LEUPHANA
UNIVERSITÄT LÜNEBURG

Künstliche Intelligenz

Analyse der Möglichkeit eines drohenden Verlusts von Autonomie, Verantwortung, Individualität und Würde

Von der Fakultät Nachhaltigkeit
der Leuphana Universität Lüneburg zur Erlangung des Grades

Doktor der Philosophie
- Dr. Phil. -

genehmigte Dissertation von Detlef Schwarting
geboren am 16. Juli 1966 in Nordenham

Eingereicht am: 27. Juli 2023

Erstgutachter: Prof. Dr. Dr. Nils Ole Oermann,
Leuphana Universität Lüneburg
Zweitgutachter: Prof. Dr. Markus Reihlen,
Leuphana Universität Lüneburg
Drittgutachter: Prof. em. Dr. Wilhelm Vossenkuhl,
Ludwig-Maximilians-Universität München

Als Dissertation eingereicht unter dem Titel:

Künstliche Intelligenz:
Analyse der Möglichkeit eines drohenden Verlusts von
Autonomie, Verantwortung, Individualität und Würde

Kurzfassung

Diese Dissertation verfolgt einerseits das Ziel der umfassenden Sichtung der technischen, psychologischen, philosophischen und neurobiologischen Grundlagen der Künstlichen Intelligenz und damit die Etablierung einer epistemischen und ontologischen Eingrenzung der Technologie und andererseits das Ziel der moralphilosophischen Beurteilung und Risikobewertung. Die zentrale Hypothese unterstellt der Technologie, die überhaupt nur durch die rasanten Fortschritte der Natur- und Ingenieurwissenschaften und damit durch die Aufklärung ermöglicht wurde, eine Unterwanderung der zentralen Prinzipien und Werte der Aufklärung. Zu befürchten ist eine selbstverschuldete Unmündigkeit des Menschen durch die Künstliche Intelligenz. Wesentlich in der Argumentation ist das Aufgeben von menschlicher Autonomie zugunsten einer Technologie, die selbst nicht autonom sein kann, die Verlagerung von Verantwortlichkeiten in einen neu entstandenen rechtsfreien Raum, der Verlust der menschlichen Individualität und letztlich die Preisgabe der Menschenwürde durch den Übergang in ein „*Reich ohne Notwendigkeit*“[1], aus dem es kein Entkommen gibt. Eine Ethik der Künstlichen Intelligenz muss, so die normative Schlussfolgerung, die Wahrung der menschlichen Autonomie, Verantwortlichkeit, Individualität und Würde in den Vordergrund stellen, sodass ein individueller und kollektiver Rückfall des Menschen in die Unmündigkeit verhindert wird.

Schlagwörter: Künstliche Intelligenz, Aufklärung, Mündigkeit, Menschenwürde, Verantwortung, Autonomie, Freiheit, Individuum

[1] Jonas (1979), S. 364

Abstract

This dissertation pursues, on the one hand, the goal of a comprehensive review of the technical, psychological, philosophical, and neurobiological foundations of artificial intelligence and thus the establishment of an epistemic and ontological delimitation of the technology and, on the other hand, the goal of a moral-philosophical assessment and risk evaluation. The central hypothesis assumes that the technology, which itself was only made possible by the rapid progress of the natural and engineering sciences and thus by the Enlightenment, subverts the central principles and values of the Enlightenment. At stake is a self-inflicted immaturity of man through artificial intelligence. Essential in the argumentation is the abandonment of human autonomy in favor of a technology that cannot itself be autonomous, the shifting of responsibilities into a newly created lawless space, the renouncement of human individuality and ultimately an abandoning of human dignity through the transition into a "*realm without necessities*[2]" from which there is no escape. An ethics of artificial intelligence must, so the normative conclusion, put the preservation of human autonomy, responsibility, individuality, and dignity in the foreground. This is the only way to prevent a relapse into individual and collective immaturity.

Keywords: Artificial Intelligence, Enlightenment, Maturity, Dignity, Responsibility, Autonomy, Freedom, Individuality

[2] Vgl. Jonas (1979), S. 364; Übersetzung DS

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Abkürzungsverzeichnis

ACM	Association of Computing Machinery
AGI	Artificial General Intelligence (künstliche allgemeine Intelligenz, „starke KI“)
ANI	Artificial Narrow Intelligence („schwache KI”)
AI	Artificial Intelligence
AIGC	AI-Generated Content
API	Application programming interface (Schnittstelle zur Anwendungsprogrammierung)
ASI	Artificial Superintelligence (“Künstliche Superintelligenz”)
ASSC	Association of Scientific Study of Consciousness
ASW	Außersinnliche Wahrnehmung
ATP	Alltagspsychologie
AWS	Autonomous Weapons Systems
BIS	Berliner Intelligenzstrukturmodell
BMF	Bewusstes mentales Feld
CA	Capability Approach (“Fähigkeitsansatz”)
CNN	Convolutional Neural Network (neuronales Faltungsnetzwerk)
CRA	Chinese Room Argument
ESP	Extra sensory perception (außersinnliche Wahrnehmung, ASW)
GG	Grundgesetz der Bundesrepublik Deutschland
GOFAI	Good old-fashioned artificial intelligence
GPS	General Problem Solver, eines der ersten KI-Programme
GPT	Generative Pre-trained Transformer (generative KI auf Basis vortrainierter Transformatoren)
GPU	Graphics Processing Unit (Graphikprozessor)
ICRC	International Committee of the Red Cross (“Internationales Komitee vom Roten Kreuz”, IKRK)
IDSIA	Istuto Dalle Molle di Studi sull Intelligenza Artificiale (Dalle-Molle-Forschungsinstitut für künstliche Intelligenz)
IHL	International Humanitarian Law (Humanitäres Völkerrecht)
IHRL	International Human Rights Law (Internationales Menschenrechtsgesetz)
IIT	Integration Information Theory of Consciousness
KI	Künstliche Intelligenz
KNN	Künstliches Neuronales Netz (Engl.: „artificial neural network“)
KPI	Key Performance Indicator
KU	Kritik der Urteilskraft (I. Kant)

LAR	Lethal Autonomous Robots/Robotics
LAWS	Lethal Autonomous Weapon Systems
LLM	Large Language Model (großes Sprachmodell)
MECE	Mutually Exclusive and Combined Exhaustive
NCC	Neural Correlates of Consciousness
NGO	Non-Governmental Organisation (“Nichtregierungsorganisation”)
NLP	Natural Language Processing (Verarbeitung natürlicher Sprache)
PAP	Principle of Alternative Possibilities (Grundsatz der alternativen Möglichkeiten)
PPO	Proximal Policy Optimization (Optimierung von Richtlinien durch sukzessive Näherung)
RLHF	Reinforcement Learning from Human Feedback
RM	Reward Model
SFT	Supervised Fine-Tuning
SOI	Structure of Intellect
SVM	Support-Vector-Machine
TNGS	Theory of Neuronal Group Selection
TT	Turing-Test
TTT	Total Turing-Test
v.u.Z.	vor unserer Zeitrechnung

1 Einleitung

Die Künstliche Intelligenz (KI) ist zum allgegenwärtigen und oftmals beherrschenden Thema in den Diskursen zur technologischen, wirtschaftlichen und gesellschaftlichen Entwicklung geworden und gilt dabei entweder als Heilsbringer oder als Bedrohung, sowohl in den Medien als auch in diversen Disziplinen der Wissenschaft[3].

Es kann festgehalten werden, dass die KI neben vielen Vorteilen und Chancen für die Menschheit auch verschiedene bereits existierende oder potenzielle Risiken birgt, die noch nicht ganzheitlich verstanden sind. Grundsätzlich wäre eine allumfassende Untersuchung zu fordern, die Nutzen und Gefahren abwägt und im Blick auf zukünftige Entwicklungen beurteilt[4].

Tatsächlich greifen viele Diskussionen zum Thema singuläre Aspekte der Technologie oder einzelne Anwendungen heraus und leiten normative Urteile ab[5]. Zugrunde gelegte Annahmen, vor allem zu den technischen Zusammenhängen und Wirkungsweisen, sind oft nicht klar. Insbesondere nähern sich viele Untersuchungen dem Thema aus einer monodisziplinären Perspektive. Nicht wenige Geisteswissenschaftler urteilen über die Technologie, ohne sie und ihre Funktionsweise wirklich zu kennen, und kommen zu weitreichenden Schlussfolgerungen, wie zum Beispiel den Postulaten vom „*Ende der Demokratie*“[6] oder dem „*Technologischen Totalitarismus*“[7]. Ingenieure und Informatiker bedienen sich ihrerseits häufig der Begriffe aus Philosophie oder Rechtswissenschaften, ohne sie korrekt anzuwenden, und ziehen teils fragwürdige Schlussfolgerungen. Als Beispiel ein Zitat von Ray Kurzweil:

> *„Die Singularität erlaubt es uns, diese Beschränkungen unserer biologischen Körper und des Gehirns zu überschreiten. Wir gewinnen Kraft über unsere Schicksale. Unsere Sterblichkeit wird in unseren Händen liegen. Wir werden in der Lage sein, so lange zu leben, wie wir wollen (eine etwas andere Feststellung als die Aussage, dass wir ewig leben werden). Wir werden das menschliche Denken vollständig verstehen und seine Reichweite erheblich ausdehnen. Am Ende des Jahrhunderts wird der nichtbiologische Teil unserer Intelligenz Aberbillionen Mal leistungsfähiger sein als die reine menschliche Intelligenz.“*[8]

---

[3] Vgl. Ouchchy et al. (2020); Standing Committee of the One Hundred Year Study of Artificial Intelligence (2021), S. 33f

[4] Vgl. Heinlein Huchler (2022); Zitat in “*How to worry about artificial intelligence”, The Economist.* 20. April 2023: „*This powerful technology poses new risks, but also offers extraordinary opportunities. Balancing the two means treading carefully. A measured approach today can provide the foundations on which further rules can be added in future. But the time to start building those foundations is now.*”

[5] Vgl. Bartneck Lütge (2019); Boddington (2017); Misselhorn (2018); Nida-Rümelin Weidenfeld (2018); Vallor (2016)

[6] Vgl. Hofstetter (2016), Titel: „*Das Ende der Demokratie. Wie die künstliche Intelligenz die Politik übernimmt und uns entmündigt*“

[7] Vgl. Schirrmacher (2015); Augstein (2017)

[8] Russell Norvig (2004), S. 1196-97; im Original: Kurzweil (2005), S. 9

In dieser Aussage stecken diverse Irrtümer, Fehlannahmen und begriffliche Unschärfen, die neben anderen in dieser Arbeit geklärt werden sollen.

Generell ist die Liste der falsch und inkonsistent verwendeten Termini lang, beginnend beim Begriff der Autonomie, den Ingenieure anders verstehen als Philosophen.

In dieser Arbeit soll die Frage geprüft werden, ob sich der Mensch aufgrund der sich immer weiter ausbreitenden KI in nahezu allen Bereichen des individuellen und kollektiven Lebens in eine "*selbstverschuldete Unmündigkeit*" begibt. Diese aus dem Aufsatz „*Was ist Aufklärung?*"[9] von Immanuel Kant stammende Formulierung ist bewusst gewählt. Die Aufklärung verfolgte u.a. das Ziel, den Menschen zum „Selbstdenken" zu befreien („*Sapere aude! Habe Mut, dich deines eigenen Verstandes zu bedienen!*"[10]) und ermöglichte die freie Entfaltung der Wissenschaften, die einen nie zuvor erreichten technologischen und gesellschaftlichen Fortschritt anstieß und letztlich auch zur Entwicklung der Computertechnologie und speziell der Künstlichen Intelligenz führte.

Wie Steven Pinker es darstellt:

> *"Provoked by challenges to conventional wisdom from science and exploration, mindful of the bloodshed of recent wars of religion, and abetted by the easy movement of ideas and people, the thinkers of Enlightenment sought a new understanding of the human condition. The era was a cornucopia of ideas, some of them contradictory, but four themes tie them together reason, science, humanism, and progress."*[11]

Die vier miteinander verwobenen Begriffe Vernunft, Wissenschaft, Humanismus und Fortschritt fassen nach Pinker bis heute das mit das durch die Aufklärung Erreichte zusammen. Die Aufklärung impliziert den mündigen selbstdenkenden Menschen. Ausgerechnet die Künstliche Intelligenz, die in hervorragender Weise illustriert, zu welchen Fortschritten die vernunftbegabte Wissenschaft in der Lage ist, verursacht nun im Konzert mit anderen Entwicklungen die sukzessive Einschränkung der Mündigkeit des Menschen, womit sich aufklärungskritische Annahmen und Vorahnungen erfüllen.

Es soll geprüft werden, ob das Individuum immer weniger zu mündigem, unabhängigem, kritisch-reflektierendem Denken fähig ist und unwissentlich bei der Realisierung einer neuen Normalität mitwirkt, indem es die Systeme der KI mit Daten füttert und sich dann mit zunehmender Routine auf die Ergebnisse verlässt und darauf vertraut, dass sie bequem und nützlich sind[12]. Einige Publikationen weisen darauf hin und verwenden ähnliche Begrifflichkeiten: „*Wie die künstliche Intelligenz die Politik übernimmt und uns*

9 Kant (1784b), S. 20-27

10 Kant (1784b), S. 20

11 Pinker (2018), S. 7

12 Galloway (2017), S. 171ff

entmündigt“[13], „*Selbstermächtigung in der digitalen Weltordnung*“[14] und „*Das Zeitalter des Überwachungskapitalismus*“[15]. Darauf soll in der vorliegenden Arbeit eingegangen werden. Vor allem soll die Argumentation klarer in der Aufklärung[16] und auch in der Aufklärungskritik[17] verankert werden.

Die Haupthypothese dieser Arbeit lautet also:

> **Der Mensch begibt sich mit der KI (erneut) in eine selbstverschuldete Unmündigkeit – die durch die Aufklärung freigesetzten Mechanismen wenden sich gegen sie selbst.**

Vier unterstützende Hypothesen sind zunächst zu untersuchen. Erstens gibt der Mensch im Rahmen der Entwicklung und Einführung der Künstlichen Intelligenz Autonomie an ein Medium ab, das selbst nicht autonom sein kann und niemals über Willensfreiheit und Autonomie verfügen wird, sondern im philosophischen Sinne determiniert, fremdbestimmt und damit heteronom ist. Von einer Selbstgesetzgebung der KI kann keine Rede sein.

Aus dem „Nicht-anders-Können“ ergibt sich zweitens eine Verantwortungslücke. Die KI wird nicht von Gründen geleitet und handelt daher auch nicht im (rechts-) philosophischen Sinne, so dass sie auch keine Verantwortung übernehmen kann. Aufgrund der Intransparenz der tatsächlichen Abläufe in den Systemen der KI kann die Verantwortlichkeit auch nicht bis zu verantwortlichen menschlichen Individuen zurückverfolgt und eindeutig zugeordnet werden. So entsteht eine Form des rechtsfreien Raums, eine Verantwortungslücke.

Drittens verliert der Mensch in einer zunehmend von der KI dominierten Welt immer mehr und unwiderruflich seine Urteilsfähigkeit und Verantwortlichkeit als Person und Individuum, sein Urrecht (gem. Fichte) als alleinige Ursache in der Sinnenwelt, seine Mündigkeit und damit die wichtigste Errungenschaft der Aufklärung und letztlich (und viertens) seine Menschenwürde. Er begibt sich aus einer Kombination von Überforderung und Bequemlichkeit in ein „*Reich ohne Notwendigkeiten*“[18], verliert seinen Subjektcharakter, nimmt Demütigung und Unterdrückung hin und gibt wichtige Fähigkeiten preis, die ein Leben in Würde ermöglichen.

Dieses Argument stützt sich auf das Grundkonzept der Aufklärung, wonach die Menschenwürde untrennbar verbunden ist mit der menschlichen Selbstbestimmung, Autonomie sowie der Erkenntnis und Gestaltung unserer Lebenswelt durch den Menschen für

---

[13] Hofstetter (2016)

[14] Augstein (2017)

[15] Zuboff (2018b)

[16] Schneiders (1997); Fleischacker (2013)

[17] Horkheimer (1944); Horkheimer Adorno (1946); Helbing (2019)

[18] Jonas (1979), S. 364

den Menschen. Eine wichtige Voraussetzung hierfür ist, so die Aussage in dieser Arbeit, das Subjektsein des Menschen.

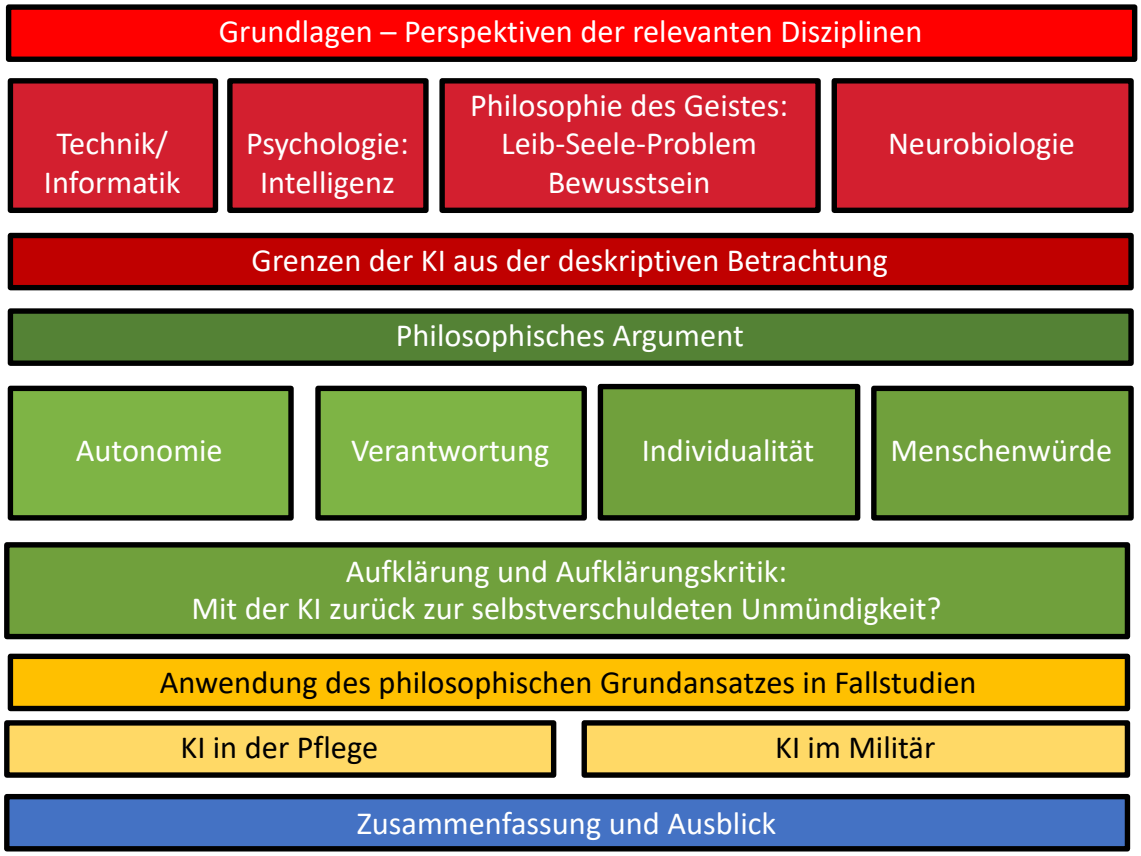

Abbildung 1: Strukturübersicht

Auf dem Weg zur Diskussion der Haupt- und unterstützenden Hypothesen soll in einem ersten Schritt im Sinne einer Propädeutik das Fundament einer soliden und wissenschaftlichen Bestandsaufnahme der Arbeitsergebnisse, Diskurse und Entwicklungen in den relevanten Disziplinen, insbesondere in der Technik, Psychologie, Geistesphilosophie und den Neurowissenschaften, gelegt werden. Dabei sollen auch Irrwege und Sackgassen der Entwicklung und des wissenschaftlichen Diskurses nachgezeichnet und beurteilt werden.

Damit einhergehend sollen einige weitverbreitete Irrtümer zur KI klar als falsch erkannt und widerlegt werden. Der erste Irrtum geht auf einen alten Menschheitstraum zurück: der Vorstellung, *es gebe eine Maschinenintelligenz, die der natürlichen Intelligenz des Menschen in jeglicher Hinsicht vergleichbar oder gar überlegen sei.*[19] Es zeichnet sich klar ab, dass das gesamte Spektrum der menschlichen Intelligenz bzw. des menschlichen Denkens außerhalb der Reichweite der Algorithmen der KI liegt. In anderen Worten: Die

[19] Vgl. Kurzweil (2005): „*The singularity is near*"; Tegmark (2017), S.87: *"Die herkömmliche Weisheit unter KI-Forschern lautet, Intelligenz laufe letztlich auf Informationen und Rechnen hinaus und habe nichts mit Fleisch, Blut oder Kohlenstoffatomen zu tun. Das heißt, es gibt keinen vernünftigen Grund, warum Maschinen eines Tages nicht mindestens so intelligent sein können wie wir."*

menschliche Intelligenz umfasst deutlich mehr als das, was Algorithmen der KI zu leisten in der Lage sind oder werden sein.

