).

3.2 INITIALE VERTRAUENSBILDUNG

Es gibt verschiedene relevante Ankerpunkte der initialen Vertrauensbildung. Zunächst sollte das mentale Modell bzw. die Einstellung des Nutzers mit den Signalen des KI-Systems übereinstimmen. Diese Übereinstimmung mit bisherigen Erfahrungen und Erwartungen kann vom Hersteller wie folgt beeinflusst werden. Der erste Ankerpunkt ist meist das Branding der Anwendung bzw. des Herstellers. Durch die Namensgebung der Technologie oder Marke kann dem Nutzer eine Vertrauensquelle nahegelegt werden und so dessen Vertrauenserwartung beeinflusst werden. Beim Sprachassistenzdienst Alexa bspw. wurde die personenspezifische Vertrauensquelle gewählt, welche durch die Namensgebung die zugehörigen Vertrauensmuster anstößt. Der zweite Ankerpunkt ist in der Regel das Erscheinungsbild, denn das menschenähnliche Äußere vieler KI-basierten Produkte hat einen Grund. Je größer diese Ähnlichkeit ist, desto leichter ist es für den Nutzer eine emotionale Bindung aufzubauen, denn umso stärker wird das gleichnamige Vertrauensmuster der Repräsentation / Ähnlichkeit stimuliert. Speziell die äußere Erscheinung inklusive Stimmen von Robotern, Sprachassistenten oder Navigationsgeräten sprechen das Vertrauensmuster der Repräsentation / Ähnlichkeit an. Dessen Einfluss ist im Marketing, der Auswahl von Influencern und bei der Weiterentwicklung von KI Anwendungen zu beobachten (Vgl. Corves & Schön, 2020).

Neben der äußeren Erscheinung lassen wir uns häufig von Vorurteilen beeinflussen. Vorurteile und Stereotypen sind nicht grundsätzlich negativ. Zu beachten ist allerdings, dass wir diese häufig unbewusst aufbauen und dadurch bestimmte Erwartungen hegen. Das Image von KI wurde durch moderne Science-Fiction Bücher und Filme stark beeinflusst. Auch wenn wir uns während des Konsums bewusst sind, dass es eine fiktive Vorstellung ist, in der ein Computer die Weltherrschaft an sich reißt, so können wir häufig nicht vorhersehen, wie sich KI-Anwendungen künftig weiterentwickeln und wo deren Grenzen liegen werden. Speziell in Deutschland sind die Nutzer für Datenmissbräuche sensibilisiert. Barton, Bender & Marqure bezeichnet Datenmissbräuche sogar als wichtigsten Disruptor des digitalen Vertrauens. Hinzu kommt ein Informationsmangel in Bezug auf Daten- und Aussagequalitäten von KI. Dadurch entsteht häufig Angst, die das initiale Vertrauen in die KI schwächt. Beim Aufbau eines Vertrauensverhältnisses muss dieses Element also gesondert berücksichtigt werden, denn die Vertrauensmuster Sicherheit und Transparenz gelten mittlerweile als wesentliche Hygienefaktoren (Vgl. Barton, Bender & Marque, 2017).

Neben unseren eigenen Erfahrungen und Einflüssen haben wir heutzutage die Möglichkeit auf Erfahrungen anderer zuzugreifen. Bewertungsportale können sowohl einen positiven als auch negativen Einfluss auf das initiale Vertrauensniveau haben (Vgl. Siau & Wang, 2018). Die Faktoren Repräsentation, Image und Bewertungen von anderen Nutzern beziehen sich alle auf die direkte oder erwartete Leistungsfähigkeit einer KI-Anwendung und sind eher der affektiven Ebene zuzuordnen. Dem gegenüber stehen die eher kognitiv orientierten Prozessfaktoren: Transparenz, Erklärungsfähigkeit und Erprobbarkeit. Das Vertrauensmuster der Übereinstimmung kann nicht automatisch bedient werden, da KI als neue Technologie gilt, deren Funktionsweise noch nicht bekannt ist. Auf affektiver Ebene ist also kein positives Bauchgefühl vorhanden. In diesem Fall verlangt Vertrauen in der Regel ein gewisses Grundverständnis und eine Nachvollziehbarkeit. Häufig hilft es den Nutzern, wenn sie verstehen, wie eine KI-Anwendung programmiert ist und welche Funktionen sie hat. Der Prozess der Entscheidungsfindung und des Verhaltens einer KI erfolgt durch Wahrscheinlichkeiten, sind diese transparent kommuniziert oder die Datenqualität gekennzeichnet fällt es uns leichter diese zu akzeptieren (Vgl. Corves & Schön, 2020; Siau & Wang, 2018). Ist diese durch Transparenz entstandene Erklärbarkeit nicht vorhanden leidet die Vertrauenswürdigkeit. Neben der theoretischen Erklärbarkeit spielt die tatsächliche, haptische Anwendung eine Rolle – die Erprobbarkeit. So fällt es jüngeren Generationen leichter in KI-Anwendungen zu vertrauen, weil sie diese von klein auf ausprobieren konnten. Ein Kind muss sich nicht aktiv dafür entscheiden der KI zu vertrauen und ein Gerät zu kaufen. Haben die Eltern eine Alexa wird das Kind diese automatisch ausprobieren und bei positiven Erfahrungen ein initiales Vertrauen gegenüber Anwendungen dieser Art entwickeln. Studien haben ergeben, dass Widerstand gegenüber neuen Technologien hauptsächlich bei Personengruppen auftritt, die diese noch nie zuvor getestet oder genutzt haben (Vgl. Siau & Wang, 2018).

