3 Zentrale Begriffe und philosophische Grundlagen

3.1 Künstliche Intelligenz: Begriffliche Analyse

Der Begriff der Künstlichen Intelligenz wurde und wird keineswegs immer einheitlich verwendet: Seine Bedeutung hat sich im Laufe der Jahre verändert und unterscheidet sich sowohl innerhalb als auch zwischen verschiedenen Berufsgruppen und Disziplinen. Eine Unterscheidung, welche in der KI-Forschung selbst, mehr noch aber in der Reflexion sowie der medialen und öffentlichen Debatte über KI eine große Rolle spielt, ist die Unterscheidung zwischen sogenannter *schwacher* und *starker* KI.[65] Während sich insbesondere die aktuelle Forschung innerhalb der Informatik vorrangig mit Fragen der schwachen KI beschäftigt, sind öffentliche Debatten ebenso wie manche Fachdiskurse regelmäßig geprägt von Visionen einer starken – also menschenähnlichen oder gar menschliche Fähigkeiten übertreffenden – KI, die mit Sorgen ebenso wie mit Hoffnungen behaftet ist. Um die hinter den jeweiligen Begriffsverwendungen stehenden substanziellen Annahmen über die Grundlagen menschlicher wie Künstlicher Intelligenz herauszuarbeiten, ist eine Klärung der Begriffe notwendig. Dabei zeigt sich, dass unterschiedliche Einschätzungen hinsichtlich der Wahrscheinlichkeit und dem Erwünschtsein einer starken KI auch von unterschiedlichen Konzeptualisierungen des mit KI umschriebenen Phänomenbereichs abhängen.

Neben der heute häufig im Kontext der Reflexion auf KI vornehmlich verwendeten Unterscheidung von starker versus schwacher KI, sind in der Geschichte der KI und auch in der gegenwärtigen internationalen Debatte andere Unterscheidungen in Gebrauch, um unterschiedliche Formen oder Grade der Annäherung künstlicher an menschliche Intelligenz zu beschreiben. Dies sind insbesondere die nachfolgend noch näher betrachteten Unterscheidungen von *spezi-*

[65] Nida-Rümelin, J. (2022): Überblick über die Verwendung der Begriffe starke & schwache Künstliche Intelligenz. In: Chibanguza, K.; Kuß, C.; Steege, H. (Hg.): Handbuch Künstliche Intelligenz. Recht und Praxis automatisierter und autonomer Systeme. Baden-Baden, 75-90; Nida-Rümelin, J. (2022): Digitaler Humanismus – philosophische Aspekte Künstlicher Intelligenz. In: Chibanguza, K.; Kuß, C.; Steege, H. (Hg.): Handbuch Künstliche Intelligenz. Recht und Praxis automatisierter und autonomer Systeme. Baden-Baden, 29-40.

eller versus *allgemeiner/ genereller* KI sowie *enger* KI versus *breiter* KI Darüber hinaus werden in den letzten Jahren, vor allem im Kontext des Transhumanismus[66] sowie der Diskussionen um Singularität[67], die Begriffe *Artificial General Intelligence* oder *Super-Intelligence* verwendet, die im Folgenden allerdings keine Rolle spielen werden.

Den jeweiligen Unterscheidungen liegt trotz einzelner Bedeutungsunterschiede das Bemühen zugrunde, menschliche Intelligenz als Maßstab der Bestimmung Künstlicher Intelligenz heranzuziehen. Über diesen Maßstab glaubt man zu verfügen, weil Intelligenztests eine (wenn auch in Grenzen) erfolgreiche Operationalisierung des Begriffs der Intelligenz erlauben (vgl. Abschnitt 3.2.1). Auf dieser Grundlage erscheint es möglich, menschliche Intelligenz zu simulieren bzw. das Vorliegen von Intelligenz in Maschinen identifizieren und messen zu können. Eine Simulation menschlicher Intelligenz stand lange im Zentrum der Forschung zu KI und der Reflexion über ihre Möglichkeiten und Grenzen. Auch wenn Methoden der KI für ganz andere Zwecke verwendet werden, beispielsweise die Nutzung zur Auswahl passender Werbung, so gilt seit der Veröffentlichung von Turings „Can Machines Think?“ im Jahre 1950 und seiner Formulierung des berühmten Turing-Tests (vgl. Abschnitt 2.1) die äußere Ununterscheidbarkeit zwischen menschlichen und maschinellen kognitiven und operativen Leistungen weithin als „Lackmustest“ für das Vorliegen Künstlicher Intelligenz.

Die Unterscheidung zwischen *spezieller und allgemeiner/genereller KI* wurde in den 1950er-Jahren eingeführt, um die damals aktuelle Forschung zu KI von der Vision einer menschenähnlichen KI abzugrenzen. Während Maschinen zumeist für spezielle Tätigkeiten konstruiert sind, kennzeichnet menschliche Intelligenz eine Vielfalt von Kompetenzen und ein weites Spektrum unterschiedlicher Zwecksetzungen. Man erhoffte sich vom Computer von Anbeginn das Potenzial einer „universellen Maschine“, die alle Aufgaben lösen könne, die sich binär, zum Beispiel als eine Abfolge von Nullen und Einsen, darstellen ließen.[68] Die Hoffnung war, dass der Computer eine generelle Künstliche Intelligenz erreichen könnte, die über das Lösen spezifischer Aufgaben weit hinausgeht. Ein Ziel der frühen Forschung war die Entwicklung des Computers zum General Problem Solver (GPS). Der GPS sollte basierend auf menschlichen Heuristiken

[66] Unter Transhumanismus werden Positionen verstanden, wonach die digitalen Technologien es uns erlauben die Beschränktheiten der menschlichen Existenzform zu überwinden und wir diese Möglichkeiten nutzen sollten, um alte Menschheitsträume zu verwirklichen, wie die einer durch Verbindungen von Gehirn und Computer (brain computer interfaces) erzeugten Vervielfachung der Intelligenz oder der individuellen Fortexistenz in Form einer Software-Kopie des eigenen Gehirns.
[67] Der Begriff der Singularität verweist auf einen möglichen Umschlagpunkt in der Zukunft, ab dem Künstliche Intelligenz menschliche Intelligenz in jeder Hinsicht übertreffen und sich fortan unkontrolliert selbst weiterentwickeln könnte.
[68] Moor, J. H. (1985): What is Computer Ethics? In: Metaphilosophy 16 (4), 266-275.

Problemlösestrategien für ausreichend formalisierbare Probleme in verschiedenen Kontexten liefern. Er kann als Vorläufer heutiger Expertensysteme gelten. Die Komplexität der Aufgaben, die mit dem GPS gelöst werden konnten, blieb allerdings beschränkt. Die Vision einer allgemeinen KI jedoch hat bis heute Bestand.

Das Begriffspaar *speziell* und *allgemein* enthielt zumeist sowohl explizite Annahmen über die Breite des Fähigkeitenspektrum einer KI als auch Erwartungen zur grundlegenden Qualität der zukünftig daraus resultierenden Künstlichen Intelligenz (bis hin zum ontologischen Status). Entsprechend wurde im Verlauf weiter differenziert. Etabliert sind inzwischen die Unterscheidungen zwischen *enger* und *breiter* KI sowie *schwacher* und *starker* KI. Die Forschung zu KI ist heute zumeist auf einen klar umrissenen Anwendungsbereich begrenzt, etwa die Interpretation von Röntgenaufnahmen, und somit ein Fall von enger KI. Doch auch hier besteht die Vision einer breiten Ausweitung des Spektrums. Dabei ist zu betonen, dass der Unterscheidung zwischen engen und breiten Formen der KI, ebenso wie die Begriffe *speziell* und *allgemein*, nicht als Gegensätze zu verstehen sind, sondern jeweils Endpunkte eines Fähigkeitenspektrums beschreiben.

Ein aktueller Bereich, in dem die Frage nach breiter KI verhandelt wird, umfasst Sprachverarbeitungssysteme wie das im vorherigen Kapitel bereits erwähnte GPT-3 und ChatGPT. Solche Systeme produzieren Texte, bei denen oft schwer oder unmöglich zu erkennen ist, ob sie von einem Menschen oder einer Maschine verfasst wurden. Erste prägnante Beispiele entfachten eine intensive philosophische und auch öffentliche Debatte dazu, ob diese Form der Textproduktion nur eine Verbreiterung des Fähigkeitenspektrums darstellt, oder bereits als ein Übergang zur oder gar eine Manifestation einer generellen oder starken KI gelten könne – wenn auch nur mit Blick auf das Sprachvermögen.[69]

Die Verwendung der verschiedenen Begriffe zur Charakterisierung Künstlicher Intelligenz als spezifischer, enger, oder schwacher Intelligenz einerseits sowie allgemeiner, breiter oder starker Intelligenz andererseits, verweist nicht nur auf Differenzen zwischen den beiden jeweiligen

[69] Der Titel eines Artikels des MIT Technology Reviews (Heavon 2020) fasst diese Debatte mit Blick auf GPT-3 zusammen: Heavon, W. D. (2020): OpenAI's new language generator GPT-3 is shockingly good – and completely mindless. In: MIT Technology Review. https://www.technologyreview.com/2020/07/20/1005454/openai-machine-learning-language-generator-gpt-3-nlp/ [22.12.2022]. Der Titel spielt auf die Kritik von J. Searle an; vgl. dazu Abschnitte 3.1 und 3.4.2. Mit Blick auf jüngere Chatbots löste 2022 zudem der ehemalige Google-Mitarbeiter Blake Lemoine eine Debatte aus, als er postulierte, der von Google entwickelte Chatbot *LaMDA* sei bei Bewusstsein – ein Vorstoß, der später zu seiner Entlassung führte. Tiku, N. (2022): The Google engineer who thinks the company´s AI has come to life. In: Washington Post. https://www.washingtonpost.com/technology/2022/06/11/google-ai-lamda-blake-lemoine/ [31.01.2023].

Arten bzw. Polen von Intelligenz. Dahinter verstecken sich, insbesondere beim Begriffspaar der schwachen und starken KI, vielmehr auch unterschiedliche Verständnisse von Intelligenz sowie unterschiedliche Positionen hinsichtlich der Kernfrage, ob es qualitative und kategorische oder nur quantitative und prinzipiell überwindbare Unterschiede zwischen menschlicher und Künstlicher Intelligenz gibt.

Wichtig für die jeweilige Beantwortung dieser Frage ist zum einen die Differenz hinsichtlich der Breite bzw. Enge des Fähigkeitenspektrums der Künstlichen Intelligenz. Die meisten Anwendungen Künstlicher Intelligenz entfalten ihre jeweilige Leistung auf klar umrissenen, engen Gebieten oder Domänen wie beispielsweise dem Spielen von Schach oder Go. Hier sind sie im direkten Vergleich Menschen inzwischen klar überlegen. Sprachproduktionssysteme wiederum sind zwar ebenfalls auf einen Kompetenzbereich beschränkt (sprachliche Ein- und Ausgabe), jedoch mittlerweile nicht mehr auf eine Domäne, *über* die gesprochen wird; hier erfolgt also eine jedenfalls funktionelle Verbreiterung des Fähigkeitenspektrums. Dennoch fehlt auch ihnen jedwedes sprachliche Verständnis über die Bedeutung der rezipierten oder produzierten Worte, operieren sie doch allein auf Basis der Wahrscheinlichkeit von Wortkombinationen.

Zum anderen hatte der Philosoph John Searle schon 1980 gegen den Turing-Test die These aufgestellt, dass die bloße Ununterscheidbarkeit von menschlicher und maschineller Sprachperformanz nicht ausreicht, um ein Text*verständnis* anzunehmen.[70] Hier geht es also zusätzlich darum, ob es jenseits einer quantitativen Differenz auch einen kategorialen Unterschied zwischen Mensch und Maschine gibt, bzw. darum, ob Intelligenz an bestimmte mentale Voraussetzungen geknüpft ist, welche über die bloße Simulation von Verständnis hinausgehen. Anders formuliert ergibt sich also die Frage, ob Intelligenz in allgemeiner oder starker Form jemals vollumfänglich Maschinen zukommen kann oder ob dafür spezifisch menschliche Eigenschaften Voraussetzung sind.[71]

Die Differenzen hinter den Begriffspaaren schwache versus starke, enge versus breite und spezielle versus allgemeine KI lassen sich vor diesem Hintergrund wie folgt zusammenfassen:

---

[70] Searle, J. (1980): Minds, Brains and Programs. In: Behavioral and Brain Sciences 3 (3), 417–457, (DOI: 10.1017/S0140525X00005756). Searles Gedankenexperiment und seine Kritik werden ausführlicher im Abschnitt 3.4 Anthropologie behandelt.

[71] Die Frage tierlicher Intelligenz wird im Rahmen dieser Stellungnahme ausgeklammert; vgl. dazu Huber, L. (2021): Das rationale Tier: Eine kognitionsbiologische Spurensuche. Berlin; sowie zur grundsätzlicheren Frage nach mentalen Zustände bei Tieren: Deutscher Ethikrat (2020): Tierwohlachtung – Zum verantwortlichen Umgang mit Nutztieren. Berlin.

Zum einen geht es um die *Breite der Fähigkeiten*, über die eine KI verfügt, sowohl graduell innerhalb von Domänen (z. B. Sprachverarbeitung) als auch bereichsübergreifend (z. B. Sprache und Motorik, Situationserfassung). Zum anderen geht es um die Antwort auf die Frage, ob die *Simulation von Intelligenz* mit *Intelligenz* gleichzusetzen ist, oder ob es einen kategorischen Unterschied zwischen der *Simulation von Verständnis* und *genuinem* (z. B. sprachlichem) *Verständnis* gibt, der für „echte" Intelligenz essenziell ist, über die jedenfalls die in dieser Stellungnahme diskutierten Systeme nicht verfügen.[72]

Am Beispiel der Sprachproduktionssysteme lassen sich die vorgestellten unterschiedlichen Verständnisse von KI noch einmal veranschaulichen. Bereits die Antwort auf die Frage, ob sie Beispiele für breite bzw. allgemeine KI sind, hängt von den jeweiligen Annahmen ab: Zwar sind einzelne Sprachproduktionssysteme nicht auf eine Domäne beschränkt und haben somit ein durchaus breites Funktionenspektrum. Um allerdings die Anforderungen an breite bzw. allgemeine KI zu erfüllen, würde gemeinhin verlangt, dass das System nicht nur im Bereich der Sprache menschliche Kompetenz (nahezu) perfekt simuliert, sondern dies eben auch zeitgleich in (allen) anderen Bereichen vermag, die gemeinhin im menschlichen Kontext als intelligentes Verhalten klassifiziert werden, wie beispielsweise koordinierte Bewegung im Raum und so weiter. Tatsächlich deuten bestimmte, auch bei sehr guten Sprachproduktionssystemen auftretende Fehlleistungen darauf hin, dass die hohe Leistungsfähigkeit nicht auf einem inhaltlichen Verständnis der Texte beruhen kann. Allgemein intelligentes Verhalten ist bei den gegenwärtigen Systemen schon funktional noch in weiterer Ferne und bedürfte zudem auch noch des Einbaus in einen physischen humanoiden Roboter. In einem humanoiden Roboter mit perfekten Bewegungsfähigkeiten und einer menschenähnlichen Mimik und Gestik würden manche ein Beispiel breiter oder gar starker KI sehen, wenn er in der Lage wäre, alle menschlichen kognitiven Fähigkeiten perfekt zu simulieren. Andere würden hingegen bestreiten, dass damit eine Form starker KI vorliegt, da auch eine perfekte Simulation nicht garantiere, dass ein solcher humanoider Roboter mentale Zustände aufweist, über Einsichts- und Urteilsfähigkeit sowie über emotive Einstellungen wie Hoffnungen und Ängste verfüge.

Die unterschiedlichen Konzeptionen hinter den diversen Begrifflichkeiten zur KI gehen auch auf verschiedene grundlegende anthropologische Theoriemodelle zurück (vgl. Abschnitt 3.4). Aus behavioristischer Sicht ist die Unterscheidung zwischen Simulans und Simulandum nicht

[72] Umstritten blieb, ob diese Aussage sich lediglich auf die aktuellen Systeme und die der absehbaren Zukunft bezieht, oder grundsätzlich zu verstehen ist.

sinnvoll, da diese epistemisch nicht unterscheidbar seien. In anderen Konzeptionen[73] jedoch werden mentale Zustände realistisch, das heißt als Merkmale der ontologischen Ausstattung der Welt interpretiert. In solchen Konzeptionen wird an einer kategorischen Unterscheidung zwischen Simulation und Realisierung festgehalten, auch wenn diese Differenz epistemisch nicht unmittelbar zugänglich ist. Entscheidend für den Unterschied zwischen menschlicher und Künstlicher Intelligenz ist demnach das Vorhandensein bestimmter mentaler Eigenschaften wie beispielsweise Verständnis oder Bewusstsein. In dieser Stellungnahme werden die Begriffe zur Charakterisierung der unterschiedlichen Formen Künstlicher Intelligenz in der unten stehenden Weise verwendet. Es wird hierbei vorausgesetzt, dass die Unterscheidung zwischen enger und breiter KI quantitativer bzw. gradueller Natur ist, die Entstehung einer starken KI jedoch einen qualitativen Sprung bedeuten würde:

Enge KI: KI-Anwendungen, die menschliche Fähigkeiten in einer Domäne simulieren bzw. Verfahren wie maschinelles Lernen verwenden, um spezifische Aufgaben zu erfüllen oder spezifische Probleme zu lösen. Nahezu alle derzeit verwendeten KI-Anwendungen fallen in diese Kategorie.

Breite KI: Breite KI-Anwendungen erweitern das Spektrum ihrer Anwendbarkeit über einzelne Domänen hinaus. Sprachproduktionssysteme wie etwa GPT-3 können als Beispiele für breiter werdende KI gelten, da sie zwar nicht domänenspezifisch, jedoch weiterhin auf sprachliche Ein- und Ausgabe beschränkt sind. Eine mögliche Zukunftsvision breiter KI wären Systeme, die solche Sprachkompetenzen mit weiteren kognitiven oder motorischen Kompetenzen zusammenführen, etwa durch Einbau in weitentwickelte Roboter.

Starke KI: Der Begriff der starken KI wird für die Vision einer Künstlichen Intelligenz verwendet, die jenseits der möglicherweise perfekten Simulation menschlicher Kognition auch über mentale Zustände, Einsichtsfähigkeit und Emotionen verfügen würde.

3.2 Intelligenz und Vernunft

3.2.1 Intelligenz

Die im vorigen Abschnitt vorgestellten unterschiedlichen Deutungen von Künstlicher Intelligenz werden seit mindestens den Siebzigerjahren des 20. Jahrhunderts von kontroversen Diskussionen der Frage begleitet, was Computer können und nicht können bzw. demnächst können

[73] Z. B. phänomenologische Positionen oder solche der intentionalistischen Semantik.

und nicht können werden.[74] Um diese Frage zu beantworten, müsste zunächst geklärt werden, von welchen Vorstellungen hinsichtlich der menschlichen Intelligenz dabei ausgegangen wird. Dazu wird jedoch in den sich damit beschäftigenden Wissenschaften, insbesondere der Psychologie, Philosophie und Informatik keine einheitliche Antwort angeboten.

Aus psychologischer Perspektive ist Intelligenz als ein hypothetisches Konstrukt aufzufassen, das als solches zwar verbal umschrieben werden kann, zum Beispiel im Sinne von Verstehen, Urteilen und Schlussfolgern[75] oder zielgerichtetem Handeln, rationalem Denken und effektiver Auseinandersetzung mit der Umwelt[76], aber nicht beobachtbar ist, sondern anhand von Indikatoren in relevanten Aspekten operationalisiert werden muss. In diesem Sinne sind Intelligenztests als Situationen aufzufassen, in denen Menschen Verhalten zeigen können, das vor dem Hintergrund eines theoretischen Vorverständnisses oder einer zugrunde gelegten Definition als mehr oder weniger „intelligent“ bezeichnet werden kann. Dabei können die gewählten Operationalisierungen – auf der Ebene von Teilkomponenten und Subskalen wie auf der Ebene von Testitems – sehr unterschiedlich sein. Sie bewähren sich im Kontext der Bestimmung von Gütekriterien, auf deren Grundlage abgeschätzt werden kann, inwieweit die Durchführung, Auswertung und Interpretation der Testung als objektiv, verlässlich (reliabel) und valide angesehen werden können.

Intelligenz wird hier im Sinne eines Abweichungsquotienten verstanden; der Mittelwert des (normalverteilten) Merkmals liegt in der Grundgesamtheit per definitionem bei 100, die Standardabweichung bei 15.[77] Dies hat Auswirkungen auf die Testkonstruktion, denn bei der Auswahl potenzieller Testitems muss berücksichtigt (geschätzt) werden, wie sich diese in der Grundgesamtheit verteilen, mit anderen Items korrelieren und zwischen verschiedenen Fähigkeitsniveaus unterscheiden. Bei der Validierung von Intelligenztests interessieren Korrelationen mit anderen Verfahren und Merkmalen, die aus theoretischer Perspektive als Indikatoren der Ausprägung verwandter und nicht verwandter Konstrukte betrachtet werden können und deshalb eine bestimmte Höhe aufweisen bzw. diese nicht überschreiten sollten. Darüber hinaus

[74] Dreyfus, H. L. (1992): What Computers still can’t do: a critique of artificial reason. 2., überar. Auflage. Cambridge (MA).
[75] Binet, A.; Simon, T. (1904): Methodes nouvelles pour le diagnostiqc du niveau intellectuel des anormaux. In: L'Année Psychologique 11, 191–244 (DOI:10.3406/psy.1904.3675).
[76] Wechsler, D. (1944). The Measurement of Adult Intelligence. Baltimore (DOI: 10.1037/11329-000).
[77] Das bedeutet, dass ca. 68% der Menschen einen IQ zwischen 85 und 115 sowie ca. 95% einen IQ zwischen 70 und 130 aufweisen.

ist die Möglichkeit, auf der Grundlage aktueller Messwerte zukünftige Leistungen bzw. interindividuelle Unterschiede in relevanten Merkmalen (z. B. Erfolg versus Misserfolg in Schule und Beruf) zu prognostizieren, von großem Interesse.