Der zweite Irrtum besteht in der Auffassung, die KI könne aus sich selbst heraus und völlig eigenständig (also nicht als Werkzeug des Menschen) kreativ und erkenntniserweiternd sein. Es wird gezeigt werden, dass die KI zu einer „transformativen Kreativität", in der neues Wissen „aus dem Nichts heraus" entsteht, nicht in der Lage ist. Immerhin kann sie als Werkzeug der Wissenschaft dazu dienen, Hypothesen für weitere wissenschaftliche Arbeit zu generieren.

Im dritten Irrtum geht es um die Willensfreiheit. Einige Wegbereiter der KI hegen die Vorstellung, Roboter und Agenten der KI verügten über einen „freien Willen" (z.B. John McCarthy[20]). Es wird gezeigt werden, dass die KI immer vollständig synthetisch determiniert ist und niemals von ihrem Algorithmus abweichen kann. Das heißt nicht, dass sie immer analytisch determiniert ist, ihre Aktionen immer prognostizierbar und vergangenheitsunabhängig sind. Jede Kontingenz wird mit den Methoden der Wahrscheinlichkeitsrechnung überwunden. Die KI verfügt über keinen freien Willen.

Der vierte Irrtum bildet den Gegenpart des dritten: Einige Vertreter aus der Hirnforschung behaupten, dass es den freien Willen des Menschen nicht gebe und der Mensch, genauso wie jeder Computer, vollständig determiniert sei[21]. Auch dies wird überzeugend widerlegt werden.

Der fünfte Irrtum betrifft den Computer-Funktionalismus bzw. den reduktiven Physikalismus als die einzig plausible Lösung des Leib-Seele-Problems[22]. Einige Gedankenexperimenten und Argumente sollen präsentiert werden, die dieses Verständnis komplett zurückweisen.

Der sechste Irrtum betrifft das menschliche Bewusstsein und die Vorstellung, dass es mit all seinen phänomenalen und intentionalen Komponenten „naturalisierbar" sei. Die induktive – und damit falsifizierbare – Widerlegung dieses Irrtums wird zeigen, dass sämtliche bisherigen Versuche der Naturalisierung[23] oder Digitalisierung des Bewusstseins erfolglos geblieben sind.

Gemäß dem siebten Irrtum gebe es ein klares und differenziertes Verständnis des menschlichen Bewusstseins und liefere konkrete Ansätze für den Nachbau des menschlichen

---

[20] Vgl. McCarthy (2000), S. 341f: „*Free will does not require a very complex system. Young children and rather simple computer systems can represent internally 'I can, but I won't' and behave accordingly*."

[21] Vgl. Singer (2004); Prinz (2004); Roth (2004)

[22] Vgl. Putnam (1960)

[23] In Anlehnung an Zahawi (2008), S. 142 ist mit der Naturalisierung des Bewusstseins der Versuch gemeint, das Bewusstsein mit nicht-bewussten Mechanismen und Prozessen zu erklären, so wie die Naturalisierung der Intentionalität die Erklärung mit nicht-intentionalen Mechanismen und Prozessen meint

Gehirns auf einem anderen Substrat[24]. Die Untersuchungen werden zeigen, dass es trotz immenser Anstrengungen in den Naturwissenschaften keine empirische Erklärung für das Zustandekommen des menschlichen Bewusstseins gibt.

Diese sieben Irrtümer betreffen auf einer übergreifenden Ebene zwei fundamentale Kernfragen zu den Grenzen der Künstlichen Intelligenz: diejenige der menschlichen Freiheit in allen Ausprägungen und Konsequenzen und diejenige des menschlichen Bewusstseins. Auf diese beiden Kernfragen wird in der Arbeit wiederholt eingegangen.

Im zweiten Schritt erfolgt der Übergang von der Darstellung der epistemischen und ontologischen Basis zu normativen Schlussfolgerungen in den Bereichen Autonomie/Freiheit, Verantwortung, Individualität/Person und Menschenwürde und schließlich die Prüfung der aufklärungskritischen Haupthypothese. Wichtig wird an dieser Stelle der Rückgriff auf aufklärungskritische Argumente der letzten zweihundert Jahre sein. Die grundsätzliche Überlegung wird zu einer Gratwanderung der Wahrung der unbestreitbaren positiven Effekte der Aufklärung bei gleichzeitiger nachhaltiger Sicherung des Freiheitsmodells der Aufklärung, insbesondere der Mündigkeit des Menschen.

Abschließend wird in zwei konkreten Fallstudien die Anwendung der KI untersucht: Zum einen in der Kranken- und Altenpflege und zum anderen in Waffensystemen des Militärs.

Methodisch soll in dieser Arbeit eine Herangehensweise des wissenschaftlichen Realismus verfolgt werden mit einer klaren Abgrenzung gegenüber der ausschließlich am Empirismus orientierten erkenntnistheoretischen Position des Positivismus einerseits und derjenigen des Konstruktivismus andererseits, demzufolge alle Objekte der sozialen Realität menschliche Konstrukte seien. Damit erfolgt auch eine klare Unterscheidung vom Anthropozentrismus (wonach die materielle Welt auf ihre Bedeutung für die menschliche Welt reduziert wird) und vom Idealismus des Primats der Ideen über die materielle Welt[25].

Der Philosoph Mario Bunge charakterisiert Realismus in seiner allgemeinen Form wie in dem folgenden Zitat:

> „*the view that the external world exists independently of our sense experience, ideation, and volition, and that it can be known*"[26]

[24] Vgl. „Das Manifest der Hirnforscher", Gehirn&Geist (2004)

[25] Vgl. Reihlen et al. (2022), S. 56f

[26] Zitiert in ebd.; Originalzitat: Bunge (1993), S. 229; Ausführlich: *„Philosophical realism, or objectivism, is the view that the external world exists independently of our sense experience, ideation, and volition, and that it can be known. The first conjunct is an ontological thesis while the second is an epistemological one.* [...] *The ontological thesis of realism can be stated thus: there are things in themselves. Its epistemological companion can be restated as follows: we can know things in themselves (not just as they appear to us)."*

In der verfeinerten Form des wissenschaftlichen Realismus verbirgt sich die Grundannahme, dass wissenschaftliche Forschung – obwohl nicht unfehlbar – den besten methodischen Zugang zur Darstellung der Welt vermittle[27].

Das Wesen dieser Arbeit ist eine philosophische Reflektion der KI mit der Vorgehensweise eines problematisierenden Reviews, inspiriert von Mats Alvesson und Jörgen Sandberg:

> "*Aim of the problematizing methodology: Generating novel research questions through a dialectical interrogation of one's own familiar position, other stances, and the literature domain targeted for assumption challenging.*"[28]

> "*As a methodology, the problematization methodology also contributes to more reflective scholarship in the sense that it counteracts or supplements the domination of gap-spotting as a research ideal. As a methodology, it encourages us to produce more novel research questions and theories by actively questioning and critically scrutinizing established knowledge in academia and society"*[29]

Ziel dieser Arbeit ist es, eine *dialektische* und ergebnisoffene Befragung zum Thema Künstliche Intelligenz in Bezug auf deren Implikationen für die Menschheit vorzulegen und damit das derzeitige Denken herauszufordern, in der Hoffnung, weitere Analysen und Diskurse auszulösen. Bei allen behandelten wissenschaftlichen und erkenntnishistorischen Zusammenhängen besteht der Anspruch, diese so detailliert wie nötig für das grundsätzliche Verständnis von Lesern aus entlegenen Disziplinen und so knapp wie möglich für das ganzheitliche Bild des betrachteten Gegenstands dieser Arbeit darzustellen.

Gerade die aktuellen Diskussionen im Jahr 2023 zu der neu eingeführten Applikation ChatGPT[30] der Firma Open AI unterstreichen die Notwendigkeit eines fundierten wissenschaftlichen und philosophischen Diskurses. Generell sollte festgehalten werden, dass die Mechanismen und Funktionalitäten dieser Technologie nicht grundsätzlich neu sind. Neu ist vielmehr, dass Open AI eine leistungsfähige Applikation erstmals einer breiten Öffentlichkeit unentgeltlich zur Verfügung gestellt hat. Die Kombination der inhaltlichen Breite und Tiefe mit einer ausgeklügelten Sprachausgabe mit hoher syntaktischer Qualität hat bei vielen Kommentatoren und Benutzern sowohl Bewunderung und Begeisterung hervorgerufen als auch Nachdenklichkeit, Besorgnis und Verunsicherung. Im öffentlichen Diskurs stellen sich viele der in dieser Arbeit diskutierten Fragen. Inwieweit versteht die eingesetzte Technologie die behandelten Themen? Aus welchen Quellen bezieht sie

---

[27] Vgl. Bunge (1993), S. 231

[28] Alvesson Sandberg (2011), S. 260, Figure 1; zu den Prinzipien des problematisierenden Reviews bei Alvesson Sandberg (2020), S. 1297: „ *... problematizing review is based on four core principles: the ideal of reflexivity, reading more broadly but selectively, not accumulating but problematizing, and the concept of 'less is more'* "

[29] Alvesson Sandberg (2011), S. 267

[30] Generative Pre-trained Transformer

die Inhalte? Welche konkreten Auswirkungen ergeben sich zum Beispiel für Wissenschaft und Lehre, für den Journalismus, für Kunst und Kultur oder für die Softwareentwicklung? Welche Gefahren des Missbrauchs bestehen? Auch gibt es dazu erste kritische Stimmen aus der Philosophie, die davor warnen, dass „*unsere humane Existenz in Mitleidenschaft*“ gezogen werden könnte[31]. Selbst einige Vertreter der Technologiebranche kommentieren die Entwicklung der generativen KI und ihrer Gefahren durchaus kritisch, so zum Beispiel der Vorstandsvorsitzende von Google, Sundar Pichai[32] oder einer seiner Vorgänger, Eric Schmidt. Der Historiker Yuval Noah Harari fürchtet gar ein „Hacking des Betriebssystems der menschlichen Zivilisation“ durch generative KI[33]. In die gleiche Richtung argumentiert auch der Linguist Noam Chomsky in einem Gastbeitrag in „The New York Times“ vom März 2023, in dem er die Befürchtung äußert, die künstliche Intelligenz werde die Wissenschaft degradieren und die Ethik entwürdigen[34]. Zuletzt forderte sogar der Vorstandsvorsitzende von Open AI, Sam Altman, eine Regulierung der Künstlichen Intelligenz[35].

[31] Sadin (2023): „*Zunächst müssen wir uns klarmachen: Es geht bei der Digitalisierung um einen Vorgang, der **unsere humane Existenz in Mitleidenschaft** zieht. Es geht um die Beeinflussung von Fähigkeiten, von denen die gute Entfaltung eines jeden von uns abhängt. Nachdem wir eine Aushöhlung unserer Urteilsfähigkeit durch die zunehmende Verbreitung von digitalkapitalistischen Systemen, die unseren Alltag bestimmen, erleben, geht es nun um eine Attacke auf unser Sprachvermögen.“*

[32] Vgl. https://www.theguardian.com/technology/2023/apr/17/google-chief-ai-harmful-sundar-pichai

[33] Vgl. „The Economist” vom 28. April 2023: https://www.economist.com/by-invitation/2023/04/28/yuval-noah-harari-argues-that-ai-has-hacked-the-operating-system-of-human-civilisation

[34] Vgl. Zitat aus “Noam Chomsky: The False Promise of ChatGPT”, *The New York Times*. 8. März 2023: *“…we fear that the most popular and fashionable strain of A.I. — machine learning — will degrade our science and debase our ethics by incorporating into our technology a fundamentally flawed conception of language and knowledge.*“

[35] Vgl. „Open AI’s Sam Altman Urges AI Regulation in Senate Hearing”, *The New York Times*. 16. Mai 2023

2 KI: Definition, Geschichte und Funktion

In der Geschichte der KI haben sich immer wieder Perioden höchster Euphorie und Etappen der Stagnation und des Bedeutungsrückgangs abgelöst.[36]

Die derzeitige Etappe kann eher als euphorisch bezeichnet werden, erstens in Bezug auf die technologischen Möglichkeiten, zweitens in Bezug auf die Anwendungsfelder und drittens hinsichtlich der damit verbundenen ökonomischen Potentiale der Künstlichen Intelligenz. Die KI „*beherrscht längst unser Leben*“[37], allerdings ist dies vielen Menschen immer noch nicht bewusst. Armbanduhren, die unsere Gesundheitsdaten erfassen, Autos, die mit uns sprechen, Call-Center, die unsere Telefonanrufe entgegennehmen, Roboter, die den Rasen mähen oder Staub saugen, und Roboter in der Landwirtschaft, die Kühe melken oder Traktoren fahren. Das sind nur einige wenige Anwendungen. Neben der Euphorie bei vielen Nutzern machen sich auch Verunsicherung und sogar Angst breit, denn viele fragen sich, welche Prozesse die KI noch übernehmen wird. Ziel dieser Arbeit ist es, etwas systematischer die verschiedenen Aspekte der KI zu erkunden, um auch mehr Klarheit in die diffuse und widersprüchliche Wahrnehmung dieser Technologie zu bringen.

Ein ethisches Urteil über die Künstliche Intelligenz kann nur auf Basis eines fundierten Verständnisses der technischen Grundlagen der Technologie sowie ihrer Möglichkeiten und Grenzen erfolgen. Daher ist es das Ziel dieses ersten Kapitels, in einem interdisziplinären Ansatz – aus der Perspektive der Technik, ihrer Historie, ihrer Paradigmen und ihrer Wirkzusammenhänge – zu beschreiben, was man heute als „Künstliche Intelligenz“ bezeichnet. Philosophische Betrachtungen zum Beobachtungsgegenstand werden in diesen ersten Abschnitten bewusst ausgeklammert. Voraussetzung für die im weiteren Verlauf anzustellenden normativen Betrachtungen ist ein faktenbasiertes Verständnis der technischen Zusammenhänge ohne jegliche Metaphysik.

[36] Vgl. Haenlein Kaplan (2019), S. 7; Görz Schneeberger Schmid (2003), S. 5f; Ilkou Koutraki (2020), S. 2

[37] Mainzer (2015), S. VII

2.1 Was ist Künstliche Intelligenz?

> *„Ziel der KI ist es, Maschinen zu entwickeln, die sich verhalten, als verfügen sie über Intelligenz“*
> John McCarthy, 1955[38]

Genauso wie der Begriff der Intelligenz, um den es in einem späteren Abschnitt gehen wird, ist auch derjenige der Künstlichen Intelligenz (KI) nicht klar und einvernehmlich definiert. Schon seit den frühesten Tagen der KI-Entwicklung sind die Bandbreite der vorgelegten Definitionen sowie die Ansprüche und Erwartungen an die Möglichkeiten der KI breitgefächert.

Sehr umfassend hat sich Stuart Russell mit der Definition der KI beschäftigt. In einer 2x2 Matrix stellt er einerseits in der Horizontalen Denken und Handeln gegenüber und andererseits in der Vertikalen die Orientierung am Menschen insgesamt und nur die Rationalität desselben.

Tabelle 1: Einige Definitionen künstlicher Intelligenz, angeordnet in vier Kategorien nach Stuart Russell[39]

Menschliches Denken	***Rationales Denken***
„Das spannende, neuartige Unterfangen, Computern das Denken beizubringen, ... Maschinen mit Verstand im wahrsten Sinne des Wortes.“ (Haugeland, 1985)	*„Die Studie mentaler Fähigkeiten durch die Nutzung programmiertechnischer Modelle.“* (Charmiak und McDermott, 1985)
[Die Automatisierung von] „*Aktivitäten, die wir dem menschlichen Denken zuordnen, Aktivitäten wie beispielsweise Entscheidungsfindung, Problemlösung, Lernen ...*“ (Bellman, 1978)	*„Das Studium derjenigen mathematischen Formalismen, die es ermöglichen, wahrzunehmen, logisch zu schließen und zu agieren.“* (Winston, 1992)
Menschliches Handeln	***Rationales Handeln***
„Die Kunst, Maschinen zu schaffen, die Funktionen erfüllen, die, werden sie von Menschen ausgeführt, der Intelligenz bedürfen.“ (Kurzweil, 1990)	*„Computerintelligenz ist die Studie des Entwurfs intelligenter Agenten.“* (Poole et al., 1998)
„Das Studium des Problems, Computer dazu zu bringen, Dinge zu tun, bei denen ihnen momentan der Mensch noch überlegen ist.“ (Rich und Knight, 1991)	*„KI ... beschäftigt sich mit intelligentem Verhalten in künstlichen Maschinen.“* (Nilsson, 1998)

[38] Ertel (2008), S. 1; zweite Quelle: https://www.industry-analytics.de/ki-wann-sind-maschinen-intelligent/

[39] Struktur und Zitate übernommen aus Russell Norvig (2004), S. 23; Originalzitate nicht verifizierbar

Die umfassendste Definition setzt beim menschlichen Denken an, dem „*Ansatz der kognitiven Modellierung"*[40]. Die grundsätzliche Annahme, dass Maschinen wie Menschen denken, setzt indes voraus, dass wir detailliert verstehen, wie der Denkprozess im Menschen funktioniert. Wir werden sehen, dass wir von der Erfüllung dieser Prämisse weit entfernt sind. Empirische Erkenntnisse der Neurobiologie helfen in einigen Bereichen bei der Ausgestaltung von Algorithmen der KI, und umgekehrt befruchtet die KI-Forschung mit ihren Modellen die Arbeit der Neurobiologie.

Die etwas enger gefasste Definition orientiert sich nicht am menschlichen Denken, sondern am menschlichen Handeln. Zentral ist hier der Ansatz mit dem Turing-Test (TT), der von Alan Turing 1950 als „*Imitation Game*"[41] entwickelt und seitdem in der KI Forschung immer weiter verfeinert wurde. Ein Computer besteht den TT, wenn ein Mensch von ihm schriftliche Antworten auf seine ebenfalls schriftlich übermittelten Fragen erhält und nicht unterscheiden kann, ob diese von einem anderen Menschen stammen oder von einem Computer. Nach Stuart Russell[42] sollte ein Computer dafür vier Voraussetzungen erfüllen: Erstens die Fähigkeit zur Verarbeitung der „*natürlichen Sprache des Menschen"* (als Empfänger und Sender), zweitens eine „*Wissensrepräsentation"* für die Speicherung seines Wissens und Verständnisses, drittens ein „*automatisches logisches Schließen"*, so dass er anhand der Fragen, seines eigenen Wissens und seiner Schlussfolgerungen die Antworten entwickeln kann; und viertens ein „*Maschinenlernen, um sich an neue Umstände anzupassen sowie Muster zu erkennen und zu extrapolieren"*. In der Weiterentwicklung des TT zum „*totalen Turing Test*" kamen noch die beiden Eigenschaften der „*Computervision"* und *Robotik* hinzu.

Allgemeinere Definitionen setzen bei der Rationalität an und verwenden bewusst nicht den Menschen als Referenz. Beim Ansatz des rationalen Denkens geht es um „Denkregeln", wie zum Beispiel die von Aristoteles entwickelten Syllogismen.

Eine deutlich enger gefasste Definition fokussiert auf das rationale Handeln. Dies ist der Ansatz des rationalen Agenten, der das bestmögliche Ergebnis erzielt. Russell schreibt hierzu:

> *„Gegenüber den anderen Ansätzen hat das Konzept des rationalen Agenten zwei Vorteile. Erstens ist es allgemeiner als der Ansatz der „Denkregeln", weil korrektes Schlussfolgern nur einen Mechanismus von vielen zur Erzielung der Rationalität darstellt. Zweitens ist es eher geeignet, den Fortschritt zu befördern als Ansätze, die sich menschliches Handeln zum Vorbild nehmen oder auf menschlichen Gedanken basieren."*[43]

---

40 Russell Norvig (2004), S. 24

41 Turing (1950), S. 433

42 Alle (folgenden) Zitate, Russell Norvig (2004), S. 23f

43 Russell Norvig (2004), S. 26

Klaus Mainzer orientiert sich in seiner „Arbeitsdefinition“ ebenfalls am rationalen Handeln:

> *„Ein System heißt intelligent, wenn es selbstständig und effizient Probleme lösen kann. Der Grad der Intelligenz hängt vom Grad der Selbstständigkeit, dem Grad der Komplexität des Problems und dem Grad der Effizienz des Problemlösungsverfahrens ab.“*[44]

Rationalität in diesem Verständnis beinhaltet auch heuristische Vorgehensweisen, die Wiederholung und Verfeinerung dessen, was in der Vergangenheit schon funktioniert hat. Auch „Trial-and-Error“ gehört zu den heuristischen Verfahrensweisen. Diese von Realismus geprägte Definition von Russell und Mainzer wird bei weitem nicht von allen KI-Forschern geteilt, insbesondere nicht in Bezug auf den Anspruch für die Zukunft.

[44] Mainzer (2015), S. 3

2.2 Geschichte der KI

Wolfgang Ertel hat in seinem „*Grundkurs Künstliche Intelligenz*“[45] die wesentlichen Entwicklungspfade der KI übersichtlich dargestellt. Bei ihm sind es sieben Pfade[46]: der von Gödel und Turing begründete Pfad auf Basis der „*Prädikatenlogik“*, das „probabilistische Schließen“, das „*Lernen mit Entscheidungsbäumen*“, die „*Fuzzy-Logik*“, die „Aussagenlogik“, die „neuronalen Netze“ und die „*Support-Vektor-Maschinen*“. Auf der obersten Ebene unterscheidet er zwischen der *symbolischen und numerischen Repräsentation* die beiden Paradigmen der Künstlichen Intelligenz. Dementsprechend wird in der Literatur oft zwischen symbolischer und neuronaler KI bzw. zwischen symbolischer und subsymbolischer KI[47] unterschieden. Beim Paradigma der symbolischen KI operieren Algorithmen, Wörter, Begriffe oder andere Variablen, die für Menschen eine Bedeutung haben, nach vorgegebenen logischen Regeln und Abläufen. Das Paradigma der numerischen oder subsymbolischen KI ist inspiriert von den Neurowissenschaften und angenommenen Abläufen im menschlichen Gehirn. Die dazugehörigen Algorithmen sind mathematische Operationen, die für den Laien in der Regel nicht nachvollziehbar sind[48].