3.3 KONTINUIERLICHE VERTRAUENSBILDUNG

Ist das initiale Vertrauen erst einmal aufgebaut werden weitere Faktoren relevant, die die Basis für ein dauerhaftes Vertrauen legen bzw. dessen Bildung beeinflussen (Abbildung 6).

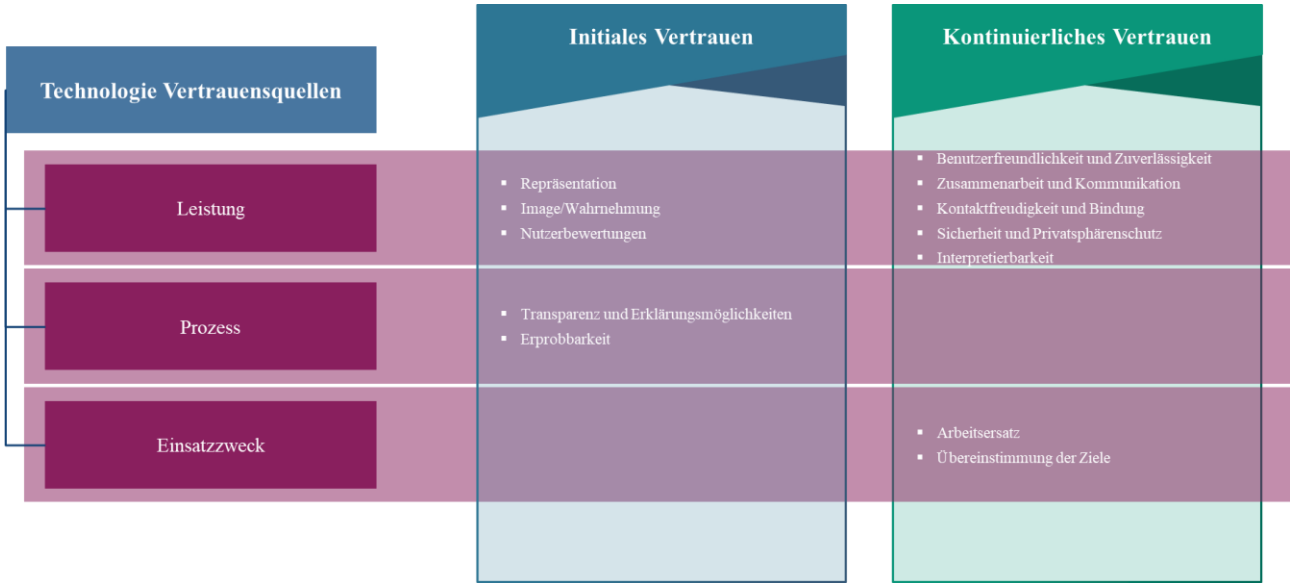

Abbildung 6: Technologische Merkmale der KI, die die Vertrauensbildung beeinflussen (eigene Darstellung in Anlehnung an Siau & Wang, 2018)

Kontinuierliches Vertrauen wird häufig auch als berechtigtes Vertrauen bezeichnet. Kontinuierliches Vertrauen wird durch kognitive und affektive Vertrauensanker entwickelt und gefestigt (Vgl. Corves & Schön, 2020), es hängt von der Leistung und dem Zweck der künstlichen Intelligenz ab. Bevor sich ein Nutzer auf die kognitive Vertrauensebene begibt, in welcher die Leistung über die Äußerlichkeiten der Anwendung hinausgeht, muss zunächst die grundsätzlich bestehende KI-Unsicherheit überwunden werden. Um dies zu erreichen, muss auf der affektiven Ebene ein Vertrauensvorschuss bestehen.