Der Wechsler-Intelligenztest (Wechsler Adult Intelligence Scale), auf dem die bekanntesten modernen Intelligenztests aufbauen, wurde 1955 von David Wechsler entwickelt und bereits 1956 in einer deutschsprachigen Version, dem Hamburg-Wechsler-Intelligenztest für Erwachsene verwendet. Die aktuelle Version des Tests besteht aus zehn Kerntests und fünf optionalen Untertests, welche unterschiedliche kognitive Fähigkeiten prüfen, die in vier Aufgabengruppen zusammengefasst werden: Sprachverständnis, wahrnehmungsgebundenes logisches Denken, Arbeitsgedächtnis und Verarbeitungsgeschwindigkeit. Hierbei ist wichtig festzuhalten, dass es keine Aufgaben gibt, die diese Dimensionen direkt messen. Vielmehr sind die verwendeten Skalen induktiv aus den im Kontext der Testkonstruktion und -normierung empirisch ermittelten Korrelationen zwischen Unterskalen und Außenkriterien abgeleitet. Das heißt, die Struktur der Intelligenz ergibt sich wesentlich induktiv aus der empirischen Erfassung und Validierung unterschiedlicher, aus dem theoretischen Vorverständnis der Testautorschaft abgeleiteten Aspekte bzw. Dimensionen kognitiver Leistungsfähigkeit (z. B. der Orientierung an einem Generalfaktormodell in der Tradition von Charles Spearman versus. einem Modell voneinander unabhängiger Primärfaktoren in der Tradition von Louis Leon Thurstone).[78]

Die Frage, ob Intelligenz eine einheitliche Fähigkeit ist oder viele Fähigkeiten umfasst, die gegebenenfalls auch voneinander unabhängig sein können, ist empirisch nicht eindeutig zu klären – die dimensionale Struktur wird auf der Grundlage von vorab festgelegten Modellannahmen bzw. Restriktionen ermittelt, die ihrerseits nicht Gegenstand einer empirischen Überprüfung werden können. Allerdings lässt sich festhalten, dass empirisch durchweg positive Korrelationen zwischen den verschiedenen Untertests bzw. Aufgabengruppen nachzuweisen sind. Im erwähnten Wechsler-Intelligenztest bedeutet dies etwa, dass höhere Werte in den Tests der Aufgabengruppe *Sprachverständnis* statistisch mit höheren Werten in den anderen drei Aufgabengruppen einhergeht, auch wenn das Ausmaß der Korrelationen zwischen den verschiedenen Tests je nach Nähe der Aufgaben variiert.[79] Diese durchweg positive Korrelation führte zur Annahme des sogenannten *Generalfaktors* der Intelligenz *g*, welcher den Anteil der

[78] Vgl. Sternberg, R. J. (Hg.) (2020): The Cambridge Handbook of Intelligence. 2. Auflage. Cambridge.
[79] Nach Deary liegt die durchschnittliche Korrelation zwischen allen Tests bei 0.45. Aufgrund ihres statistischen Ursprungs, ist Korrelation innerhalb der vier Skalen höher als zwischen den Skalen. Die höchste Korrelation von 0.74 besteht zwischen Wortschatz-Test und dem Test für Allgemeines Verständnis. Deary, I. J. (2020): Intelligence: A Very Short Introduction. Oxford.

allgemeinen Intelligenz bzw. der kognitiven Leistungen zugrunde liegenden allgemeinen geistigen Fähigkeit bezeichnet, deren Ausprägung sich – weil unidimensional – in einem einzigen Wert ausdrücken lässt, und ca. 40 Prozent der in Leistungsmessungen beobachteten Varianz erklärt. Die übrigen 60 Prozent lassen sich demnach auf unterschiedliche spezifische Fähigkeiten zurückführen.

Innerhalb der Allgemeinen Psychologie und insbesondere mit der Entwicklung der Kognitionspsychologie in den 1970er-Jahren rückte zunehmend die Analyse der Prozesse in den Fokus, die nötig sind, um die Aufgaben der Intelligenztests zu lösen. Umfangreiche Forschungen zur Bearbeitung informationsverarbeitender Aufgaben wie Informationskodierung und geteiltes Hören führten zu der Annahme, dass (verbale) Intelligenz durch die Fähigkeit zur Auswahl und Benutzung von Informationsverarbeitungsmethoden bestimmt wird.[80] Die Kernthemen der kognitionspsychologischen Forschung jedoch, nämlich Denken, Problemlösung und Entscheidungsfindung, die auch außerhalb der Disziplin der Psychologie häufig mit dem Begriff Intelligenz assoziiert werden und auf die häufig in der Entwicklung Künstlicher Intelligenz Bezug genommen wird, fallen in der Psychologie nicht unter den Intelligenzbegriff. Dies mag mit der zunehmenden Spezialisierung innerhalb der Psychologie zusammenhängen. So ist die Erfassung menschlicher Intelligenz ein Teilgebiet der Differenziellen Psychologie, die sich mit Unterschieden zwischen Menschen befasst, wohingegen die Allgemeine Psychologie sich mit den Grundlagen von Wahrnehmung, Lernen sowie insgesamt menschlicher Kognition und Emotion beschäftigt.

Verwiesen sei in diesem Zusammenhang auch auf die Unterscheidung zwischen Intelligenz und Kreativität[81], wobei Letztere in Anlehnung an Guilford, den Begründer der modernen Kreativitätsforschung, als flüssige, flexible und ursprüngliche Erzeugung von Konzepten von Lösungen für neuartige Probleme definiert werden kann.[82] Von Interesse ist hier seine Unterscheidung zwischen konvergentem und divergentem Denken.[83] Im Unterschied zum konvergenten Denken, das durch logische Schlussfolgerungen zu einer einzigen oder besten Lösung gelangt (wobei das Ergebnis mehr oder weniger vollständig durch die vorhandene Information determiniert ist), liefert das für Kreativität charakteristische divergente Denken mehrere alternative

[80] Hunt, E. et al. (1975): What does ist meant to be high verbal? In: Cognitive Psychology, 7 (2), 194-227 (DOI: 10.1016/0010-0285(75)90010-9).
[81] Kruse, A.; Schmitt, E. (2011): Die Ausbildung und Verwirklichung kreativer Potenziale im Alter. In dies. (Hg.): Kreativität im Alter. Heidelberg, 15-46; Lubart, T. I. (2018): The Creative Process: Perspectives from Multiple Domains. Paris.
[82] Guilford, J. P. (1950): Creativity. In: American Psychologist, 5(9), 444-454 (DOI: 10.1037/h0063487).
[83] De Vries, H. B.; Lubart, T. I. (2017): Scientific Creativity: Divergent and Convergent Thinking and the Impact of Culture. In: The Journal of Creative Behavior, 53 (2), 145-155 (DOI: 10.1002/jocb.184).

Lösungen, die jeweils den gegebenen Anforderungen entsprechen. Dabei gelten die Anzahl der generierten Lösungen und deren Qualität als Maß für die Ausprägung des divergenten Denkens. Neben dem divergenten Denken wurden und werden auch weitere kognitive Prozesse als zentrale Voraussetzungen für Kreativität diskutiert. Zu nennen sind hier insbesondere Fähigkeiten und Fertigkeiten im Bereich von Wahrnehmung, Problemdefinition, Einsicht, Induktion, Bildung von Analogien und ungewöhnlichen Assoziationen, die Bewertung von Ideen und die Organisation von Wissenssystemen.

Während der Begriff der Intelligenz ursprünglich auf ein Ensemble menschlicher Leistungen verweist, wie sie in klassischen Intelligenztests gemessen werden[84], hat sich der Blick auf Intelligenz in jüngerer Zeit sukzessive erweitert. So entstanden Konzepte wie die der sozialen bzw. emotionalen Intelligenz, welche einerseits den Begriff der Intelligenz auf weitere Fähigkeiten ausweiteten sowie andererseits die Wechselwirkungen zwischen Emotion und Kognition sowie den sozialen und kulturellen Aspekt von Intelligenz in den Blickpunkt rückten. Darüber hinaus entwickelte sich rund um die Stichwörter *embodied*, *embedded*, *enactive* und *extended cognition* ein Forschungsfeld, das in Philosophie, Psychologie und Robotik die Rolle des Körpers einerseits und der Umwelt andererseits für Intelligenz und kognitive Leistungen erforscht.[85]

Spätestens diese Erweiterungen werfen die Frage auf, wie die Übertragung des Intelligenzbegriffs auf technische Artefakte zu verstehen ist. Bei einigen Merkmalen wie beispielsweise der elementaren Rechenfähigkeit, dem logischen Schlussfolgern oder Gedächtnisleistungen entsteht der Eindruck, dass menschliche Fähigkeiten eindeutig auf technische Artefakte übertragen werden können. Auch diesbezüglich sind schon kritische Fragen zu stellen, zum Beispiel ob die menschliche Erinnerungsfähigkeit in gleicher Weise eine Gedächtnisleistung ist wie die Aktivierung eines technischen Speichers. Einerseits sind die quantitativen Leistungen technischer

[84] Verwiesen sei hier auf die von Thurstone faktorenanalytisch ermittelten sieben Primärfaktoren induktives Schließen, räumliches Vorstellungsvermögen, Wahrnehmungsgeschwindigkeit, Rechenfähigkeit, verbales Verständnis, assoziatives Gedächtnis und Wortflüssigkeit (vgl. Thurstone, L.L. (1938). Primary mental abilities. University of Chicago Press: Chicago. Thurstone, L.L. & Thurstone, G.W. (1941). Factorial Studies Of Intelligence. University of Chicago Press: Chicago.) sowie die Differenzierungen zwischen fluider vs. kristalliner Intelligenz bei John Horn und Raymond Catell (Horn, J.L. & Cattell, R.B. (1966). Refinement and test of the theory of fluid and crystallized intelligence. Journal of Educational Psychology, 57, 253 – 270.) und kognitiver Mechanik vs. Pragmatik bei Paul Baltes (Baltes, P. B. (1987). Theoretical Propositions of Lifespan Developmental Psychology: On the Dynamics between Growth and Decline. Developmental Psychology, 23, 611-626. http://dx.doi.org/10.1037/0012-1649.23.5.611).

[85] Clark, A (2012): Embodied, embedded, and extended cognition. In: Frankish, K.; Ramsey, W. (Hg.): The Cambridge Handbook of Cognitive Science. (DOI: 10.1017/CBO9781139033916.018).

Speicher der menschlichen Erinnerungsfähigkeit um Größenordnungen überlegen. Andererseits sortiert der Mensch seine Gedächtnisleistungen beispielsweise nach der jeweils kontextuell bestimmten Bedeutung, während ein technischer Speicher unterschiedslos je nach den technischen Vorgaben Daten aufnimmt oder nicht. Bei Intelligenzleistungen mit emotiven und kreativen Qualitäten verstärkt sich der Verdacht, dass es sich hierbei um anthropomorphe Übertragungen handelt. Die Klärung solcher Vergleichbarkeitsprobleme hängt somit wesentlich von den Kriterien ab, durch die man eine spezifisch menschliche Intelligenzleistung bestimmt sieht. Man sollte daher die Verwendung des Ausdrucks „Intelligenz“ in der Wortverbindung „Künstliche Intelligenz“ eher als eine Metapher einordnen, deren Beschreibungs- und Erklärungsfunktion genauerer Aufklärung bedarf.

3.2.2 Vernunft

Bereits lange vor der Einführung des Begriffs der Intelligenz wurde der Begriff der Vernunft verwendet, um die spezifische menschliche Fähigkeit zu kennzeichnen, sich in der Welt zu orientieren, selbstverantwortlich zu handeln und so der eigenen Lebenspraxis eine kohärente Struktur zu geben. Intelligenz ist für Vernunft eine wichtige Voraussetzung, aber keine hinreichende Bedingung. Der Begriff der Vernunft gehört zu den basalen Grundkategorien menschlicher Selbst- und Weltdeutung, die unsere Kultur seit der Antike maßgeblich geprägt haben. Schon das weite Wortfeld (griechisch: *logos, nous, dianoia, phronesis*; lateinisch: *ratio, mens, intellectus, prudentia*) deutet darauf hin, dass es sich um einen überaus komplexen Begriff handelt, der vielfältige Binnendifferenzierungen kennt und verschiedene (kognitive) Teilkompetenzen umfasst. Strukturell geht es um ein mehrdimensionales Beziehungsgefüge von Denk-, Reflexions- und Operationsformen, das in seiner Gesamtheit im Dienste einer möglichst adäquaten Wirklichkeitserschließung steht und in einen komplexen sozialen und kulturellen Kontext verwoben ist. Als Inbegriff bestimmter Ansprüche, denen wir uns im Denken, Sprechen, Erleben und Handeln unterstellen, umfasst der Vernunftbegriff unterschiedliche – propositionale und nichtpropositionale – Wissensformen und Rationalitätstypen, die von methodisch-prozeduralem Know-how über ästhetische Wahrnehmungsfähigkeit und Kreativität sowie verschiedene soziale Interaktionsfähigkeiten bis hin zu einer umfassenden Lebensführungskompetenz reichen. Von grundlegender Bedeutung für unsere Thematik ist dabei die Gegenüberstellung von *theoretischer Vernunft*, die sich auf den Erkenntnisgewinn richtet, um zu wahren empirischen oder apriorischen Urteilen zu gelangen, und *praktischer Vernunft*, die auf ein kohärentes, verantwortliches Handeln abzielt, um ein gutes Leben zu ermöglichen.

Vor allem im Blick auf den Gebrauch der *theoretischen* Vernunft, der primär auf Erkenntnisgewinn durch die Formulierung wahrer empirischer Urteile abzielt, scheinen sich zumindest prima facie einige Parallelen zur Arbeitsweise von KI-Systemen aufzudrängen. So spielen in beiden Bereichen Fähigkeiten der Informationsverarbeitung, des Lernens, des logischen Schlussfolgerns und konsistenten Regelfolgens sowie der sinnvollen Verknüpfung gespeicherter Daten eine zentrale Rolle. Bei näherer Betrachtung zeigen sich jedoch insofern gravierende Differenzen, als sich nicht nur die Arbeitsweise des menschlichen Gedächtnisses in mehrfacher Hinsicht vom technischen Speicher eines Computers unterscheidet, sondern auch die menschliche Urteilspraxis technisch nicht substituierbar ist. Auch wenn in diesem Zusammenhang die wahrheitstheoretischen Implikationen der Formulierung und Begründung deskriptiver Urteile nicht näher entfaltet werden können[86], ist doch darauf hinzuweisen, dass zumindest die bislang verfügbaren KI-Systeme die dafür relevanten Fähigkeiten des Sinnverstehens, der Intentionalität und der Referenz auf eine außersprachliche Wirklichkeit nicht besitzen.

Dieser Befund bestätigt sich auch bezüglich der uns hier besonders interessierenden *praktischen* Vernunft, die insofern noch weit komplexerer Natur ist, als ihr Ziel nicht nur in wohlbegründeten praktischen Einzelurteilen, sondern in einem möglichst richtigen und verantwortlichen Handeln besteht, das über einen langen Zeitraum aufrechterhalten wird, eine kohärente Ordnung der Praxis garantiert und damit ein insgesamt gutes Leben ermöglicht.[87] Dazu bedarf es mehrerer Einzelkompetenzen, deren Simulationsmöglichkeiten durch technische Artefakte gegenwärtig mit Blick auf die unterschiedlichen Relationen und Wechselwirkungen zwischen Mensch und Maschine (vgl. Abschnitt 4.3) kontrovers diskutiert werden. Ohne Anspruch auf Vollständigkeit seien dabei die folgenden acht Teilfähigkeiten exemplarisch besonders hervorgehoben:

Erstens braucht es ein *Verständnis* der für unsere Moralsprache konstitutiven evaluativen und deontischen Prädikatoren: Von einem vernünftigen Wesen erwarten wir, dass es über die Fähigkeit verfügt, die Bedeutung der verschiedenen, phänomenologisch gehaltvollen Ausdrücke zur Bezeichnung moralisch relevanter Güter, Werte und Haltungen sowie deontischer Prädikate zur Qualifizierung von Handlungen (wie richtig bzw. falsch) angemessen zu verstehen und situationsadäquat zu gebrauchen.

---

[86] Hier könnte man z. B. aufführen: Bovens, L.; Hartmann, S. (2004): Bayesian Epistemology. Oxford.
[87] Bormann, F.-J. (2021): Ist die praktische Vernunft des Menschen durch KI-Systeme ersetzbar? Zum unterschiedlichen Status von menschlichen Personen und (selbst-)lernenden Maschinen. In: Fritz, A. et al.(2021): Digitalisierung im Gesundheitswesen. Anthropologische und ethische Herausforderungen der Mensch-Maschine-Interaktion. Freiburg, 41-64. S. 48-51.

Zweitens wird ein *Unterscheidungs- und Einfühlungsvermögen* benötigt, um die moralisch relevanten Differenzen zwischen einzelnen moralischen Gütern, Werten, Handlungstypen und Lebensformen möglichst präzise und realitätsnah erfassen sowie anderen Menschen empathisch begegnen zu können.

Drittens muss die Fähigkeit zur *Abwägung* konfligierender Güter und Werte vorliegen: Die praktische Vernunft beinhaltet auch ein deliberatives Vermögen, das immer dann ins Spiel kommt, wenn komplexe Handlungsstrategien entwickelt werden müssen oder mehrere moralisch bedeutsame Gesichtspunkte aufgrund bestimmter ungünstiger Umstände in einer konflikthaften Beziehung zueinanderstehen. Mittels des Vermögens der Güterabwägung vermag die handelnde Person nicht nur zu erkennen, welche Güter in zeitlicher Hinsicht prioritär erstrebt oder gesichert werden müssen, um bestimmte Ziele zu erreichen, sondern welchen Gütern im Konfliktfall der Vorrang zuzuerkennen ist, um ein situativ richtiges Handeln zu ermöglichen.

Viertens bedarf es der Befähigung zum *reflektierten Umgang mit Regeln* unterschiedlicher Reichweite: Die praktische Vernunft schließt auch die Fähigkeit ein, moralische Regeln (wie z. B. Normen und Prinzipien) verstehen, korrekt anwenden und falls nötig auch weiterentwickeln zu können, um Probleme zu lösen und ein realitätsadäquates kohärentes Handeln über längere Zeiträume zu ermöglichen. Obwohl ein Großteil des menschlichen Handelns von Routinen und Konventionen bestimmt wird, gibt es auch vielfältige Herausforderungen und Konfliktsituationen, die durch ein konventionelles oder gar starr deterministisches Regelfolgen allein gerade nicht zu bewältigen sind, sondern Kreativität und einen flexibleren Umgang mit regulatorischen Vorgaben auf der Grundlage eines unvertretbaren Aktes der praktischen Urteilskraft erfordern.[88]

Fünftens wird die Fähigkeit zum *intuitiven Erfassen komplexer Handlungssituationen und Umstände* benötigt: Menschen müssen oft unter großem Zeitdruck weitreichende Entscheidungen treffen und dabei vielfältige Merkmale eines Handlungskontextes berücksichtigen.

[88] In diesem Zusammenhang ist auf einen grundsätzlichen Unterschied zwischen dem algorithmischen gegenüber einem heuristischen Regelverständnis hinzuweisen. In einem algorithmischen Verfahren bilden die Regeln den Auswahlfilter, durch den die Fälle als Kandidaten überprüft und verworfen oder zur Überprüfung im nächsten Schritt angenommen werden. In algorithmischen Verfahren steht damit das Verhältnis von Regel und Fall fest. In einem heuristischen Verfahren werden die Regeln durch die Subsumtion eines Falles pragmatisch und semantisch mit-konstituiert, sodass nicht prä-prozedural feststeht, unter welche Regel der Fall gehört.

Sechstens bedarf es eines *Urteilsvermögens,* mittels dessen Personen in der Lage sind, Entscheidungen zwischen verschiedenen Handlungsalternativen zu treffen und singuläre Handlungskonstellationen bestimmten generellen Handlungstypen zuzuordnen.[89]

Siebtens braucht es die Fähigkeit zur *Begründung* der eigenen moralischen Urteile und der ihnen korrespondierenden Praxis: Die Fähigkeit, Gründe zu geben und zu nehmen (*give and take reasons*) und sich im Urteilen und Handeln daran auszurichten, schließt neben der Bereitschaft zur kritischen Reflexion eigener partikularer Interessen auch die Fähigkeit ein, einen moralischen Standpunkt (*moral point of view*) einzunehmen, also die für die moralische Qualifikation einer Handlung relevanten Gründe aus der Dritte-Person-Perspektive zu beurteilen.