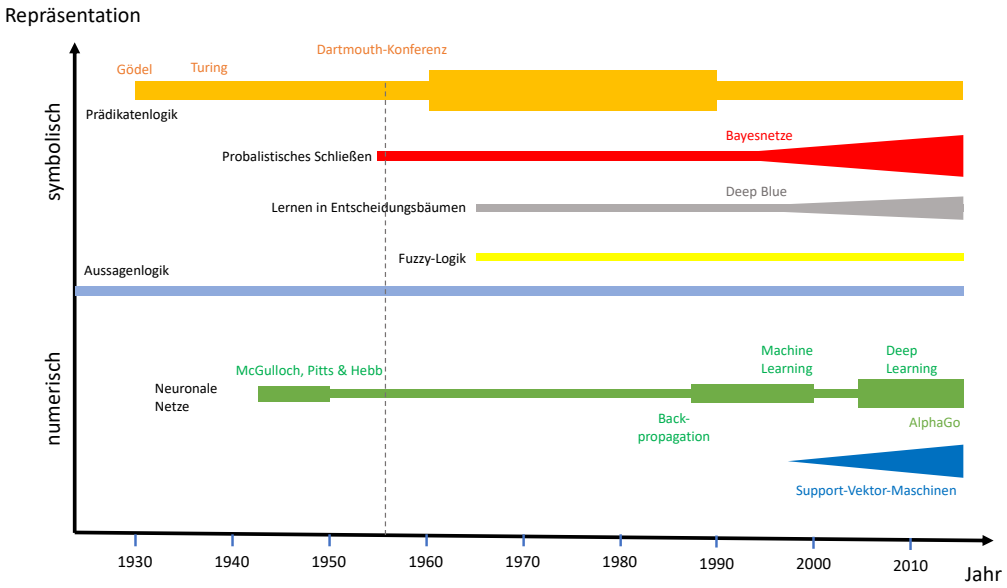

Abbildung 2: Die Geschichte verschiedener KI-Ausrichtungen[49]

In den folgenden Abschnitten wird auf einige der dargestellten Meilensteine und Teilgebiete der KI Bezug genommen.

[45] Ertel (2008)

[46] Vgl. Ertel (2008), S. 8

[47] Vgl. Mitchell (2019), S. 21f

[48] Ebd.

[49] Nachgezeichnet und angelehnt an Ertel (2008), S. 8

2.2.1 Ursprünge

Die Ursprünge der heute diskutierten Künstlichen Intelligenz liegen in den 1930er bis 1950er Jahren. Vier Meilensteine aus jener Zeit sind bis heute prägend für die Entwicklung und begriffliche Positionierung dieser Technologie: Erstens Kurt Gödels Unvollständigkeitssätze aus den 1930er Jahren; zweitens die erste Arbeit, die heute inhaltlich als KI anerkannt wird, von Warren McCulloch und Walter Pitts zur Modellierung neuronaler Netze von 1943; drittens der im vorherigen Kapitel bereits angesprochene Artikel von Alan Turing zum Denken der Maschinen aus dem Jahr 1950; und viertens die Begriffsbildung der „*Artificial Intelligence*“ am Dartmouth College von 1956.

2.2.1.1 Gödel

Systeme der Künstlichen Intelligenz wie auch alle Computer werden von Algorithmen gesteuert. Schon seit den frühen Tagen der Technologie stellte sich die Frage, „*ob alles mathematische Denken sich in Algorithmen darstellen lässt“. Auch* [viele] *„Beweise lassen sich algorithmisch formulieren. In diesem Sinne entwickelte der deutsche Mathematiker David Hilbert (1862 – 1943) das Programm, die Widerspruchsfreiheit der Mathematik – zunächst der Arithmetik – durch endliche „finite“ mathematische Operationen, zu beweisen.“*[50]

„Die Leistung des“ Mathematikers Kurt Gödel *„aus dem Jahre 1931 bestand in seinem Nachweis, dass die Widerspruchsfreiheit der Arithmetik nicht mit finiten Methoden bewiesen werden kann“*[51]. Dies formulierte er in zwei nach ihm benannten Unvollständigkeitssätzen (als Theoreme):

> ***Theorem 1****: „Es gibt in der Sprache von S (formalisierte Peano-Axiome und Typentheorie) korrekt gebildete arithmetische Prädikate F(x), deren Zutreffen/Nichtzutreffen für gewissen Werte m ∈ IN unentscheidbar ist, falls S konsistent ist.“*[52]
>
> Oder:
> *„Jedes Z umfassende formale System S mit endlich vielen Axiomen und der Einsetzungs- und Implikationsregel als einzigen Schlussprinzipien ist unvollständig, d.h. es gibt darin Sätze (und zwar Sätze aus Z), die aus den Axiomen von S unentscheidbar sind, vorausgesetzt, daß S ω-widerspruchsfrei ist. Dabei heiße ein System ω-widerspruchsfrei, wenn für keine*

---

50 Achtner (2006). S. 1; Ergänzende Anmerkung: Hilbert war mindestens bis in die 1930er Jahre der Überzeugung, dass es „*keinen Ignorabismus* „gebe, weder für die Mathematik noch für die Naturwissenschaften. Zitat: „*Der wahre Grund, warum es nicht gelang, ein unlösbares Problem zu finden, besteht nach meiner Meinung darin, daß es unlösbare Probleme überhaupt nicht gibt. Statt des törichten Ignorabismus heiße es im Gegenteil unsere Losung: wir müssen wissen, wir werden wissen*.“. Damit stellt er sich direkt gegen den berühmten Ausspruch von Emil Du Bois-Reymond: **Ignoramus et Ignorabimus** („*Wir wissen es nicht und wir werden es niemals wissen*“), Quelle: Ewers (2016), S. 57, [Hervorhebung DS]

51 Achtner (2006), S. 1f

52 Scholz (2006), S. 30

> *Eigenschaft F natürlicher Zahlen zugleich (Ex) $\overline{F(x)}$ und sämtliche Formeln F(i), i=1,2, ... usw. beweisbar sind"*[53]
>
> **Theorem 2:** *„In jedem solchen System S ist insbesondere die Aussage, dass S widerspruchsfrei ist (genauer die mit ihr äquivalente arithmetische Aussage, welche man erhält, indem man die Formeln ein-eindeutig auf natürliche Zahlen abbildet), unbeweisbar."*[54]

In Kürze heißt dies, dass nicht alle wahren arithmetischen Aussagen beweisbar sind und dass auch die Widerspruchsfreiheit nicht bewiesen werden kann. Nobelpreisträger Roger Penrose schloss daraus weiter, „*dass sich Verständnis und Einsicht nicht auf ein System von Rechenregeln reduzieren lassen"*[55]. Es gibt nach Gödel, so schlussfolgert Penrose, Aussagen der Arithmetik, deren Wahrheit sich nur durch „*menschliche Intuition und Einsicht"* erschließen lassen, die beide „*nicht auf ein System von Regeln"* reduziert werden können[56]. Für Penrose ist Gödels Satz die Grundlage für seine Behauptung, „*am menschlichen Denken sei mehr ‚dran', als ein Computer, wie wir den Begriff ‚Computer' heute verstehen, je erreichen kann"*[57]. Daraus ergab sich nach Ertel eine „*schmerzhafte Grenze formaler Systeme"*[58], intelligenter Programme und der KI bis in die heutige Zeit.

---

53 Gödel (1932), S. 234

54 Gödel (1932), S. 234; Vgl. für ähnliche Formulierung auch Gödel (1931), S. 196; Anmerkung: Im Originalaufsatz von Gödel ist dies „Satz XI" in einem umfassenden Beweis. Gödel positioniert seine Erkenntnisse im gleichen Aufsatz (S. 173) in der Einleitung sehr deutlich: „*Die Entwicklung der Mathematik in der Richtung zu größerer Exaktheit hat bekanntlich dazu geführt, daß weite Gebiete von ihr formalisiert wurden, in der Art, daß das Beweisen nach einigen wenigen mechanischen Regeln vollzogen werden kann. Die umfassendsten derzeit aufgestellten formalen Systeme sind das System der Principia Mathematica (PM) einerseits, das Zermelo-Fraenkelsche (von J. v. Neumann weiter ausgebildete) Axiomensystem der Mengenlehre andererseits. Diese beiden Systeme sind so weit, daß alle heute in der Mathematik angewendeten Beweismethoden in ihnen formalisiert, d.h. auf einige wenige Axiome und Schlußregeln zurückgeführt sind. Es liegt also die Vermutung nahe, daß diese Axiome und Schlußregeln dazu ausreichen, alle mathematischen Fragen, die sich in den betreffenden Systemen überhaupt formal ausdrücken lasse, auch zu entscheiden. Im folgenden wird gezeigt, daß dies nicht der Fall ist, sondern daß es in den beiden angeführten Systemen sogar relativ einfache Probleme aus der Theorie der gewöhnlichen ganzen Zahlen gibt, die sich aus den Axiomen nicht entscheiden lassen".*

55 Penrose (1994a), S. 82

56 Dazu sinngemäßes Zitat von Penrose in Landgrebe Smith (2022), S. 208-209: "*The inescapable conclusion seems to be: Mathematicians are not using a knowably sound calculation procedure in order to ascertain mathematical truth. We deduce that mathematical understanding – the means whereby mathematicians arrive at their conclusions with respect to mathematical truth – cannot be reduced to blind calculation."*;
Im Originalzitat in Penrose (1994b), S. 133 etwas weniger bestimmt: "*Let us then accept the apparently inescapable implication of the Gödel(-Turing) argument: mathematicians do not simply ascertain mathematical truth by means of knowably sound calculational procedures. There remain the possibilities that they might use unknowable or unsound calculational procedures – or, as is my own belief, that they simply do not just use calculational procedures when they ascertain truth. With regard to the calculational possibilities, I should point out that mathematicians certainly don't think that they are using unknowable or unsound procedures in order to ascertain mathematical truth!"*

57 Penrose (1994a), S. 82

58 Ertel (2008), S. 8

Obwohl die Gödelschen Unvollständigkeitssätze allgemein anerkannt sind und als wahr gelten, trifft dies nicht im gleichen Maße auf die oben zitierten Schlussfolgerungen von Penrose und anderen Wissenschaftlern, den Antimechanisten, zu, die ähnlich argumentiert haben, wie z.B. der Philosoph John Randolph Lucas und der Mathematiker Hao Wang. Eine andere Gruppe von Gegnern der Antimechanisten hat sich davon nicht beeindrucken lassen und bestreitet die Schlussfolgerung aus Gödels Theoremen, wonach der menschliche Geist allen Computern überlegen sei. Elke Brendel hat in einer Bestandsaufnahme[59] die wichtigsten Argumente der „*Antimechanisten*“[60] und deren Gegner zusammengetragen. Sie fasst zwei Strategien der Gegner der antimechanistischen Position zusammen:

> *„Die **erste Strategie*** [Hervorhebung durch EB] *besteht darin, das vorausgesetzte Maschinenverständnis als für eine mögliche Gleichsetzung mit dem menschlichen Geist von vornherein inadäquat anzusehen. Da der menschliche Geist offensichtlich die Bedingungen eines formalen Systems nicht erfüllt, so lautet das Argument, können Gödels Unvollständigkeitstheoreme überhaupt nicht ins Spiel gebracht werden. Gödels Theoreme tragen somit nichts zur Frage bei, ob Maschinen denken können oder nicht, da sie nur auf formale Systeme anwendbar sind, die man auf den menschlichen Geist ohnehin nicht übertragen kann.*
>
> *In der **zweiten Strategie*** [Hervorhebung durch EB] *wird zwar die Gleichsetzung des menschlichen Geistes mit dem in den Gödel-Theoremen spezifizierten formalen System akzeptiert, jedoch wird zu zeigen versucht, dass eine solche „Maschine“ dieselben Erkenntnisse über den Gödel-Satz zu erlangen vermag wie der menschliche Geist, so dass zumindest in dieser Hinsicht zwischen Menschen und Maschinen kein prinzipieller Unterschied besteht. Ein Unterschied wäre nur dann gegeben, wenn man den Maschinenbegriff zusätzlich derart einschränken und abschwächen würde, dass er aus fast schon trivialen Gründen den kognitiven Leistungen der Menschen niemals entsprechen kann.*
>
> *In beiden Argumentationsstrategien wird also nicht versucht, mit Gödel die Gegenposition, d.h. eine Version des Mechanismus, zu verteidigen. Vielmehr soll gezeigt werden, dass die Gödel Theoreme für die Auseinandersetzung zwischen Mechanisten und Antimechanisten nicht signifikant sind.“*[61]

Mit einer überzeugenden Argumentation widerlegt Brendel beide Strategien, allerdings ohne damit die antimechanistische Position vollständig und für alle Zeiten zu bestätigen. In Kürze: Mit der Aufgabe des Anspruchs einer Annäherung an das menschliche Denken über die formale Mathematik ohne Alternative gerät die gesamte Diskussion in eine Sackgasse. Richtig ist, dass Menschen widersprüchlich argumentieren und oft zu logischen Fehlschlüssen und Inkonsistenzen neigen. Allein die Tatsache, dass wir genau dies immer

[59] Vgl. Brendel (2006)

[60] Antimechanisten gehen davon aus, dass der Mentalismus dem Mechanismus grundsätzlich überlegen sei.

[61] Brendel (2006), S. 44f

wieder feststellen und korrigieren, widerlegt den Ansatz der Gegner der Antimechanisten. Menschliche Intelligenz korrigiert menschliche Intelligenz.

Brendel stellt heraus, dass Gödel ein formales *System T*[62] unterstellt, das „*insbesondere die Arithmetik der natürlichen Zahlen enthalten und konsistent sein müsse"*[63] und dem „*die klassische deduktive Logik zugrunde liegt*. *Zu nicht-deduktiven Schlussformen, wie etwa in induktiven oder abduktiven Schlüssen ist Gödels System zunächst nicht fähig"*. *Das System*, so fährt sie fort, „*ist darüber hinaus geschlossen, d.h. insbesondere nicht in der Lage, neue Axiome und Schlussregeln zu erzeugen"*.

Darauf aufbauend argumentiert sie, es sei „*unplausibel*, *dass diese Bedingungen* und Einschränkungen *auch für das menschliche Denken gelten*. Dies sei *psychologisch unrealistisch oder stelle sehr starke Idealisierungen dar".* Vier Beispiele:

- Konsistenz sei im menschlichen Erkenntnissystem nicht gegeben; es existierten zu viele „*Widersprüche"*, konträre Modelle und „*inkonsistente Theorien"* in der Wissenschaft[64].
- Das menschliche Erkenntnissystem enthielte einerseits niemals die „*vollständige Arithmetik der natürlichen Zahlen"*.
- Andererseits: „*menschliches Räsonieren"* gehe weit über „*rein deduktive Schlüsse"* hinaus. Menschen nehmen dazu eine metatheoretische Perspektive ein, die – Stand heute – Maschinen nicht einnehmen kann[65].
- *„Ein formales System, das den menschlichen Geist modellieren soll, wäre darüber hinaus ein offenes System, das in der Lage ist, aufgrund neuer Erfahrungen*

---

62 Bei Gödel im Original: S; Vgl. Gödel (1931), S. 234

63 Dieses und die folgenden Zitate: Brendel (2006), S. 45

64 Ergänzung bei Brendel (2006), S.47/48: „*Auf die maschinelle Sprechweise übertragen versteht sich der Mensch (zumindest was seine logischen und mathematischen Schlussfähigkeiten angeht) also eher als* ***fallible und manchmal unkorrekt funktionierende Maschine*** *und nicht als formal inkorrekte Maschine."* Sie zitiert Lucas (1961), S. 121: "*The fact that we are all sometimes inconsistent cannot be gainsaid, but from this it does not follow that we are a tantamount to inconsistent systems. Our inconsistencies are mistakes rather than set policies. They correspond to the occasional malfunctioning of a machine, not its normal scheme of operation. Witness to this we eschew inconsistencies when we recognize them for what they are.* […] *This is surely a characteristic of the mental operations of human beings: they are selective, they do discriminate between favored – true – and unfavored – false – statements: when a person is prepared to say anything and is prepared to contradict himself without any qualm or repugnance, then he is adjudged to have "lost his mind". Human beings, although not perfectly consistent, are not so much inconsistent as fallible.",* [Hervorhebung DS]

65 Dazu Ergänzung bei Brendel (2006), S. 51: „*Wenn wir somit einen streng formalistischen Maschinenbegriff aufgeben und einer Maschine auch die Möglichkeit einräumen, ihre Systemgrenezen durch die Entwicklung eines Metasystems zu erweitern, wäre sie somit in der Lage, ihre eigene Semantik zu formalisieren.* ***Dass diese ‚Reflexionsleistung' in Form einer ‚Trnszendierung' des eigenen Systems durch den Übergang in ein Metasystem aus prinzipiellen Gründen nur Menschen vorbehalten sein soll, lässt sich jedoch mit den Argumenten der Antimechanisten nicht rechtfertigen****.",* Hervorhebung DS]

> *neue Axiome und Schlussregeln hinzuzufügen bzw. alte, die sich als unbrauchbar erwiesen haben, zu streichen"*[66].

Vor diesem Hintergrund kann festgehalten werden, dass für Maschinen, die innerhalb der oben beschriebenen Einschränkungen (Konsistenz, Arithmetik der natürlichen Zahlen, Logik ausschließlich auf Basis von Deduktion und geschlossenes System) operieren, die antimechanistischen Schlussfolgerungen aus den Unvollständigkeitssätzen gelten. Ausserhalb dieser Einschränkungen ist Gödel (noch) nicht anwendbar.

Hingegen ist nicht ausgeschlossen, dass es eines Tages Maschinen geben könnte, die z.B. „*aufgrund der induktiven Kraft empirischer Evidenz* [...] *an die Konsistenz ihres zugrunde gelegten formalen Systems ‚glauben' und die die Fähigkeit besitzen, eine metatheoretische Perspektive einzunehmen und auf diese Weise die Semantik des Ursprungssystems formalisieren können"*[67]. Dies sei aber höchst unwahrscheinlich, wie Brendel in ihrem Schlusswort selbst zusammenfasst:

> *„Antimechanisten, die einfach behaupten, dass der ‚Glaube' an die Konsistenz des arithmetischen Systems und das ‚Einsehen' der Wahrheit des Gödel-Satzes* ***auf spezifischen kognitiven Fähigkeiten der Menschen*** *beruhen, die Maschinen prinzipiell nicht besitzen können, begehen eine petitio principii*[68]*. Ob Maschinen diese Fähigkeiten erlangen können und vielleicht sogar eines Tages Konferenzen über philosophische Implikationen der Gödel-Theoreme abhalten werden,* ***ist natürlich äußerst fraglich*** [Hervorhebung DS], *lässt sich jedoch mittels der metalogischen Resultate der Gödel Theoreme* ***nicht auf apriorischem Wege entscheiden"*** [Hervorhebung durch EB und **DS**].

Im Sinne des wissenschaftlichen Realismus soll davon ausgegangen werden, dass aufgrund der bewiesenen Gödel-Theoreme Maschinen dem menschlichen Geist nicht vollständig entsprechen können. Die Frage, ob die KI-Forschung *„Systeme entwickeln kann, die über Bewusstsein, Verstehen, Denken,* [Intuition] *und andere kognitiven Fähigkeiten des Menschen verfügen"*[69], lässt sich nach Brendel nicht mittels der Gödelschen Unvollständigkeitssätze entscheiden.

In einfachen Worten lässt sich das Ganze wie folgt zusammenfassen:

1. Die Mechanisten gehen davon aus, dass mit Algorithmen das mathematische Denken des Menschen nachgebildet werden könne.
2. Dazu müsste die Zahlentheorie vollständig und widerspruchsfrei axiomatisiert und deren Widerspruchsfreiheit beweisbar sein.
3. Gödel widerlegt mit seinen Unvollständigkeitssätzen beides.

---

66 Brendel (2006), S. 46

67 Dieses und die folgenden Zitate (einschließlich Blockzitat): Brendel (2006), S. 52

68 Petitio Principii: Verwendung eines unbewiesenen, erst noch zu beweisenden Satzes als Beweisgrund für einen anderen Satz

69 Brendel (2006), S. 40 - 41

4. Daher ist gezeigt, dass mit uns bekannten Maschinen das menschliche mathematische Denken (in einer idealisierten Form) nur unvollständig und nicht widerspruchsfrei nachgebildet werden kann.
5. Menschliches Denken geht allerdings über mathematisches Denken und reine deduktive Logik hinaus; darauf lässt sich Gödel nicht anwenden.