Nutzer beurteilen die Leistung daraufhin, indem die Zuverlässigkeit der Aufgabenerfüllung und die Benutzerfreundlichkeit überprüft werden. Somit sind unerwartete Ausfallzeiten oder Abstürze ein großes Problem, da sie einem dauerhaften Vertrauen im Wege stehen. Die Benutzerfreundlichkeit lässt sich neben der intuitiven Bedienbarkeit um die Dimensionen der Kollaboration und Kommunikation erweitern – Nutzern ist es wichtig, dass die Zusammenarbeit und Kommunikation mit der KI-Anwendung reibungslos und einfach ablaufen. Neben den technischen Faktoren spielt auch die emotionale und soziale Bindung eine wichtige Rolle. Erkennt ein Roboterhund oder ein Sprachassistenten seinen Besitzer wieder stärkt das die Bindung und damit auch das Vertrauen. KI-Anwendungen benötigen enorme Datenmengen, um ihre Funktionen zu ermöglichen, dies birgt gewisse Risiken in Bezug auf Daten- und Betriebssicherheit. Diese beiden Faktoren beeinflussen das Vertrauen der Nutzer sehr stark. Wie schon bei der initialen Vertrauensbildung erwähnt, sind die meisten ML-Modelle undurchschaubar. Um dieses Problem vollumfänglich zu lösen, ist es auch bei der kontinuierlichen Vertrauensbildung notwendig ein Verständnis für die Prozesse und Generierung der Ergebnisse zu schaffen. Der Maschine muss die Möglichkeit oder ein Modell gegeben werden, ihre Schlussfolgerungen oder Aktionen immer

wieder auch auf Basis des aktuellen Ergebnisses zu erklären und nicht nur initial die grobe Vorgehensweise zu schildern. Siau und Wang betiteln diese Notwendigkeit als Interpretierbarkeit was aus Sicht der kontinuierliche Vertrauensbildung das Gegenstück zur initialen Quelle „Transparenz und Erklärbarkeit“ darstellt. Generell ist hinzuzufügen, dass speziell bei leistungsbezogenen Faktoren das Technologieverständnis des Nutzers eine große Rolle spielt bzw. dessen Erwartungen beeinflusst (Vgl. Siau & Wang, 2018).

Über die Leistungsfaktoren hinaus sollten zwei weitere Faktoren berücksichtigt werden, die sich auf den Einsatzzweck der KI-Anwendung beziehen. Häufig stellt sich die Frage, ob KI-Anwendungen künftig Arbeitsplätze ersetzen können. Mittlerweile geht es allerdings schon darüber hinaus und somit ist die Frage nicht mehr ob Arbeitsplätze ersetzt werden, sondern welche und wie viele. Durch diese Entwicklung entsteht wiederum Angst gegenüber KI-Anwendungen. Da diese Angst absolut berechtigt ist, es aber dennoch sinnvoll ist sich bspw. repetitive oder auch gefährliche Tätigkeiten von Maschinen übernehmen zu lassen kann dieser Angst lediglich durch die Aussicht auf Ausbildungen und Umschulungen entgegengewirkt werden. Die Menschen müssen sich umorientieren und ihre Aufgaben neu definieren – hilft man ihnen dabei, kann man das Vertrauen in die KI und deren Sinnhaftigkeit stärken. Die soeben dargestellte Bedrohlichkeit von KI-Anwendungen kann ebenfalls umgangen werden, indem sichergestellt wird, dass die Ziele der Anwendung mit denen der Menschen, die sie nutzen übereinstimmen. Zielkongruenz gilt als Voraussetzung für die Beständigkeit von Vertrauen (Vgl. Siau & Wang, 2018). Ethik und Governance der KI sind an dieser Stelle wichtige Stichpunkte, die im Rahmen dieser Arbeit allerdings nicht im Fokus stehen und daher nicht weiter detailliert werden.