Achtens muss die Fähigkeit zur Affekt- und Impulskontrolle vorliegen, um die jeweils gefällten praktischen Urteile auch handlungswirksam werden zu lassen. Gerade bei der Verfolgung anspruchsvoller Ziele, die vielfältige Vorarbeiten und einen langen Atem verlangen, ist es wichtig, die erforderliche Willensstärke aufzubringen und zumindest solchen spontanen Affekten, Neigungen und Impulsen zu widerstehen, die den langfristigen Erfolg der jeweiligen Bemühungen gefährden oder sogar verunmöglichen können.

Die Unterscheidung der verschiedenen Teilkompetenzen der Vernunft ist für unsere Thematik aus zwei Gründen bedeutsam: Erstens ist es durchaus möglich, dass es partielle Überschneidungen des Kompetenzprofils moderner KI-Systeme mit dem komplexen Phänomen menschlicher Vernunft gibt, was insbesondere im Bereich des Regelfolgens und der Weiterentwicklung vorgegebener Algorithmen der Fall sein dürfte. Zweitens ist zu berücksichtigen, dass die hier genannten, für den Bündelbegriff der „praktischen Vernunft" konstitutiven Fähigkeiten nicht einfach im Sinne isolierter Einzelelemente beziehungslos nebeneinanderstehen. Vielmehr ist von vielfältigen Wechselwirkungen, Rückkopplungen und Bedingungsverhältnissen zwischen ihnen auszugehen. Sie bilden einen integralen Bestandteil einer komplexen menschlichen Natur, die im Sinne einer leib-seelischen Einheit zu verstehen ist. Menschliche Vernunft ist stets als *verleiblichte Vernunft* zu begreifen (vgl. Abschnitt 3.4.3). Nur so ist zu erklären, dass praktische Überlegungen überhaupt handlungswirksam werden können. Zurückzuweisen ist eine Deutung,

[89] In Anspielung auf die durch das richterliche Judiz zu leistende intellektuelle Aufgabe hat Kant für heuristische Verfahren dieser Art den Begriff der *Urteilskraft* geprägt. Das algorithmische Verfahren ordnet Kant der bestimmenden Urteilskraft zu, der Domäne des Verstandes (vgl. Kant, I. (1781): Kritik der reinen Vernunft- B 360f.). Das in den praktischen Disziplinen wie Pädagogik, Jurisprudenz oder Ökonomik zugrunde zulegende Verfahren der Suche nach der angemessenen Passung von Regel und Fall zeichnet Kant als reflektierende Urteilskraft aus, die zur Domäne der praktischen Vernunft gehört z. B. Kant, I. (1790): Kritik der Urteilskraft. AA 385.

die versucht, vernünftige Vollzüge aus einer rein individualistischen Perspektive zu rekonstruieren. Da jeder Mensch Teil einer sozialen Mitwelt und kulturellen Umgebung ist, die sich nachhaltig auf seine Sozialisation auswirkt, müssen auch überindividuelle kulturelle Faktoren in die Deutung der praktischen Vernunft einbezogen werden. Ein angemessenes Verständnis insbesondere des praktischen Vernunftgebrauchs ist eng mit unserem basalen Selbstverständnis als handlungsfähige Personen verbunden. Da technische Artefakte in immer neuen Formen in die Handlungswelt der Menschen integriert werden, mit Menschen interagieren oder sogar Teilfunktionen menschlichen Handelns übernehmen, ist es wichtig, zunächst den Handlungs- und Verantwortungsbegriff zu klären.

3.3 Handlung und Verantwortung

3.3.1 Handlung

Auch wenn im Alltag gelegentlich alle möglichen Ereignisse als Handlungen bezeichnet werden, fassen die Normwissenschaften Ethik und Jurisprudenz und auch die Psychologie den Handlungsbegriff oft enger.[90] Dabei wird angenommen, dass Menschen in der Lage sind, aktiv, zweckgerichtet und kontrolliert auf die Umwelt einzuwirken und dadurch Veränderungen zu verursachen. Das bedeutet, dass nicht jedes menschliche Tun, das auf die Umwelt einwirkt, als Handlung zu verstehen ist, sondern nur solches, das zweckgerichtet, beabsichtigt und kontrolliert ist.[91] Unterstellt man, dass Maschinen nicht zweckgerichtet operieren, also keine Absichten haben, dann ist die Zuschreibung von Handlungen in Bezug auf Maschinen in diesem engen Sinne nicht möglich.

Für den Menschen, der im engen Sinn handelt, hat sich auf dem Hintergrund einer langen zurückreichenden Begriffsgeschichte[92] im Rahmen einer umfassenden Theorie praktischer Vernunft der Begriff der *Person* eingebürgert.[93] Personen sind Akteure, die Verantwortung für ihr

[90] In der Psychologie auch, vgl. z. B. die verschiedenen (etablierten) Handlungsphasenmodelle.
[91] Hier wird eine Form der Handlungserklärung unterstellt, die die Warumfrage (im Sinne von „Warum hast du das getan?“) durch Angabe der Absicht bzw. Zwecke der Handlung beantwortet sieht (intentionale, teleologische Handlungserklärung, „Intentionalismus“). Ihr steht eine Form der Handlungserklärung gegenüber, die die Angabe der die Handlung auslösenden Ursachen als Antwort vorsieht (kausale Handlungserklärung, „Kausalismus“). Ob eine Handlung intentionalistisch oder kausalistisch erklärt werden sollte, ist keine Frage der Wahrheit / Falschheit der Erklärung, sondern eine Frage der Kontextadäquatheit der Handlungsdeutung. Für die Normwissenschaften steht die Frage der Absicht bzw. der Zweck der Handlung im Vordergrund der Betrachtung, ohne dass die Möglichkeit einer kausalen Handlungserklärung in Abrede gestellt wird. Vgl. Horn, C.; Löhrer, G. (Hg.) (2010): Gründe und Zwecke. Texte zur aktuellen Handlungstheorie. Berlin.
[92] Fuhrmann, M. et al. (1989): Person. In: Ritter, J.; Gründer, K. (Hg.): Historisches Wörterbuch der Philosophie. Band 7: P-Q. Basel, Sp. 269-338 (DOI: 10.24894/HWPh.5339).
[93] Quante, M. (2007): Person. Berlin.

Verhalten tragen, die die kognitiven Bedingungen erfüllen, um Handlungsoptionen zu erkennen und die Gründe deliberieren können, also über ein hinreichendes Maß theoretischer und praktischer Vernunft verfügen.

Der Handlungsbegriff ist auch deswegen so bedeutsam in der Diskussion um maschinelle Fertigkeiten und Künstliche Intelligenz, da dieser Diskurs sich etwa seit der Jahrtausendwende von der Konzentration auf möglicherweise kognitive Kompetenzen abgewandt hat und inzwischen zunehmend auf praktische Kompetenzen konzentriert.[94]Nicht in welcher Weise, wenn überhaupt, Maschinen denken, sondern in welchem Sinne Maschinen handeln können, steht verstärkt im Vordergrund. Diese Verschiebung der Aufmerksamkeit geht Hand in Hand mit technischen Entwicklungen hin zu maschinellen Systemen, die nicht lediglich auf hohe Informationsverarbeitungskapazität setzen, um menschliches Handeln zu unterstützen, sondern jenes teilweise gar ersetzen können oder sollen.

Automatisierte oder algorithmische Entscheidungssysteme (ADM-Systeme, abgekürzt von *automated decision making* oder *algorithmic decision making*) zum Beispiel erstellen auf Basis von Berechnungen Prognosen darüber, wie geeignet eine sich bewerbende Person für eine Stelle ist oder mit welcher Wahrscheinlichkeit Menschen Kredite zurückzahlen oder straffällig werden (vgl. Kapitel 8). Auch wenn diese Systeme oftmals nur der Unterstützung der menschlichen Entscheidung dienen, so können Entscheidungen auch komplett an jene delegiert werden. Damit stellt sich die Frage, in welchem Sinne solche maschinellen Vollzüge außerhalb des obigen engen Handlungsbegriffs doch in bestimmten Kontexten als Handlungen in einem weiteren Sinne wahrgenommen werden können oder berücksichtigt werden müssen. Daran anknüpfend gibt es einen Diskurs, ob und inwieweit zunehmend eigenständige, das heißt ohne menschliches Zutun funktionierende maschinelle Systeme als „Agenten" in der Folge für ihr „Handeln" verantwortlich gemacht werden können, etwa mit Blick auf Fragen der Haftung (vgl. Abschnitt 2.2.5).[95]

Den folgenden Überlegungen liegt mit Blick auf die oft schillernde Verwendung zentraler ethisch relevanter Begriffe wie *Handlung* oder *Verantwortung* im Kontext der zeitgenössischen KI-Debatte die Annahme zugrunde, dass wenigstens drei Klassen von Entitäten terminologisch

[94] Dreyfus, H. L. (1992): What Computers still can't do. 2., überar. Auflage. Cambridge (MA), London.
[95] Vgl. dazu etwa Hilgendorf, E. (2014): Robotik im Kontext von Recht und Moral. Baden-Baden; Hilgendorf, E (2020): Digitalisierung, Virtualisierung und das Recht. In: Kasprowicz, D.; Rieger, S. (Hg.): Handbuch Virtualität. Wiesbaden, 405–424; Ebers, M. et al. (Hg.) (2020): Künstliche Intelligenz und Robotik. München.

klar gegeneinander abzugrenzen sind. Dies wären erstens Pflanzen und Tiere, die zwar in vielfältiger Weise auf ihre Umwelt reagieren können, in ihrem jeweiligen Repertoire aber doch so begrenzt sind, dass der ausgeführte enge Handlungsbegriff auf sie nicht anwendbar ist.[96] Zweitens sind Menschen in dem Maße im engen Sinn als handlungsfähig zu bezeichnen, als sie dazu imstande sind, absichtlich Veränderungen zu bewirken, wobei solche Handlungen nicht nur Taten, sondern auch deren bewusstes und absichtliches Unterlassen umfassen können.[97] Drittens gibt es technische Artefakte unterschiedlicher Komplexitätsgrade, deren jeweilige Vollzüge oder Operationen zwar Veränderungen in der Welt bewirken und flexibel mit anspruchsvollen Herausforderungen der menschlichen Lebenswelt umgehen können.[98] Da sie diese Veränderungen aber nicht absichtlich herbeiführen, haben sie selbige daher auch nicht in einem moralischen und rechtlichen Sinne zu verantworten (vgl. Abschnitt 3.3.2).

Auch wenn es zwischen diesen drei Klassen von Entitäten unterschiedliche Arten von Wechselwirkungen (vgl. Abschnitt 4.3) geben kann, scheint es sinnvoll, den Handlungsbegriff im engen Sinne Menschen vorzubehalten, um inflationären Ausweitungen des Akteur-Status zu vermeiden und konzeptionelle Grenzziehungen zu ermöglichen. Entscheidend ist demnach das Konzept der Handlungsurheberschaft bzw. Autorschaft, das auf die universelle menschliche Handlungserfahrung verweist, sich selbst und andere im Hinblick auf bestimmte Ereignisse und Zustände als Urheber anzusehen.[99] Die Fähigkeit zur Handlungsurheberschaft kann als Grundlage von Autonomie betrachtet werden, also dafür, dass handelnde Menschen ihre Handlungen nach Maximen ausrichten können, die sie sich selbst setzen. Diese Konzeption schließt jedoch nicht aus, dass Handlungen mitunter auch durch Befolgen von Autoritäten und Traditionen aus-

[96] Ausnahmen werden für einige hochentwickelte Tiere diskutiert, darunter z. B. Primaten und Rabenvögel; vgl. etwa Huber, L. (2021): Das rationale Tier: Eine kognitionsbiologische Spurensuche. Berlin.
[97] Für Unterlassungen, nicht aber für Nicht-Handeln, kann man getadelt oder verurteilt werden. Das Unterlassen bezieht sich demgemäß auf ein bestimmtes Handlungsschema (z. B. des Helfens) und ist nicht nur die unbestimmte Negation der Ausführung eines Handlungsschemas. Unterlassungshandlungen sind somit „Ereignisse in der Welt" und als solche ohne Weiteres mögliche Ursachen von Wirkungen (Folgen) in der Welt. Siehe etwa Roxin, C.; Greco, L. (2020): Strafrecht, Allgemeiner Teil. Band I: Grundlagen. Der Aufbau der Verbrechenslehre. 5. Auflage. München, 335 ff.; Bottek, C. (2014): Unterlassungen und ihre Folgen. Handlungs- und kausaltheoretische Überlegungen. Tübingen.
[98] Rammert, W.; Schulz-Schaeffer, I. (2002): Technik und Handeln – wenn soziales Handeln sich auf menschliches Verhalten und technische Artefakte verteilt. Berlin. https://nbn-resolving.org/urn:nbn:de:0168-ssoar-1107 [04.01.2023], 11-64.
[99] Deutscher Ethikrat (2017). Big Data und Gesundheit – Datensouveränität als informationelle Freiheitsgestaltung. Berlin, 175-178; Nida-Rümelin, J. (2020): Eine Theorie praktischer Vernunft. Berlin, 376-408.

geführt werden. Wo dies geschieht, setzt das Konzept der Handlungsurheberschaft jedoch voraus, dass Menschen ihr eigenes Dasein in ein Verhältnis zu solchen Bestimmungen setzen können, etwa durch Überwindung, Widerstand oder Nachgeben.

Aus einer Handlung können neben den beabsichtigten Folgen auch nicht beabsichtigte, aber der handelnden Person erkennbare Folgen erwachsen. Auf diese Weise erscheint es möglich, auch fahrlässiges Tun zu erfassen, das im Kontext von KI eine große Rolle spielt.[100] Des Weiteren finden Handlungen umgeben von anderen raumzeitlichen Ereignissen statt, die als Umstände der Handlung für deren moralische und rechtliche Bewertung von Bedeutung sein können. Auch werden Handlungen zwar methodisch primär individuellen Akteuren zugesprochen; das schließt aber kollektive Handlungen nicht aus, bei denen die einzelnen Personen von vornherein in einem Kontext der Koordination agieren, ihre Handlungen auf Kooperation bezogen sind und durch Kommunikation gestützt werden.

Der hier verwendete, eng gefasste Handlungsbegriff schließt nicht aus, dass Technologie erheblichen Einfluss auf menschliches Handeln oder die menschliche Handlungserfahrung haben kann. Technik beeinflusst und verändert Gesellschaft; und gleichzeitig beeinflusst Gesellschaft die Technikentwicklung und den Technikeinsatz. Gerade die in den letzten Jahren stark zunehmende Durchdringung der menschlichen Lebenswelt mit informationstechnisch immer leistungsfähigeren Maschinen, die mit anspruchsvoller Sensorik und Motorik sowie vernetzt arbeiten, führt zu hybriden, sozio-technischen Konstellationen, in denen Menschen und Maschinen eng verwoben sind und auf komplexe Weise interagieren. Dies kann das Verhalten und die Handlungen von Menschen stark beeinflussen und ihre individuellen Freiheitsspielräume und Kontrollmöglichkeiten einschränken. Zudem können fortgeschrittene und mit flexiblen, selbstlernenden Algorithmen arbeitende maschinelle Systeme menschliches Tun zum Teil so gut imitieren, dass sie wie intentionales menschliches Handeln erscheinen, was weitere ethische Fragen aufwirft (vgl. Kapitel 4).

Auch mit Blick auf diese sozio-technische Komplexität bis hin zu Unterscheidungsschwierigkeiten erscheint es sinnvoll, an einem engen Handlungsbegriff, der an das zentrale Kriterium der Intentionalität gebunden ist, festzuhalten. Dieses Intentionalitätskriterium ist zudem entscheidend für die Möglichkeit der Zuschreibung von Verantwortung im Kontext von Mensch-

[100] Ein Beispiel liefert etwa der Bereich des autonomen Fahrens: Unterläuft dem Programmierer ein Fehler, der später zur Schädigung von Verkehrsteilnehmern führt und war dies für ihn vorhersehbar und vermeidbar sowie aus Gründen des überwiegenden Schutzes der Interessen der anderen Personen auch von Rechts wegen zu vermeiden, liegt hierin ein fahrlässiges Fehlverhalten, für das er zur Verantwortung zu ziehen ist. Hevelke, A.; Nida-Rümelin, J. (2015): Responsibility for Crashes of Autonomous Vehicles: An Ethical Analysis. In: Science and Engineering Ethics 21, 619-630.

Maschine-Interaktionen in zunehmend komplexer sozio-technischer Vernetzung. Im Hinblick auf digitale Techniken handeln Menschen nicht jederzeit in vollem Maße autonom, sondern verstehen oft weder die technischen Zusammenhänge, noch haben sie immer umfassende Informationen oder hinreichende Wahlfreiheit, um bewusste Entscheidungen im Umgang mit der Technik treffen zu können. Solche Konstellationen können vielfältig ethische Bedeutung entfalten, wenn es um Fragen der Verantwortung geht.

3.3.2 Verantwortung

Fragen der Verantwortung und Verantwortlichkeiten knüpfen an den Handlungsbegriff an. Die Rolle menschlicher Verantwortung für die kausale Verursachung von Handlungsfolgen kann in vielfacher Weise thematisiert, reflektiert, diskutiert und modifiziert werden. Bei der Verantwortung für die Resultate von Zuschreibungshandlungen in sozialen Kontexten geht es vor allem darum zu klären, welche Verantwortung von welchen Akteuren übernommen werden *soll*, um einer zunehmenden Verantwortungsdiffusion entgegenzuwirken.[101] Verantwortungszuschreibung und -verteilung erfolgen zu dem Zweck, Praxisfelder wie den Straßenverkehr, den Schulbetrieb oder den Umgang mit KI in verschiedenen Anwendungsbereichen so zu strukturieren und gegebenenfalls rechtlich zu regulieren, dass sich dadurch eine möglichst „gute Praxis“ entfalten kann.

In den Zuschreibungen wird der Kreis der verantwortungsfähigen Individuen abgegrenzt, der Stellenwert der kausalen Verursachung geregelt und es werden Kriterien festgelegt, welche Voraussetzungen Menschen erfüllen müssen, um ihnen Verantwortung zuschreiben zu können. Somit stellt sich die Frage, wer für was direkt oder indirekt Verantwortung übernehmen kann oder soll, wenn Individuen, Gruppen und Institutionen aus und in verschiedenen Bereichen wie im Privatleben, in der Forschung, Wirtschaft und Politik sowohl miteinander als auch mit maschinellen Systemen und insbesondere KI-Systemen zusammenwirken.

Verantwortung kann als Konzept einer vielfachen Relation rekonstruiert werden. Im Kontext dieser Stellungnahme erscheint eine fünfstellige Relation angemessen:[102] Wer ist für was, gegenüber wem, vor wem und unter welcher Norm verantwortlich? Allerdings sprechen wir auch

---

[101] Vgl. Grunwald, A. (2021): Der homo responsibilis. Nachdenklicher Gang durch den Garten aktueller Erzählungen. In: ders. (Hg.): Wer bist du, Mensch? Transformationen menschlicher Selbstverständnisse im wissenschaftlich-technischen Fortschritt. Freiburg, 216-239.

[102] Loh, J. (2016): Strukturen und Relata der Verantwortung. In: Heidbrink, L. et al.(Hg.): Handbuch Verantwortung. Kiel, Berlin, Wien.

von einer verantwortlichen Person; in diesem Falle ist der Begriff einstellig, und der Zusammenhang zwischen dem einstelligen moralischen Begriff der verantwortlichen Person und den unterschiedlichen, meist mehrstelligen Kriterien der Verantwortungszuschreibung ist eine eigene philosophische und rechtstheoretische Thematik. Ohne die genuine individuelle moralische Verantwortung wären auch die weitergehenden Ausdifferenzierungen gegenstandslos.[103]

Demnach kann man erstens ganz grundsätzlich beim Verantwortungssubjekt (*wer*) ansetzen, das Verantwortung übernehmen kann. Verantwortungssubjekte tragen Verantwortung, als Einzelperson oder Mitglieder eines Kollektivs, beispielsweise einer Institution. Davon zu unterscheiden ist zweitens das Verantwortungsobjekt (*was*), für das Verantwortung übernommen wird, zum Beispiel Handlungen sowie deren Gründe und Folgen.[104] Als drittes Relationselement werden die vom Handeln des Verantwortungssubjektes (direkt oder indirekt) Betroffenen benannt (*gegenüber wem*). Das vierte Relationselement bildet die Instanz, vor der die Verantwortung übernommen wird (*vor wem).* Das Gewissen als Inbegriff der praktischen Vernunft, andere Personen oder auch eine staatliche Rechtsgemeinschaft sind hier denkbar. Für eine normative Stellungnahme ist zudem ein fünftes Relationselement bedeutsam, nämlich die Regel oder das Prinzip, dem eine verantwortliche Praxis gerecht werden sollte, zum Beispiel das Prinzip, andere nicht zu schädigen (*unter welcher Norm).*[105]

Vor allem das dritte Relationselement – die Betroffenen – ist nicht immer leicht einzugrenzen. Ungeachtet dessen ist deren Berücksichtigung und Einbezug gerade in Anbetracht des steigenden Einsatzes von KI-Systemen in vielen Gesellschafts- und Lebensbereichen von zentraler Bedeutung. An dieses Desiderat knüpfen sich auch Forderungen nach Transparenz und Nachvollziehbarkeit, welche eine Voraussetzung für die Beteiligung und Berücksichtigung von Betroffenen darstellen.