2.2.1.2 McCulloch, Pitts & Hebb

Der Neurophysiologe Warren McCulloch war einer der frühen Begründer der kognitiven Wissenschaften und einer der ersten Theoretiker der Künstlichen Intelligenz[70]. Er beschäftigte sich intensiv mit Kybernetik, war Mitbegründer der „American Society of Cybernetics" und beeinflusste den Begründer der Kybernetik 2. Ordnung, Heinz von Foerster. Er war der Ansicht, Geist und Maschinen seien gleichwertig:

> „*Alles, was wir über Organismen lernen, lässt uns zu dem Schluss kommen, dass sie nicht nur Maschinen ähneln, sondern dass sie Maschinen sind. Von Menschen geschaffene Maschinen sind keine Gehirne, aber Gehirne sind eine sehr schlecht verstandene Variante von Rechenmaschinen. Kybernetik half uns die Mauer zwischen der großen Welt der Physik und dem Ghetto des Geistes einzureissen*."[71]

Schon 1943 beschäftigten sich McCulloch und der Logiker und kognitive Psychologe Walter Pitts in ihrem Aufsatz „*A logical calculus of the ideas immanent in nervous activity*"[72] mit einer Computertheorie des Geistes und des Gehirns. Auf Basis der „*Kenntnis des grundsätzlichen Aufbaus und der Funktion der Neuronen im Gehirn, einer formalen Analyse der Aussagenlogik durch* [Bertrand] *Russell und Whitehead und der Programmiertheorie von Turing"*[73] schlugen sie ein Modell künstlicher Neuronen vor:

> *"Because of the „all-or-none" character of nervous activity, neural events and the relations among them can be treated by means of propositional logic. It is found that the behavior of every net can be described in these terms, with the addition of more complicated means for nets containing circles; and that for any logical expression satisfying certain conditions, one*

---

70 Vgl. Ramage Shipp (2009), S. 22

71 Zitiert in Ramage Shipp (2009), S. 22; Übersetzung DS; Ausführliches Originalzitat: McCulloch (1965), S. 163: *„Everything we learn of organisms leads us to conclude not merely that they are analogous to machines. Man-made machines are not brains, but brains are a very ill-understood variety of computing machines.* ***Cybernetics has helped to pull down the wall between the great world of physics and the ghetto of the mind****."*;

Ergänzende Anmerkung: Foersters Position, dass Geist und Maschine gleichwertig seien, wird NICHT von allen Kybernetikern geteilt; Vgl. Günther (1957), S. 180: "***Ein Mechanismus erzeugt kein Bewusstsein****, auch nicht, wenn sein Arbeitsrhythmus transklassisch ist*"; sowie Wiener (1952), S. 198: „*Die Maschine aber, die wie der Flaschengeist lernen kann und aufgrund ihres Lernens Entscheidungen fällen kann, wird durchaus nicht gebunden sein, Entscheidungen zu treffen, wie wir sie getroffen hätten oder wie sie für uns annehmbar wären. Denn* ***der Mensch, der das Problem seiner Verantwortung blindlings auf die Maschine abwälzt,*** *sei sie nun lernfähig oder nicht,* ***streut seine Verantwortung in alle Winde*** *und wird sie auf den Schwingen des Sturmwinds zurückkommen sehen"*; Hervorhebungen DS

72 McCulloch Pitts (1943), S. 115 ff

73 Russell Norvig (2004), S. 39

> *can find a net behaving in the fashion it describes. It is shown that many particular choices among possible neurophysiological assumptions are equivalent, in the sense that for every net behaving under one assumption, there exists another net which behaves under the other and gives the same results, although perhaps not in the same time."*[74]

Dies ist eine – in der Geschichte der KI – deutliche Beschreibung der Funktionsweise und des Anspruchs neuronaler Netze, die bis heute gilt. In ihrem Papier entwickelten die beiden Wissenschaftler einen Ansatz, mit dem alle logischen Verknüpfungen (UND, ODER, NICHT usw.) in einem Netz aus verbundenen Neuronen realisiert werden können[75].

Die McCulloch-Pitts-Neuronen können den Wert 0 oder 1 annehmen und zu Netzen verbunden werden. Einige Neuronen sind „erregend", d.h. sie stimulieren andere Neuronen. Für diese erregende Stimulation der neuronen wird auch der Begriff des „Feuerns" verwendet. Andere Neuronen sind hemmend, d.h. sie reduzieren das Feuern der mit ihnen verbundenen Neuronen. Ist die Summe der erregenden Eingabewerte abzüglich der Summe der hemmenden Eingabewerte grösser als ein bestimmter Schwellwert, feuert das Neuron[76].

Der kanadische Psychologe und Biologe Donald Olding Hebb knüpfte 1949 an dieses Konzept an und entwickelte eine Regel zum Zustandekommen des Lernens in neuronalen Netzwerken mit einer „*einfache*[n] *Aktualisierungsregel für die Veränderung der Verbindungsstärken zwischen den Neuronen. Seine Regel, heute auch als Hebb'sches Lernen bezeichnet, ist bis heute ein einflussreiches Modell geblieben."*[77] Verkürzt formuliert, meint Hebb: Je häufiger ein bestimmtes Neuron gleichzeitig mit einem anderen Neuron aktiv ist, umso intensiver werden die beiden Neuronen aufeinander reagieren – ganz nach dem Motto: „*what fires together, wires together – was zusammen feuert, verbindet sich".*[78]

---

[74] McCulloch Pitts (1943), S. 115

[75] Vgl. Russell Norvig (2004), S. 39

[76] Nilsson (2014), S. 16

[77] Russell Norvig (2004), S. 39

[78] Nach Keysers Gazzola (2014) wird diese Formulierung oft irrtümlicherweise Daniel Hebb direkt zugeschrieben, tatsächlich taucht sie erstmalig 1992 bei Carla Shatz auf (Shatz (1992), S. 64); dabei bezieht sie sich aber klar auf Hebbs Erkenntnisse: „*In a sense, then,* ***cells that fire together wire together****. The timing of action-potential activity is critical in determining which synaptic connections are strengthened and retained and which are weakened and eliminated. Under normal circumstances, vision itself acts to correlate the activity of neighboring retinal ganglion cells, because the cells receive inputs from the same parts of the visual world. What is the synaptic mechanism that strengthens or weakens the connections? As long ago as 1949,* ***Donald O. Hebb of McGill University*** *proposed the existence of special synapses that could execute the task. The signal strength in such synapses would increase whenever activities in a presynaptic cell (the cell supplying the synaptic input) and in a postsynaptic cell (the cell receiving the input) coincide. Clear evidence showing that such '****Hebb synapses****' exist comes from studies of the phenomenon of long-term potentiation in the hippocampus. Researchers found that the pairing of pre-synaptic and post-synaptic activity in the hippocampus can cause incremental increases in the strength of synaptic transmission between the paired cells. The strengthened state can last from hours to days."* [Hervorhebungen DS]

2.2.1.3 Turing

„*Können Maschinen denken?*“ – mit dieser provokanten Frage leitete der Mathematiker Alan Turing 1950 seinen Aufsatz „*Computing Machinery and Intelligence*“[79] ein, in dem er seine Vision der prinzipiellen Möglichkeiten der Maschinenintelligenz darstellte und das bereits erwähnte „*Imitation Game*“ vorschlug. Weiterhin entwickelte er eine Liste von neun Gegenargumenten und Einwänden gegen seine Vision, mit der er vor mehr als siebzig Jahren bemerkenswert viel Weitsicht bewies:

1. ***Der theologische Einwand***
   Gemäß dieser Position sei Denken das Privileg des von Gott geschaffenen Menschen mit seiner unsterblichen Seele. Kein Tier und auch keine Maschine sollte gemäß dieser Argumentation dieses Privileg besitzen.
2. ***Der „Kopf im Sand“ Einwand***
   Bei diesem Argument handelt es sich um eine diffuse Risikoaversion: „*The consequences of machines thinking would be too dreadful. Let us hope and believe that they cannot do so.”*[80]
3. ***Der mathematische Einwand***
   Hier bezieht sich Turing auf die bereits oben angesprochenen Unvollständigkeitssätze von Gödel und seine weiter unten beschriebene Church-Turing-These. Als Gegenargument wendet er ein, dass es ähnliche grundsätzliche Ein- und Beschränkungen der menschlichen Intelligenz und Kognition geben könne.
4. ***Der Einwand unter Verweis auf das Bewusstsein***
   Dieser Einwand beschäftigt die Forschung bis heute und spielt auch eine große Rolle in dieser Arbeit. Turing hierzu:
   “This argument is very well expressed in Professor Jefferson’s Lister Oration for 1949, from which I quote. ‘Not until a machine can write a sonnet or compose a concerto because of thoughts and emotions felt, and not by the chance fall of symbols, could we agree that machines equal brain – that is, not only write it but know what it had written. No mechanism should feel (and not merely artificially signal, an easy contrivance) pleasure at its successes, grief when it valves fuse, be warmed by flattery, be made miserable by its mistakes, be charmed by sex, be angry or be depressed when it cannot get what it wants.”
   Er selbst tut dieses Argument als solipsistisch und destruktiv ab und verweist auf die Zeit nach dem bestandenen „Imitation Game“.
5. ***Einwand der vielen Einschränkungen und Behinderungen***
   Dieser Einwand stützt sich darauf, dass Kritiker immer menschliche Fähigkeiten finden können, die die Maschine nicht aufweist. Turing liefert selbst einige

[79] Turing (1950), S. 433ff

[80] Turing (1950), S. 444ff, gilt auch für alle folgenden Zitate in dieser Auflistung

Beispiele:
"Be kind, resourceful, beautiful, friendly, have initiative, have a sense of humor, tell right from wrong, make mistakes, fall in love, enjoy strawberries and cream, make someone fall in love with it, learn from experience, use words properly, be the subject of its own thought, have as much diversity of behavior as a man, do something really new." Mit einigen dieser Beispiele setzt sich Turing auseinander, trotzdem fehlt das allumfassende Gegenargument.

6. ***Der Einwand von Lady Lovelace***
   Hier dreht es sich um ein mehr als hundert Jahre vorher vorgebrachtes Argument gegen die von Lady Lovelace und Charles Babbage entwickelte „Analytical Engine"[81]: „*The Analytical Engine has no pretensions to originate anything. It can do whatever we know how to order it to perform*". Lady Lovelace traut den Maschinen keinerlei Innovationen und Überraschungen zu. Turings Gegenargument ist ebenfalls sehr weitsichtig. Er erwartete eine Reihe von neuen Einsichten und Konsequenzen aus Datenanalysen mit allgemeinen Prinzipien:
   "A natural consequence of doing so is that one then assumes that there is no virtue in the mere working out of consequences from data and general principles."
7. ***Der Einwand von der (analogen) Kontinuität in Nervensystemen***
   Hierbei handelt es sich um das auch noch heute bekannte Argument des Unterschiedes zwischen analog und digital. Turings Antwort entspricht der heutigen: es kommt auf die Auflösung an.
8. ***Der Einwand der menschlichen Intuition*** (frei übersetzt, im Original: „*The Argument from Informality of Behavior*")
   Auch diese Überlegung spielt in unserer Gegenwart noch eine Rolle: Wie geht eine Maschine, die darauf programmiert ist, bei Grün über die Straße zu gehen und bei Rot zu stoppen, damit um, wenn beide Farben leuchten? 1950 war man noch sehr weit vom autonomen Fahren entfernt, trotzdem beschäftigt uns genau diese Frage bis heute und wohl noch für einige Zeit.
9. ***Der Einwand von den außersinnlichen Wahrnehmungen*** (ESP = extra-sensory-perception)
   Dieser aus heutiger Perspektive etwas ungewöhnliche Einwand setzt voraus, dass es so etwas wie ESP oder den „siebten Sinn" (im Englischen seltsamerweise „6th sense") gibt. Andererseits geht er davon aus, dass Maschinen Zugriff

[81] "***Analytical Engine**, generally considered the first computer, designed and partly built by the English inventor **Charles Babbage** in the 19th century (he worked on it until his death in 1871). While working on the Difference Engine, a simpler calculating machine commissioned by the British government, Babbage began to imagine ways to improve it. Chiefly he thought about generalizing its operation so that it could perform other kinds of calculations. By the time funding ran out for his Difference Engine in 1833, he had conceived of something far more revolutionary: a general-purpose computing machine called the Analytical Engine*." (Quelle: Britannica; https://www.britannica.com/technology/Analytical-Engine)

> auf alle anderen Sinne haben, nicht aber auf die außersinnlichen Wahrnehmungen. Diese Diskussion findet sich auch in der heutigen Geistesphilosophie nicht wieder. Turings Empfehlung klingt recht pragmatisch: "*put the competitors into a "telepathy-proof room*".

Turing hatte sich zusammen mit Alonzo Church bereits einige Jahre zuvor mit der Church-Turing-These und beim sogenannten Halteproblem für die Entwicklung bzw. das Verständnis der Grenzen der KI verdient gemacht.

Die Church-Turing-These lautet:

> *„Die Klasse der Turing-berechenbaren Funktionen* [Anmerkung: also mit einer Turing-Maschine berechenbar] *stimmt mit der Klasse der intuitiv berechenbaren Funktionen überein."*[82]

Verständlicher formuliert:

> *„Alles was intuitiv berechenbar ist, d.h. alles, was von einem Menschen berechnet werden kann, das kann auch von einer Turingmaschine berechnet werden. Ebenso ist alles, was eine andere Maschine berechnen kann, auch von einer Turingmaschine berechenbar."*[83]

Oder im Umkehrschluss: „*Was eine Turingmaschine nicht berechnen kann, kann auch kein Mensch* [mit Algorithmen] *berechnen"*[84].[85]

Gemäß dieser These kann es keine Rechnermodelle geben, die prinzipiell mehr können als die klassische Turing-Maschine[86], was wiederum bedeuten würde, dass auch

---

82 Hoffmann (2011), S. 308

83 Quelle: Formale Grundlagen der Informatik; http://fgi1-skript.de/die-church-turing-these/

84 Ebd.

85 Vgl. die Erläuterung von Jobst Landgrebe und Barry Smith in Landgrebe Smith (2022), S. 115: "*How, then, are we to provide a precise formal characterization of what it is to be computable in this intuitive sense? This question was answered more than 80 years ago by Alonzo Church and Alan Turing. Working independently, they formulated their respective definitions using different terms and methods. These definitions were, however, subsequently proved to be mathematically equivalent, and the resultant Church-Turing thesis, which has remained unchallenged ever since, states that only algorithms that can be formulated as a sequence of elementary recursive functions are computable. The most important example of non-computability in elementary mathematics was given by Turing in his first paper on what we now call Turing machines. He was addressing in this paper the problem of whether there are any purely mathematical yes-no questions that can never be answered by computation (the halting problem), which is equivalent to the Entscheidungsproblem described by Church. Turing proved that this problem cannot be solved when 'by computation' is taken as meaning 'by a universal computation machine' (which is modelled by what we call today the universal Turing machine). This means that there is no general possibility to determine whether a syntactically well-formed mathematical proposition can be proven or falsified.* ***From this it follows, for example, that first-order logic is not decidable."***

86 „*Turing beschreibt eine hypothetische Rechenmaschine mit einem Lese- und Schreibkopf, der sich auf einem unbegrenzt langen Band mit Zeichen aus einem endlichen Alphabet schrittweise vor- und zurückbewegt und Zeichen einliest, löscht sowie neue Zeichen auf das Band schreibt. Dieser Als Turing Maschine bekannte Automat verarbeitet allein unter dem Gesichtspunkt ihrer Form bzw. Struktur und ist in der Lage, jede algorithmisch berechenbare Funktion in endlich vielen Schritten zu berechnen*", Quelle: Müller (2015), S. 1; Online: Müller, Jean. (2015). Künstliche Intelligenz. 10.1515/wsk.15.0.kunstlicheintelligenz

Quantencomputer, von denen man einen „Quantensprung“ in der Leistungsfähigkeit erwartet, dieser Beschränkung unterliegen. Neuere Forschungsergebnisse (Shor-Algorithmus[87]) haben diese Interpretation der Church-Turing-These jedoch widerlegt. Es gibt einzelne Aufgabenstellungen, die mit „*klassischen deterministischen Algorithmen“*[88] bei exponentiell wachsendem Rechenaufwand und mit einem Quantenalgorithmus unter deutlich langsamer wachsendem Rechenaufwand gelöst werden können, „*nämlich nur polynomial“*[89]:

> *„Die logisch-mathematischen Grundlagen sind allerdings davon nicht betroffen: Probleme, die prinzipiell nicht entscheidbar sind, bleiben auch bei Quantenalgorithmen unentscheidbar*.“[90]

Generell sollte die Fähigkeit des menschlichen Verstandes nicht auf „Berechnen“ reduziert werden. Hierzu schreiben Landgrebe und Smith:

> *“When models are run on computers, they take the form of algorithms. To be executable by a computer, an algorithm must be computable. We shall see that* ***the set of computable algorithms is a subset of all the algorithms that can be formulated in mathematical terms****. But what does it mean to say that an algorithm is ‘computable’? On the intuitive meaning of this term it means that the algorithm can be applied to an input to yield an output effectively, which means automatically, without any contribution – decision, intention, insight, intuition, ingenuity, fiddling about – from a human being.”*[91], Hervorhebung DS

Diese Erkenntnis steht im Einklang mit Gödels Unvollständigkeitssätzen.

Das sogenannte „Halteproblem“ beschreibt, dass es NICHT entscheidbar ist, ob die Ausführung eines Algorithmus ohne Eingriff von außen zu einem Ende gelangt. Zum Beispiel ist damit auch gezeigt, dass es kein System geben kann, mit dem Computerprogramme endgültig verifiziert werden können, im Sinne einer bewiesenen Fehlerfreiheit. Bis heute ist die Verifikation von Softwaresystemen ein aufwändiger Prozess, der niemals abschließend und vollständig durchgeführt werden kann. Diese ewig währende inhärente Einschränkung der Computertechnologie und KI zeichnete sich also bereits in der Frühphase dieser neuen Technologie klar ab.

Damit verbleibt die Frage, ob eine Turing-Maschine ein Gehirn simulieren kann, also alle Funktionen und Operationen nachvollziehen kann, die das Gehirn leistet. Dazu der

---

87 *Beim im Jahre 2000 erstmals angewandten Shor-Algorithmus handelt es sich um eine Quanten-Fouriertransformation, mit der Primfaktoren effizient bestimmt werden können. Dieses funktioniert ausschließlich mit Quantencomputern. Es ist kein klassischer Polynomialzeitalgorithmus für Turing-Maschinen bekannt, der ähnlich effizient wäre. Diese Erkenntnis, wenn wissenschaftlich bewiesen, würde dann auch die Church-Turing These widerlegen*. Vgl.: Homeister (2005), S. 231

88 Dieses und folgende Zitate: Mainzer (2014), S. 126

89 Ein Polynom summiert die Vielfachen von Potenzen einer Variablen

90 Mainzer (2014), S. 126

91 Landgrebe Smith (2022), S. 115

folgende Eintrag aus der Stanford Encyclopedia of Philosophy mit einem ähnlichen Urteil wie Elke Brendel in der Gödel-Diskussion:

> *"Any device or organ whose internal processes can be described completely by means of (what Church called) effectively calculable functions can be simulated exactly by a Turing machine (providing that the input into the device or organ is itself computable by Turing machine). But any device or organ whose mathematical description involves functions that are not effectively calculable cannot be so simulated. As Turing showed, there are uncountably many such functions. It is an open question whether a completed neuroscience will need to employ functions that are not effectively calculable."*[92]

Letzten Endes weiß die Wissenschaft – wie wir später noch sehen werden – zu wenig über die konkreten funktionalen Abläufe im menschlichen Gehirn und insbesondere im Bewusstsein, um die Frage beantworten zu können, ob dort Funktionen ablaufen, die nicht Turing-berechenbar sind. Es ist hingegen etabliertes Wissen, dass es mathematische Probleme bzw. Modelle gibt, die von Turing-Maschinen nicht entscheidbar sind[93]. Daran werden auch neue Erkenntnisse zur Funktionsweise des menschlichen Gehirns nichts ändern. Die KI wird vieles leisten, was Menschen nicht vollbringen, umgekehrt aber vieles schuldig bleiben, was der menschliche Geist zustande bringt.

2.2.1.4 Dartmouth Conference

Am Dartmouth College in Hanover in New Hampshire trafen sich 1956 einige Wissenschaftler um John McCarthy, Marvin Minsky, Nathaniel Rochester, Claude Shannon und Herbert Simon zu einem mehrwöchigen Seminar, das mit „Artificial Intelligence" (Künstliche Intelligenz) betitelt wurde; dies markierte die erstmalige Verwendung des Begriffs. Auf der Konferenz befasste man sich mit Computern, die mehr können sollten als nur mit Zahlen zu rechnen, also zum Beispiel sprechen, lesen oder Brettspiele spielen. Das Konferenzprogramm zeugte von ebenso viel Ambition wie Weitsicht:

> *"We propose that a 2-month, 10-man study of artificial intelligence be carried out during the summer of 1956 at Dartmouth College in Hanover, New Hampshire. The study is to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it. An attempt will be made to find how to make machines use language, form abstractions and concepts, solve kinds of problems now reserved for humans, and improve*

[92] Stanford Encyclopedia of Philosophy: https://plato.stanford.edu/entries/church-turing/#SimuThes

[93] Hierzu Landgrebe Smith (2022), S. 116f: "*It is important to note that any mathematical model that runs on a Turing machine can only model comprehensively and adequately what we have called logic systems.* […] *... the vast majority of useful applications of mathematics can be formulated as Turing-computable algorithms – but not all of them can.* […] *... full first-order logic (FOL) (and higher-order logic) models are not computationally decidable, both of these are essential in mathematics and useful in philosophy, linguistics, and other areas. Most mathematical proofs cannot be stated without-first-order logic."*

themselves. We think that a significant advance can be made in one or more of these problems if a carefully selected group of scientists work on it together for a summer. "[94]

Nur wenige konkrete Ergebnisse wurden im Rahmen der Konferenz erarbeitet. Der Wert der Veranstaltung lag eher in der grundsätzlichen Richtungsdiskussion, in der Etablierung eines langfristigen Forschungsprogrammes und im Aufbau eines Netzwerks unter den relevanten Experten und Wissenschaftlern.