4 LÖSUNGSANSÄTZE

4.1 VOICE USER INTERFACES

Der Smart Home Markt boomt: Statista prognostiziert einen Anstieg der Smart Home Haushalte in Europa um 100% von 50 Millionen (2021) auf 100 Millionen im Jahre 2025 – knapp 20 Millionen davon in Deutschland. Die Klassifizierung als „Smart Home Haushalt“ setzt zwar keine künstliche Intelligenz im Sinne eine Optimierung durch ML oder DL voraus, allerdings finden diese Technologien zunehmen ihren Weg in die bestehenden Anwendungen. Außerdem gibt es ein spezielles KI-Element, das sich in über zwei Drittel aller Smart Homes findet: Die Steuerung mittels Sprachassistent. Der in Deutschland mit Abstand beliebteste Assistent ist der Amazon Dienst „Alexa“, der sich inzwischen auf verschiedensten Smart Speaker Produkten etabliert hat. Die Maßnahmen zur initialen Vertrauensbildung sind hier sehr plakativ und haben sich inzwischen auf zahlreiche andere Anwendungen übertragen. Das erste eher allgemeine Element ist die Bezeichnung als „Sprachassistent“. Es handelt sich also vorrangig um einen Assistenten, der einen per Sprachbefehlt in allen möglichen Szenarien zur Seite stehen soll. Gemeinsam mit dem Namen „Alexa“ wird eine Personifizierung aufgebaut, die dem initialen Vertrauensmuster „Repräsentation“ zugeordnet werden kann. Die Tatsache, dass die

Sprachassistenten Teil des eigenen Zuhauses sind stellt den Schutz der Privatsphäre und der persönlichen Daten selbstverständlich besonders in der Vordergrund.

Bei Amazon arbeitet das „Alexa Trust-Team“ kontinuierlich daran nicht nur das initiale, sondern vor allem auch das kontinuierliche Vertrauen des Endnutzers in den Dienst zu stärken. Alexa kann zum Beispiel komplett ausgeschaltet oder stumm geschaltet werden. Gespräche, die man mit dem Assistenten führen, können eingesehen und gelöscht werden. Dies kann sogar mittels der Sprachbefehle, „Alexa, lösche, was ich gerade gesagt habe“ oder „Alexa, lösche alles, was ich je gesagt habe“ geschehen. Das blaue Ringlicht an der Oberseite der Smart Speaker der Amazon Echo Reihe signalisiert außerdem, wann die Aufzeichnung bzw. die Übermittlung aktiv ist. Dies vermittelt Kontrolle, Bedienungsfreundlichkeit und Transparenz. (Quelle Amazon & pcket init).

4.2 VERTRAUENSBILDUNG AUTONOMES FAHREN

Autonomes Fahren ist einer der Schlüsselfaktoren für die Mobilitätsszenarien der Zukunft. Hersteller der ganzen Welt wetteifern, um die zuverlässigste Technologie mit dem Ziel möglichst früh die Testphasen zu verlassen und selbstfahrende Fahrzeuge für die breite Masse zugänglich zu machen. Im Mai 2022 ist der Mercedes-Benz AG ein Durchbruch gelungen: Als erster Autohersteller weltweit hat Mercedes-Benz die Zulassung für ein hochautomatisiertes "Level-3"-Fahrzeug erhalten. Dies bedeutet, dass man zwar abgelenkt sein darf, jedoch wach bleiben und jederzeit eingreifen können muss (Vgl. Diekmann, 2022). Aus rechtlicher Sicht ist also eine wichtige Hürde überwunden. Eine im Februar 2021 veröffentliche Statista Umfrage zeigte allerdings, dass lediglich 24 % der Befragten dem autonomen Fahren im Straßenverkehr (ausgenommen Ein- und Ausparken) ausreichend vertrauen (Vgl. Statista, 2021). Die Hersteller zeigen sich allerdings bemüht dies zu ändern. Unzählige Magazine und Testplattformen berichten von ihren Erfahrungen mit dem „Mercedes Drive Pilot“. Im Bereich der ausführlichen Nutzerbewertungen ist also bereits ein Grundstein gelegt. Auch der Name „Drive Pilot“ an sich impliziert nicht direkt, dass es eine Maschine ist, die das Fahren für den Menschen übernimmt, sondern ein „Pilot“, auf dem man sich – wie im Flugzeug schließlich auch – verlassen kann. Darüber hinaus werden immer wieder Events angeboten, auf denen man durch Virtual-Reality-Anwendungen in die Welt des Autonomen Fahrens eintauchen kann, um sich bereits im Voraus ein eigenes Bild davon zu machen, wie es sich anfühlt mit „Drive Pilot“ zu fahren und welche Entscheidungen das Auto künftig selbstständig trifft. All diese Beispiele fördern die initiale Vertrauensbildung in das autonome Fahrzeug (Vgl. Mercedes-Benz AG, 2022).