---

[103] Nida-Rümelin, J. (2011): Verantwortung. Stuttgart.

[104] Personen können auch für ein Unterlassen verantwortlich sein. Auch durch ein Unterlassen wird ein Ereignis in der Welt verursacht. Die Verantwortung kann aus einer Sonderbeziehung des Unterlassenden im Hinblick auf das dadurch beeinträchtigte Rechtsgut erwachsen („Garantenstellung"). Siehe Freund, G. (1992): Erfolgsdelikt und Unterlassen. Zu den Legitimationsbedingungen von Schuldspruch und Strafe. Köln, München. Ein „Jedermanns-Unterlassen", für das es auf keine solche Sonderbeziehung ankommt, lässt sich darüber hinaus auf allgemeine Solidaritätspflichten stützen, siehe Frisch, W. (2016): Strafrecht und Solidarität – Zugleich zu Notstand und unterlassener Hilfeleistung. In: Goltdammer's Archiv für Strafrecht 163 (3), 121-137.

[105] Damit bildet der Verstoß gegen eine Norm den Anknüpfungspunkt von Verantwortung. Normen schränken die Freiheit des Einzelnen ein und bedürfen daher der Legitimation. Im Recht kann insoweit auf den Grundsatz der Verhältnismäßigkeit (i.w.S.) zurückgegriffen werden. Zu diesem Grundsatz siehe Dechsling, R. (1989): Das Verhältnismäßigkeitsgebot: Eine Bestandsaufnahme der Literatur zur Verhältnismäßigkeit staatlichen Handelns. München.

Über diese Konstellation hinaus ist in der Verantwortungsdiskussion zum wissenschaftlich-technischen Fortschritt die epistemologische Dimension zu bedenken. Denn Handlungsfolgen sind oft nur unter hohen und nicht eliminierbaren Unsicherheiten des Wissens antizipierbar.[106] Verantwortungszuschreibung muss daher die Dimension des Handelns unter Unsicherheit und damit die Risikothematik[107] berücksichtigen. Dies lässt auch noch einmal zwischen retrospektiver und prospektiver Verantwortung unterscheiden. Beim Blick auf vergangene Handlungen kann eine nachträgliche Verantwortungsübernahme unter Umständen angezeigt sein. Vor allem die prospektive Verantwortung unterliegt den gerade konstatierten Unsicherheiten.

Um Verantwortung im Zusammenspiel von Menschen und maschinellen Systemen näher zu betrachten, kann das fünfstellige Verantwortungskonzept technikspezifisch gerahmt werden. Ausgangspunkt ist hier, dass eine Verantwortungsübernahme (als *Verantwortungssubjekt*) nur Personen als verantwortlichen Wesen möglich ist, beispielsweise den Individuen, die Technik entwickeln und herstellen, die ihren Einsatz etwa in der Politik oder Unternehmen ermöglichen und fördern, oder denjenigen, die Technologien einsetzen. Das *Verantwortungsobjekt* ist dann je nach Rolle der Verantwortungssubjekte und ihrer Handlungen zu beschreiben: zum Beispiel Planen, Erfinden, Entwickeln oder Anwenden. Zu den *Betroffenen* können sowohl die von dem technischen Angebot direkt angesprochenen Personen(gruppen) gehören, zum Beispiel Angestellte in einem Krankenhaus, die mithilfe KI-gestützter Software Entscheidungen treffen, als auch weitere Personen wie zum Beispiel diejenigen, die auf Grundlage solcher Entscheidungen Diagnosen, Therapieempfehlungen oder sonstigen medizinischen Rat erhalten. Relevante *Instanzen* und relevante *Normen* sind hierbei verknüpft. Rechtliche Verantwortung besteht in letzter Instanz gegenüber der staatlich verfassten Gemeinschaft.

Moralische Verantwortung können nur natürliche Personen übernehmen, insofern sie über Handlungsfähigkeit verfügen, das heißt in der Lage sind, aktiv, zweckgerichtet und kontrolliert auf die Umwelt einzuwirken und dadurch Veränderungen zu verursachen. Träfe dies auch auf Maschinen zu, wären auch diese verantwortungsfähig. Dann müsste Maschinen der Personenstatus zugeschrieben werden, was jedoch weder aktuell noch angesichts der in absehbarer Zukunft erwartbaren qualitativen Entwicklungen maschineller Systeme angemessen wäre. Verantwortung kann daher nicht direkt von maschinellen Systemen übernommen werden, sondern nur von den Menschen, die in je unterschiedlichen Funktionen hinter diesen Systemen stehen,

---

[106] Grunwald, A. (2013): Modes of orientation provided by futures studies: making sense of diversity and divergence. In: European Journal of Futures Research 15:30 (DOI 10.1007/s40309-013-0030-5).
[107] Nida-Rümelin, J. et al. (2012): Risikoethik. Berlin.

gegebenenfalls im Rahmen institutioneller Verantwortung. Auch wenn ein technisches System eingesetzt wird, um im Rahmen einer automatisierten Datenauswertung Schlussfolgerungen wie die Gewährung eines Kredites anzuwenden, ist es die Verantwortung des Menschen, dieses System in einer ethisch vertretbaren Weise zu entwickeln und einzusetzen.[108]

Wer nun konkret als Verantwortungsträger fungiert, kann mit dem Konzept der *Multiakteurs-verantwortung* umrissen werden.[109] Kommt es bereits zu facettenreichen Verantwortungsgefügen, wenn man von nur drei prinzipiellen Ebenen möglicher Verantwortungszuschreibung ausgeht – Individuen, Organisationen und Staat –, so entsteht ein noch komplexeres Bild, wenn man Wechselwirkungen zwischen verschiedenen Akteuren aus diesen drei Ebenen berücksichtigt. Dies gilt erst recht, wenn diese Interaktionen zumindest teilweise von algorithmischen Systemen gestützt oder vermittelt werden, die mitunter für andere Beteiligte selbst autonom und kaum durchschaubar zu agieren scheinen.

Hier stellt sich die Frage, wie Verantwortung sinnvoll zwischen unterschiedlichen Beteiligten geteilt werden kann, zum Beispiel zwischen denjenigen, die maschinelle Systeme konzipieren und entwickeln, die ihre Nutzung beauftragen oder vorantreiben, die in Nutzungsprozesse oder ihre Überwachung direkt eingebunden sind, die Ergebnisse solcher Prozesse verwenden oder von ihnen direkt oder indirekt betroffen sind oder die ihre Auswirkungen auf unterschiedlichen gesellschaftlichen Ebenen beobachten und eventuell regulierend eingreifen können. In Anlehnung an das Konzept der Datensouveränität[110] ist die geeignete Gestaltung von Multiakteursverantwortung demnach zentral für eine angemessene informationelle Freiheitsgestaltung, die den Chancen und Risiken einer zunehmend digital vernetzten und algorithmisch gestützten Welt gerecht wird. Eine solche Freiheitsgestaltung kann nur dann verantwortlich sein, wenn sie sich gleichzeitig an den gesellschaftlichen Anforderungen von Solidarität und Gerechtigkeit orientiert.

Auch für die Diskursführung über die zahlreichen Wechselwirkungen von Menschen und Maschinen und die gesellschaftlichen Auswirkungen einer zunehmenden Durchdringung der menschlichen Gesellschaft mit algorithmischen Systemen muss Verantwortung übernommen

---

[108] Datenethikkommission der Bundesregierung (2019): Gutachten der Datenethikkommission. Berlin. https://www.bmi.bund.de/SharedDocs/downloads/DE/publikationen/themen/it-digitalpolitik/gutachten-datenethikkommission.pdf;jsessionid=4A012DD4E717D4E3FA62DD51238229C3.1_cid295?__blob=publicationFile&v=7 [10.02.2023].
[109] Deutscher Ethikrat (2017): Big Data und Gesundheit – Datensouveränität als informationelle Freiheitsgestaltung. Berlin, 249 f.
[110] Deutscher Ethikrat (2017): Big Data und Gesundheit – Datensouveränität als informationelle Freiheitsgestaltung. Berlin, 252 f.

werden. Warnungen vor unkritischem Vertrauen in maschinelle Systeme, insbesondere im Falle Künstlicher Intelligenz, sollten einen Platz haben und sind Ausdruck wahrgenommener Verantwortung. Ebenso sind die Auswahl und Gewichtung bestimmter normativer Kriterien und Prinzipien im Diskurs zur Ethik von maschinellen Systemen, Algorithmen und KI selbst Gegenstand von Kontroversen.[111] Wer hat die Deutungshoheit, Werte und Normen, die im Umgang mit KI relevant sind, zu bestimmen? Wer prüft, welche Betroffenen vornehmlich in den Blick genommen werden oder wie die Lasten und Nutzen bestimmter Anwendungen verteilt sind? Es stellt sich also allgemein die Frage, wer für das fünfte Relationselement, die Bestimmung normativer präskriptiver Prinzipien, zuständig ist.

3.4 Anthropologische Aspekte des Mensch-Maschine-Verhältnisses

3.4.1 Philosophische Grundbestimmung des Menschseins

Handlung, Vernunft und Verantwortung stehen im Zentrum humanistischer[112] Philosophie. Menschen sind befähigt zur Handlungsurheberschaft und somit zur Autorschaft ihres Lebens. Sie sind frei und tragen daher Verantwortung für die Gestaltung ihres Handelns. Freiheit und Verantwortung sind zwei einander wechselseitig bedingende Aspekte menschlicher Autorschaft. Autorschaft ist wiederum an Vernunftfähigkeit gebunden. Die strafrechtlichen Kriterien für Schuldfähigkeit konvergieren mit der lebensweltlichen Praxis moralischer Zuschreibungen. Personen sind jedenfalls in wichtigen sozialen Kontexten moralisch verantwortlich. Das heißt, man erwartet von ihnen, dass sie zurechnungsfähig handeln und urteilen.[113]

Diese Trias aus Vernunft, Freiheit und Verantwortung prägt heute sowohl die lebensweltliche Moral als auch die Rechtsordnung in hohem Maße. Im Mittelpunkt steht dabei das Phänomen

[111] Jobin, A. et al. (2019): The global landscape of AI ethics guidelines. In Nature Machine Intelligence 1, 389-399 (DOI: 10.1038/s42256-019-0088-2); Rudschies, C. et al. (2021): Value Pluralism in the AI Ethics Debate – Different Actors, Different Priorities. In: International Review of Information Ethics 29. https://informationethics.ca/index.php/irie/article/view/419/396 [04.01.2023].

[112] Die unterschiedlichen Verwendungsweisen des Humanismusbegriffs in der Philosophie stimmen in einigen normativen Kernelementen überein, die zur Klärung anthropologischer Aspekte des Mensch-Maschine-Verhältnisses wichtig sind und nachfolgend näher entfaltet werden.

[113] In der Technikanthropologie, die sich mit anthropologischen Aspekten des Technischen und der Technik befasst, vor allem mit Mensch/Technik- oder Mensch/Maschine-Konstellationen, wird eine Vielzahl auch anderer Perspektiven verfolgt (Heßler, M.; Liggieri, K. (2020): Technikanthropologie. Handbuch für Wissenschaft und Studium. Baden-Baden). Hierzu gehören etwa homo faber und homo creator, trans- und posthumanistische Positionen sowie die Akteur-Netzwerk-Theorie. Die enge Verbindung ethischer Fragen zu den Konzepten von Freiheit und Verantwortung impliziert, diese in den Betrachtungen nicht eigens zu berücksichtigen, sondern die humanistische Perspektive in die Mitte zu stellen.

der Affektion durch Gründe. Praktische Gründe sprechen *für* Handlungen, sie sind per se normativ, nicht erst über den Umweg eigener Wünsche. Ein Grund spricht dafür, das zu tun, was diesen Grund erfüllt, wenn nicht andere Gründe dem entgegenstehen.[114] Theoretische Gründe sprechen *für* Überzeugungen; auch diese sind normativ. In der Regel gibt es Gründe das eine zu tun und das andere zu lassen, die gegeneinander abgewogen werden müssen. Der Konflikt von Gründen zwingt dann zur Abwägung und zur Systematisierung dieser Abwägung in Gestalt ethischer Theoriebildung.

Die menschliche Lebensform ist von reaktiven Einstellungen und moralischen Gefühlen geprägt, die von normativen Gründen begleitet sind. Wir vergeben einer Person, die uns Unrecht getan hat, wenn wir den Eindruck haben, sie habe das Unrechte ihres Tuns eingesehen und werde diese Praxis nicht fortsetzen. Wir sind dankbar, wenn wir meinen, dass eine Person etwas Gutes getan hat, ohne daraus Vorteile zu ziehen, wir nehmen etwas übel, nur dann, wenn wir die betreffende Person für voll zurechnungsfähig und in ihrem Handeln frei halten.[115] Die verobjektivierende Einstellung gegenüber anderen Menschen, die diese zum bloßen Gegenstand der Beeinflussung macht, sie gewissermaßen zu einem Teil der Umwelt degradiert, lässt sich nur für ganz spezifische Situationen – wenn überhaupt – durchhalten. Aber wenn diese verobjektivierende Einstellung ohne moralische Empfindungen zur allgemeinen Praxis würde, gäbe es die menschliche Lebensform nicht mehr. Diese ist gerade dadurch geprägt, dass wir ein Verhalten übelnehmen, wenn es uns inakzeptabel erscheint, dass wir zum Beispiel in der Lage sind zu verzeihen, wenn wir dafür Gründe haben, oder, dass wir Dankbarkeit empfinden.

Freiheit kommt insofern ins Spiel, als wir rationaliter bestimmte moralische Gefühle und reaktive Einstellungen zurückstellen, wenn wir erfahren, dass die betreffende Person in ihrem Handeln nicht frei war, was immer die Ursachen dieser Unfreiheit sind, wie beispielsweise äußerer Zwang, psychische Erkrankung oder überwältigende Angst. Diese Praxis der Zuschreibung von Freiheit und Verantwortung ist essenziell für die Grundlegung moralischer Beurteilung wie auch für moralische Gefühle und reaktive Einstellungen, und daher ist es ausgeschlossen, diese aufzugeben, wie überzeugend auch immer wissenschaftliche Theorien, die prima facie dagegen sprechen, sein mögen.[116] Es macht unsere Zugehörigkeit zur menschlichen Lebensform aus,

---

[114] Scanlon, T.M (1998): What We Owe to Each Other. Cambridge. Scanlon, T.M (2014): Being Realistic about Reasons, Oxford. Halbig, C. (2007): Praktische Gründe und die Realität der Moral. Frankfurt a. M.
[115] Vgl. den einflussreichen Ansatz von Strawson, P. F. (1964): Freedom and Resentment and other Essays. London.
[116] Siehe beispielsweise die Debatte rund um die Experimente von Libet. Libet, B. (2004): Mind Time. The Temporal Factor in Human Consciousness. Cambridge (MA), London.

dass solche moralischen Beurteilungen, Gefühle und Einstellungen unsere soziale Praxis prägen. Die Normen von Moral und Recht sind ohne die Annahme menschlicher Verantwortlichkeit und damit Freiheitsfähigkeit und Vernunftfähigkeit unbegründet. Sie würden in bloße Instrumente der Verhaltenssteuerung transformiert[117] und paradoxerweise wäre es gerade diese Transformation, die ihre Wirksamkeit für die Verhaltenssteuerung zugleich gefährden würde.

Wenn menschliche Freiheit im Sinne des Anderskönnens bestritten wird, kann an der Verantwortlichkeit menschlicher Personen nicht festgehalten werden.[118] In § 20 StGB werden die praktischen und epistemischen Bedingungen von Schuldfähigkeit dargestellt, die ohne diese anthropologischen Voraussetzungen von Freiheit und Vernunft nicht aufrechtzuerhalten wären. Ohne die Fähigkeit, sich anders entscheiden zu können, gibt es keine Handlungsurheberschaft und keine Verantwortung.

Obwohl die humanistische Perspektive nicht nur die lebensweltliche Normativität, sondern auch die juridische Deliberation von den Menschenrechten bis zum Strafrecht imprägniert und das kulturelle Fundament demokratischer Ordnungen ausmacht, wird sie doch immer wieder infrage gestellt. Zwei jüngere Formen der Kritik, die sich teilweise überlagern, stützen sich einerseits auf neurowissenschaftliche Begriffe und Paradigmen sowie andererseits auf solche aus Debatten um das Thema Künstliche Intelligenz. In den Neurowissenschaften wurden bestimmte empirische Studien, nach denen zum Beispiel das motorische Zentrum des Gehirns schon mit der Vorbereitung einer Bewegung beginnt, bevor man sich bewusst für die Ausführung der Bewegung entschieden hat[119], als Beleg dafür interpretiert, dass es Freiheit und damit menschliche Verantwortlichkeit nicht gebe. Stattdessen handele es sich dabei lediglich um – möglicherweise sozial nützliche – Illusionen. Tatsächlich lassen solche Befunde jedoch unterschiedliche Interpretationen zu und eignen sich nicht als Widerlegung menschlicher Freiheit und Verantwortlichkeit. Auch wenn alle mentalen Prozesse neurophysiologisch realisiert sind, sprechen die empirischen Befunde aus den Neurowissenschaften nicht gegen das Phänomen

---

[117] Vgl. dazu Schlick, M. (1930): Fragen der Ethik. Wien. der exemplarisch für diese anti-humanistische Auffassung steht.

[118] Es gibt allerdings in der zeitgenössischen Philosophie auch die Auffassung, Verantwortlichkeit könnte auch ohne die Bedingung der Freiheit postuliert werden, besonders prominent bei Harry Frankfurt. Kane, R. (Hg.) (2011): The Oxford Handbook of Free Will. 2. Auflage. Oxford (DOI: 10.1093/oxfordhb/9780195399691.001.0001), hierbei das Kapitel 5 „Moral Responsibility, Alternative Possibilities, And Frankfurt-Type Examples“. Siehe auch Frankfurt, H. (1971): Freedom of the Will and the Concept of a Person. In: The Journal of Philosophy 68 (1), 5-20 (DOI: 10.2307/2024717).

[119] Libet, B. (2004): Mind Time. The Temporal Factor in Human Consciousness. Cambridge (MA), London.

normativer Gründe und ihrer Rolle für menschliche Handlungsmotivation, da wir es hier mit zwei Sprachebenen zu tun haben, die sich wechselseitig nicht in die Quere kommen können.[120]

Die zweite, von der KI-Debatte inspirierte Kritik der humanistischen Anthropologie changiert zwischen einer Überwindung des Menschen in Gestalt des Transhumanismus, der mit neuen Mensch-Maschinen-Symbiosen die Reichweite menschlichen Wirkens in neue Dimensionen heben möchte und einem Maschinenparadigma, das den menschlichen Geist auf das Modell eines algorithmischen Systems reduziert. Gerade Letzteres entfaltet besondere Relevanz im Kontext dieser Stellungnahme, da es großen Einfluss auf die Interpretation der Wechselwirkungen zwischen Mensch und Maschine und deren Rückwirkungen auf das menschliche Selbstverständnis hat.

3.4.2 Der Mensch als Maschine – die Maschine als Mensch?

Der Mensch als Maschine ist eine alte Metapher, deren Ursprünge bis in die Frühe Neuzeit zurückreichen. Der mechanistische Materialismus des rationalistischen Zeitalters lässt die Welt als Uhrwerk erscheinen und den Menschen als Rädchen im Getriebe. Der große Uhrmacher ist dann der Schöpfer, der dafür gesorgt hat, dass nichts dem Zufall überlassen ist und ein Rädchen ins andere greift. Für menschliche Freiheit, Verantwortung und Vernunft ist in diesem Bild kein Platz.

Die vielleicht aktuell größte Herausforderung für das humanistische Menschenbild stellt das digital erneuerte Maschinenparadigma des Menschen dar. Das digitale Weltmodell, die Welt als umfassender Algorithmus[121], scheint als zeitgenössische Variante des Maschinenparadigmas einer Deutung der Welt als Maschine eine attraktive Interpretation anzubieten. Diese beruht auf der Unterscheidung zwischen Software und Hardware. Es handelt sich dabei um zwei Beschreibungsebenen: die der Hardware, die lediglich auf physikalische und technische Begriffe zurückgreifen muss, und die der Software, die sich wiederum in eine syntaktische und eine semantische Ebene aufteilen lässt. Die syntaktische Beschreibung beruht auf der Zeichenverarbeitung, genauer dem Vokabular und den Regeln der Zeichenverarbeitung. Die Semantik unterlegt den Zeichen und den Regeln, nach denen sie verarbeitet werden, eine Bedeutung. Im Falle von Behauptungssätzen führt diese Unterlegung zu einer Wahrheitswertverteilung; die

[120] Strawson, P.F. (1974): Freedom and Resentment and Other Essays, London. Nida-Rümelin, J. (2005): Über menschliche Freiheit. Stuttgart. Korsgaard, C.M. (1992): The Sources of Normativity. Cambridge.
[121] Nida-Rümelin, J.; Weidenfeld, N. (2018): Digitaler Humanismus. München.

Sätze werden dann als wahr oder falsch markiert; im Falle einer arithmetischen Semantik folgt die Wahrheitswertverteilung mathematischen Regeln.