Interessanterweise teilte sich dieser Expertenkreis schon damals in drei Lager. Zum ersten Lager zählten überwiegend Mathematiker, die auf die mathematische Logik und auf Deduktion als Sprache der Rationalität setzten. Die zweite Gruppe verfolgte eher induktive Ansätze, nach denen Programme Statistiken aus Daten extrahieren und Unsicherheiten mit den Methoden der Wahrscheinlichkeitsrechnung adressiert werden sollten. Eine dritte Gruppe setzte auf Inspirationen aus Biologie und Psychologie, um damit technische Lösungen zu kreieren, die sich am menschlichen Gehirn orientierten[95].

Vier der Teilnehmer, John McCarthy, Marvin Minsky, Allen Newell und Herbert Simon, wurden zu Pionieren der neuen Technologie[96] und begründeten auch das dazugehörige wissenschaftliche Netzwerk und die Infrastruktur in Nordamerika, die bis heute Bestand hat. So gründete zum Beispiel McCarthy das Stanford Artificial Intelligence Project und Minsky das MIT AI Lab[97].

2.2.2 Sommer und Winter der Künstlichen Intelligenz[98]

Nach der Dartmouth Konferenz erlebte die KI über zwei Dekaden sowohl große Erfolge als auch Niederlagen. Das von Joseph Weizenbaum Mitte der 1960er Jahre am MIT entwickelte Programm ELIZA stellte einen wichtigen Meilenstein dar. Es kommunizierte über natürliche Sprache und sollte den Turing-Test bestehen, was nicht gelang. Eine weitere frühe Errungenschaft war ab 1957 die Entwicklung des „General Problem Solver" (GPS) durch den Dartmouth-Teilnehmer und späteren Nobelpreisträger Herbert Simon, den RAND Corporation Wissenschaftler Cliff Shaw sowie Alan Nowell. Das Programm war der erste Versuch der Realisierung einer allgemeinen Problemlösungsmethode, die sehr spezifische Probleme in Teilprobleme zerlegte, die dann einzeln gelöst wurden. Das Programm an sich gilt als gescheitert, aber viele kognitionswissenschaftliche Aspekte konnten für die weitere Entwicklung der Methodik der Formalisierung und Lösung von Problemen verwendet werden.

---

94 Quelle: *A proposal for the Dartmouth summer research project on Artificial Intelligence*; McCarthy, Minsky, Rochester, Shannon; 31. August 1955

95 Vgl. Mitchell (2019), S. 20-21

96 Vgl. Mitchell (2019), S. 19; Melanie Mitchell bezeichnete sie als die „Big Four" Pioniere der neuen Technologie

97 Vgl. Mitchell (2019), S. 31

98 Vgl. Haenlein Kaplan (2019), S. 7

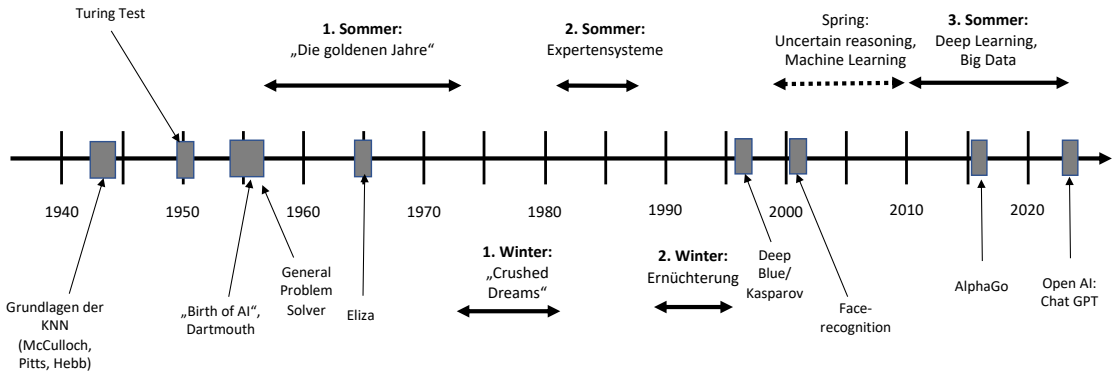

Abbildung 3: Zeitleiste KI Sommer und Winter[99]

In den 1960er Jahren wurden an amerikanischen Universitäten zahlreiche KI-Forschungsgruppen eingerichtet, die insbesondere an Sprachverarbeitung, automatischem Problemlösen und visueller Szenenanalyse arbeiteten. Auch das Verteidigungsministerium der USA unterstützte die Projekte[100].

Eine erste Welle der Euphorie bewog Marvin Minsky 1970 in einem Interview mit „Life Magazine“ dazu, die Entwicklung einer Maschine mit der Intelligenz eines durchschnittlichen Menschen binnen drei bis acht Jahren anzukündigen[101]. Diese Vorhersage und andere Hoffnungen konnten nicht realisiert werden. Es machte sich Enttäuschung breit. Ab 1973 reduzierten sowohl der amerikanische Kongress als auch die britische Regierung ihre Förderprogramme für die Künstliche Intelligenz. Die Periode des ersten „*KI Winters*“[102] begann. Er ging einher mit einem breiten Rückgang des öffentlichen und wirtschaftlichen Interesses an der Künstlichen Intelligenz. Es herrschte eine Ernüchterung, die auf nicht erfüllte Erwartungen an die Leistungen der regelbasierten Systeme zurückzuführen war. Es folgte in den 80er Jahren ein weiterer von Erwartungen an Expertensystemen getriebener Sommer mit einer erneuten Ernüchterung gegen Ende des vergangenen Jahrhunderts. Der (vorläufige) Durchbruch (und derzeitige „3. KI-Sommer“) wurde seit den späten 1990er Jahren mit „Machine Learning“ und „Deep Learning“ eingeleitet.

2.2.3 Der Durchbruch

Ein Grund für den fehlenden Fortschritt in den 1970er und zu Beginn der 80er Jahre lag darin, dass man in den frühen Systemen ELIZA und GPS und deren Weiterentwicklungen einen hierarchischen Problemlösungsansatz der symbolischen KI verwendete, der auf das vorgegebene Expertenwissen der Programmierer angewiesen war und nicht selbst lernen

99 Übersetzt, nachgezeichnet und ergänzt nach Ilkou Koutraki (2020), S. 2, Figure 1

100 Vgl. Görz Schneeberger Schmid (2003), S. 5

101 Vgl. Görz Schneeberger Schmid (2003), S. 7

102 Görz Schneeberger Schmid (2003), S. 8

konnte. Man sprach ausdrücklich von sogenannten „Expertensystemen“, die ausgezeichnet für „Top-Down“-Ansätze, „if-then“ Verzweigungen und eindeutig definierte Regeln geeignet waren. Ein Beispiel war das IBM Deep Blue Schachprogramm, das 1997 in der Lage war, den Schachweltmeister Garri Kasparov zu besiegen. Das Programm konnte bis zu „*30 Milliarden Positionen pro Zug“* durchrechnen und auf eine „*Datenbank mit 700.000 Großmeisterspielen“*, ein „*Eröffnungsbuch mit etwa 4000 Positionen*“ sowie eine „*Endspieldatenbank mit gelösten Positionen*“ mit fünf bis sechs Figuren zugreifen[103].

> *„Der Erfolg von Deep Blue verstärkte den allgemein herrschenden Glauben, dass der Erfolg von Spiele spielenden Computern hauptsächlich der immer leistungsfähigeren Hardware zu verdanken sei – eine Ansicht, die von IBM bestärkt wurde.“*

Deswegen wurde diese Vorgehensweise auch als „Brute-Force“ Ansatz bezeichnet[104]. Expertensysteme versagten in Situationen, in denen eine Formalisierung wie beim Schachspiel nicht möglich war. Sie konnten keine Gesichter erkennen oder zwischen verschiedenen Gegenständen unterscheiden. Für derartige Aufgaben war es notwendig, dass das System externe Daten korrekt interpretiert und davon lernt und das Gelernte zur Zielerreichung bei gleichzeitiger flexibler Anpassung an das Umfeld anwenden kann. Insofern handelte es sich bei Expertensystemen nicht um wirkliche „Künstliche Intelligenz“.

Der alternative Pfad der **künstlichen neuronalen Netzwerke** (KNN, auf Englisch: „artificial neural networks“) ergab sich auf Basis der frühen Arbeiten von McGulloch, Pitts und insbesondere Hebbs mit seiner Hebb‘schen Lernregel zum Nachbau von Neuronen des menschlichen Gehirns. Lange Zeit war die Leistungsfähigkeit der Computer nicht ausreichend, um derartige Netzwerke zu realisieren.

Der amerikanische Psychologe Frank Rosenblatt entwickelte Ende der 1950er Jahre „*die erste neuronale Netzwerkmaschine, die Mustererkennung mit neuronenähnlichen Einheiten bewerkstelligen sollte“*[105], das nach ihm benannte „Rosenblatt Perceptron“. Dies war der erste Versuch, die Hebb’sche Regel in einem neuronalen Netz für Maschinenlernen anzuwenden. Anfangs äußerte sich Rosenblatt im „New Yorker“ höchst euphorisch:

> *“Our success in developing the perceptron means that for the first time a non-biological object will achieve an organization of its external environment in a meaningful way. That’s a safe definition of what a perceptron can do. My colleague disapproves of all the loose talk one hears nowadays about mechanical brains. He prefers to call our machine a self-organizing system, but, between you and me, that’s precisely what any brain is.”*[106]

[103] Russell Norvig (2004), S. 232

[104] Vgl. Manhart (2022)

[105] Mainzer (2015), S. 105

[106] Zitiert aus dem „New Yorker“ in Christian (2020), S. 19; die Originalquelle konnte nicht verifiziert werden

Das Rosenblatt Perceptron erwies sich allerdings als langsam und besaß eine beschränkte, komplizierte Lernfähigkeit, so dass es trotz der anfänglichen Erwartungen zunächst nicht weiter genutzt wurde[107].

Erst in den 1980er Jahren griff man die frühen Ideen aus den 1940er und 1950er Jahren und einige der weitergehenden Konzepte der Realisierung neuronaler Netze (wie z.B. Backpropagation[108]) wieder auf und entwickelte sie sukzessive weiter in Richtung Deep Learning; dieses fand seine Anwendung in der Gesichts-, Bild- und Handschrifterkennung und 2015 auch im Google Programm AlphaGo, mit dem der Weltmeister im Brettspiel Go besiegt werden konnte. Dieses Teilgebiet der KI nennt man Konnektionismus; genauso wie die Expertensysteme erzielte es jedoch lange Zeit nicht den erhofften Durchbruch:

> *„Die neuronalen Netze konnten zwar eindrucksvolle Fähigkeiten lernen, aber es war meist nicht möglich, das gelernte Konzept in einfache Formeln oder logische Regeln zu fassen. Die Kombination von neuronalen Netzen mit logischen Regeln oder menschlichem Expertenwissen bereitete große Schwierigkeiten.“*[109]

Etliche weitere vielversprechende Ansätze wurden entwickelt. Hierzu zählte einerseits das probalistische Schließen mit Hilfe von Bayes-Netzen, die in Diagnose- und Expertensysteme der Medizin für das logische Schließen und Entscheiden auf Basis bedingter Wahrscheinlichkeiten und der Wahrscheinlichkeitstheorie eingesetzt wurden. Und andererseits erforschte man in den letzten zwei Jahrzehnten mathematische Verfahren und Ansätze für die Mustererkennung im Machine Learning, wie z.B. dasjenige der Support-Vektor-Maschinen.

In der industriellen Regelungstechnik wurde in den 1970er Jahren (durch Lotfi Zadeh) eine Logik entwickelt, die nicht auf die beiden Werte 0 und 1 beschränkt ist, sondern auch beliebige Werte dazwischen zulässt, die sogenannte „Fuzzy-Logic“. Es werden unscharfe Mengen definiert, zu denen Objekte wie in der klassischen Logik vollständig zugehörig sein können oder nur zu einem bestimmten Grad, d.h. ein Objekt X kann zu 40% der Menge A zugehörig sein und zu 60% der Menge B. Es werden demgemäß Zugehörigkeitsfunktionen definiert, die jedem Objekt den Grad seiner Zugehörigkeit zuordnen. Ergänzend wurden logische Operationen für unscharfe Mengen entwickelt, als Verallgemeinerung der klassischen zweiwertigen Logik (0 oder 1; wahr oder nicht wahr). In der KI ist diese Art der Logik hilfreich bei Sprach- und Schrifterkennung.

Aus der Synthese von Logik und neuronalen Netzen ergaben sich „Hybride Systeme“, die effiziente Suchalgorithmen erlaubten.

---

[107] Vgl. Russell Norvig (2004), S. 44f und S. 880

[108] Backpropagation, auf deutsch: Fehlerrückführung; ein Algorithmus für das Einlernen von künstlichen neuronalen Netzen

[109] Ertel (2008), S. 10

Für die Auswertung großer Datenbanken und die Extraktion von Wissen hat sich die weitere Disziplin des „Data Minings“ entwickelt. Mit maschinellen Lernmethoden werden Muster und Auffälligkeiten identifiziert. Bei großen dafür ausgewerteten Datenmengen aus vielen verschiedenen Quellen oder aus umfassenden Datenbanken spricht man oft von „Big Data“. In diesem Teilbereich der Informatik werden teilweise auch Methoden und Ansätze der KI angewandt, es geht insgesamt allerdings weit darüber hinaus.

Bis zum heutigen Tag hat sich in der KI noch keine integrierte, universell verfügbare Technologie entwickelt, sondern eher ein „Werkzeugkasten“ mit unterschiedlichen Ansätzen und Modulen, die für spezifische Anwendungen maßgeschneidert und zusammengestellt werden. Das hat sich auch mit der Einführung der generativen KI nicht geändert.

2.3 Paradigmen und Wirkzusammenhänge

In diesem Abschnitt sollen einige für die weitere philosophische Diskussion wichtige Wirkmechanismen erläutert werden. Beide Paradigmen, symbolische und subsymbolische KI, werden dabei berücksichtigt. Aus dem Bereich der symbolischen KI wird die Funktionsweise eines wissensbasierten Expertensystems und das probabilistische Schließen in Bayes-Netzen vorgestellt. Aus dem Bereich der subsymbolischen bzw. neuronalen KI mit numerischer Repräsentation stehen der Konnektionismus mit dem Beispiel des künstlichen neuronalen Netzes und das maschinelle Lernen einschließlich der Fehlerrückführung im Vordergrund. Die Darstellung der Ansätze für umfassende Datenanalysen, auch „Big Data" genannt, erfolgt anschließend[110]. Zum Abschluss werden noch einige Überlegungen zum „Black-Box-Problem" insbesondere in der neuronalen KI angestellt.

2.3.1 Symbolische und sub-symbolische KI

Wie schon im Kapitel 2.2 beschrieben, unterscheiden wir zwei Paradigmen der Künstlichen Intelligenz: die symbolische und die sub-symbolische KI. In den ersten Jahrzehnten der Entwicklung der Technologie, insbesondere während des ersten KI-Sommers, der von den 1950er Jahren bis in die 1970er Jahre anhielt, war das Paradigma der symbolischen KI dominierend. Problemlösungen und logische Schlussfolgerungen standen im Vordergrund. Die sub-symbolische KI hatte zwar ihren Ursprung in den 1940er Jahren mit den Arbeiten von McCulloch, Pitts und Hebb, gewann allerdings erst Ende der 1980er und zu Beginn der 1990er Jahre an Bedeutung. Beide Paradigmen sollen in der Folge kurz vorgestellt und verglichen werden[111].

Symbolische Methoden werden oft als „***Good Old Fashioned Artificial Intelligence (GOFAI)***"[112] beschrieben. Die Algorithmen der KI orientieren sich an von Menschen vorgegebenen kognitiven Abläufen und sind daher in der Regel für Menschen eingängig und nachvollziehbar. Eindeutige von menschlichen Programmierern vorgegebene Symbole werden mit den Methoden der Prädikatenlogik und der Mathematik manövriert. Es kommen auch Entscheidungsbäume und strukturierte Argumente zum Einsatz. Das menschliche Wissen ist in den Strukturen der Prozesse und in Datenbanken abgelegt, auf die die Technologie zurückgreift. Görz et al. schreiben dazu:

> *„Gegenstand der ‚symbolischen KI' sind folglich nicht das Gehirn und Prozesse des Abrufs von Gedächtnisinhalten, sondern vielmehr die **Bedeutung**, die sich einem Prozess aufgrund symbolischer Beschreibungen zuordnen lässt."*[113] Hervorhebung DS

[110] In dem Zusammenhang wird auch die Funktionsweise der generativen KI dargestellt

[111] Vgl. Ilkou Koutraki (2020), S. 1f; hier und im weiteren Verlauf dieses Abschnitts

[112] Ilkay Koutraki (2020), S. 2

[113] Görz et al. (2021), S. 12

Eine typische Anwendung ist das Expertensystem, das an späterer Stelle noch vorgestellt werden wird. Ein wesentliches Merkmal der symbolischen Methode ist die Transparenz der Abläufe und deren Nachvollziehbarkeit für den Menschen. Die Programme sind in der Regel modular aufgebaut. Die klar definierten Symbole erlauben den Transfer von Daten (und Wissen) zwischen den Modulen und Programmen. Mit den Systemen können zwar großen Datenmengen verarbeitet werden, allerdings stößt die Technologie bei verrauschten und dynamischen Daten aus der realen Welt an ihre Grenzen. So ist es beispielsweise nahezu ausgeschlossen mit der symbolischen KI eine Handschrifterkennung zu programmieren. Zu viele Abweichungen der individuellen Schreibweise von einer „Normschrift“ müssten im Programm vorgesehen werden, was sich trotz vieler Versuche als sehr fehlerhaft im praktischen Einsatz herausgestellt hat.

Sub-symbolische Methoden etablieren die Regeln und Abläufe, die dem Input den Output zuweisen selbst. Die dafür erforderlichen Verknüpfungen und Beziehungen sind außerordentlich komplex und für den Menschen, auch den Programmierer im Einzelfall nur mit großem Aufwand nachvollziehbar. Nur die grundsätzlichen Strukturen des Systems sind vom menschlichen Programmierer vorgegeben. Diese Strukturen orientieren sich am menschlichen Gehirn und werden deshalb auch Künstliche Neurale Netzwerke (KNN) (auf Englisch: Artificial Neural Networks, ANN) genannt. Statistische Methoden kommen zum Einsatz, was auch den Umgang mit Unsicherheiten und verrauschten Daten erleichtert. Probleme liegen in der Intransparenz der internen Abläufe und den daraus erwachsenen Risiken und der hohen Abhängigkeit von Trainingsdaten.

Tabelle 2: Attribute der symbolischen und sub-symbolischen KI[114]

Symbolische KI	Sub-symbolische KI
Symbole	Zahlen
Logisch	Assoziativ, inhaltsorientiert
Serielle Datenverarbeitung	Parallele Datenverarbeitung
Logische Abläufe und Schlussfolgerungen	Lernend aus der Vorgabe von Soll-Output zum Input
Starr und statisch	Dynamisch und anpassungsfähig
Konzepterstellung und -erweiterung durch den Menschen	Konzepterstellung durch den Menschen und deren Verallgemeinerung durch die Maschine
Abstraktion der Modelle (erforderlich für das Programm)	Anpassen der inneren Strukturen und Variablen an Daten
Menschlicher Eingriff und Kontrolle	Lernen von Daten
Komplexe Programme, begrenzte Daten	Einfache Programme, „big data“
Wörtliche / präzise Eingabe	Verrauschte/unvollständige Eingabe

In der neueren Zeit hat man begonnen die beiden Paradigmen in sogenannten „*In-between Methods*“[115] zu kombinieren, was den Ausgleich der komplementären Vor- und Nachteile gestattet.

In den folgenden Abschnitten sind die Wirkungsweisen einiger zentraler Technologien der symbolischen KI (wissensbasierte Expertensysteme, Bayes’sche Netzwerke) und der sub-symbolischen KI (Konnektionismus/Künstliche Neuronale Netze, Maschinelles Lernen, Support-Vektor Maschinen, Big Data) dargestellt.

2.3.2 Wissensbasierte Expertensysteme

Vereinfacht dargestellt handelt es sich bei wissensbasierten Expertensystemen um Programme, die Wissen über ausgewählte Gebiete sammeln und speichern und aus dem Wissen deduktiv Schlussfolgerungen ableiten und für Probleme Handlungsempfehlungen entwickeln. Das Wissen des Expertensystems ist sehr spezifisch für die zu bearbeitende Aufgabenstellung und nicht allgemein und umfassend.