Andere Hersteller fokussieren sich hingegen explizit auf die kontinuierliche Vertrauensbildung im Innenraum des Fahrzeugs während der Fahrt. NIO ist ein asiatischer Automobilhersteller und nutzt die Schnittstellengestaltung zwischen KI und Fahrer. So wurde beispielsweise eine kleine Kugel namens »Nomi« mit einem Emoji-ähnlichen Gesicht in den Fahrzeuginnenraum eingebaut. Dadurch werden Emotionen durch menschliche Eigenschaften geweckt, die auch auf die Bedürfnisse der Nutzer wie ein persönlicher Begleiter eingehen können. Nomi stellt individuell anpassbare Informationen über Gefahren und Handlungen des Fahrzeugs

bereit. Als weiteres Kommunikationsmittel werden auch optische Signale eingesetzt, zum Beispiel mit farbigen Lichtsignalen auf einer großen Fläche im Innenraum. Ein weiterer wesentlicher Aspekt ist die Wohlfühlatmosphäre im Fahrzeuginnenraum. Designer beachten bei der Gestaltung des Innenraums die Verwendung hellen Farben und runden Formen in der Farb- und Materialgestaltung und streben die Schaffung einer Wohlfühlatmosphäre an, da sich diese positiv auf das Vertrauen auswirken (Vgl. Nio, 2020). Hyundai geht sogar noch einen Schritt weiter und setzt KI ein, um die Gefühlswelt des Fahrers bestmöglich aufzunehmen und die Situation im Innenraum dynamisch daran anzupassen. Hierfür ist in das Auto eine so genannte „Emotion Adaptive Vehicle Control“ integriert, die mit Hilfe von KI die Emotionen des Fahrers überwacht, indem Mimik, Atem- und Herzfrequenz analysiert werden. Entsprechend der gesammelten Daten ändert die KI Beleuchtung, Musik und den Duft im Fahrzeug, was mögliche Ängste reduzieren soll. Die Testphase läuft bereits im Rahmen des Projekts „Little Big e-Motion“: Selbstfahrende Miniautos von Hyundai mit KI-Unterstützung wurden an ein Kinderkrankenhaus in Barcelona gespendet. Das Auto soll die Angst bei der Fahrt vom Krankenhausbett zum Behandlungsraum reduzieren (Vgl. Hyundai Motor Company, 2020).

4.3 Die EY Trusted AI Platform

Das Beispiel der EY Trusted AI Plattform betrachtet die Vertrauensbildung aus einer weiteren Perspektive. Statt nur den Anwender bzw. Endverbraucher zu berücksichtigen, stellt sich die Vertrauensfrage vor allem auch im B2B Bereich. Hier spielt vor allem das wirtschaftliche Risiko eine große Rolle. Es ist auch hier wieder davon auszugehen, dass bei der zu entwickelnden KI die eingesetzten ML- oder DL-Prozesse inkl. den möglichen Parameter durch den Anwender oder ggf. auch den Entwickler nicht ganzheitlich genug erfasst werden können, um das Risiko mit der nötigen Sicherheit einzuschätzen. Die EY Trusted AI Plattform bietet Erkenntnisse zu Risikofaktoren. Aus der Kombination verschiedenster Faktoren berechnet die Plattform das Restrisiko eines KI-Systems. Außerdem werden auch soziale und ethische Aspekte miteinbezogen. Durch die Analyse können Maßnahmen für die Reduktion des Restrisikos abgeleitet werden. Durch die quantitative eines KI-Systems soll eine umfassende Transparenz hinsichtlich Risikotreibern, deren Auswirkungen und möglicher Designänderungen ermöglicht werden. Dies ist ein wichtiger Schritt bei der Entwicklung eines robusten Risikomanagementsystems und soll die initiale Vertrauenswürdigkeit unterschiedlichster KI Systeme transparent machen bzw. erhöhen (Vgl. Ernst & Young Global Limited, 2022).

5 Limitationen

Vertrauen ist ein sehr komplexes Konstrukt und KI-Anwendungen werden stetig verbessert, weiterentwickelt oder mittels neuer Ansätze erschaffen. Für den Umfang der vorliegenden Arbeit wurde die KI Trust Journey auf die Phasen Vertrauensvorschuss, initiales Vertrauen und kontinuierliches Vertrauen beschränkt. Darüber hinaus muss allerdings beachtet werden, dass speziell im KI-Umfeld die bereits angesprochenen Vorurteile nicht außer Acht gelassen werden dürfen. Erweitert man die KI Trust Journey können diese der Phase des