Die Beschreibung (und Erklärung) von Softwaresystemen als Hardware ist geschlossen: Jeder Vorgang (Ereignis, Prozess, Zustand) lässt sich als kausal determiniert durch den vorausgegangenen Zustand der Hardware eindeutig bestimmen. Als Modell auf den Menschen übertragen heißt dies, dass die physikalisch-physiologische „Hardware“ wie ein algorithmisches System mit einer durch Genetik, Epigenetik und sensorische Stimuli eindeutig festgelegten zeitlichen Zustandsfolge von Eigenschaften funktioniert, die durch mentale Termini beschrieben wird und damit bedeutungsvolles Reden und Handeln ermöglicht. Das humanistische Menschenbild und damit die normativen Grundlagen von Moral und Recht erweist sich dann als pure Illusion bzw. kollektive menschliche Selbsttäuschung.[122]

Schon in der ersten Digitalisierungswelle nach dem Zweiten Weltkrieg erwies sich interessanterweise nicht das eben geschilderte materialistische Maschinenparadigma, sondern eine animistische Variante als wirkungsmächtiger. Das animistische Paradigma geht gewissermaßen den umgekehrten Weg der Interpretation: Anstatt den menschlichen Geist und mentale Zustände als Epiphänomene materieller Prozesse in einer physikalisch geschlossenen Welt zu interpretieren und das Materielle mechanistisch zu beschreiben, wird nun im Kontext des Turing-Tests (vgl. Abschnitt 2.1) das algorithmische System mit mentalen Eigenschaften ausgestattet, sofern es in seinem äußeren (Ausgabe-) Verhalten demjenigen von Menschen hinreichend (das heißt verwechselbar) ähnelt.

Entsprechend war in der ersten Phase der Diskussion um Künstliche Intelligenz ab den Siebzigerjahren des 20. Jahrhunderts die Frage, ob Computer denken können, leitend. Für die Fragerichtung ist die kontroverse Diskussion um die Interpretation des von Turing vorgeschlagenen Kriteriums paradigmatisch.[123] Wie zuvor bereits dargelegt, kann nach Turing die Frage, ob technische Artefakte „denken“ können, dadurch entschieden werden, dass eine Person (für sie verdeckten) Menschen und Geräten beliebige Fragen stellt. Wenn in einer größeren Zahl von Durchgängen mit wechselnden Fragenden und wechselnden Menschen/Geräten die Antworten zu einem hinreichend großen Anteil (z. B. 50%) nicht eindeutig Mensch oder Gerät zugeordnet

---

[122] Singer, W. (2004): Verschaltungen legen uns fest. Wir sollten aufhören, von Freiheit zu sprechen. In: Geyer, C. (Hg.): Hirnforschung und Willensfreiheit. Zur Deutung der neusten Experimente. Berlin, 30-65. Dennett, D. et al. (2007): Neuroscience and Philosophy: Brain, Mind, and Language. New York.

[123] Turing, A. M. (1950): Computing Machinery and Intelligence. In: Mind LIX (236), 433-460 (DOI: 10.1093/mind/LIX.236.433); vgl. die kritische Darstellung bei Mainzer, K. (1995): Computer – Neue Flügel des Geistes? Berlin, 113 f.

werden können, gibt es nach Turing keinen Grund, technischen Artefakten weniger Denkvermögen zuzuschreiben als Menschen.

Die in Diskursen rund um KI teilweise verbreitete Tendenz, im Anschluss an den Turing-Test eine äußerliche Ununterscheidbarkeit von menschlicher und maschineller Performanz pauschal mit der Annahme von Intelligenz und Denkvermögen solcher Maschinen gleichzusetzen, auf diese Weise die Differenz zwischen dem Simulierten und dem Simulierenden einzuebnen und die menschliche Vernunft damit tendenziell für maschinell ersetzbar zu halten, ist kein Zufall, sondern das Ergebnis bestimmter theoretischer Vorannahmen insbesondere *behavioristischer* und *funktionalistischer* Art.[124] Schon der klassische Behaviorismus[125] hatte sich zu Beginn des 20. Jahrhunderts in dem Bemühen, das menschliche Verhalten auf der Grundlage präzise beschreibbarer Reiz-Reaktion-Schemata zu erklären und die Psychologie damit in eine exakte Wissenschaft zu verwandeln, im Grunde einer Black-Box-Methode bedient, die das Innenleben derart beschriebener Organismen komplett ausblendet.

Der Text von Alan Turing ist zweifellos vom Logischen Behaviorismus inspiriert, der in den Nachkriegsjahren die zeitgenössischen Debatten insbesondere in der britischen Philosophie zunehmend prägte[126]und nach dem sich mentale Zustände ontologisch auf Verhaltensdispositionen reduzieren lassen, also auf die Neigung eines Organismus, sich unter bestimmten Bedingungen auf eine bestimmte Weise zu verhalten. Ein mentaler Zustand wie Schmerz ist demnach lediglich ein Verhaltensmuster, etwa die Veranlagung zu schreien oder zu weinen, wenn man sich verletzt hat. Auch Turings Text identifiziert den Sinn eines sprachlichen Ausdrucks nicht etwa mit der Intention der sprechenden Person, sondern mit den empirischen Verhaltensmustern, die mit einer Äußerung üblicherweise einhergehen. Die Paradoxa, die den Logischen Behaviorismus unglaubwürdig machen, gelten auch für die Turing'sche Variante: Auch wenn wir die Bedeutung eines Satzes, wie „Diese Person hat Schmerzen" lernen, indem wir darauf achten, welches Verhalten jeweils darauf hinweist, dass sie Schmerzen hat, so kann schon deshalb die

[124] Bormann, F.-J. (2021): Ist die praktische Vernunft des Menschen durch KI-Systeme ersetzbar? Zum unterschiedlichen Status von menschlichen Personen und (selbst-)lernenden Maschinen. In: Fritz, A. et al. (Hg.): Digitalisierung im Gesundheitswesen. Anthropologische und ethische Herausforderungen er Mensch-Maschine-Interaktion. Freiburg, 41-64, 51 ff.
[125] Watson, J. B. (1925): Der Behaviorismus. New York.
[126] Wittgenstein, L. (1953): Philosophische Untersuchungen. Oxford, Malden. Ryle, G. (1949)*:* The Concept of Mind. London. Einen guten Zugang zur damaligen *ordinary language philosophy* vermittelt Savigny, E (1973): Zur Philosophie der normalen Sprache. Frankfurt a. M.

Bedeutung von „Schmerzen haben“ nicht lediglich ein Verhaltensmuster sein, weil „Superspartaner“, die keine Schmerzen zeigen, dann auch keine Schmerzen haben könnten.[127] Obwohl sich der behavioristische Theorieansatz schon bald als zu eng erweisen sollte und in den folgenden Jahrzehnten verschiedene Transformationen erfuhr, fand er mit dem sich seit den 1950er-Jahren ausbreitenden Funktionalismus, nach dem mentale Zustände funktional vollständig erfasst werden, eine Nachfolgetheorie, die der aufkeimenden Kognitionspsychologie und Computerwissenschaft noch besser entsprach, weil Computer in dieser Lesart ein geeignetes Modell für mentale Prozesse sein können.[128]

Der Reiz funktionalistischer Ansätze besteht zunächst darin, dass sie sich gegenüber den ontologischen Implikationen des Leib-Seele-Problems neutral verhalten, das heißt sie umgehen die Frage nach der Beziehung zwischen Körper und Geist sowie die Frage, wo und wie sich in diesen das denkende und fühlende Subjekt verorten lässt. Der Funktionalismus plädiert dafür, die Frage nach der Seinsart mentaler Zustände zugunsten der genauen Beschreibung ihrer Funktion aufzuheben. Durch die These der *multiplen Realisierung*[129], nach der bestimmte mentale Ereignisse, Eigenschaften oder Zustände durch ganz unterschiedliche physikalische Ereignisse, Eigenschaften oder Zustände realisiert werden können, schien es zudem möglich, auch Computern mentale Zustände zuzuschreiben, obwohl sie keine biologischen Strukturen besitzen.[130] Indem der Funktionalismus durch die ausdrückliche Anerkennung theorierelevanter innerer Zustände eines Systems nicht nur das Blackbox-Prinzip des Behaviorismus überwand, sondern mit der funktionalen Interpretation solcher Zustände auch die Integration biologischer und maschineller Entitäten in eine umfassende einheitliche Theorie zu ermöglichen schien, bahnte er insofern auch einer animistischen Deutung von KI-Systemen den Weg, als diese nun umso

[127] Putnam, H. (1965): Brains and Behaviour. In: Butler, R. J. (Hg.): Analytical Philosophy, Band 2. Oxford, 24-36.
[129] Vgl. Putnam, H. (1980): Philosophy and Our Mental Life. In: Block, N. (Hg.): Readings in Philosophy of Psychology, Vol. I § 7. Cambridge (MA), London, 134-143 (DOI: 10.4159/harvard.9780674594623.c11). Von dierser Position distanziert sich Putnam zu späterer Zeit, siehe dazu: Putnam, H. (1992): Renewing Philosophy. Cambridge (MA).
[129] Für die unterschiedlichen Entwicklungsstufen dieser These vgl. Bickle, J. (1998):Multiple Realizability. In: Stanford Encyclopedia of Philosophy, Summer 2020 Edition.
https://plato.stanford.edu/archives/sum2020/entries/multiple-realizability/ [04.01.2023]. Sowie Polger, T.; Shapiro, L. (2016): The Multiple Realization Book. New York.
[130] Putnam, H. (1967): Psychological Predicates. In: Capitan, W. H.; Merrill, D. D. (1967): Art, Mind, and Religion. Pittsburgh, 37–48. (dt. Übersetzung: Die Natur mentaler Zustände, In: Metzinger, T. (2007): Grundkurs Philosophie des Geistes 2: Das Leib-Seele-Problem. Paderborn, 372–385) sowie Fodor, J. (1974): Special Sciences (or: The disunity of science as a working hypothesis). In: Synthese 28 (2), 97–115 (DOI: 10.1007/BF00485230).

leichter als handlungsfähige Agenten mit einem mentalen Innenleben vorgestellt werden konnten, denen man zutraute, die menschliche Vernunft irgendwann einmal ersetzen zu können.

Obwohl der Funktionalismus aufgrund dieser Vorteile zunächst viel Zuspruch sowohl in der analytischen Philosophie des Geistes als auch in der KI-Forschung fand, wurden schon bald Einwände gegen dieses Theoriemodell vorgetragen. Eine erste gewichtige Kritik einer funktionalistischen Betrachtung des Mentalen, die den menschlichen Geist als Rechenmaschine[131] fasst, deren innere Zustände allein von ihrer Funktion im Sinne einer kausalen Verknüpfung von Eingabe und Ausgabe bestimmt werden, stammt von Thomas Nagel. Seines Erachtens lässt unsere gewöhnliche Auffassung mentaler Phänomene eine solche reduktionistische Sichtweise allein schon deswegen nicht zu, weil das Mentale neben seiner bloßen Funktion auch durch ein bestimmtes *phänomenales Bewusstsein*[132] geprägt sei, das in dieser Beschreibung verloren gehe. Nagel leugnet weder, „daß bewußte mentale Zustände und Ereignisse Verhalten verursachen, noch, daß man sie funktional charakterisieren könnte", sondern bestreitet lediglich, „daß derartiges eine vollständige Analyse ergibt".[133]

Ein Wesen kann nur *als* dieses Wesen mentale Zustände haben. Daher kann von den *Empfindungsqualitäten* des phänomenalen Bewusstseins nicht abgesehen werden, die allein aufgrund *äußeren* Verhaltens nicht zugänglich sind. Es fühlt sich für uns immer auf eine ganz bestimmte Art und Weise an, ein erlebendes, denkendes und handelndes Subjekt zu sein. Diese besonderen Qualitäten sind kein flüchtiges Beiwerk mentaler Zustände, sie gehören vielmehr insofern konstitutiv zu allen unseren Erfahrungen, als wir die uns umgebende Welt prinzipiell gar nicht anders erleben können als aus der Perspektive eines solchen phänomenalen Bewusstseins.

Dieses phänomenale Bewusstsein setzt dem Vermögen, die Qualität des Erlebens oder die mentalen Zustände anderer Lebewesen zu beurteilen, gewisse Grenzen. Dies illustriert Nagel am Beispiel des Empfindens einer Fledermaus, deren Orientierung aufgrund spezifischer sensorischer Besonderheiten (wie Radar oder Echolotortung) ganz anders strukturiert ist als beim Menschen. Als Menschen können wir zwar versuchen, uns vorstellen, wie es ist, sich auf gänzlich andere Weise im Raum zu orientieren, wir bleiben dabei aber immer unserer eigenen, spezifisch

---

[131] oder – um eine Kritik von Ned Block aufzugreifen – wie ein Getränkeautomat: Block, N. (1979): Troubles with Functionalism. In: Minnesota Studies in the Philosophy of Science 9, 261–325.
[132] Zur Debatte um die Bedeutung des phänomenalen Bewusstseins vergleiche auch die Textsammlungen von Heckmann, H.-D.; Walter, S. (2006*)*: Qualia. Ausgewählte Beiträge. Paderborn und Metzinger, T. (2009): Grundkurs Philosophie des Geistes 1: Phänomenales Bewusstsein. Paderborn.
[133] Nagel, T. (1981): Wie ist es, eine Fledermaus zu sein? In: Bieri, P. (Hg.): Analytische Philosophie des Geistes. Königstein im Taunus, 261–275, 262.

menschlichen Weise des Erlebens verhaftet, ohne jemals Zugang zu den besonderen Qualitäten der mentalen Zustände einer Fledermaus zu erhalten. Unsere subjektive Perspektive bleibt unüberwindlich.[134] Vor diesem Hintergrund dieses auch als Qualia-Argument[135] bezeichneten Gedankengangs basiert die funktionalistisch inspirierte Mensch-Computer-Analogie auf einer fragwürdigen Reduktion, die nur gelingen kann, „wenn die artspezifische Betrachtungsweise von dem, was reduziert werden soll, ausgeklammert wird".[136]

Gegen eine funktionalistische Interpretation Künstlicher Intelligenz hat der Philosoph John Searle ein Argument entwickelt, das als das meistdiskutierte in der zeitgenössischen Philosophie gilt: *The Chinese room*. Es verweist auf ein Gedankenexperiment, in dem eine Person in einem Zimmer sitzt, in das durch einen Schlitz jeweils Fragen in chinesischer Schrift gereicht werden. Die Person reicht Antworten auf diese Fragen ebenfalls durch den Schlitz heraus. Wenn die Antworten hinreichend plausibel erscheinen, mag man vermuten, dass die Person im chinesischen Zimmer des Chinesischen mächtig ist. Nun stellt sich aber heraus, dass jede der eingereichten Fragen eine Ziffer trägt und die Person über vorgefertigte Antworten und eine Tabelle mit Zuordnungen verfügt, sodass sie lediglich eine Antwort heraussuchen muss, die eine Ziffer trägt, die der Ziffer der Fragestellung zugeordnet ist.

Die Person beherrscht die chinesische Sprache nicht, und auch das Zimmer als Ganzes mit einer Eingabe- und Ausgabefunktion beherrscht diese Sprache nicht. Aber zweifellos muss es irgendjemanden geben, der des Chinesischen mächtig ist und daher in der Lage war, den Fragen mithilfe der Ziffern Antworten so zuzuordnen, sodass der Eindruck entsteht, dass die Person im chinesischen Zimmer Chinesisch versteht. Die Analogien zu Softwaresystemen liegen auf der Hand. Es handelt sich um Zuordnungsregeln (genauer um Algorithmen), die lediglich für die Programmierung und den Gebrauch der digitalen Maschine Bedeutung haben, aber nicht das Softwaresystem selbst zu einer semantischen Maschine machen. Dieses verfügt nicht über Bedeutungen, es versteht nichts, es entscheidet nichts. Softwaresysteme verfügen nicht über eine Semantik, es handelt sich nicht um semantische Maschinen.[137]

---

[134] Nagel, T. (1981): Wie ist es, eine Fledermaus zu sein? In: Bieri, P.(Hg.): Analytische Philosophie des Geistes. Königstein im Taunus, 261–275, 262.
[135] von lat. *qualis* „wie beschaffen"
[136] Nagel, T. (1981): Wie ist es, eine Fledermaus zu sein? In: Bieri, P. (Hg.): Analytische Philosophie des Geistes. Königstein im Taunus, 261–275, 269.
[137] Searle, J. R. (1980): Minds, Brains and Programs. In: Behavioral and Brain Sciences 3 (3), 417-457. Eine frühere Zurückweisung des sogenannten Funktionalismus stammt von Block, N. (1978): Troubles with Functionalism. In Savage, C. W. (Hg.): Perception and Cognition. Minneapolis, 9-261.

3.4.3 Verleiblichte Vernunft

Die Zurückweisung funktionalistischer Maschinenparadigmen lenkt den Blick auf die bereits im Qualia-Argument angedeutete besondere Qualität menschlicher Vernunft und deren Bedeutung für das menschliche Selbstverständnis. Menschliche Vernunft ist leibliche Vernunft. Diese Einsicht wendet sich gegen den in der abendländischen Tradition lange herrschenden Gedanken eines Dualismus zwischen Vernunft und Natur, Körper und Geist, der den Menschen als Naturwesen auf der einen und als Vernunftwesen auf der anderen Seite begreift. Solch dualistische Vorstellungen wurden sowohl durch Erkenntnisse in der Evolutionsbiologie als auch in der Hirnforschung und in den Kognitionswissenschaften infrage gestellt, die stattdessen auf die Relevanz des Leiblichen für die Bestimmung menschlicher Intelligenz und auf die Bedeutung unbewusster Prozesse für die Entwicklung höherer geistiger Leistungen verweisen.

Mit Maurice Merleau-Pontys Unterscheidung von zweierlei Arten des Körpers kann dies veranschaulicht werden.[138] Er unterscheidet den lebendigen handelnden *Leib* vom rein physikalischen *Körper*. Die Fähigkeit, soziale Bindungen einzugehen, sich in andere hineinzuversetzen, wird ermöglicht dadurch, dass der Mensch Leib ist und nicht nur einen Körper hat. Mit dem Leib ist der empfindende Organismus gemeint, mit seinem Vermögen, zu fühlen und sich zu bewegen. Wir sind an diesen Leib gebunden, in allem, was wir denken und tun. Er ist daher Ausgangspunkt und Bestandteil jeder Wahrnehmung und Empfindung. Als solcher ist er Voraussetzung für unser In-der-Welt-Sein und zugleich dafür, eine Welt zu haben und eine Beziehung zu anderen herzustellen.

Kognitive Fähigkeiten sind in ihrem Entstehungs- und Vollzugsprozess also an Sinnlichkeit und Leiblichkeit gebunden. Dies hat Konsequenzen für unser Verständnis vom Gehirn als Erkenntnisorgan und von der Vernunft als Erkenntnisvermögen – und damit auch für das Verständnis unseres Zugangs zur Realität. Wesentlich ist dabei, dass das in der Verkörperung des Gehirns eingeschlossene Naturverhältnis einer leiblichen Vernunft des Menschen seine Sozialität impliziert und seine Kulturalität bestimmt. Im menschlichen Leib sind Sozialität und Kulturalität von Anfang an angelegt, vor aller Entwicklung eines reflexiven und sprachlich vermittelten Bewusstseins.[139] Denn leibliche Vernunft vollzieht nicht nur einen kognitiven

---

[138] Vgl. dazu Merleau-Ponty, M. (1966): Phänomenologie der Wahrnehmung. Übersetzt und mit einem Vorwort von Boehm, R. Berlin.

[139] Vgl. hierzu Fuchs, T. (2013): Verkörperung, Sozialität und Kultur. In: Breyer, T. et al. (Hg.): Interdisziplinäre Anthropologie, Leib – Geist – Kultur. Heidelberg, 9-261, 11 ff. Vier Erscheinungsformen des Leibes können diese Anlage der Vermittlung zwischen der Natur- und der Kulturseite des Menschen plausibel machen: (i) ein mit der Umwelt vertrauter Leib, der sich vor allem im Umgang mit kulturellen Gegenständen entwickelt, (ii) ein

Informationsaustausch, sondern mit ihr spielen auch Kommunikation und Kooperation eine Rolle.[140] Beides sind Faktoren, die von Kindheit an entscheidend sind für jene bewussten Prozesse, in denen sich die Kulturfähigkeit bildet, die dem Menschen als sozusagen „zweite Natur“[141] zuwächst. Für das menschliche Gehirn bedeutet das, dass es mit all seinen sich entfaltenden Fähigkeiten von Anfang an in biologisch-organische wie in sozial-kulturelle Entwicklungsprozesse eingebunden ist.

Ein solches Verständnis des Menschen geht nicht von einer blutleeren „reinen“ Vernunft aus, sondern begreift auch die Vernunft als immer schon leiblich eingebunden und sozial wirksam. Damit ist die Frage nach dem Praktischwerden der Vernunft, das heißt nach den normativen Orientierungen und nach der Motivation moralischen Handelns, keine zweite Frage, sondern sie begleitet alles Denken, das als solches Lebensgestaltung nicht nur ermöglicht, sondern immer bereits vollzieht.