Zwei Arten von Wissen sind in Expertensystemen abgelegt. Bei der ersten Art des Wissens geht es um die Fakten des gewählten Anwendungsbereiches, die man in Lehrbüchern, Zeitschriften oder Enzyklopädien findet. Viel schwieriger darzustellen ist die

[114] Vgl. Ilkou Koutraki (2020), S. 3, Tabelle 1; übersetzt und angepasst durch DS

[115] Ilkou Koutraki (2020), S. 3

zweite Art von Wissen: das heuristische Wissen, das ein menschlicher Experte, wie zum Beispiel ein Arzt oder Ingenieur, über Jahre entwickelt. Es beinhaltet ein spezifisches Urteilsvermögen, das der Mensch aufgrund gemachter Erfahrung herausbildet. Speziell geschulte „Wissensingenieure" programmieren das heuristische Wissen in Form von „Wenn-dann" Regeln. Hierbei handelt es sich um regelbasierte Expertensysteme, die auch als Entscheidungsbäume dargestellt werden können. Alternativ gibt es auch fallbasierte Expertensysteme, in denen die Wissensbasis aus einer großen Falldatenbank besteht. In der Anwendung sucht das System ähnliche Fälle, die für eine Diagnose oder Empfehlung herangezogen werden. Insbesondere bei der medizinischen Diagnostik, bei der Fehlersuche in der Technik oder im juristischen Bereich wird dies eingesetzt.

2.3.3 Bayes'sche Netze

Sowohl menschliche Entscheidungen als auch diejenigen von Maschinen erfolgen häufig in Situationen der Unsicherheit. Für den Umgang mit diesen Unsicherheiten werden Methoden der Wahrscheinlichkeitstheorie angewandt. Ein „*grundlegender Begriff in der Wahrscheinlichkeitstheorie ist derjenige der bedingten Wahrscheinlichkeit"*[116]. Wahrscheinlichkeiten sind oftmals abhängig von der Erfüllung bestimmter Vorbedingungen.

Hierfür ein Beispiel: Die Wahrscheinlichkeit für das Auslösen einer Alarmanlage in einer Wohnung ist abhängig vom Vorliegen eines Einbruchs. Wenn ein Einbruch geschieht, sollte die Wahrscheinlichkeit des Auslösens der Anlage hoch sein, und wenn kein Einbruch vorliegt, niedrig, idealerweise nahe Null. Es gibt also eine bedingte Wahrscheinlichkeit bei vorliegendem Einbruch von 99% oder mehr (hoffentlich nicht geringer) und eine bedingte Wahrscheinlichkeit von – wie hier angenommen – 1%, wenn kein Einbruch vorliegt. Hier stellt sich nun die Frage: Wie hoch ist bei Auslösen der Alarmanlage die Wahrscheinlichkeit, dass tatsächlich ein Einbruch stattgefunden hat, wenn gleichzeitig die Wahrscheinlichkeit für einen Einbruch bei 0,1% an einem beliebigen Tag liegt?

Mit dieser Art von Fragen hat sich bereits im späten 18. Jahrhundert Thomas Bayes beschäftigt und die nach ihm benannte Theorie der Berechnung bedingter Wahrscheinlichkeiten etabliert (Satz von Bayes[117]). Die Auflösung für das obige Beispiel lautet: ca. 9%[118]. Damit ist berechnet, dass bei Auslösen des Alarms mit 91%iger Wahrscheinlichkeit ein

---

[116] Harrach (2014), S. 128

[117] „**Satz von Bayes**: *Theorem aus der Wahrscheinlichkeitsrechnung zur Berechnung bedingter Wahrscheinlichkeiten. Bezeichnen w(A) und w(B) für das Eintreten der stochastisch abhängigen Ereignisse A und B ,und bezeichnet weiterhin w(B I A) die (bedingte) Wahrscheinlichkeit für das Ereignis B unter der Bedingung, dass A eingetreten ist, so gilt für die umgekehrte bedingte Wahrscheinlichkeit w(A I B) nach dem Bayes-Theorem: w(A | B) = (w(B | A) * w(A)) / w(B)"*; (Quelle: Gabler Online Wirtschaftslexikon)

[118] W(Einbruch | Alarm) = (W(Alarm | Einbruch) * W(Einbruch))/(W(Alarm)
Mit W(Alarm) = W(Einbruch)*W(Alarm | Einbruch)+W(Kein Einbruch)* (Alarm | kein Einbruch)
ergibt sich: W(Alarm) = 0,01098 = ca. 1,1% und
W(Einbruch | Alarm) = (0,99 * 0,001)/0,01098 = 0,00099/0,01098 = 0,090164 = **ca. 9%**

Fehlalarm vorliegt. Deswegen ist es auch nicht sinnvoll, sofort die Polizei zu informieren, wenn die Alarmanlage auslöst.

In der medizinischen Diagnostik, bei der Suche nach technischen Problemen in Produktionsanlagen oder Flugzeugen oder in Warenkorbanalysen beim Online Versandhandel sind diese Arten von bedingten Wahrscheinlichkeiten über mehrere Ebenen kaskadiert. Dafür werden dann Bayes'sche Netze eingesetzt, oftmals kombiniert mit maschinellem Lernen und/oder als Bestandteil von Expertensystemen. Sie konstruieren eine hochdimensionale Wahrscheinlichkeitsverteilung, basierend auf lokalen probalistischen Regeln.

2.3.4 Konnektionismus/Künstliche Neuronale Netze[119]

Wie schon im vorherigen Kapitel angesprochen, besteht die grundsätzliche Idee bei den künstlichen neuronalen Netzen (KNN) darin, dass sie das menschliche Gehirn simulieren. Sie bestehen aus Knoten, die Neuronen entsprechen, und aus Kanten, die Synapsen entsprechen. Aufgrund der Vernetzung der Knoten und Kanten spricht man auch vom Konnektionismus. In der folgenden Skizze ist ein KNN schematisch dargestellt.

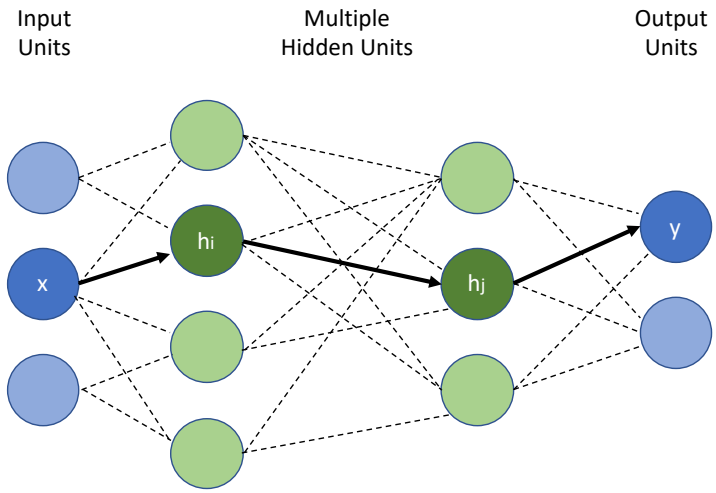

Abbildung 4: Skizzenhafte Darstellung eines KNN[120]

In der einfachen Form werden drei Typen von Knoten (Neuronen) unterschieden, hier „Units" genannt:

- *Input-Units enthalten die Eingangsdaten; dabei kann es sich zum Beispiel um die Pixel eines Bildes oder Laborwerte in einer medizinischen Anwendung handeln.*

[119] Darstellung in diesem Abschnitt einschließlich der Formeln zitiert aus Buxmann Schmidt (2019), Seiten 13-15

[120] Buxmann Schmidt (2019), Seite 14

- *Output-Units enthalten die Ausgangsdaten, beispielsweise die Klassifikation „Hund“ oder „Katze“ bei einer Bilderkennung oder Zahlen und Buchstaben bei der Schrifterkennung.*
- *Hidden-Units sind dazwischen angeordnet und bilden das Innere des Künstlichen Neuronalen Netzwerks, das aus mehreren Ebenen bestehen kann.*[121]

Das Grundprinzip ist in Abbildung 3 dargestellt. In jeder Ebene sind mehrere Units enthalten. Knoten sind durch Kanten miteinander verbunden. Die Kante ij liegt zwischen den zwei Knoten i und j. Ihr ist das Gewicht w_{ij} zugeordnet. Die Gesamtheit aller Gewichte der Kanten des Netzwerks repräsentiert das erlernte Wissen. Die mathematische Darstellung erfolgt in Matrizen.

„Der Input, den ein Neuron von anderen erhält, hängt von dem Output des sendenden Neurons bzw. der sendenden Neuronen und den Gewichten entlang der Kanten ab. Bezeichnet $Output_i$ das Aktivitätslevel eines sendenden Neurons i, so lässt sich der Input, den ein Neuron j erhält, mit der folgenden Formel ausdrücken:

$$Input_j = \sum (Output_i \times w_{ij}) + b_j$$

Der Output eines Neurons wiederum basiert auf dem Input und einer Aktivierungsfunktion. Bei dieser Aktivierungsfunktion a sind verschiedene Funktionstypen denkbar – im einfachsten Fall ist sie linear.

$$Output_i = a\,(Input_i)\text{“}$$[122]

Wie oben beschrieben: Das „Wissen des KNN“ wird durch die Gewichte wij repräsentiert. Diese werden beim überwachten Lernen auf Basis von Lernregeln in dem Sinne modifiziert und angepasst, dass der festgestellte Fehler bei der Anwendung des Systems bei Trainingsdaten minimiert wird. Das dafür übliche Verfahren ist die sogenannte Fehlerrückführung (Backpropagation) . In einem Autoadaptionsverfahren werden die Gewichte iterativ in einem fortwährenden Vergleich des Outputs des KNN in der Ausgangsschicht mit der Vorgabe angepasst. Aus der Abweichung des Outputs vom Soll ergibt sich ein Fehler. Der anteilige Beitrag der Hidden Units zu diesem Fehler wird bestimmt und die entsprechenden Gewichte werden so lange angepasst bis nur noch ein vertretbarer oder kein Fehler mehr auftritt.

Im Folgenden ist die Handschrifterkennung in einem neuronalen Netz erläutert:

> *„Betrachten wir dazu ein einfaches Beispiel: ein neuronales Netz, welches handschriftlich notierte Ziffern erkennen soll. Ein solches Netz wird in unserem einfachen Beispiel als ein ‚multilayer perceptron‘ aufgebaut. Die erste Schicht enthält z.B. 784 Neuronen, welche einem Gitter mit der Auflösung von 28x28 Pixeln entspricht. Dieses Gitter wird unter die*

[121] Vgl. Buxmann Schmidt (2019), S. 14

[122] Buxmann Schmidt (2019), S. 14f

handschriftlich notierte Ziffer gelegt, so dass für jede Box der jeweilige Licht- und Schattenwert bestimmt werden kann. Jedes der 784 Neuronen weist daher einen Wert zwischen 0 und 1 auf; dieser Wert entspricht dem Farbwert (von 0 für weiß bis 1 für schwarz). Die letzte Schicht weist 10 Neuronen auf, denen die zu erkennende Zahlenwerte (0-9) entsprechen. Diese letzte Schichte kann ebenfalls wieder Werte zwischen 0 und 1 aufweisen; dem entspricht die Wahrscheinlichkeit, mit der der Zahlenwert der handschriftlichen Ziffer (von 0-9) korrekt erkannt wurde. Dazwischen finden sich, in diesem simplen Beispiel, zwei so genannte ‚hidden layers' mit jeweils 16 Neuronen. Jedes Neuron der einen Schicht ist mit allen Neuronen der nächsten Schicht verbunden. Zudem werden so genannte Gewichte und Biaswerte eingeführt. In diesem Netz ergeben sich so 13.002 Parameter. Der Lernvorgang besteht darin, die Werte für die Parameter so zu verändern, dass die handschriftlichen Ziffern bestmöglich in ihren Zahlenwerten erkannt werden. Dafür werden wiederum Trainingsdaten verwendet, bei denen der Zahlenwert annotiert ist; der lernende Algorithmus kann so entsprechend der Abweichungen sein Modell verbessern, bis es den richtigen Zahlenwert erkennt (überwachtes Lernen)."[123]

Wie beschrieben, sind die Neuronen in unterschiedlichen Schichten organisiert. Bei der Analyse von Photographien würde die erste Schicht zwischen hellen und dunklen Pixeln unterscheiden, die nächste erkennt Kanten, die dritte horizontale und vertikale Linien, eine weitere würde Augen erkennen und schließlich ein menschliches Gesicht. Wichtig ist, dass dies nicht in die Struktur hineinprogrammiert wird, sondern sich aus dem iterativen Lernprozess ergibt.[124]

Insbesondere für die komplexe Bilderkennung ist neben dem bereits beschriebenen Prozess der Fehlerrückführung und iterativen Anpassung der Parameter die Technologie der Faltungsnetze (CNN; Convolutional Neural Networks) bedeutsam. Mit diesem Verfahren werden durch geeignete „Faltungen"(„Convolutions") in den Bildern sich wiederholende Strukturen erkannt, wie zum Beispiel Linien, Kanten oder Farbtupfer. Diese werden dann in einer weiteren Ebene des KNN zu Basis-Strukturen wie Kurven, Kreisen und anderen Formen kombiniert. Mit jeder Ebene steigt der Abstraktionsgrad bis hin zur Erkennung eines komplexen Objektes oder eines Gesichts[125].

2.3.5 Maschinelles Lernen

Als „Maschinelles Lernen" bezeichnet man in der Künstlichen Intelligenz Methoden, die es erlauben, Muster und Zusammenhänge in Datensätzen zu erkennen und darauf basierend Klassifizierungen oder auch Vorhersagen zu treffen. Im Englischen spricht man von

[123] Kaminski (2020), S. 158

[124] Vgl. Jones (2014), S. 147: "*The strategy called for simulated neurons to be organized into several layers. Give such a system a picture and the first layer of learning will simply notice all the dark and light pixels. The next layer might realize that some of these pixels form edges; the next might distinguish between horizontal and vertical lines. Eventually, a layer might recognize eyes, and might realize that two eyes are usually present in a human face."*

[125] Quelle: https://jaai.de/convolutional-neural-networks-cnn-aufbau-funktion-und-anwendungsgebiete-1691/

„Machine Learning“ oder bei größerer Komplexität und höherer Tiefe der genutzten Strukturen des neuronalen Netzes von „Deep Learning“. Es ist also nicht zu verwechseln mit E-Learning oder Lernen mit Maschinen. Informatiker sprechen auch von der Musterbildung autoadaptiver Systeme. Damit sind Algorithmen gemeint, „*die ihre Struktur auf Basis von Sensordaten auf eine bestimmte Weise adaptieren*“[126]. Maschinelles Lernen ist also kein maschinengestütztes Lernen von Menschen, sondern ein Autoadaptionsprozess, der Strukturvorschläge und -veränderungen hervorbringt, die ihrerseits in ihrer Anwendung die Ergebnisse hervorbringen.

> *„Maschinelles Lernen […] beschreibt Algorithmen, die Rohdaten erhalten und darin Reizkonstellationen identifizieren, die eine bestimmte Regelmäßigkeit aufweisen.“*[127]

Der Paradigmenwechsel des maschinellen Lernens gegenüber Expertensystemen besteht darin, dass der Softwareprogrammierer nicht mehr das Expertenwissen für die Nutzung im System codieren muss und die „wenn-dann“ („if-then“) Entscheidungen in Baumstrukturen vordenken muss. Das System wird auf Basis von Erfahrungen trainiert; die konkrete Umsetzung in den internen Strukturen der KI (i.d.R. des künstlichen neuronalen Netzes) erfolgt durch das System selbst, z.B. über Fehlerrückführung (Backpropagation).

Ein Beispiel dieses Ansatzes ist das Identifizieren von Objekten in Bildern. Für das Erkennen von bestimmten Tieren auf Bildern muss der Softwareentwickler nicht mehr konkrete Merkmale verschiedener Tiere (z.B. Hunde, Katzen und Mäuse) einprogrammieren (z.B. Größe, Kopfform, Pfoten, Krallen, Fell, usw.), sondern er trainiert das System mit unterschiedlichen Bildern und Beispielen. Damit erlernt das System selbst, wie sich diese Tiere von anderen Objekten und untereinander unterscheiden. Grundsätzlich ähnelt dies dem Lernen bei kleinen Kindern, wenn man ihnen Objekte zeigt und dazu den Namen oder Gattungsbegriff nennt.

In der KI-Literatur[128] werden üblicherweise drei Arten von maschinellem Lernen unterschieden: ***überwachtes Lernen*** („*supervised learning*“), ***unüberwachtes Lernen*** („*unsupervised learning*“) und ***verstärkendes Lernen*** („*reinforcement learning*“). Beim überwachten Lernen lernt das System aus vorhandenen Trainingsdaten, die vorab eingegeben werden. Beim obigen Tier-Beispiel würde man mehrere tausend Tierfotos bereitstellen und das System damit anlernen. Dann wird in mehreren Iterationen die Präzision und Treffgenauigkeit mit Testdaten überprüft. Die Kategorien, im Beispiel Hund, Katze und Maus, wird vorgegeben und kann vom System nicht mehr verändert werden. Beim unüberwachten Lernen werden keine Kategorien vorgegeben, das System findet selbst Kategorien und Muster, was bei Tierbildern problematisch und in der medizinischen Diagnostik oder im Marketing durchaus vorteilhaft sein kann. Beim verstärkenden Lernen soll

[126] Harrach (2014), S. 21

[127] Ebd.

[128] Vgl. Russell Norvig (2004), S. 811; Ertel (2008), S. 301 und 313

für ein gegebenes Problem eine optimale Strategie erlernt werden. Dafür wird eine Anreiz- und Belohnungsfunktion eingesetzt. Dies wird zum Beispiel in Navigationssystemen oder auch bei Robotern genutzt.

Für das Verständnis dieses Lernens ist zu betonen, dass das System nicht etwa die Bedeutung dessen lernt, was es erfasst, sondern lediglich eine Ordnungsstruktur schafft, wie im folgenden Zitat von Andreas Kaminski beschrieben wird:

> *„In allen betrachteten Fällen besteht der Lernvorgang aber darin, für gegebene Daten eine Funktion zu finden. Diese Funktion entspricht* [...] *einem Modell, das die Ordnung der gegebenen Daten beschreibt. Diese Ordnung weist einen Zeitindex auf, es handelt sich um eine Art Hypothese, dass das Modell nicht nur die vergangenen, sondern auch die zukünftigen Daten beschreibt. Aufgrund dieser Zeitlichkeit ist die Modellbildung dynamisch, was die Dynamik der Transformation kennzeichnet. Einige Unterschiede zwischen den genannten Lernstrategien gehen darauf zurück, wie dies erfolgt: Der Hypothesenraum wird von den verschiedenen Lernstrategien anfänglich auf unterschiedliche Weise entworfen bzw. später dann unterschiedlich angepasst"*[129]

Aus dem Lernprozess entwickelt das KNN für sich in seiner Struktur eine Ordnung oder auch Hypothesen für Regelmäßigkeiten und Zusammenhänge. Dies gilt insbesondere für komplexere „Deep Learning" Ansätze des verstärkenden Lernens. Die drei maßgeblichen Pioniere dieser Ausprägung, Yann LeCun, Yoshua Bengio und Geoffrey Hinton, fassen die Technologie in einem gemeinsamen Aufsatz wie folgt zusammen:

> *"Deep learning allows computational models that are composed of multiple processing layers to learn representation of data with multiple levels of abstraction. These methods have dramatically improved the state-of-the art in speech recognition, visual object recognition, object detection and many other domains such as drug discovery and genomics. Deep learning discovers intricate structure in large data sets by using the backpropagation algorithm to indicate how a machine should change its internal parameters that are used to compute the representation in each layer from the representation in the previous layer. Deep convolutional nets have brought about breakthroughs in processing images, video, speech and audio, whereas recurrent nets have shone light on sequential data such as text and speech."*[130]

Für die weitere Entwicklung des Deep Learnings rechnen LeCun et al. mit großen Fortschritten des unüberwachten Lernens, weil – so argumentieren sie – auch menschliches und tierliches Lernen unüberwacht stattfindet, allein durch das Beobachten der Strukturen der Welt, ohne den Namen eines jeden Objekts zu kennen[131]. Auch knüpfen sie hohe Erwartungen an die Kombination von Maschinenlernen mit Logik („*From Machine Learning to Machine Reasoning*").

[129] Kaminski (2012), S. 75

[130] LeCun Bengio Hinton (2015), S. 436

[131] Vgl. LeCun Bengio Hinton, S. 442

2.3.6 Support-Vektor-Maschinen

Beim maschinellen Lernen wird oftmals das mathematische Verfahren der Support-Vektor-Maschine (SVM) eingesetzt. Eine SVM erlaubt die Klassifizierung von Objekten (linear und nichtlinear), insbesondere in der Bild-, Text- oder Handschrifterkennung.

Technisch betrachtet, handelt es sich bei der Support-Vektor-Maschine um eine mathematische und statistische Methodik, die in einem Algorithmus realisiert ist, der mit einem sogenannten „*Large Margin Classifier*" Grenzen zwischen verschiedenen Klassen von Objekten ermittelt und dabei einen möglichst breiten Bereich anstrebt, der frei von Objekten ist. Beim Einsatz für *nichtlineare Klassifizierung* (wenn die Grenze zwischen zwei Gruppen von Objekten durch eine nichtlineare Kurve beschreibbar ist) wird der sogenannte *Kernel Trick* genutzt. Durch mathematische Operationen werden zusätzliche Dimensionen definiert, in denen die nichtlineare Grenze zwischen den Objekten linear wird und Objekte besser separierbar sind. Diese zusätzlichen Dimensionen nennt man Hyperebenen[132].