Misstrauens bzw. de Mangel an Vertrauen zugeordnet werden. In dieser Phase gelten Datenmissbrauch, öffentliche Skandale und eine fehlende Kennzeichnung zu den Vertrauensankern. Ist diese Phase überwunden entsteht eine sogenannte Situationsoffenheit, welche den Grundstein für den darauffolgenden Vertrauensvorschuss bildet. Um eine Situationsoffenheit zu erreichen, wird häufig auf den Vertrauensanker der Transparenz gesetzt, denn dadurch ist es dem Nutzer erst möglich einen Vertrauensvorschuss zu gewähren. Das Misstrauen ist weitaus komplexer zu erläutern als das Vertrauen selbst, weshalb die zwei vorangestellten Phasen Misstrauen und Situationsoffenheit in diesem Paper zunächst ausgeklammert wurden und dennoch nicht vernachlässigt, werden dürfen (Vgl. Corves & Schön, 2020).
Die Relevanz von KI-Anwendungen ist mittlerweile unbestritten, ebenfalls die Notwendigkeit von Vertrauen in die Algorithmen. Wann eine KI-Anwendung allerdings vertrauenswürdig ist, lässt sich nicht zweifelsfrei sagen. Die Messbarkeit von Vertrauen in KI ist als ein sehr wichtiger Faktor, welcher aktuell noch nicht näher beleuchtet werden kann (Vgl. Engel, 2021).

6 FAZIT & AUSBLICK

Ein wichtiger Faktor, um die Entwicklungen im Bereich KI von zentraler Stelle zu begleiten, steuern und einzuordnen sind Richtlinien und rechtliche Rahmenbedingungen. Dies schafft eine solide Basis für ein Grundvertrauen und ein Gefühl von Sicherheit, das Vorurteilen und pauschalen Ängsten entgegenwirkt. Auf EU-Ebene hat die EU-Kommission einen rechtlichen Rahmen erarbeitet, um eine KI als „vertrauenswürdig“ einzustufen oder ggf. eben auch nicht. Genauer: Die Kommission schlägt neue Vorschriften vor, um sicherzustellen, dass KI-Systeme, die in der EU verwendet werden, sicher, transparent, ethisch, unparteiisch und unter menschlicher Kontrolle sind. Dies erfolgt auf zwei Ebenen: Auf der ersten Ebene erfolgt eine Risikoeinschätzung auf vier Stufen: Unacceptable Risk, High Risk, Limited Risk, Minimal Risk.

- **Unacceptable Risk:** Alles, was als eindeutige Bedrohung angesehen wird, wird verboten: von der behördlichen Bewertung des sozialen Verhaltens (Social Scoring) bis hin zu Spielzeug mit Sprachassistent, das Kinder zu riskantem Verhalten verleitet.
- **High Risk:** Systeme, die in einem High Risk Umfeld agieren, werden im Rahmen der zweiten Ebene einer detaillierte Konformitätsuntersuchung unterzogen und müssen eine Reihe zusätzlicher Kriterien erfüllen, um zugelassen zu werden. (Beispiele für High Risk Umfelder wären: Kritische Infrastrukturen, Zentrale private und öffentliche Dienstleistungen, Strafverfolgung oder Grenzkontrolle).
- **Limited Risk:** Für KI-Systeme wie Chatbots gelten minimale Transparenzverpflichtungen, die es den mit ihnen interagierenden Nutzern ermöglichen sollen, fundierte Entscheidungen zu treffen. Die Nutzer können dann entscheiden, ob sie die Anwendung weiter nutzen oder nicht.
- **Minimal Risk:** Kostenlose Nutzung von Anwendungen wie KI-gestützten Videospielen oder Spamfiltern. Unter diese Kategorie, in der die neuen Vorschriften nicht greifen, fällt die große Mehrzahl der

KI-Systeme, weil diese Systeme nur ein minimales oder kein Risiko für die Bürgerrechte oder die Sicherheit darstellen.

Solche Richtlinien sind angesichts der steigenden Integration unterschiedlicher KI-Anwendungen in unseren Alltag enorm wichtig, um einerseits natürlich die Sicherheit auch unerfahrener Anwender zu gewährleisten, andererseits aber auch das Grundvertrauen in eine KI-Anwendung auf dem Markt zu stärken (Vgl. Europäische Komission, 2022).

Auch in Deutschland gibt es ähnliche Bemühungen. Auf Bundesebene beschäftig sich das Bundesministerium für Wirtschaft und Klimaschutz (BWMK) im Rahmen des Projekts „Foresight“ unter anderem mit der Frage: „Wie weit darf eine KI-Unterstützung gehen, ehe sie zur Bevormundung wird?“. Dabei wird gezielt der Mensch in den Mittelpunkt gesetzt, um jederzeit die Möglichkeit des Eingreifens sowie die finale Entscheidungsgewalt des Anwenders sicherzustellen. Im Rahmen dieses Projekts wurde ein Ethik-Kodex für vertrauenswürdige KI ausgearbeitet, welcher auch weiterhin kontinuierlich weiterentwickelt werden soll. Insgesamt umfasst dieser Kodex aktuell sieben ethische Indikatoren (Vgl. BMWK, 2022):