Bereits die praktische Einsicht, dass bestimmte normative Gründe für eine Handlung sprechen, wird damit *handlungswirksam*. Damit wäre der entscheidende Aspekt einer umfassenden Theorie praktischer Vernunft, nämlich die Frage, wie sich moralische Überzeugungen in Handlungen überführen lassen, berührt. Bezüglich dieser Frage plädiert der Philosoph John McDowell dafür, in diesem Kontext nicht entweder die Vernunft oder aber die subjektiven Einstellungen und Strebungen zum alles beherrschenden Faktor der Handlungsverursachung zu stilisieren, sondern zu akzeptieren, dass in der menschlichen moralischen Erfahrung diese beiden (das heißt die kognitive und die appetitive Dimension) immer schon unauflöslich miteinander verschränkt seien.[142]

Entscheidend ist dabei, dass leibliche Erfahrungsstrukturen einhergehen mit der Fähigkeit, sich in andere hineinversetzen und sich mitteilen zu können, das heißt mit einer Prosozialität, auf deren Grundlage sich die Fähigkeit zu geteilter Intention entwickeln und Empathie und Motivation initiieren kann.

Insofern nun das Gehirn kein isolierter Gegenstand ist, sondern eingelassen ist in Erfahrungen gemeinsamer Praxis, in der sich körperlich-biologisches und kulturell-soziologisches Erleben

---

„passiv-affizierbarer“ Leib der affektiv mit anderen verbunden ist, (iii) ein „mimetisch-resonanter“ Leib, der durch Nachahmung in grundlegende Kommunikation mit anderen eingebunden ist, bis er so als (iv) kooperativ kultivierter Leib zum Körper für andere wird, indem er Haltungen und Rollen übernimmt, die ihm somit zur „zweiten Natur“ werden. Vgl. ebd. 26f.

[140] Schmitz, H. (1990): Der unerschöpfliche Gegenstand. Grundzüge der Philosophie. Bonn.

[141] McDowell, J. (2001): Geist und Welt. Frankfurt a. M., 109. McDowell, J. (2009): Zwei Arten von Naturalismus. In: ebd.: Wert und Wirklichkeit. Aufsätze zur Moralphilosophie. Frankfurt a. M., 30-73.

[142] McDowell, J (2002): Interne und externe Gründe. In: ebd.: Wert und Wirklichkeit. Aufsätze zur Moralphilosophie, Frankfurt a. M., 156-178. 177.

verschränken, entwickeln sich in solch sozialer Praxis ein Bedeutungswissen und ein Wissen um die Perspektivität von Erkenntnis, die reflexiv auf die Relationalität von Wissen und Erkenntnis verweist. Denn diese Perspektivität des Wissens erschließt sich insbesondere durch dessen In-Relation-Stehen zum eigenen Leib, durch die sogenannte „Eigenleiberfahrung". So wird mit dem Rekurs auf die verleiblichte Vernunft deutlich, dass zur menschlichen Intelligenz unabdingbar Reflexivität hinzugehört. Diese setzt menschliche Erkenntnis instand, zu unterschiedlichen Perspektiven Stellung nehmen und urteilen zu können.

Grenzen der Formalisierbarkeit und Simulierbarkeit menschlicher Vernunft

Mit der Reflexivität des Bewusstseins ist das Verstehen- und Vermittelnkönnen angesprochen, mit anderen Worten die hermeneutische Dimension, die sich auch in der Unterscheidung und Anwendung verschiedener Wissensformen darstellt und die ein besonderes Charakteristikum menschlicher Intelligenz bildet. Diese hermeneutische Dimension von Wissen ist aber nur begrenzt formalisierbar oder simulierbar und bezieht sich auf den Sinn und die Bedeutung menschlichen Erkennens und Handelns. Die Aneignung menschlicher Erfahrung ist immer mit Deutungsprozessen verbunden und setzt immer ein Beteiligtsein, ein Engagement voraus.[143] Die Art und Weise, *wie* wir wissen (*knowing how*), ist eine eigene Kompetenz, die sich nicht durch bloßes Sachwissen (*knowing that*) abbilden lässt.

Auch hier spielt der Leib eine wichtige Rolle, denn er ermöglicht ein Handeln, das allein mittels bewusster Planung und Berechnung so nicht möglich wäre. In der leiblichen Verfasstheit gründet daher auch die Nichtsimulierbarkeit des Denkens. Mit der leiblichen Verankerung des Bewusstseins, die eine Komplexität von Hintergrunderfahrungen mit sich führt, die Voraussetzung für alle bewussten Prozesse der Planung und Entscheidung sowie deren Begründung bilden, stößt die Entwicklung von Künstlicher Intelligenz an ihre Grenzen.

Der Sachverhalt solch leiblich verfassten Hintergrundwissens bedeutet mithin eine Grenze des rationalistischen Versuchs, Wissen vollständig in formalisierte Regeln zu überführen und künstlich nachzubilden. Es zeigt sich in der leiblichen Verschränkung von kognitiven und emotional appetitiven Momenten im Vernunftvollzug dann vielmehr die Relevanz des Nichtformalisierbaren. Es geht im Begreifen nicht mehr nur darum, das *Was* – die Fakten – zu begreifen,

[143] Meyer-Drawe, K. (2001): Leiblichkeit und Sozialität. 3., unveränderte Auflage. München.

sondern darum, *wie* wir verstehen. Und dies wird entscheidend durch unsere leiblichen Vollzüge und Fähigkeiten bestimmt, die vorbegrifflich und unausgesprochenen unser Verhalten mitbestimmen.

Menschliches Wissen ist insofern eingebettet in einen Horizont des Nichtwissens. Im Raum individueller und sozio-kultureller Erfahrung wird deutlich, dass nicht nur die Klarheit logischen Schließens, sondern auch die Vagheit und Offenheit menschliches Denken auszeichnet. Gerade Vagheit und Unbestimmtheit des Wissens sind Voraussetzung für Kreativität und Intuition, die ein Handeln unter der Bedingung von Ungewissheit ermöglichen, mit der Menschen situativ auf konkrete Herausforderungen reagieren und Verantwortung übernehmen können. Für menschliche Intelligenz ist kennzeichnend, dass sie sich auf plötzliche Situationen einstellen kann, um in Erlebnisgegenwart Entscheidungen für oder gegen Zukunftsszenarien zu treffen.

Wesentlich ist daher für menschliche Intelligenz und deren Verantwortungsfähigkeit auch das Erleben von und der Umgang mit Zeit. Entscheidungen werden in Gegenwart erlebt und in solchem Gegenwartserleben bewusst gehalten. Auch dieses Gegenwartserleben ist in der Leiblichkeit der Vernunft verankert. Veranschaulicht werden kann dies an der Bedeutung des menschlichen Gedächtnisses. Das menschliche Gedächtnis funktioniert nicht wie ein Speicher, der einen gedanklichen Bestand bildet und abrufbar wäre. Vielmehr ist es durch Prozesse des Erinnerns und Vergessens ausgezeichnet. Was jeweils im Moment erinnert – oder vergessen – wird, hängt von der je konkreten leiblichen Verfasstheit und den sozialen Bezügen, in denen der Mensch steht, ab. Erinnern ist damit nicht gleichbedeutend mit dem Abrufen einer Information. Es ist vielmehr ein hermeneutischer Akt, mit dem sich ein Erfahrungsraum[144] vergegenwärtigt.

3.5 Fazit

Aus den vorherigen Überlegungen lassen sich einige entscheidende Aspekte kognitiver Leistungen und Operationen von Menschen und Maschinen zusammenfassen. Das Kognitive ist im Falle menschlicher Intelligenz unauflöslich mit den kognitiven und emotiven, ästhetischen und ethischen, technischen und gestalterischen, sozialen und individuellen sowie zeitlichen Dimensionen der menschlichen Lebenswelt verbunden. Menschliche Intelligenz zeigt sich nicht nur in kognitiv kohärentem Urteil, sondern auch in einer kohärenten Praxis. Diese ist gründegeleitet und Ausdruck von akzeptierten Werten und Normen, die nicht beliebig zur Disposition stehen.

[144] Meyer-Drawe, K. (2001): Leiblichkeit und Sozialität. 3., unveränderte Auflage. München.

Der Mensch ist durch die Fähigkeit, Gründe zu geben und zu nehmen und sich im Urteil und im Handeln an diesen zu orientieren, als Spezies charakterisiert. Veränderungen im normativen Gefüge der eigenen Praxis bedürfen der Begründung und bedrohen im Grenzfall die persönliche Integrität und Identität. Ein hinreichend entwickeltes lebensweltliches Orientierungswissen ist Voraussetzung für eine intelligente Praxis. Damit sich dieses aufbaut, muss die betreffende Person die Fähigkeit haben, Wichtiges von Unwichtigem zu unterscheiden und normative Grenzen zu akzeptieren.

Menschliche Intelligenz ergibt sich zudem nicht allein aus dem Orientierungsbedarf des Individuums in einer natürlichen Welt, sondern ist das Ergebnis sozialer Interaktion. Von Geburt an hängt das Wohlbefinden menschlicher Wesen und hängen deren Entwicklungschancen vom Austausch mit anderen Menschen ab. Der intelligente Umgang mit den Herausforderungen der Welt ist nicht das Ergebnis eines fortgesetzten Puzzlespiels, sondern im Wesentlichen Folge der Einbettung der eigenen individuellen Praxis in den größeren sozialen und kulturellen Zusammenhang. Mit dem Erwerb der Sprache können Kinder auf Gründe reagieren, sich von Gründen affizieren lassen und diese selbst auf ihr eigenes Handeln applizieren. Das kulturelle Wissen wird über diese Praxis von einer Generation auf die nächste übertragen, immer wieder veränderten Bedingungen angepasst und bettet das einzelne Individuum in die menschliche Lebensform ein. Im Ergebnis verweben sich dann die Einheit der Person mit der Einheit des Wissens und der Einheit der menschlichen Lebensform. Die einzelne Person zerfällt nicht in Funktionalitäten, sondern wird zusammengehalten durch Gründe, die ihre theoretische und praktische Lebensorientierung bestimmen. Das Individuum wird zur Person und zum Handelnden. Die Identität der Person äußert sich in einer kohärenten Praxis, die von stabilen Gründen geleitet ist. Diese integriert unterschiedliche Aspekte menschlicher Existenz – kognitive, emotive, soziale, ethische, ästhetische, technische und gestalterische.

Es ist fraglich, ob eine derart gründegeleitete, multidimensional bestimmte und soziokulturell eingebettete kohärente Praxis selbst für komplexe maschinelle Systeme jemals plausibel sein könnte. Softwaresysteme leisten Beachtliches. In vielen Bereichen sind sie menschlichen Fähigkeiten bei Weitem überlegen. Aber sie verfügen nicht über ein Analogon zu menschlicher Intelligenz. Es wird der Softwareentwicklung der Zukunft vermutlich in wachsendem Umfang gelingen, menschliche Fähigkeiten zu simulieren und in vielen Fällen zu übertreffen. Das sollte uns aber nicht dazu verführen, ihnen personale Eigenschaften zuzuschreiben, die für genuine menschliche Existenz essenziell sind.

Trotz dieser kategorialen Unterschiede von Mensch und Maschine beeinflussen Mensch und Maschine einander fortwährend. Menschen entwickeln zu bestimmten Zwecken Technologien, die auf die Handlungsmöglichkeiten von Menschen zurückwirken, indem sie jene verändern, erweitern oder vermindern. Diese Mensch-Technik-Relationen und ihre ethische Relevanz genauer zu bestimmen, ist Gegenstand des folgenden Kapitels.

4 Mensch-Technik-Relationen

4.1 Einleitung

Das Verhältnis zwischen menschlichem Handeln, der Verfügbarkeit von Technik und technischen Prozessen ist für die Ethik hoch relevant, denn in diesem Verhältnis können sich Einfluss- und Randbedingungen für Autonomie und Freiheit des Menschen und damit die Möglichkeit der Zuschreibung von Verantwortung auf durchaus komplexe Weise ändern. Dies gilt vor allem und auf spezifische Weise bei KI-Systemen. Es ist daher im Rahmen einer ethischen Analyse und Beurteilung relevant, das Zusammenspiel von Mensch und Technik bzw. von menschlichem Handeln und technischen Prozessen differenziert zu erfassen. Menschen entwickeln und gestalten Technik und nutzen technische Produkte und Systeme oder darauf aufbauende Dienstleistungen als Mittel zum Zweck. Gleichzeitig wirken diese häufig zurück und beeinflussen menschliche Handlungsmöglichkeiten, von der Eröffnung neuer Optionen und der Vergrößerung von Freiheitsgraden bis hin zur Anpassungserzwingung. In diesem Kapitel soll es darum gehen, in welcher Art und Weise verschiedene Mensch-Technik-Relationen die Handlungsmöglichkeiten des Menschen erweitern oder vermindern können, bis hin zur Ersetzung menschlicher Handlungen durch maschinelle Vollzüge. Damit verbunden ist die Frage, wie sich die Spielräume für die Entfaltung menschlicher Autorschaft und die Übernahme von Verantwortung jeweils verändern.

Zum einen geht es darum, dass Tätigkeiten, die vormals (allein) von Menschen durchgeführt wurden, graduell an technische Systeme delegiert werden. Dies reicht vom Delegieren einfacher Tätigkeiten über das Automatisieren komplexer Tätigkeiten oder ganzer Funktionen bis hin zur vollständigen Ersetzung des Menschen durch Technik. Der Begriff „*Ersetzen*“ beschreibt hier also den Endpunkt einer vollständigen Delegation. Zum anderen geht es um Rückwirkungen dieses mehr oder minder umfassenden Delegierens auf menschliche Akteure, das heißt um Fragen, inwiefern jenes Delegieren Handlungsfähigkeit, Möglichkeiten, Fertigkeiten und Kompetenzen von Menschen *erweitert* oder *vermindert*.

Die drei Begriffe des Erweiterns, Verminderns und Ersetzens dienen in diesem Kapitel als analytische Matrix. Sie werden in den folgenden Kapiteln auf ausgewählte Sektoren bezogen, um ein differenziertes Bild der Veränderungen durch KI und ihrer ethisch relevanten Aspekte zu gewinnen.

4.2 Technikdeterminismus versus Sozialkonstruktivismus

Mensch-Technik-Relationen sind Gegenstand vieler Disziplinen. Verortet zwischen Informatik und Psychologie, beschäftigt sich insbesondere das Feld der Human-Computer-Interaction bzw. Computer-Human-Interaction mit dem Verständnis und der Gestaltung von Mensch-Maschine-Schnittstellen. In den Geistes- und Sozialwissenschaften haben insbesondere die Wissenschafts- und Technikforschung, die Science and Technology Studies (STS), die Techniksoziologie und die Technikphilosophie Konzepte und Theorien zur begrifflichen Analyse von Mensch-Technik-Relationen bereitgestellt. Das Verhältnis von Menschen und Gesellschaft zur Technik wurde vielfach entlang der Deutungslinie zwischen sozialem Konstruktivismus und technologischem Determinismus beschrieben.[145] Dahinter steht die Frage nach dem letztlich treibenden Faktor: Folgen Technikgestaltung im Einzelnen und damit auch der technische Fortschritt als Prozess eher menschlich gesetzten Zwecken oder eher einer Eigendynamik, der sich Mensch und Gesellschaft letztlich unterordnen und anpassen müssen. Auch wenn es keine einheitliche Verwendungsweise dieser beiden Deutungen gibt und auch wenn viele Ansätze keine der Extrempositionen vertreten, sondern sich an unterschiedlichen Stellen zwischen Sozial- und Technikdeterminismus verorten, ist eine kurze Erläuterung illustrativ und inhaltlich für diese Stellungnahme wichtig.

In der bereits seit den 1920er-Jahren vertretenen technikdeterministischen Sichtweise wird eine Eigenlogik im technischen Wandel vermutet, die Mensch und Gesellschaft zur Anpassung nötigt. Während sich einzelne Techniken auf menschliche Zwecke zurückführen lassen, folge die gesamte Technologieentwicklung einer inneren und damit nicht oder kaum beeinflussbaren Dynamik. Als Treiber hinter dieser vermuteten Eigendynamik wird immer wieder auf ökonomische Verhältnisse und insbesondere den wirtschaftlichen Wettbewerb zwischen Unternehmen, aber auch den Wettbewerb zwischen Staaten und Weltregionen um vordere Plätze in der technologischen Forschung und Entwicklung hingewiesen. Der auf diese Weise zustande kommende, sozusagen blinde technische Wandel wirke sodann mit seinen Produkten auf die Gesellschaft ein und führe zu Anpassungsnotwendigkeiten, die von konkreter Akzeptanz einzelner Techniken bis hin zur Adaptation an letztlich technologisches Denken reichen.[146]

---

[145] Grunwald, A. (2007): Technikdeterminismus oder Sozialdeterminismus: Zeitbezüge und Kausalverhältnisse. In: Dolata, U.; Werle, R. (Hg.): Gesellschaft und die Macht der Technik. Sozioökonomischer und institutioneller Wandel durch Technisierung. Frankfurt a.M., New York, 63–82.

[146] Rapp, F. (1978): Analytische Technikphilosophie. Freiburg; Ropohl, G. (1982): Kritik des technologischen Determinismus. In: Rapp, F.; Durbin, P. T. (Hg.): Technikphilosophie in der Diskussion. Braunschweig, 3–18. Grunwald, A. (2019): Technology Assessment in Practice and Theory. Abingdon.

In der sozialkonstruktivistischen Sichtweise dagegen treten Technologien nicht eigendynamisch oder zwangsläufig auf den Plan, sondern sind das Ergebnis komplexer und sozial situierter Entwicklungs- und Gestaltungspraktiken bzw. von Ko-Konstruktions-Prozessen unter Mitwirkung zahlreicher Akteure. Die Technikgeneseforschung[147] hat herausgearbeitet, nach welchen Mechanismen Technik durch Entscheidungsprozesse aus ersten Ideen über Entwicklungsprogramme, Experimente und Prototypen bis zum letztendlichen Ergebnis entsteht. Beispielsweise wurde die Rolle von gesellschaftlichen Leitbildern in diesen Prozessen untersucht. [148] Sozialkonstruktivistisch gesehen werden Algorithmen, Roboter, digitale Dienstleistungen oder Geschäftsmodelle für digitale Plattformen von Menschen in möglicherweise langwierigen Entscheidungsprozessen und Handlungssträngen erfunden, entworfen, hergestellt und eingesetzt sowie weiterentwickelt und an neue Umgebungen angepasst. Die „Macher" der Digitalisierung arbeiten in der Regel in Unternehmen, Forschungsinstitutionen oder Geheimdiensten mit bestimmten Agenden, Interessen und Geschäftsmodellen. Wenn *andere* Personen und Institutionen mit *anderen* Werten, Perspektiven und Interessen mitgestalten könnten, würden beispielsweise KI-Systeme mit anderen Eigenschaften entstehen, als wenn man diese den einschlägigen Konzernen mit deren Interessen und Geschäftsmodellen überlässt. Diese Sicht eröffnete Gestaltungsmöglichkeiten und motivierte partizipative Ansätze in der Technikgestaltung wie beispielsweise das Constructive Technology Assessment.[149]

Die grobe Skizze dieser beiden Positionen macht die jeweiligen blinden Flecken deutlich. Weder darf die Bedeutung gesellschaftlicher Hintergründe oder spezifischer Entscheidungen bei der Entwicklung von Technik ignoriert werden, so etwa in Unternehmen oder der öffentlich geförderten Forschung, noch kann abgestritten werden, dass Technologien auf gesellschaftliche Realitäten und menschliche Handlungsmöglichkeiten zurückwirken. Daher erscheint es ratsam, Technikdeterminismus und Sozialkonstruktivismus als Pole eines empirisch vielfältigen und

[147] Bijker, W. E. et al. (1987): The Social Construction of Technological Systems: New Directions in the Sociology and History of Technology. Cambridge (MA), London; Weyer, J. et al. (1997): Technik, die Gesellschaft schafft. Soziale Netzwerke als Ort der Technikgenese. Berlin. Weingart, P. (1989): Technik als sozialer Prozess. Frankfurt a.M.
[148] Dierkes, M. et al. (1992): Leitbild und Technik. Zur Entstehung und Steuerung technischer Innovationen. Berlin, Frankfurt a. M., New York.
[149] Rip, A. et al. (1995): *Managing Technology in Society. The Approach of Constructive Technology Assessment.* London.

differenzierten Spektrums zu begreifen, die den Blick auf unterschiedliche Aspekte von Technikentwicklung und Mensch-Technik-Relationen werfen.[150] Entsprechend sind weder die technikdeterministische noch die konstruktivistische Betrachtung der Mensch-Technik-Relation falsch, beide sind jedoch unterkomplex. Sie werden empirisch unzutreffend, wenn sie in ihrer jeweiligen Perspektive verabsolutiert werden. Die Mensch-Technik-Relation unterliegt vielmehr von Grund auf einem Ko-Konstruktions-Verhältnis und kann als Ko-Evolution beschrieben werden.[151] Soziale Kontexte und normative Kriterien auf der einen und Technologien auf der anderen Seite entwickeln sich weiter in gegenseitiger Wechselwirkung. Die Verfügbarkeit von Technik beeinflusst Handlungsmöglichkeiten und deren Realisierung, aber auch die Bedingungen und Möglichkeiten menschlicher Weltwahrnehmung, wodurch sich Lebensstile und Einstellungen verändern können. Umgekehrt entstehen, wie dies die Technikgeneseforschung in vielen empirischen Studien belegt hat, neue Techniken vor dem Hintergrund von sozialen Befindlichkeiten, normativen Kriterien und Zukunftsentwürfen.