Die ersten Ideen zu dieser Methodik kamen bereits in den 1930er Jahren auf. Viele Jahre später wurden SVM durch Wladimir Wapnik und Alexei Jakowlewitsch Tscherwonenkis entwickelt. Vermehrt zum Einsatz kommen sie beim maschinellen Lernen seit den 1990er Jahren.

2.3.7 Big Data

Wenn es eines alternativen Begriffs für das mit „Big Data" bezeichnete Phänomen bedurft hätte, dann wäre dies das Wort „Muster"[133]. Bei Big Data geht es darum, mit den Methoden der KI, insbesondere mit maschinellem Lernen, Muster in großen Datenmengen zu erkennen. Für den Prozess der Findung von Mustern und Strukturen in großen Datenmengen bzw. ausgeprägten Zusammenhängen zwischen Daten spricht man auch von Data Mining. In welchem Umfang Big Data ein Teilgebiet der KI ist, bleibt auch unter den Experten umstritten[134]. Die Analyse großer Datenmengen für sich allein genommen ist noch kein Attribut der KI. Erst die Verwendung etwa von künstlichen neuronalen Netzen und der Einsatz von maschinellem Lernen rechtfertigen dies.

Beispiele für Big Data und den Einsatz von Data Mining finden sich zuhauf: die Analyse von Schadensfällen bei Versicherungen, die Bestimmung der Kreditwürdigkeit von Bankkunden, die Auswertung diagnostischer Daten in der Medizin (z.B. Radiologie oder Labordiagnostik) und die Auswertung von Suchbegriffen in Internetsuchmaschinen. Immer geht es um die Identifikation von Regelmäßigkeiten, die grundsätzlich auch ohne

[132] Vgl. auch BIGDATA INSIDER „Was ist eine Support Vector Machine?" und Russell Norvig (2004), S. 863; Vgl. auch Glossar dieser Arbeit

[133] Vgl. auch: Nassehi (2019)

[134] Vgl. Kersting Meyer (2017)

maschinelles Lernen von Menschen erkannt werden könnten, was aber allein aufgrund des stark wachsenden Umfangs der verfügbaren und einzubeziehenden Datenmengen nicht realistisch ist.

In Zeiten der Rechnerallgegenwart (auch „ubiquitäres Computing“ bzw. „Ubiquitous Computing“[135]) nimmt die Zahl der eingesetzten Computer und der verwendeten Sensoren in allen Bereichen der Gesellschaft signifikant zu und die damit produzierten Daten sogar exponentiell. Hierzu zählt insbesondere das Internet einschließlich der sozialen Netzwerke und des sogenannten „Internet der Dinge“ (z.B. Alarm- und Sicherheitstechnik, allerlei Sensorik am menschlichen Körper, sogenannte „Wearables“, „smarte“ Haushaltsgeräte, Computersysteme in Automobilen und Systeme in der öffentlichen Infrastruktur). Es entstehen Audio-, Video und Bilddaten, Labordaten, Sensordaten, oftmals verknüpft mit Geodaten (GPS).

Big Data hilft beim Erkennen von Mustern und Regelmäßigkeiten in den Daten, was wiederum dazu beiträgt, Hypothesen für mögliche Ursache-Wirkung-Zusammenhänge zu bestimmen, zum Beispiel in der Medizin oder in den Sozialwissenschaften. Es geht dabei weder um Deduktion noch um Induktion, sondern um Abduktion. Deduktion ist der „wasserdichte Schluss“ im Sinne der Prädikatenlogik, der hingegen nicht wissenserweiternd ist. Induktion ist dagegen wissenserweiternd, aber nach Karl Popper falsifizierbar. Abduktion ist nicht zwingend wissenserweiternd, unterstützt aber die Entwicklung von Hypothesen für mögliche Zusammenhänge. Begründet wurde das heutige Verständnis der Abduktion durch Charles Sanders Peirce:

> *„Abduktion ist der Prozess, eine erklärende Hypothese zu bilden. Es ist die einzige Operation, die irgendeine Idee einführt; denn Induktion determiniert nur einen Wert und Deduktion entwickelt nur die notwendigen Folgen aus einer reinen Hypothese. Deduktion beweist, dass etwas sein muss; Induktion zeigt, dass etwas tatsächlich wirkt; Abduktion legt nur nahe, dass etwas sein kann.* […] *Doch jeder einzelne Punkt einer wissenschaftlichen Theorie, die heute feststeht, wird der Abduktion verdankt.“*[136]

Beispiele für wissenschaftliche Theorien, die zunächst über abduktive Schlüsse entdeckt wurden, gibt es vielfach: die Entdeckung des Penicillins durch Alexander Fleming oder die der Röntgenstrahlung durch Wilhelm Conrad Röntgen, aber auch die Entdeckung der Higgs-Teilchen mit Big Data Methoden im CERN[137].

[135] Vgl. Lyytinen Yoo (2002), S. S. 64: Ubiquitous Computing = High Level of Embeddedness + High Level of Mobility

[136] Peirce (1973), S. 115

[137] Vgl. Mainzer (2014), S. 233

Christian Wadephul schreibt in seinem Aufsatz „*Führt Big Data zur abduktiven Wende in den Wissenschaften*?“[138] unter Bezugnahme auf Hans-Jörg Rheinberger[139] davon, dass sich neben den bisherigen *drei Säulen wissenschaftlicher Erkenntnisgewinnung* nämlich *Theorie, Experiment und Simulation* eine *vierte Säule datenbasierter Methoden*[140] etabliert habe. Trotzdem kommt er zu dem Schluss, dass *Big Data kein Theorieersatz sein könne*[141]. Der oben schon zitierte Logiker Peirce schrieb – ohne Big Data gekannt zu haben – in seinen „*Maximen des Pragmatismus*“[142]:

> *„Der dritte Schleifsteinsatz* [Anmerkung DS: Peirce formuliert in seinen Maximen drei „Schleifsteinsätze“] *lautet, dass abduktives Schließen allmählich, ohne scharfe Trennungslinie, in ein Wahrnehmungsurteil übergeht; oder mit anderen Worten, unsere ersten Prämissen, die Wahrnehmungsurteile, müssen als ein extremer Fall abduktiven Schließens angesehen werden, von dem sie sich dadurch unterscheiden, dass sie absolut jenseits von Kritik sind. Die abduktive Vermutung kommt uns blitzartig. Sie ist ein Akt der Einsicht,* ***obwohl von außerordentlich trügerischer Einsicht****.“*[143] [Hervorhebung DS]

Für die technische Realisierung von Big Data Untersuchungen wurde zum Beispiel ein sogenannter *MapReduce-Algorithmus* entwickelt, der zunächst die Gesamtdaten in Datenpakete unterteilt, in denen parallel identische Suchprozesse ablaufen, deren Ergebnis dann in *Zwischenergebnislisten zusammengefügt* werden, die schließlich ihrerseits konsolidiert werden.[144]

„Das abduktive Schlussverfahren“, so fasst Armin Nassehi zusammen, „*ermöglicht es* […] *einem wahrnehmenden Operator, selbst die Struktur des Objekts zu identifizieren und sich an früheren Wahrnehmungen zu orientieren“:*

> *„Solche lernenden Systeme verbessern ihre Genauigkeit durch Erfahrung, sind aber darin nicht auf eindeutige Ergebnisse festgelegt. Das abduktive Verfahren führt stets zu stochastischen Lösungen, zu Annäherungen, zu hypothetischen Lösungen. Es beinhaltet immer ein Rest Unbestimmtheit, der in streng deduktiven und induktiven Verfahren nicht*

---

138 Wadephul (2016), Seiten 1 - 14

139 Wadephul bezieht sich auf: Rheinberger (2007), S. 123 - 124; ergänzendes Zitat aus gleicher Quelle (S. 123): „*Im Zeitalter der Molekularbiologie machte eine Spur nur Sinn und gewann nur Bedeutung als Datum, wenn man sie zu einem vermuteten Sachverhalt in Beziehung setzen konnte, einem möglichen Faktum im gängigen Sprachgebrauch der Wissenschaften. Man suchte also nach datenfähigen Spuren, um Fakten darzustellen und abzusichern. Heute macht ein Datum eher Sinn und gewinnt auch nur an Bedeutung, wenn man strukturiert darauf zugreifen kann. Es ist in diesem Zusammenhang geradezu von einer sich gegenwärtig vollziehenden epistemologischen Revolution gesprochen worden: Man ist, so lautet die Behauptung, von der* ***hypothesengeleiteten zur datengeleiteten Forschung*** [Hervorhebung DS] *übergegangen. Das heißt: Spuren werden nicht mehr im Lichte von Phänomenen generiert, sondern als Daten gepoolt, um gegebenenfalls noch unbekannten, neuen Fakten ans Licht zu verhelfen.“*

140 Vgl. Wadephul (2016), S. 1

141 Vgl. Wadephul (2016), S. 13

142 Peirce (1973), Seiten 122-123

143 Ebd.

144 Vgl. Mainzer (2014), S. 234

> *gegeben ist, weil dort gewissermaßen unwandelbare, also axiomatische „Gesetze" in Form von konkreten Festlegungen vorliegen."*[145]

In Bezug auf die induktiven Verfahren ist die Aussage von Armin Nassehi einzuschränken. Nach Karl Popper sind induktive Wahrheiten immer falsifizierbar, also vorläufig wie die abduktiv erworbenen Erkenntnisse. In Bezug auf die Abgrenzung zwischen deduktiven und abduktiven Verfahren ist Nassehi in der Tat zuzustimmen. Damit ist die „*abduktive Logik von deep learning systems* […] *letztlich auf die Vorläufigkeit aller Lösungen aufgebaut"*[146]:

> *„Solche Lösungen sind Lösungen in einer bestimmten Gegenwart, die durch den nächsten Schritt erweitert, korrigiert, bestätigt oder verworfen werden, was einer abduktiv arbeitenden Software letztlich niemals so etwas wie eine Letztentscheidung ermöglicht – eine solche wird letztlich dadurch erzwungen, dass sie operativ getroffen werden muss."*

Dies ist „*eine heuristische Form des Lernens"*[147].

Echtes Lernen erfordert allerdings die Kombination der erkannten Muster mit erklärenden kausalen Zusammenhängen. Judea Pearl warnt vor einem von Statistik und einer „Machine Learning Culture" dominierten datenzentrierten Denken, das rationale Entscheidungen ausschließlich aus Daten ableitet[148]:

> *„ 'Finding a needle in a haystack is difficult, and it is probably impossible if you haven't seen a needle before.' Most machine learning researchers today have not seen a needle (i.e., a causal model drawing inferences on interventions and counterfactuals); an educational hindrance that needs to be corrected in order to hasten the discovery of the learning principles we aspire to uncover."*[149]

Nach Pearl kann echtes Lernen und echte Wissensgenerierung nur aus einer Balance aus Daten und von Menschen gemachten Modellen der Datengenerierung (und angenommenen kausalen Zusammenhängen) erfolgen:

> *„A hybrid strategy balancing ‚data-fitting' with 'data-interpretation' better captures the stages of knowledge compilation that the evolutionary processes entail."*[150]

Die Wissenschaft wird auch in Zukunft auf Probleme, Gedanken, Theorien und von Menschen entwickelte Experimente aufbauen[151].

---

145 Nassehi (2019), S. 235

146 Nassehi (2019), S. 241f; einschließlich des folgenden Blockzitats

147 Nassehi (2019), S. 241f

148 Vgl. Pearl (2021), S. 80

149 Pearl (2021), S. 82

150 Pearl (2021), S. 82

151 Vgl. Frické (2015), S. 660; Kitchin (2014)

2.3.8 Generative AI

Mittels einer Kombination der drei Lernverfahren (überwacht, unüberwacht und verstärkend) konnte eine neue Kategorie von KI-Systemen entwickelt werden, die man als „Generative AI“ bezeichnet[152]. Darunter versteht man Systeme, die Inhalte generieren können, wie zum Beispiel Texte, Bilder, Videos, Audios oder Softwarecodes[153]. Es werden also nicht nur vorgegebene Daten ausgewertet und klassifiziert, sondern auch neue Inhalte produziert bzw. generiert, die einer (vorher entwickelten) Klassifizierung genügen. Die Verwendung der Vokabel “generieren“ ist problematisch, da diese Systeme keinesfalls mit Kreativität und Sachverstand neue Inhalte hervorbringen, sondern ausschließlich mit statistischen Modellen, die auf Grundlage von Basisdaten, die dem System zur Verfügung gestellt werden, den wahrscheinlichsten Output erzeugen. Die Systeme haben weder ein Verständnis für die ihnen gestellten Fragen noch für die Antworten.

Das folgende Beispiel soll die grundsätzliche Funktionsweise illustrieren. Zum Zwecke der Gesichtserkennung wurden Programme und Modelle entwickelt, die nach einem Anlernen mit einzelnen Fotografien von Personen diese auf Bildern von Überwachungskameras erkennen können. Genau diese Art von Modellen kann allerdings auch genutzt werden, um mögliche bzw. erwartbare Bilder von Personen zum Beispiel nach fünf oder zehn Jahren Alterung zu erzeugen. Die Modelle dienen also dazu, neue Bilder zu erzeugen, die die gleichen Merkmale aufweisen, jedoch – in diesem Beispiel – einem Alterungsprozess unterworfen sind.

Eine ähnliche Anwendung wäre die Erstellung von Landkarten anhand von Satellitenbildern. Existierende Applikationen aus dieser Kategorie sind die Systeme ChatGPT (Textgenerator), DALL-E (Bildgenerator auf Basis von Text) von Open AI oder BERT (Spracherzeugung) von Google.

Die viel diskutierte Applikation ChatGPT von OpenAI gehört zur Kategorie der Large Language Models (LLM). Sie ist in der Lage, Texte unterschiedlicher Schreibstile und für diverse Zwecke mit hoher Präzision, Detailgenauigkeit und Kohärenz zu generieren. Trainiert wird das neuronale Netz des Systems mit unzähligen von Menschen erstellten Texten aus dem Internet, aus Büchern und aus den Medien. Ziel des Trainings ist die Erzeugung bzw. Generierung von Texten, die Bezug nehmen auf die Text-, Satz- und Wortstrukturen und abgeleiteten Statistiken der Trainingsdaten. Auf Basis von „Eingabeaufforderungen“ (Prompts) erstellt das System Texte, die den Trainingstexten entsprechen sollten. Dreh- und Angelpunkt der Funktionsweise des Modells ist das Lernen der statistischen Sprachstrukturen wie z.B. Wortsequenzen und Muster der Wortverwendung. Wie oben bereits angesprochen, geht es nicht um die Inhalte oder Bedeutungen der Texte.

[152] Vgl. Cao et al. (2023)

[153] Vgl. McKinsey (2023)

Trainiert wird das System in drei Schritten, wie in Abbildung 5 für den Vorläufer von ChatGPT GPT-3 dargestellt.

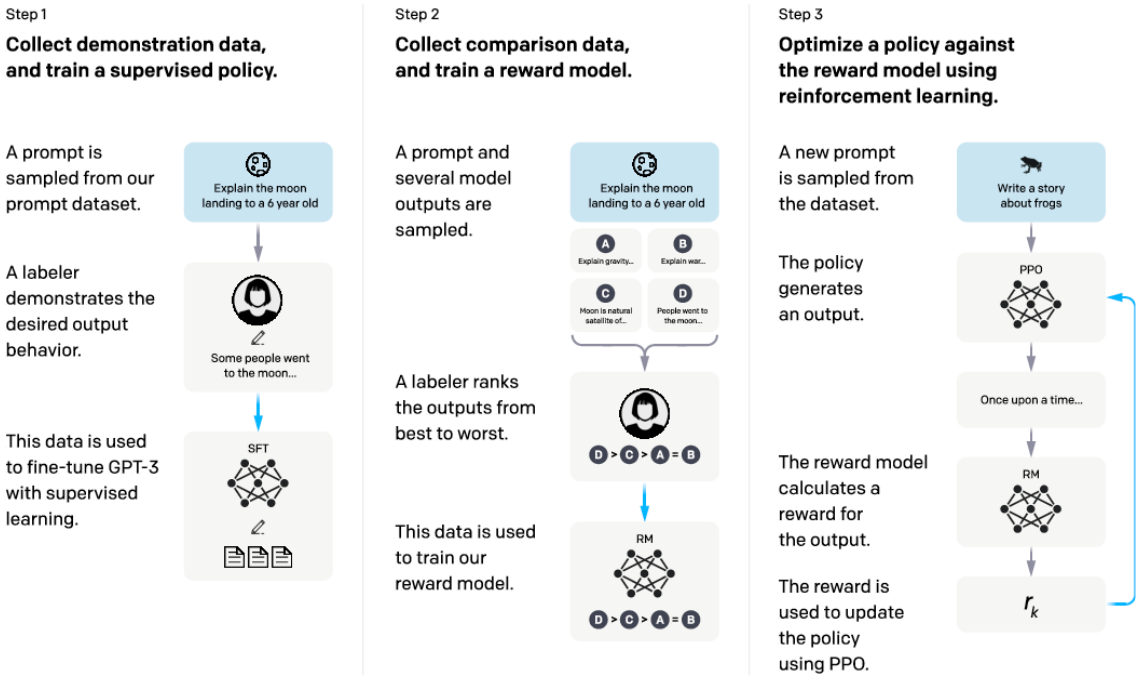

Abbildung 5: Drei Schritte der Optimierung des GPT-3 Modells[154]

Im ersten Schritt werden in einem arbeitsaufwändigen Prozess für eine Vielzahl von Eingabeaufforderungen (engl. prompts) Musterantworten von Menschen geschrieben, die dann für die weitere Verfeinerung der Parameter des Modells in einem Prozess des überwachten Lernens verwendet werden. Das dabei erstellte Modell ist das sogenannte Supervised Fine-Tuning (SFT) Model.

Dieses Modell wird dann in einem zweiten Schritt genutzt, um für die gleichen und neu formulierte Eingabeaufforderungen eine Reihe von verschiedenen Textausgaben zu produzieren. Menschliche Experten reihen diese Mustertexte nach Qualität und Wahrheitsgehalt. Diese Daten werden dann dafür verwendet, das Anreizmodell (RM, Reward Model) des Systems zu trainieren.

Dieses Anreizmodell wiederum wird in einem dritten Schritt des verstärkenden Lernens (Reinforcement Learning) genutzt, um Ausgaben des Modells zu bewerten und dann iterativ zu optimieren. Diese Schleife nennt sich Proximal Policy Optimization (PPO)[155].

Grundsätzlich funktioniert ChatGTP nach einem sehr ähnlichen Prinzip wie der Algorithmus der Autovervollständigung von Kurznachrichten, wie er bei den meisten

[154] Entnommen aus Ouyang et al. (2022), Figure 2, S. 3

[155] Vgl. Ouyang et al. (2022)

Smartphones zum Einsatz kommt. Bei der Eingabe einer mit „Ich wünsche Dir einen“ beginnenden Nachricht bietet das Programm auf Basis einer Analyse vorheriger Nachrichten die drei Optionen „guten“, „schönen“ und „erfolgreichen“ an. Wenn man dann „guten“ eingibt, sind die weiteren Optionen „Tag“, „Start“ und „Rutsch“. Hätte man „schönen“ eingegeben, erschienen die Optionen „Tag“, „Abend“ und „Sonntag“. All diese Vorschläge resultieren ausschließlich aus einer Statistik der Wortreihenfolgen und nicht aus einem Verständnis dessen, was dessen Benutzer sagen möchte. Bei ChatGPT sind der Algorithmus und die statistischen Analysen deutlich komplexer, trotzdem gilt das gleiche Grundprinzip. Deswegen lässt sich nicht vermeiden, dass manche Textvorschläge schlicht falsch oder komplett unangemessen sind oder gar verletzend, diskriminierend oder beleidigend. Große Probleme sind die Vorurteile (Biases), die sich oft aus entsprechend vorgeprägten Trainingstexten ergeben, und sogenannte Halluzinationen, die dann auftreten, wenn ein Text zwar syntaktisch korrekt ergänzt wird, allerdings ohne jeden Wahrheitsgehalt.

2.4 Das Black-Box-Problem

Ein Problem, das im Laufe dieser Arbeit an verschiedenen Stellen noch aufgegriffen werden soll, besteht in der Tatsache, dass die Deep Learning Technologie im heutigen Entwicklungsstand bereits mit sehr hoher Präzision Objekte identifiziert, Diagnosen erstellt und Entscheidungen vorbereitet, aber immer noch weit davon entfernt ist, unfehlbar zu sein. Es finden sich Beispiele aus dem Bereich des autonomen Fahrens, bei denen selbststeuernde Fahrzeuge Hindernisse komplett übersehen oder falsch klassifiziert haben. Ähnliches stellt man bei KI-Anwendungen fest, die in den Personalabteilungen Lebensläufe auf Eignung bzw. Nichteignung „screenen". Aufgrund von nichtrepräsentativen Beispielen in den Lernprozessen könnte sich eine systematische Benachteiligung bestimmter Gruppen von Bewerbern ergeben, wie zum Beispiel nach Geschlecht, Hautfarbe oder sexueller Orientierung. In keinem der Fälle kann die KI erklären, wie sie zu einer bestimmten Entscheidung gekommen ist. Auch spezifische Fehlerquellen im künstlichen neuronalen Netz lassen sich damit nicht identifizieren. Das ist eine signifikante Herausforderung auch für das juristische Nachverfolgen und die Zurechnung von Verantwortung und Schuld bei Unfällen oder Fehlentscheidungen.