- Vorrang menschlichen Handelns und menschlicher Aufsicht
- Technische Robustheit und Sicherheit
- Schutz der Privatsphäre und Datenqualitätsmanagement
- Transparenz und Erklärbarkeit
- Vielfalt, Nichtdiskriminierung und Fairness
- Gesellschaftliches und ökologisches Wohlergehen
- Rechenschaftspflicht

Wie so oft im technischen Umfeld ist es auf dem Feld der künstlichen Intelligenz so, dass die Innovation wesentlich schneller ist als die generelle Akzeptanz der potenziellen Anwender oder gar die Entwicklung rechtlicher Rahmenbedingungen. Um diese natürlich entstandene Lücke zu füllen, sind die Anstrengungen der EU und dem BMWK von großer Bedeutung. Sie liefern eine wichtige Basis für den weiteren Vertrauensaufbau durch die Anbieter, wie er in dem vorigen Kapitel der vorliegenden Arbeit beschrieben wurde. Diese Entwicklung steckt allerdings noch in den Kinderschuhen und die KI-Anwendungen entwickeln sich selbstverständlich auch mit rasantem Tempo weiter. Deshalb ist der gesamte Vertrauensaufbau als iterativer Prozess zu betrachten, den es immer wieder zu hinterfragen und anzupassen gilt.

7 LITERATURVERZEICHNIS

Reference list

Allen, J. (1998) ‘AI Growing Up: The Changes and Opportunities’, *AI Magazine* (19), p. 13.

Apt, W. and Priesack, K. (2019) ‘KI und Arbeit – Chance und Risiko zugleich.’, in Wittpahl, V. (ed.) *Künstliche Intelligenz.* Berlin, Heidelberg: Springer Berlin Heidelberg.

Barton, R., Bender, M. and Marque, F. (2017) ‘Trust in the digitale age’.

BMWK (2022) *Ethische Leitlinien für Künstliche Intelligenz: Neue KI-Methoden für Smart-Living-Anwendungen: Das BMWK-geförderte Projekt "Foresight" rückt bei der Entwicklung den Menschen in den Fokus*. Available at: https://www.bmwk.de/Redaktion/DE/Schlaglichter-der-Wirtschaftspolitik/2021/09/11-ethische-leitlinien-fur-kunstliche-intelligenz.html.

Buchkremer, R. (2020) ‘Natural Language Processing in der KI: Eine Erfassung der aktuellen Patente- und Literatursituation’, in Buchkremer, R., Heupel, T. and Koch, O. (eds.) *Künstliche Intelligenz in Wirtschaft & Gesellschaft: Auswirkungen, Herausforderungen & Handlungsempfehlungen.* (FOM-Edition). Wiesbaden: Springer Gabler.

Corves, A. and Schön, E.-M. (2020) ‘Digital Trust für KI-basierte Mensch-Maschine-Schnittstellen’, in Boßow-Thies, S., Hofmann-Stölting, C. and Jochims, H. (eds.) *Data-driven Marketing: Insights aus Wissenschaft und Praxis*. Wiesbaden: Springer Gabler.

Dettmann, U. and Kopecz, J. (2020) ‘Moralische Maschinen – Zur ethischen Ununterscheidbarkeit von Mensch und Maschine’, in Buchkremer, R., Heupel, T. and Koch, O. (eds.) *Künstliche Intelligenz in Wirtschaft & Gesellschaft: Auswirkungen, Herausforderungen & Handlungsempfehlungen.* (FOM-Edition). Wiesbaden: Springer Gabler.

Diekmann, T. (2022) *Freihändig über die Autobahn: Fahrzeuge mit Autopilot*.

Döbel, I. *et al.* (2018) ‘Maschinelles Lernen. Eine Analyse zu Kompetenzen, Forschung und Anwendung’, *Fraunhofer-Gesellschaft zur Förderung der angewandten Forschung e.V.*

Engel, S. (2021) ‘Künstliche Intelligenz in Unternehmen skalieren – die Rolle von Vertrauen’, in Lichtenthaler, U. (ed.) *Künstliche Intelligenz erfolgreich umsetzen: Praxisbeispiele für integrierte Intelligenz.* Wiesbaden: Springer Gabler.

Ernst & Young Global Limited (2022) *Die EY Trusted AI Platform*.

Europäische Komission (2022) *Künstliche Intelligenz – Exzellenz und Vertrauen*. Available at: https://ec.europa.eu/info/strategy/priorities-2019-2024/europe-fit-digital-age/excellence-trust-artificial-intelligence_de.

Hyundai Motor Company (2020) *Hyundai Motor's Mini '45' EV Puts Emotions in Motion*. Available at: https://www.hyundai.news/eu/articles/press-releases/hyundai-motors-mini-45-ev-puts-emotions-in-motion.html.