Gerade im Kontext der Künstlichen Intelligenz sind die Arbeiten der Anthropologin Lucy Suchman von großer Bedeutung. Ihr Buch „Human-Machine Reconfigurations“[152] liefert eine Reflexion und Kontextualisierung ihrer Studien der KI-Forschung in den 1980ern, welche 1987 unter dem Titel „Plans and Situated Action“ veröffentlicht wurden.[153] Suchman kritisiert das Planungsmodell von Interaktion, das einem Großteil der damaligen Forschung zugrunde liegt, und schlägt einen Perspektivenwechsel in der Betrachtung der Mensch-Maschine-Relation vor, der Einsichten aus den Sozialwissenschaften Rechnung trägt. Danach ist menschliches Handeln auf vielfältige Weise sozial situiert und beeinflusst, ohne vollständig determiniert zu sein. Sie argumentiert, dass Menschen sinnvoll handeln, indem sie auf der Grundlage ihrer sozialen und ökologischen Ressourcen häufig weniger planen als improvisieren. Sie kritisiert also die theoretisch-konzeptionellen Grundlagen des Designs interaktiver technischer Systeme als aus anthropologischer Sicht unangemessen, weil menschliches Handeln ständig aus dynamischen Interaktionen mit der materiellen, insbesondere technischen, und der sozialen Welt konstruiert und rekonstruiert werde.

---

[150] Dolata, U.; Werle, R. (2007): Gesellschaft und die Macht der Technik. Sozioökonomischer und institutioneller Wandel durch Technisierung. Frankfurt a. M., New York.
[151] Rip, A. (2007): Die Verzahnung von technologischen und sozialen Determinismen und die Ambivalenzen von Handlungsträgerschaft im „Constructive Technology Assessment“. In: Dolata, U.; Werle, R. (Hg.): Gesellschaft und die Macht der Technik. Sozioökonomischer und institutioneller Wandel durch Technisierung. Frankfurt a. M., New York, 83–106.
[152] Suchman, L. A. (2007): Human-Machine Reconfigurations. Plans and Situated Actions, 2. Auflage. Cambridge.
[153] Suchman, L. A. (1987). Plans and situated actions: The problem of human-machine communication. Cambridge.

Technikphilosophie und -soziologie haben die zunehmende Komplexität der Mensch-Maschine-Relation in unterschiedlichen Theorien gedeutet und zugespitzt. In der Technikphilosophie wird Technik häufig nicht mehr als Ensemble technischer Objekte verstanden, sondern als Medium, mit dem sich menschliches Handeln und Verhalten vollzieht. Während die einzelnen Elemente dieses Mediums instrumentellem Zweck-Mittel-Denken entstammen, stelle ihre Gesamtheit eine *Zweite Natur* dar, die Randbedingungen und Erfolgsbedingungen für weiteres menschliches Leben setzt und auch Weltsicht und das Problemlösen beeinflusst.[154] Als die bereits technologisch orientierten Menschen, zu denen sie im Rahmen vieler Technisierungsprozesse geworden sind, werden sie zum Beispiel dazu neigen, Herausforderungen von Kommunikation oder Sicherheit als Probleme anzusehen, die primär technologisch zu lösen sind. Somit ist neue Technologie oft bereits das Ergebnis einer technologischen Art und Weise, wie Menschen die Welt sehen und sich zu ihr in Beziehung setzen.

Techniksoziologisch sind hier vor allem Ansätze der Ko-Evolution von Technik und Gesellschaft zu nennen.[155] In ihnen sind die sozialkonstruktivistischen Motive der Gestaltung aufgenommen, jedoch wird ihnen die vielfältige Rückwirkung einmal entwickelter und verfügbarer Technik auf Mensch und Gesellschaft zur Seite gestellt, zu denen beispielsweise die großen Infrastruktursysteme – wie jene der Mobilität – geeignetes Illustrationsmaterial liefern. Zunächst gestaltet nach Zweck-Mittel-Erwägungen unter Berücksichtigung vielfältiger gesellschaftlicher Belange beispielsweise aus Wirtschaft, Bürgerschaft und Umweltschutz, werden sie nach Fertigstellung zu Randbedingungen menschlicher Entscheidungen, zum Beispiel in Bezug auf die Wahl des Wohnortes oder die Ansiedlung von Betrieben. Die Akteur-Netzwerk-Theorie[156] sowie unterschiedliche Sichtweisen innerhalb der Technikphilosophie[157] haben die Gedanken von Ko-Konstruktion und Ko-Evolution erweitert und teils die technischen Objekte aufgrund ihres Einflusses auf den Menschen als Ko-Akteure (Aktanten) definiert (siehe unten).

---

[154] Hubig, C. (2006): Die Kunst des Möglichen I. Technikphilosophie als Reflexion der Medialität. Bielefeld.
[155] Rip, A. (2007): Die Verzahnung von technologischen und sozialen Determinismen und die Ambivalenzen von Handlungsträgerschaft im „Constructive Technology Assessment". In: Dolata, U.; Werle, R. (Hg.): Gesellschaft und die Macht der Technik. Sozioökonomischer und institutioneller Wandel durch Technisierung. Frankfurt a. M., New York, 83–106.
[156] Latour, B. (2007): Eine neue Soziologie für eine neue Gesellschaft. Einführung in die Akteur-Netzwerk-Theorie. Frankfurt a. M.; bzw. die englische Version: Latour, B. (2005): Reassembling the Social. An Introduction to Actor-Network-Theory. Oxford.
[157] Hubig, C. (2006): Die Kunst des Möglichen I. Technikphilosophie als Reflexion der Medialität. Bielefeld; Ihde, D. (1990): Technology and the lifeworld. Bloomington.

4.3 Mehrstufige Mensch-Technik-Wechselwirkungen

Die zunehmende Komplexität der Mensch-Technik- bzw. Mensch-Maschine-Relation verändert auch deren Wahrnehmung. Digitale Technik, insbesondere KI-gesteuerte Systeme wie Produktionsroboter, „autonome" Fahrzeuge, Therapieprogramme oder Schachcomputer sind Beispiele, in denen die vormals klaren Unterscheidungen von Mensch und Technik weniger eindeutig zu werden scheinen. Androide Roboter erscheinen menschenähnlich, Hilfesuchende interagieren mit Therapieprogrammen, als ob es sich um menschliche Therapiekräfte handeln würde, und der Schachcomputer scheint die Partie gewinnen „zu wollen". Die Anthropomorphisierung digitaler Technik ist in der Umgangssprache weit fortgeschritten. Sie zeigt sich darin, dass KI und Robotern Fähigkeiten wie Denken, Lernen, Entscheiden oder Emotionalität zugeschrieben werden, wodurch sie scheinbar in die Gemeinschaft der denkenden, lernenden, entscheidenden und fühlenden Menschen aufgenommen werden.

Phänomenologisch geht damit einher, dass sich durch „autonom" werdende KI-gestützte Technik Subjekt-Objekt-Verhältnisse zwischen Mensch und Technik verändern. Im traditionellen Bild gestalten und nutzen menschliche Subjekte technische Objekte. Bereits mit traditioneller Software, mehr noch mit KI, kommt es jedoch zu neuen Konstellationen. In vernetzten Systemen haben Menschen teils die Subjekt-, teils aber auch die Objektrolle inne. Wenn einerseits Entscheidungen über Menschen an Softwaresysteme delegiert werden, beispielsweise hinsichtlich der Gewährung von Krediten oder Sozialleistungen, werden Menschen zu Objekten der „Entscheidungen" dieser Systeme, die hier auftreten, als ob sie Subjekte seien. Andererseits kann die Subjektrolle von Menschen durch gute Software zur Entscheidungsunterstützung erhöht werden, beispielsweise wenn diese qualitativ hochwertige, diskriminierungsfreie und nachvollziehbare Informationen liefern, welche die Qualität menschlicher Entscheidungen und deren Begründbarkeit verbessern. Verschiebungen in den Subjekt-Objekt-Rollen zwischen Mensch und Technik müssen daher differenziert betrachtet werden. Sie hängen einerseits vom Ausmaß und diversen technischen und organisationalen Details ab; andererseits – und dies ist von besonderer ethischer Relevanz – manifestieren sie sich bei verschiedenen Personengruppen auf unterschiedliche Weise.

Die Gestaltung der Software und der damit operierenden Maschinen gibt jeweils die Alternativen vor, innerhalb derer Menschen handeln können. Optionen, die in dem Design nicht vorge-

sehen sind, werden ausgeschlossen. Algorithmen und Maschinen regulieren somit menschliches Handeln.[158] Derartige Prozesse finden in der Digitalisierung seit Jahrzehnten statt, werden aber durch KI-Systeme verschärft. Menschliche Akteure erleben dadurch einerseits eine Verminderung ihrer Autorschaft über das eigene Handeln und fühlen sich zunehmend eingeschränkt und fremdbestimmt. Andererseits werden KI-Systeme zielgerichtet von Akteuren eingesetzt, um die eigenen Handlungsmöglichkeiten zu erweitern. Ein Beispiel hierfür sind ADM-Systeme, die Klassifikationen und Prognosen vornehmen und beispielsweise durch das Errechnen von Risikoscores Menschen bei der Entscheidungsfindung unterstützen (vgl. Beispiele in Teil II). Es ist immer wieder ein erwünschter Effekt erweiterter Autorschaft, wenn menschliche Entscheidungen dadurch auf eine sachlichere Grundlage gestellt werden. Auf der anderen Seite jedoch droht das Risiko, dass Menschen den Ergebnissen der KI-Systeme, auch wenn diese nur als Vorschläge unterbreitet werden, einfach blind folgen. Dann würde die Person, die eine Entscheidung trifft, eher *reagieren* als aus eigener Einsicht heraus *agieren*, was ihre Autorschaft vermindern würde (*automation bias*). Neben den bereits vielfach diskutierten Herausforderungen an eine transparente und rechtssichere Zuschreibung von Verantwortung in komplexen Mensch-Technik-Systemen, etwa beim „autonomen" Fahren[159], ist von anthropologischer und ethischer Relevanz, inwieweit die digitalen Systeme Menschen unterstützen und dadurch die Möglichkeit der Entfaltung der menschlichen Fähigkeiten erweitern oder durch technische, von den Herstellern der Systeme vorgegebene Handlungsschemata diese Entfaltung behindern und vermindern.

Die angedeutete Komplexität neuer Mensch-Maschine-Wechselwirkungen und die erwähnten Verschiebungen in Subjekt-Objektrollen haben nicht nur in der öffentlichen Debatte, sondern auch in der Wissenschaft zu einer Aufweichung eines strikten Dualismus zwischen Mensch und Maschine geführt. So wird vor allem in der Wissenschafts- und Technikforschung (Science and Technology Studies) sowie der Techniksoziologie seit Jahrzehnten über die „Handlungsträgerschaft" von technischen Systemen diskutiert. Einige dieser Positionen, insbesondere die Akteur-Netzwerk-Theorie[160], die einen sehr schwachen Handlungsbegriff propagiert und diesen auf viele Entitäten ausweitet, stehen dabei in deutlicher Spannung mit der im vorigen Kapitel

[158] Orwat, C. et al. (2010): Software als Institution und ihre Gestaltbarkeit. In: Informatik-Spektrum 33, 626-633, 626.
[159] Ethik-Kommission (2017): Automatisiertes und vernetztes Fahren. Endbericht. Berlin. https://www.bmvi.de/SharedDocs/DE/Publikationen/DG/bericht-der-ethik-kommission.pdf?__blob=publicationFile [18.01.2023].
[160] Latour, B. (2005): Reassembling the Social. An Introduction to Actor-Network-Theory. Oxford; Law, J.; Hassard, J. (1999): Actor Network and After. Oxford.

ausführten philosophischen Handlungstheorie, die einen anspruchsvollen Handlungsbegriff beschreibt und diesen auf menschliche Akteure beschränkt. Jenseits disziplinärer Einzelperspektiven stellen sich in Anbetracht der zunehmenden Verschränkung und wechselseitigen Beeinflussung von Menschen und Maschinen unter anderem folgende Fragen:

- Reicht das technische Vokabular zur Beschreibung von KI-Systemen noch aus, um Phänomene komplexer Interaktionen von Mensch und Maschine zu beschreiben?
- Inwiefern kann bzw. sollte davon gesprochen werden, dass Maschinen handeln oder mithandeln können? Sollte also Maschinen die Rolle von Akteuren zugesprochen werden und wenn ja, unter welchen Bedingungen und mit welchen Implikationen?
- Kommt es zu Verschiebungen in den Möglichkeiten menschlicher Autorschaft und wenn ja, in welchen Richtungen?

Die Akteur-Netzwerk-Theorie beantwortet diese Fragen, indem sie technischen Systemen den Status von Akteuren mit eigenen Dynamiken zuspricht und von hybriden Handlungszusammenhängen zwischen Mensch und Maschine ausgeht.[161]Die Akteur-Netzwerk-Theorie als Beobachtungstheorie ohne spezifische normativ-anthropologische Prämissen hilft, vermeintlich autonome Wirkungen von Techniken, Artefakten oder Sachen und deren gesellschaftsverändernde Potenzial zu erkennen. So kann es gelingen, Phänomene aus der Beobachterperspektive in den Blick zu nehmen, die mit einem starken, die Autonomie betonenden Handlungsbegriff tendenziell ausgeblendet werden. Freilich bleibt die Frage nach dem Zusammenhang zwischen Handlung und Verantwortung hier offen.

Der Techniksoziologe Werner Rammert, einer der Pioniere der Thematisierung möglicher Handlungsträgerschaft von Technik, schlägt einen Mittelweg zwischen einer anspruchsvoll normativen Vorstellung von Handeln und der Vorstellung eigenmächtigen maschinellen Agierens vor. Stattdessen soll von einer verteilten Handlungsträgerschaft zwischen Mensch und Maschine gesprochen werden, um die Vorstellung zu vermeiden, Technik sei etwas außerhalb des Sozialen Stehendes.[162] So geht Rammert wie die Akteur-Netzwerk-Theorie von hybriden, sozio-technischen Konstellationen aus, in denen Menschen und Maschinen auf komplexe Weise wechselwirken. Das Handeln des Menschen in dieser Perspektive sieht er zwar von technischen Prozessen beeinflusst, jedoch nicht als determiniert an. Der Einfluss der Technik kann sich in

[161] Latour, B (2007): Eine neue Soziologie für eine neue Gesellschaft. Einführung in die Akteur-Netzwerk-Theorie. Frankfurt a. M.
[162] Rammert, W.; Schulz-Schaeffer, I. (2002): Technik und Handeln- wenn soziales Handeln sich auf menschliches Verhalten und technische Artefakte verteilt. In: Dies. (Hg.): Können Maschinen handeln? Frankfurt a. M., New York, 11-64.

beiden Richtungen auswirken: Individuelle Freiheitsspielräume und die Entfaltung der menschlichen Autorschaft des eigenen Lebens können sowohl erweitert als auch vermindert.

Vor diesem Hintergrund schlägt Rammert vor, die Wechselwirkung von Mensch und Technik dreistufig zu beschreiben, um sowohl die Komplexität dieser Wechselwirkungen empirisch zu erfassen als auch die Zuschreibung von Verantwortung auf Menschen zu begrenzen. Als Stufe 1 nennt er *Kausalität* im Sinne von, Veränderung bewirken zu können. Stufe 2 beschreibt er als *Kontingenz* mit der Bedeutung, auch anders agieren zu können. Stufe 3 schließlich ist durch *Intentionalität* gekennzeichnet, was beinhaltet, das eigene Verhalten deuten und steuern zu können. Kausalität und Kontingenz charakterisieren nach Rammert nicht nur die menschliche Intervention, sondern auch die von Technologien. Algorithmen „wählen" zwischen Alternativen; automatisierte Entscheidungssysteme „bestimmen" einen Risikoscore, auf dessen Basis entweder ein Mensch eine Entscheidung trifft, beispielsweise über eine Kreditvergabe, oder die Software „entscheidet" sogar selbst, welche Bewerbungen bereits vorab aussortiert und der Personalabteilung erst gar nicht angezeigt werden. Algorithmen wirken hier auf mannigfaltige und höchst komplexe Weise, ohne dass ihnen dafür Intentionalität unterstellt werden kann. Intentionalität nämlich, die höchste Stufe des „Agierens", ist dem Handeln von Menschen vorbehalten. Daher kann nach Rammert nur ihnen Verantwortung zugeschrieben werden.

Die genannten Schwierigkeiten bei der Zuschreibung von Verantwortung haben auch den Technikphilosophen Luciano Floridi motiviert, eine Theorie „verteilter Moralität"[163] zu entwickeln, die auf die Komplexität digitaler Mensch-Maschine-Systeme zugeschnitten ist. Basierend auf früheren Arbeiten[164] unterscheidet er zwischen moralischer Handlungsfähigkeit (*moral agency)* und moralischer Verantwortlichkeit (*moral responsibility)*.[165] Während er – wie Rammert – Intentionalität als notwendig für Verantwortlichkeit erachtet, ist sie keine Bedingung für Handlungsfähigkeit. Für Letzteres reichen Interaktivität, Autonomie und Adaptivität aus. KI-Algorithmen bzw. verteilte Mensch/KI-Systeme können demnach für ihre „Handlungen" zwar *accountable*, aber nicht *responsible* sein, da ihnen die Intentionalität fehle. Die moralischen Eigenschaften der Ergebnisse verteilter Handlungsträgerschaft *emergieren* aus den einzelnen Elementen, ohne dass ihnen eine Intention zugrunde liegt.

---

[163] Floridi, L. (2012): Distributed morality in the information society. In: Science and Engineering Ethics 19(3), 727-74. (DOI: 10.1007/s11948-012-9413-4).
[164] Floridi, L.; Sanders, J. W. (2004): On the morality of artificial agents. In: Minds and Machine 14, 349-379.
[165] Da die üblichen deutschen Übersetzungen nicht immer genau zu passen scheinen, sind hier die englischen Originalbegriffe erhalten.

Auch der Versuch von Floridi, die Interaktionen von Mensch und Maschine angemessen zu beschreiben, lässt die Frage der Zuschreibung von Verantwortung offen.[166] Es gibt moralisch bedeutsame Aktionen von KI-Systemen, insofern als moralisch problematische Resultate durch KI-Systeme verursacht werden. Den Systemen kann dafür aber keine Verantwortung zugeschrieben werden. Verantwortung muss daher anders geregelt werden, zum Beispiel durch Zuschreibung von Verantwortung an relevante Institutionen bzw. Organisationen. Dies können beispielsweise die Betreiber dieser Systeme sein, da sie aufgrund ihrer intentionalen Entscheidung zu deren Einsatz einschließlich der diffusen Verteilung von *agency* zugestimmt haben und damit – auch im Sinne von Floridi – verantwortlich sind. Der von der Europäischen Union geplante Artificial Intelligence Act (AI Act) wird folglich die Verantwortlichkeit konkret zuweisen müssen, um wirksam zu regulieren.

Die hier vorgestellten techniksoziologischen und -philosophischen Ansätze, die sich um ein differenziertes Verständnis der zunehmend komplexen Mensch-Technik-Wechselwirkungen in Bezug auf KI-Systeme bemühen, können mit den anthropologischen Positionen, die in Kapitel 3 dargelegt wurden, durchaus in Konflikt geraten. Entsprechend der dort ausgeführten anspruchsvollen philosophischen Handlungstheorie können Maschinen nicht handeln, weil sie nicht über Intentionen verfügen, und kommen daher als genuine Akteure nicht infrage, jedenfalls nicht nach gegenwärtigem Stand der Entwicklung. Aber auch, wenn technischen Systemen keine Handlungsfähigkeit und damit Verantwortung zugeschrieben werden kann, haben sie Einfluss auf menschliches Handeln. Menschliches Handeln ist weder völlig autonom noch völlig sozial oder technisch determiniert, sondern in zunehmendem Maß soziotechnisch situiert. Auch in der Digitalisierung und der KI ist dies empirisch durch zahlreiche Studien belegt, insbesondere zum sogenannten *Automation Bias* bzw. zu den Effekten von *Nudging*.[167] Diese Effekte sind höchst bedeutsam für ethische Fragen, zeigen sie doch, dass technologische Entwicklungen menschliche Handlungsfähigkeit beeinflussen und menschliche Autonomie und Autorschaft sowohl erweitern als auch vermindern können.

---

[166] Floridi, L. (2016): Faultless responsibility: on the nature and allocation of moral responsibility for distributed moral actions. In: Philosophical Transaction A 374: 20160112 (DOI: 10.1098/rsta.2016.0112).

[167] Unter Nudging versteht man die Formatierung einer Entscheidungssituation ohne an den Handlungsalternativen etwas verändern, sodass erwünschtes Verhalten wahrscheinlicher wird. Vgl. Thaler, R. H.; Sunstein, C. R. (2018): Nudge. Wie man kluge Entscheidungen anstößt. Übersetzt von Bausum, C. 13. Auflage. Berlin.

4.4 Erweitern und Vermindern menschlicher Autonomie und Autorschaft

Angesichts der konstatierten intensiven und komplexen Wechselwirkungen zwischen Mensch und Technik stellt sich die Frage nach den Folgen des digital-technischen Fortschritts für die Bedingungen gelingenden Handelns und die Möglichkeiten, menschliche Autorschaft zu entfalten. Bisherige Erfahrungen zeigen Ambivalenzen und Dialektiken von ethischer Relevanz.[168] Neue Technologien sollen einerseits und vor allem, so die Erzählung spätestens seit Francis Bacon und der Europäischen Aufklärung, die Menschen von den Zwängen der Natur und der Tradition emanzipieren, Freiheitsräume durch neue Handlungsoptionen eröffnen und damit die Entfaltungsmöglichkeiten menschlichen Handelns erweitern. Entsprechende Effekte zeigen sich in der Tat im raschen Fortschritt der digitalen Technologien: globale Kommunikation in Echtzeit, schnelle Information, Mustererkennung durch Big Data, Effizienzsteigerung und Beschleunigung der Produktion, neue Dienstleistungen und Geschäftsmodelle, bessere medizinische Diagnosen und Therapien, Roboter und Algorithmen als künstliche Assistenten, selbstfahrende Autos, Minenräumroboter und vieles mehr. Speziell die KI eröffnet Möglichkeiten menschliches Handeln zu verbessern, so etwa durch Mustererkennung in großen Datenmengen für medizinische oder behördliche Zwecke, durch darauf aufbauende verbesserte Prognosemöglichkeiten, zum Beispiel zur Ausbreitung von Infektionskrankheiten oder für Prognosen in der Polizeiarbeit (*predictive policing*) , durch neue Möglichkeiten individualisierter Information und Werbung, aber auch durch Anwendungen im Bildungsbereich. Technik ist zentraler Teil menschlichen Lebens und gesellschaftlicher Vollzüge zumindest in den industrialisierten Regionen der Welt geworden und hat in vielen Fällen eindeutig positive Folgen in dem Sinne gezeitigt, dass die Möglichkeiten menschlicher Autorschaft erweitert wurden – jedenfalls für den Teil der Erdbevölkerung, der, vor allem im Globalen Norden, Zugang zu ihren Vorteilen hat.

Im Rahmen der Diffusion von Technik und Innovationen in die Gesellschaft, ihrer Nutzung und Veralltäglichung kommt es jedoch häufig zu Sekundäreffekten, die als negativ wahrgenommen werden. Zu den nicht intendierten Folgen wie Umweltproblemen und sozialen Verwerfungen zählen auch Begrenzungen menschlicher Entfaltungsmöglichkeiten. In die Erweiterung menschlicher Autonomie und Autorschaft im technischen Fortschritt ist ihre simultane Verminderung oft entweder bereits eingeschrieben oder entwickelt sich empirisch kontingent. Erweiterung und Verminderung sind häufig ineinander verschränkt, betreffen jedoch meist unterschiedliche Beteiligte in unterschiedlicher Weise, so etwa diejenigen, die Entscheidungen

[168] Grunwald, A. (2022): Technikfolgenabschätzung- Eine Einführung. Baden-Baden.

treffen, und diejenigen, die von diesen Entscheidungen betroffen sind. Während die Facetten der Erweiterung offensiv kommuniziert werden und oft auch deutlich sichtbar sind, etwa aufgrund neuer Fähigkeiten von IT-Systemen, ist ihre Kehrseite, die nicht immer aber immer wieder auftretende simultane Verminderung menschlicher Entfaltungsmöglichkeiten, oft nicht gut erkennbar.

Verminderungen qualifizierter Handlungsformen und Entfaltungsmöglichkeiten im Rahmen der Nutzung von Technik können etwa in folgenden Richtungen auftreten:

(1) Entstehende Abhängigkeiten: Mit dem Erfolg von Technik sind moderne Gesellschaften von ihrem reibungslosen Funktionieren abhängig geworden. Dies beginnt individuell mit der Abhängigkeit von Computer und Auto und reicht gesellschaftlich bis hin zur vollständigen Abhängigkeit vom Funktionieren der Energieversorgung und der weltweiten Datenkommunikationsnetze. Die individuelle wie kollektive Abhängigkeit von digitalen Technologien, insbesondere vom Internet, durchzieht sämtliche gesellschaftlichen Prozesse: ohne Internet keine funktionierende Weltwirtschaft, keine Finanztransaktionen, Kollaps internationaler Logistikketten, Zusammenbruch der öffentlichen wie privaten Kommunikation. Entsprechend führt jede Übertragung von Zuständigkeiten beispielsweise durch den Einsatz von Software zur Entscheidungsunterstützung, zu einer gewissen Abhängigkeit. Abhängigkeit jedoch vermindert menschliche Entfaltungsmöglichkeit, da sie Sachzwänge zum Weiterbetrieb der technischen Systeme nach sich zieht und die Vulnerabilität der Gesellschaft-gegenüber technischem Versagen und intendierten Störungen (z. B. Hacking) steigert.

(2) Anpassungsdruck: Technik nötigt zur Anpassung. Das ist auf der Ebene konkreter technischer Objekte wie Maschinen trivial; es müssen beispielsweise Bedienungsanleitungen beachtet werden, um die Technik sachgerecht nutzen zu können. Digitale Technik jedoch reguliert und ändert subtil menschliches Handeln und Verhalten. Softwaresysteme steuern vielfach explizit oder implizit Verhalten.[169] So strukturieren privat geführte digitale Infrastrukturen die politische Kommunikation, sortieren Suchmaschinen die Welt mithilfe der von ihnen gesetzten Filter und strukturieren Online-Plattformen Geschäftsprozesse. Dahinter steht beispielsweise die schlecht durch Daten belegbare Befürchtung, dass menschliches Denken und Handeln durch fortschreitende Anpassung an Softwaresysteme allmählich nach deren Anforderungen und Vorgaben reguliert und immer stärker im technischen Sinn normiert werden könnte. Menschliche

[169] Bowker, G. C.; Star, S. L. (1999): Sorting Things Out – Classification and Its Consequences. Cambridge (MA); Nguyen, C. T. (2021): How Twitter gamifies communication. In: Lacey, J. (Hg.): Applied Epistemology. Oxford, 410-436 (DOI: 10.1093/oso/9780198833659.003.0017).

Autorschaft würde leise und unbemerkt, sozusagen durch allmähliche Gewöhnung, unkritische Übernahme algorithmischer Vorschläge (Automation Bias) und Anpassung an technische Voreinstellungen, ausgehöhlt.

(3) Verschließen von Optionen: Immer wieder werden mit dem Eröffnen neuer Handlungsspielräume andere, bis dato etablierte Optionen abgewertet oder ganz verschlossen. In der Innovationstheorie gilt dies als „schöpferische Zerstörung".[170] Dies ist einerseits der normale Gang von Transformation und Wandel. Andererseits aber stürzen Innovationen vorhandene Anerkennungs- und Wertstrukturen durch disruptive Effekte um und ziehen Gewinner wie auch Verlierer nach sich. Von den neuen Optionen profitieren häufig andere Personen und Gruppen als die, die dann im Verschließen der traditionellen Optionen zu Verlierern des Wandels werden. Zum Verschließen von Optionen menschlicher Entfaltung durch technischen Fortschritt führen unterschiedliche Mechanismen. So werden Infrastruktursysteme häufig faktisch machtförmig, indem sie Lebensformen außerhalb dieser Systeme benachteiligen oder unmöglich machen. Beispielsweise wird mittlerweile häufig die Nutzung eines Smartphones vorausgesetzt, um an bestimmten Lebensvollzügen teilnehmen zu können. Diese Form der Verschließung von Optionen kann verschiedene Bevölkerungsgruppen unterschiedlich treffen und Gerechtigkeitsprobleme mit sich bringen, wie dies zum Beispiel unter dem Stichwort digitale Spaltung (*digital divide*) diskutiert wird. Ein anderer Mechanismus besteht in allmählicher Gewöhnung als Folge der oben erwähnten Anpassung. Technik, gerade die Digitaltechnik, macht vielfach das Leben angenehm und komfortabel. Sobald Routinehandlungen in Beruf oder Freizeit daran adaptiert wurden, gehört diese Technik so zum Leben, dass es ohne diese Technik oft kaum noch vorstellbar ist. Alternative Optionen verschließen sich, vermeintliche Sachzwang-Argumente erwecken den Anschein der Alternativlosigkeit, sind jedoch nur Ausdruck der schleichend eingetretenen Pfadabhängigkeit durch allmähliche Anpassung und Gewöhnung.

Diese Mechanismen, in denen Optionen menschlichen Handelns sich verschließen, können im Rahmen der Diffusion neuer Technologien in die Anwendung auftreten, ohne dass Intentionen von Akteuren dahinterstehen. Es geht nicht um eine Verminderung menschlicher Autorschaft durch bewusste Delegation vormals menschlicher Handlungsvollzüge an KI-Systeme, sondern um Effekte, die schleichend und teilweise unbewusst durch Verhaltensänderungen entstehen. Das Ersetzen als Endpunkt des Delegierens vormals menschlich ausgeübter Tätigkeiten an tech-

[170] Schumpeter, J. A. (2018): Kapitalismus, Sozialismus und Demokratie (9. Aufl.). Tübingen, S. 113 ff.

nische Systeme erfolgt jedoch intentional. Es betrifft Funktionen und Tätigkeiten, die technomorph beschrieben und sodann von KI-gesteuerten Systemen übernommen werden können, im Idealfall funktionsäquivalent oder „besser".[171] Motivationen, menschliche Tätigkeiten durch KI-Systeme zu ersetzen, sind zum Beispiel die Effizienzsteigerung behördlicher Funktionen, die Kostensenkung in der industriellen Produktion, die Routinisierung diagnostischer Auswertungen in der Medizin oder die Ermöglichung automatisierter Überwachung in Echtzeit.

Bezogen auf die Möglichkeit menschlicher Autorschaft stellt sich die Ersetzung menschlicher Tätigkeiten zunächst als Resultat menschlicher Entscheidungen dar. So wird beispielsweise in Unternehmen oder Behörden aus unterschiedlichen Gründen die Entscheidung getroffen, bestimmte Tätigkeiten, die zuvor von Menschen durchgeführt wurden, an maschinelle Systeme zu übertragen. Dies können beispielsweise ADM-Systeme in unterschiedlichen Anwendungen sein, etwa in der Medizin, im Sicherheitsbereich oder im Sozialwesen. Diese Übertragung ist für sich genommen ein Ausdruck der Wahrnehmung menschlicher Autorschaft in bestimmten institutionellen Kontexten und unter entsprechenden Randbedingungen. Die zentrale ethische Frage ist, ob und wie diese Übertragung die Möglichkeiten *anderer* Menschen beeinflusst, vor allem jener, über die entschieden wird. Es stellt sich hier also die Frage, wie die Delegation von Tätigkeiten an Technik die Handlungsmöglichkeit und Autorschaft der Betroffenen beeinflusst. Dies stellt sich in unterschiedlichen Anwendungsfeldern auf je andere Weise dar.

Bereits mehrfach hat sich damit gezeigt, dass mit einer erwünschten Erweiterung der Möglichkeiten menschlicher Autorschaft oft simultan eine Verminderung verbunden ist, häufig in Bezug auf andere Aspekte und Felder von Autorschaft bzw. andere Menschen. Auf jeden Fall verschieben KI-Systeme die Möglichkeiten der Wahrnehmung menschlicher Autorschaft. Dies betrifft unterschiedliche Bevölkerungsgruppen auf unterschiedliche Weise und weist daher eine soziale Dimension mit ethischen Fragen auf. Bei der Betrachtung von Chancen und Risiken etwa von Entscheidungsunterstützungssystemen im Sicherheitsbereich ist zu berücksichtigen, *für wen* es sich hier jeweils um Chancen oder Risiken, um Erweiterungen oder Verminderungen der Autorschaft handelt. Damit sind hier Aspekte sozialer Gerechtigkeit und Macht involviert. In der Digitalisierung und speziell bei der Entwicklung und Nutzung von KI ist grundsätzlich zu fragen, wer die die entsprechenden Prozesse bzw. der konkreten Software-Applikationen und KI-Algorithmen jeweils gestaltet und ob – und wenn ja mit welcher Legitimation – sie in

[171] Das Wort „besser" suggeriert, dass es Verbesserung „als solche" gebe, und ignoriert, dass „besser" semantisch grundsätzlich nur im Zusammenhang mit normativen Kriterien des „besser" sinnvoll ist – und diese Kriterien können umstritten und kontrovers sein.

die Autonomie und Autorschaft derjenigen, die diese Produkte nutzen, oder weiterer Betroffener eingreifen. Auch jenseits der intentionalen Manipulation durch die Gestalter sind Effekte von Beeinflussung, Gewöhnung und Abhängigkeiten bis hin zu digitalem Mediensuchtverhalten zu beobachten, in denen offenkundig menschliche Autorschaft eingeengt wird.

Die sich hier andeutenden ethischen Herausforderungen sind mit epistemologischen Herausforderungen verbunden. Allmähliche Verschiebungen, wie etwa die erwähnten Gewöhnungsprozesse an technisch normierte Handlungsmuster (Automation Bias), sind oft nur schwer aufzudecken und empirisch zu belegen. Es besteht das Risiko verspäteter Entdeckung, zu einem Zeitpunkt, zu dem möglicherweise nur noch schwer beeinflussbare Pfadabhängigkeiten bereits eingetreten sind, schlimmstenfalls ein Point of no Return überschritten wurde. Zwischen der hohen Relevanz dieser an die Dialektik von Herr und Knecht gemahnenden Situation und der epistemologisch schwierigen Nachweislage ist die Bewusstmachung möglicher ethisch bedenklicher Zukunftsentwicklungen dieses Typs eine Herausforderung. Denn angesichts starker Gegenwartspräferenzen vieler Akteure ist sie mit den bekannten Problemen vorsorgeorientierter Kommunikation konfrontiert. Von der einen Seite droht der Vorwurf der Übertreibung, Dramatisierung oder gar Technikfeindlichkeit, von der anderen der Vorwurf der Verharmlosung. Hohe epistemologische Unsicherheit macht vorsorgeorientierte Kommunikation anfällig für Ideologie, interessengeleitete Statements und Spekulation.

4.5 Fazit

Die ethische Analyse und Beurteilung des Einsatzes von KI-Systemen bedürfen über die begriffliche, anthropologische und handlungstheoretische Vergewisserung (vgl. Kapitel 3) hinaus eines genaueren Blicks auf die sich mit der Digitalisierung verändernden Konstellationen zwischen Mensch und Technik. Im Rahmen der philosophischen Handlungstheorie können Maschinen nicht handeln und kommen als genuine Akteure mit Verantwortung nicht infrage. Dennoch haben sie Einfluss auf menschliches Handeln, das in modernen Gesellschaften in zunehmendem Maß soziotechnisch situiert ist. Erfahrung und empirische Forschung zeigen, dass Technik einerseits von Menschen als Mittel nach Zwecken gestaltet wird, dass aber andererseits neue Technik und darauf aufbauende Innovationen oder Dienstleistungen menschliches Handeln und Verhalten beeinflussen. Ethisch relevant ist insbesondere, wie sich diese Wechselwirkungen auf die Möglichkeiten menschlicher Autorschaft und Verantwortungsübernahme auswirken und wie diese sich angesichts der zunehmenden Verbreitung von KI-Systemen verändern.

KI-Systeme können in vielen Feldern menschliche Handlungen und Entscheidungen unterstützen, dadurch zu besseren Ergebnissen beitragen und damit menschliche Autorschaft erweitern. Vor allem die durch Algorithmen eröffnete Möglichkeit, in großen Datenmengen (Big Data) Muster zu erkennen, die den Menschen ansonsten verborgen wären, ist die Basis für den unterstützenden Einsatz der KI zum Beispiel in der medizinischen Diagnostik, im Bildungsbereich aber auch im Medienbereich und in der Verwaltung. Gerade im Bereich der Sozialen Medien zeigt sich hier ein Phänomen, das als *Hypernudge*[172] beschrieben wurde: Datenbasierte algorithmische Systeme kuratieren dynamisch hochgradig personalisierte Informationsumgebungen, denen man sich nur schwer entziehen kann. Menschliche Autorschaft kann also durch KI nicht nur erweitert, sondern auch vermindert werden, entweder durch intendierte Delegation von Entscheidungen an automatische Systeme oder durch allmähliche Effekte der Gewöhnung und Anpassung an datengenerierte Empfehlungen von KI-Systemen.

KI-Systeme verschieben die Möglichkeiten der Wahrnehmung menschlicher Autorschaft. Die grundsätzlich erwünschte Erweiterung von Autorschaft ist häufig simultan mit einer Verminderung in Bezug auf andere Aspekte von Autorschaft bzw. andere Akteure verbunden. Insbesondere sind verschiedene Akteursgruppen in unterschiedlicher Weise betroffen. Die ethische Analyse von Chancen und Risiken etwa von ADM-Systemen in der öffentlichen Verwaltung (vgl. Kapitel 8) muss darauf achten, *für wen* es zu Erweiterungen oder Verminderungen der Autorschaft kommt, etwa in der Differenz von Entscheidern und Betroffenen. Es sind also mit dem Einsatz von KI-Systemen auch Fragen von Gerechtigkeit und Autonomie- bzw. Machtverteilung involviert. Speziell ist zu fragen, ob und wie die jeweiligen Gestalter der entsprechenden Prozesse bzw. der konkreten Software-Applikationen und KI-Algorithmen in die Autonomie und Autorschaft derjenigen eingreifen, die diese Produkte nutzen oder anderweitig von ihnen betroffen sind.

Weiterhin sind psychologische Effekte zu beachten, die spezifisch für digitale Instrumente und insbesondere für KI-Systeme sind. Hier ist vor allem der Automation Bias zu nennen. Menschen vertrauen, so empirische Untersuchungen, algorithmisch erzeugten Ergebnissen und automatisierten Entscheidungsprozeduren häufig mehr als menschlichen Entscheidungen. Vermutlich spielen dabei verbreitete Objektivitätsunterstellungen gegenüber Daten und Rechenverfahren eine Rolle, während menschliche Urteile tendenziell als subjektiv wahrge-

[172] Yeung, K. (2017): 'Hypernudge': Big Data as a mode of regulation by design. In: Information, Communication and Society 20 (1), 118-136 (DOI: 10.1080/1369118X.2016.118671).

nommen werden. Gerade bei Entscheidungen, die mit einer großen prognostischen Unsicherheit konfrontiert sind und zugleich gravierende Auswirkungen haben, besteht die latente Tendenz, den datenbasierten algorithmischen Auswertungen mehr zu vertrauen. Damit wird Verantwortung – zumindest unbewusst – auf diese „Quasi-Akteure" delegiert. Dieser Bias zugunsten der algorithmischen Verfahren kann beispielsweise dazu führen, dass auch bei einer handlungstheoretisch korrekten Organisation von Entscheidungsprozessen, in denen ein KI-System normativ strikt auf die Rolle der Entscheidungsunterstützung begrenzt und dem Mensch, der die Entscheidung trifft, die Verantwortung zugeschrieben wird, das KI-System allmählich in die Rolle des eigentlichen „Entscheiders" gerät und menschliche Autorschaft und Verantwortung ausgehöhlt werden. Bisweilen wird versucht, dieser Gefahr vorzubeugen, indem bei Verwendung eines Entscheidungsunterstützungstools ein entsprechender Warnhinweis gegeben wird. Eine weitere denkbare Vorkehrung wäre die Verpflichtung der entscheidenden Fachkräfte, die etwaige Übernahme des algorithmischen Entscheidungsvorschlages – etwa mit Verweis auf die eigene erfahrungsbezogen intuitive oder kollegial erörterte Prognose – ausdrücklich zu begründen. Auf jeden Fall bedarf dieser Überlappungsbereich normativer Regulierung und empirisch-psychologischer Effekte besonderer Aufmerksamkeit in den ethischen Analysen zu den Anwendungsfeldern in den folgenden Kapiteln.

Menschliche Entscheider haben kaum eine Möglichkeit, die epistemische Evidenz der Korrelationen und Muster kritisch zu beurteilen, sondern sind vielfach darauf angewiesen, sie so zu nehmen, wie sie von den Systemen bereitgestellt werden. Sie unterliegen damit einem verborgenen Nudging durch die Art und Weise, wie die KI-Systeme zu ihren Ergebnissen kommen, und werden in bestimmte Richtungen des Entscheidens gedrängt. Mögliche Einseitigkeiten, etwa auf Basis der Datenlage, sowie daraus möglicherweise resultierende Diskriminierungen geraten aus dem Blick und menschliche Autorschaft wird entleert.

In den Abwägungen zwischen Erweiterung und Verminderung menschlicher Autorschaft in ihrer sozialen Verteilung sind bereits auf einer abstrakten ethischen Ebene mehrere Dimensionen zu berücksichtigen. Erstens bedarf die Übertragung menschlicher Tätigkeiten auf KI-Systeme der Transparenz gegenüber den Betroffenen. Sie sollten darüber informiert sein, auf welche Weise Entscheidungen zustande kommen, von denen sie dann betroffen sind. Dies hat zweitens mit der Klarstellung von Verantwortungszuschreibungen zu tun. Um ein Verantwortungs- und gegebenenfalls auch Haftungsvakuum zu verhindern, muss die Verantwortungszuschreibung etwa über die Betreiber der Systeme oder die menschlichen Akteure, die die Übertragung an sie beschlossen haben, geregelt werden. Drittens bedarf es der Sicherstellung der Nachvollziehbarkeit in Bezug auf das zweckhafte Funktionieren der KI-Systeme. Viertens müssen mögliche

nicht intendierte Folgen wie beispielsweise schleichend einkehrende Abhängigkeiten von den KI-Systemen oder allmähliche Aushöhlung menschlicher Autorschaft sorgfältig beobachtet werden, um gegebenenfalls rechtzeitig korrigierend eingreifen zu können.