Jung, M. and Garrel, J. von (2021) 'Mitarbeiterfreundliche Implementierung von KI -Systemen im Hinblick auf Akzeptanz und Vertrauen', *TATuP - Zeitschrift für Technikfolgenabschätzung in Theorie und Praxis*, 30(3), pp. 37–43. doi: 10.14512/tatup.30.3.37

Kahneman, D. (2012) *Schnelles denken, langsames Denken*: Siedler Verlag.

KPMG (2021) *Startups und Künstliche Intelligenz - Innovation trifft Verantwortung: Öffentliche Meinung zur Künstlichen Intelligenz in Deutschland und USA im Jahr 2020*. Available at: https://de.statista.com/statistik/daten/studie/1306685/umfrage/oeffentliche-meinung-zur-kuenstlichen-intelligenz-in-deutschland-und-usa/.

Lämmel, U. and Cleve, J. (2020) *Künstliche Intelligenz.: Wissensverarbeitung - Neuronale Netze*. 5th edn.: Carl Hanser Verlag.

Lee, J.D. and See, K.A. (2004) 'Trust in automation: designing for appropriate reliance', *Human Factors*, 46(1), pp. 50–80. doi: 10.1518/hfes.46.1.50_30392

Leopold, H. (2022) 'Warum wir der künstlichen Intelligenz nicht blind vertrauen dürfen – fünf Ansätze um künstliche Intelligenz zu beherrschen', in Altenburger, R. and Schmidpeter (eds.) *CSR und Künstliche Intelligenz:* Springer.

McAfee, A. and Brynjolfsson, E. (2017) 'Machine, platform, crowd: Harnessing our digital future', *WW Norton & Company*.

Mele, C., Spena, T.R. and Peschiera, S. (2018) 'Value Creation and Cognitive Technologies: Opportunities and Challenges', *Journal of Creating Value*, 4(2), pp. 182–195. doi: 10.1177/2394964318809152

Mercedes-Benz AG (2022) *Mercedes-Benz DIVE PILOT: Zukunft erfahren: Mit dem aktuellen Höchstmaß an Aisstenz und Automatisierung Ihre Ziele sicher und entspannter erreichen*.

Nio (2020) *https://www.nio.com/blog/nomi-worlds-first-vehicle-artificial-intelligence*. Available at: https://www.nio.com/blog/nomi-worlds-first-vehicle-artificial-intelligence.

Petermann, F. (2013) *Psychologie des Vertrauens*. 4th edn. (Hogrefe eLibrary). Göttingen: Hogrefe. Available at: https://elibrary.hogrefe.de/book/99.110005/9783840924156.

Saßmannshausen, T.M. and Heupel, T. (2020) 'Vertrauen in KI – Eine empirische Analyse innerhalb des Produktionsmanagements', in Buchkremer, R., Heupel, T. and Koch, O. (eds.) *Künstliche Intelligenz in Wirtschaft & Gesellschaft: Auswirkungen, Herausforderungen & Handlungsempfehlungen.* (FOM-Edition). Wiesbaden: Springer Gabler.

Seitz, J., Willbold, K. and Haiber, R. (2022) *AI value creation Studie: Potenziale und Hindernisse von AI business use cases in Unternehmen*. Kappel-Grafenhausen: Digipolis Verlag. Available at: http://nbn-resolving.org/urn:nbn:de:bsz:31-epflicht-1968238.

Siau, K. and Wang, W. (2018) 'Building Trust in Artificial Intelligence, Machine Learning, and Robotics', *Cutter Business Technology Journal*, 2018(31).

Silver, D. *et al.* (2016) 'Mastering the game of Go with deep neural networks and tree search', *Nature* (529), pp. 484–489.

Statista (2021) 'Umfrage zum Vertrauen gegenüber autonomen Autos in Deutschland nach Gründen 2021: Würdest du autonomen Autos vertrauen?'.

TÜV (2019) *Sicherheit und künstliche Intelligenz: Was empfinden Sie, wenn Sie an KI denken?* Available at: https://de.statista.com/statistik/daten/studie/1309442/umfrage/umfrage-zur-einstellung-gegenueber-kuenstlicher-intelligenz/.

Wang, P. (2019) 'On Defining Artificial Intelligence', *Journal of Artificial General Intelligence*, 10(2), pp. 1–37. doi: 10.2478/jagi-2019-0002

Waymo LLC (2022) *Seeing the road ahead: Our Company*. Available at: https://waymo.com/company/.

Widuckel, W. *et al.* (2015) *Arbeitskultur 2020: Herausforderungen und Best Practices der Arbeitswelt der Zukunft*. Wiesbaden: GABLER.