Die Problematik ist aber noch viel grundsätzlicher: In der Entscheidungstheorie hängen Entscheidungen sehr eng mit Gründen zusammen. Entscheidungen, die eine KI nach den hier diskutieren Prozessen der künstlichen neuronalen Netzwerke und des maschinellen Lernens trifft, beruhen auf Strukturmodellen, die von lernenden Algorithmen selbst gebildet werden und intransparent sind, also von Personen nicht nachzuvollziehen sind[156]. Insbesondere wenn die KI auf Basis von Big Data Analysen zu Schlüssen kommt, besteht die Gefahr, dass sie die menschlichen Vorurteile („Biases") aus der Datenbasis fortschreibt oder sogar verstärkt. Das vom Programmierer in der Regel nicht beabsichtigte Fortschreiben von Vorurteilen („biases") ist das eine Problem in den Untiefen der Deep Learning Black Box. Das absichtliche Herbeiführen von falschen Klassifizierungen in künstlichen neuronalen Netzen ist das andere Problem. In der KI Forschung nennt man dies die „*Black Box Attack*"[157] . Kleinste Manipulationen an den Trainingsdaten oder physischen Objekten können systematische Fehlklassifikationen provozieren. Dies kann verhängnisvolle Auswirkungen haben, wenn dadurch zum Beispiel Verkehrsschilder falsch abgelesen oder gedeutet werden.

Ein weiteres Problem ergibt sich aus der Tatsache, dass der Autoadaptionsprozess in vielen (nicht allen) Anwendungen niemals zum Abschluss kommt. Damit ist nicht sichergestellt, dass nachdem bei einem System nach A und B als Input und L und M als Output folgte, dies beim nächsten Mal immer noch so ist. Vielleicht folgen X und Y. Die

[156] Vgl. Kaminski (2020), S. 153

[157] Bhagoji et al. (2018), S.1

grundsätzliche Forderung der empirischen bzw. experimentellen Wissenschaftstheorie nach einer Reproduzierbarkeit ist also oftmals nicht erfüllt.

Der Kybernetiker Heinz von Foerster hat für diese Art von Technologie den Begriff der „*nichttrivialen Technologie*“ im Gegensatz zur „*trivialen Technologie*“ geprägt. In seinem Modell übersetzt ein technischer Apparat oder eine triviale Technik einen Input X mit einer Transformationsregel R in einen Output Y. Die Transformationsregel genügt bei einer trivialen Maschine nach von Foerster vier Anforderungen[158]:

1. Sie ist “*synthetisch determiniert*“, d.h. es gibt eine klar definierte Regel.
2. Diese kann man sich aus dem Vergleich von Output und Input erschließen, d.h. sie ist auch „*analytisch determiniert*“.
3. Sie ist „*vergangenheitsunabhängig*“, d.h. man kann die Maschine also beliebig oft mit dem gleichen Input befüllen und erzielt immer den gleichen Output.
4. Damit ist die Maschine „*prognostizierbar*“.

Maschinen sind nach Foerster „*nichttrivial“*, wenn sich die Transformationsregel verändert und deren Wandel nicht nachvollzogen werden kann[159]. Sie sind ebenfalls „*synthetisch determiniert“*, hingegen „*analytisch nicht-determinierbar*, *vergangenheitsabhängig* und *nicht vorhersagbar“*[160]. Dies ist bei lernenden Maschinen der Fall. Sie sind „*mathematisch opak (intransparent). Personen können dann das Zustandekommen der Inferenzen und damit der Entscheidung nicht nachvollziehen“*[161].

Foerster schreibt zu den nichttrivialen Maschinen:

> *„Nimmt man nun auch noch alle die anderen unangenehmen Ärgerlichkeiten dieser Maschinen hinzu, nämlich ihre Abhängigkeit von ihrer Geschichte und ihre Unvorhersagbarkeit, dann werden unsere Anstrengungen, alle Ungewissheiten in unserer Umwelt zu beseitigen oder zu unterdrücken, sehr verständlich. Wenn wir eine Maschine kaufen, dann wollen wir, dass sie genauso arbeitet, wie wir dies wünschen. […] Wir wollen triviale Maschinen. Wenn sie trotzdem nichttriviale Tendenzen zeigt, ein Auto zum Beispiel nicht starten will usw., dann rufen wir einen Trivialisierungsexperten, um die Situation zu bereinigen.“*[162]

Andreas Kaminski und Andreas Gelhard führen diesen Gedanken noch weiter zu seinem Begriff der *informellen Technisierung*, der nicht nur auf maschinelles Lernen, sondern auf die KI insgesamt anwendbar ist:

> *„Es sind drei teilweise in einem Zusammenhang stehende Ebenen benannt, welche Technik informell werden lassen. „Nutzer“ können sich nicht mehr in einem sinnvollen Bezug zur Technik setzen, weil sie entweder nicht wissen, (1) dass sie es überhaupt mit Technik zu tun*

158 Vgl. Foerster (1993b), S. 245f; Kaminski Gelhard (2014), S. 66

159 Vgl. Foerster (1993b), S. 251; Kaminski Gelhard (2014), S. 8

160 Kaminski Gelhard (2014), S. 67

161 Kaminski (2020), S. 153

162 Foerster (1993b), S. 251f

> *haben, (2) dass sie mit ihr interagieren oder (3) welches die Effekte der Technik sind und welche Regeln ihnen zugrunde liegen.“*[163]

Kaminski und Gelhard zitieren Niklas Luhmann mit dem auf dem ersten Blick stark vereinfachenden Satz „*Technik sei das, was kaputt gehen kann*“[164]. Die Unterscheidung zwischen „heil“ und „kaputt“ ist bei informeller Technik nicht mehr möglich, was eine Vielzahl von Konsequenzen hat, die im weiteren Verlauf dieser Arbeit noch zu untersuchen sind.

2.5 Superintelligenz und Singularität

> *„Let an ultraintelligent machine be defined as a machine that can far surpass all the intellectual activities of any man however clever. Since the design of machines is one of these intellectual activities, an ultraintelligent machine could design even better machines; there would then unquestionably be an 'intelligence explosion,' and the intelligence of man would be left far behind. Thus the first ultraintelligent machine is the last invention that man need ever make."*[165]
>
> Irving John Good

Vielfach diskutiert wird die Frage[166], ob es eine Künstliche Intelligenz gibt bzw. jemals existieren wird, die das gesamte Spektrum der menschlichen Intelligenz abdeckt, die sogenannte „Artificial General Intelligence“. In diesem Zusammenhang fällt oft der Begriff der Singularität. In der Mathematik bezeichnet man mit der Singularität Punkte einer Kurve, an denen die Bildung einer Tangente nicht möglich ist, z.B. bei der Funktion $f(x)=1/x$ bei $x=0$. In der Physik liegt die Singularität vor, wenn in der Nähe eines schwarzen Loches aufgrund der unendlichen Gravitationskraft die klassischen Beschreibungen und Zusammenhänge der Physik nicht mehr gelten. Die Singularität der Künstlichen Intelligenz beschreibt das von einigen Forschern erwartete Phänomen der Intelligenzexplosion oder der Superintelligenz. Erstmals verwendet wurde der Begriff der Singularität in diesem Zusammenhang durch den amerikanischen Mathematiker Vernor Vinge 1993 in seinem Aufsatz „*The coming technological singularity: how to survive in the post-human era*“. Im Abstract schrieb er:

> „*Within thirty years, we will have the technological means to create superhuman intelligence. Shortly after, the human era will be ended.*“[167]

Bei Vinge und auch schon im Eingangszitat von Good kommt die zentrale Idee zum Ausdruck, dass Maschinen, die intelligenter als Menschen seien, dem Menschen auch beim Design von (zukünftigen) Maschinen überlegen seien. Das wiederum würde bedeuten,

---

[163] Kaminski Gelhard (2014), S. 12

[164] Kaminski Gelhard (2014), S. 15; Originalzitat bei Luhmann (1990), S. 263

[165] Good (1965), S. 31ff

[166] Vgl. Callaghan et al. (2017); Kurzweil (2005); Shanahan (2021); Vinge (1993)

[167] Vinge (1993), S. 245

dass zukünftige Maschinen noch intelligenter sein werden als ihre Entwickler. Dadurch ergibt sich eine Sequenz von immer intelligenteren Maschinen, die die Menschen weit hinter sich lassen werden[168].

Im Wesentlichen verbergen sich dahinter drei Grundannahmen:

1. Es existiert eine KI, die der humanen Intelligenz entspricht, nämlich die oben erwähnte „Artificial General Intelligence" oder auch „Human Level Intelligence".
2. Die Maschinenintelligenz kann sich selbst optimieren und multiplizieren und ihre Leistungsfähigkeit sukzessive steigern (Intelligence explosion). Gleichzeitig nimmt auch die Geschwindigkeit der eingesetzten Hardware exponentiell zu (Speed explosion).
3. Es bestehen keine Ressourcen- oder Energiebeschränkungen für die Maschinenintelligenz sowie keine fundamentalen physikalischen Grenzen der Maschinenleitungsfähigkeit.

Eine Kombination der drei Annahmen legt die Schlussfolgerung nahe, dass es einen Zeitpunkt geben wird, an dem die Leistungsfähigkeit der konsolidierten Maschinenintelligenz derjenigen der biologischen menschlichen Intelligenz entspricht. Von dem Moment an wird die Maschinenintelligenz die biologische Intelligenz weit hinter sich lassen, und zwar mit exponentiell wachsendem Abstand. Die Validität dieser Erwartung hängt vom Wahrheitsgehalt der drei Annahmen ab. Auch ohne eine Diskussion der zweiten und dritten Annahme widersprechen viele Experten bereits der ersten Grundannahme. David Chalmers schrieb dazu:

> *„To determine whether there might be an intelligence explosion, we need to better understand what intelligence is and whether machines might have it."*[169]

Dies wird das Thema in Kapitel 3, 4 und 5 sein.

[168] Vgl. Chalmers (2010), S. 7f

[169] Chalmers (2010), S.11; Chalmers kam am Ende des gleichen Artikels zu einer positiven Schlussfolgerung, die der Autor dieser Arbeit explizit NICHT teilt: „*Will there be a singularity? I think that is certainly not out of the question, and that the main obstacles are likely to be obstacles of motivation rather than obstacles of capacity. How should we negotiate the singularity? Very carefully, by building appropriate values into machines, and by building the first AI and AI+ systems in virtual worlds. How can we integrate into a post-singularity world? By gradual uploading followed by enhancement if we are still around then, and be reconstructive uploading followed by enhancement if we are not."*

2.6 Resümee

> *„Die KI ist nicht intelligent im engeren Sinne eines sinnverarbeitenden Systems. Sie ist allenfalls intelligent in dem Sinne, dass sie eine so hohe Komplexität von Rekombinationsmöglichkeiten von Daten verarbeiten kann, dass sie als Black Box immer unsichtbarer wird und deshalb eine zurechnungsfähige handlungsfähige Maschine wird. Am Ende arbeitet sie aber nur das ab, wofür sie konzipiert wurde, selbst wenn sie zu Ergebnissen kommt, die nicht unmittelbar mitkonzipiert wurden.“*[170]
>
> Armin Nassehi

Die Künstliche Intelligenz hat sich seit ihren Anfängen um die Mitte des vorherigen Jahrhunderts in Wellen von Fortschritten und Rückschlägen zu einer der wichtigsten, wenn nicht sogar zu der wichtigsten Zukunftstechnologie dieses beginnenden Jahrtausends entwickelt. Insbesondere wurde mit dem Verlassen der reinen symbolischen KI und der Entwicklung der numerischen sub-symbolischen KI und damit den künstlichen neuronalen Netzen, dem Maschinenlernen und dem „Deep Learning“ ein Durchbruch erzielt. Grundlage des großen, fast schon exponentiell zu nennenden Fortschritts der letzten zwei Jahrzehnte sind vier Komponenten: erstens der Paradigmenübergang von symbolisch zu subsymbolisch, zweitens die weiterhin exponentielle Steigerung von Rechenkapazität und -geschwindigkeit, drittens die ebenfalls exponentiell wachsende Verfügbarkeit von Daten (Big Data), die ein Deep Learning erst ermöglichen, und viertens die breite Ausweitung der Anwendung der KI in fast allen Lebens- und Tätigkeitsbereichen des Menschen.

Wichtig ist die Erkenntnis, dass es sich bei der Künstlichen Intelligenz um eine *nichttriviale Technologie* handelt, die *synthetisch determiniert*, allerdings explizit *nicht analytisch determiniert* ist, so daß sie *vergangenheitsabhängig* und *nicht vorhersagbar*[171] ist. Die KI wird immer auch eine Black Box sein, woraus sich weitreichende Konsequenzen insbesondere für das Thema der Verantwortung ergeben.

Aus der rein technischen Betrachtung zeichnen sich bereits die fundamentalen Grenzen der Technologie ab. Die KI ist über ihre Algorithmen sowohl in der symbolischen als auch in der sub-symbolischen Ausprägung vollständig determiniert. Das sich aus den Analysen großer Datenmengen ergebende Muster hängt von der Qualität und Quantität der zur Verfügung gestellten Daten ab. Es erscheint zweifelhaft, ob die KI je ein Verständnis von der Bedeutung der operierten Symbole oder der festgestellten Muster entwickeln kann.

---

[170] Nassehi (2019), S. 260

[171] Vgl. Abschnitt 2.4 sowie Foerster (1993b), S. 251; Kaminski Gelhard (2014), S.8

3 Intelligenz

Nachdem im vorherigen Abschnitt Definition, Historie und Wirkzusammenhänge der Künstlichen Intelligenz im Überblick dargelegt wurden, soll es in diesem Kapitel um das „Original" gehen, die natürliche Intelligenz des Menschen. Ebenso wenig wie die Definition der KI offensichtlich ist, so ist auch der Begriff der Intelligenz sowohl für Laien als auch für die damit befassten Wissenschaftler keineswegs eindeutig und klar. Viel ist zu dem Thema in den letzten zwei Jahrhunderten geforscht und publiziert worden. Die unterschiedlichen Ansätze und Modelle sollen hier vorgestellt werden. Von besonderem Interesse für die Gesamtfragestellung in dieser Arbeit sind zwei Teilaspekte oder Dimensionen der Intelligenz: die moralische und die kreative Intelligenz. Wenn bestimmt werden soll, ob die KI zu einem moralischen Urteil in der Lage ist oder ob sie menschenähnliche Kreativität hervorbringen kann, ist die Perspektive der Psychologie auf diese beiden Unterthemen relevant. In dem Zusammenhang soll auch der Frage nachgegangen werden, ob die Intelligenztheorien der Psychologie die gesamte Breite der menschlichen Intelligenz abdecken.

3.1 Was ist Intelligenz?

> *"What is intelligence? Intelligence is hard to define and descriptions are generally beset with paradoxes. Thus intelligence is attributed to those who have to think because they do not know a lot, and to those who know a lot and so do not have to think."* [172]
>
> Richard L. Gregory

Das wichtigste Anwendungsfeld der Informatik ist heutzutage die „Künstliche Intelligenz" und „*das am besten erforschte Merkmal der Psychologie"*[173] ist die Intelligenz, so behauptet Detlef Rost. Joachim Funke spricht von der Intelligenz als dem „*zentralen Konstrukt der modernen Psychologie*"[174]. Insofern müsste man glauben, dass es zumindest innerhalb dieser beiden Disziplinen oder gar übergreifend eine allgemein anerkannte Definition der Intelligenz gebe. Das ist definitiv nicht der Fall. Der amerikanische Psychologe und Intelligenztheoretiker Robert J. Sternberg wird dazu wie folgt zitiert:

> *"Viewed narrowly, there seem to be almost as many definitions of intelligence as there were experts asked to define it*."[175]

Eine allgemein gültige *„explizit-verbale Definition von Intelligenz"* gibt es nicht[176]. Bestenfalls existieren umfassende Sammlungen von Einzelzitaten zum Thema und deren Synthese. Ein Beispiel dazu: Shane Legg und Marcus Hutter vom Schweizer KI-Institut

[172] Gregory (1994), S. 13
[173] Rost (2013), S. 11
[174] Funke (2022), S. 87
[175] Zitiert in Legg Hutter (2007), S. 2
[176] Rost (2013), S. 39

IDSIA[177] haben eine Reihe von Definitionen aus Enzyklopädien, von Psychologen und KI Forschern gesammelt und daraus eine eigene knappe Definition abgeleitet:

> *"Intelligence measures an agent's ability to achieve goals in a wide range of environments."*[178]

Auf eine etwas ausführlichere Definition haben sich 52 bekannte Intelligenzforscher im Jahre 1994 geeinigt:

> *"A very general mental capability that, among other things, involves the ability to reason, plan, solve problems, think abstractly, comprehend complex ideas, learn quickly and learn from experience. It is not merely book learning, a narrow academic skill, or test-taking smarts. Rather, it reflects a broader and deeper capability for comprehending our surroundings – 'catching on', 'making sense' of things, or 'figuring out' what to do."*[179]

Im Vordergrund steht hier erstens die erfolgreiche Zielerreichung, zweitens die Interaktion mit dem Umfeld und drittens die Anpassungsfähigkeit.

Die Intelligenztheorie beschäftigt sich einerseits theoretisch und empirisch mit den inhaltlichen Dimensionen der Intelligenz und andererseits mit der Mess- und Erfassbarkeit der Intelligenz im Sinne des Intelligenzquotienten (IQ). Im Rahmen der Aufgabenstellung dieser Arbeit geht es primär um die Theorie der Intelligenz und deren inhaltliche Ansätze und weniger um Letzteres.

Detlef Rost unterscheidet *drei Gruppen von Intelligenztheorien*[180]:

1. Vorwiegend „*a priori Modelle"*
   Das sind „*geisteswissenschaftlich-konzeptuelle Ansätze"*, die oftmals spekulativen Charakter haben und häufig auch als „*vor-empirisch*" gekennzeichnet werden, da sie Praxiserfahrungen und Fallbeispiele zur Illustration der Theorie heranziehen, jedoch nicht zur umfassenden Bestätigung. Ein wichtiges Beispiel für diesen Ansatz ist die „*Theorie der Multiplen Intelligenzen*" von H. Gardner, auf die später noch eingegangen werden soll.
2. Vorwiegend „*a posteriori Modelle"*
   A posteriori Modelle setzen auf empirische Analysen und erklären die erfahrungswissenschaftlichen Erkenntnisse „*geisteswissenschaftlich-hermeneutisch"*. Beispiele hierfür sind hier die Generalfaktortheorie von Robert Spearman, die primäre Gruppenfaktortheorie von Thurstone und die Theorie der fluiden und kristallinen Intelligenz von Cattell.

[177] IDSIA: Istuto Dalle Molle di Studi sull Intelligenza Artificiale (Dalle-Molle-Forschungsinstitut für künstliche Intelligenz)

[178] Legg Hutter (2007), S. 9

[179] Zitiert aus Funke (2022), S. 88; Originalzitat: „Mainstream science on intelligence", Wall Street Journal, 13. Dezember 1994, A18

[180] Rost (2013), S. 39 und alle folgenden Zitate

3. Ansätze, „*die dazwischen liegen“*
   Hierbei handelt es sich um Intelligenztheorien, die ihren Ausgangspunkt in der „*begrifflich-logischen Klassifikation intellektueller Fähigkeiten“* haben, die dann aber empirisch untermauert werden. Das „*Structure-of-Intellect-Model*“ (SOI) von Guilford ist hierfür ein wichtiges Beispiel, jedoch auch das *Berliner Intelligenzstrukturmodell* (BIS) von Jäger.

Eine Auswahl dieser Modelle soll im Folgenden vorgestellt werden.

3.1.1 Generalfaktortheorie (Spearman)

Die bereits im Jahr 1904 vorgestellte Intelligenztheorie von Charles E. Spearman baut auf vorherige eindimensionale Modelle zur Beschreibung der kognitiven Leistungsfähigkeit auf[181]. In seinen empirischen Arbeiten war dem Psychologen aufgefallen, dass Schulleistungen in Fächern, die augenscheinlich nichts oder nur wenig miteinander zu tun haben (z.B. Mathematik und Sprachen), dennoch positiv miteinander korrelieren[182]. Diese Spearman'sche *„Hypothese der positiven Mannigfaltigkeit (positive manifold) aller intelligenten Leistungen“*[183] ist unter den Experten unumstritten. In Spearmans „*Generalfaktortheorie der Intelligenz*“ wurde der übergreifende Faktor der „*generellen Intelligenz*“ (abgekürzt g) definiert, der die aufgaben- oder fachübergreifende Korrelation abbildet. Für aufgabenspezifische Abweichungen sieht er spezifische Faktoren vor (s_1, s_2, ...,s_j, ... s_n). Damit werden die Leistungen in einem Test mit gleichwertigen Aufgaben durch jeweils zwei Faktoren bestimmt: die generelle Intelligenz, die in allen Tests feststellbar ist, und eine umgrenzte Fähigkeit, die für den jeweiligen Aufgabentyp spezifisch ist (s_j). Die s-Faktoren für die verschiedenen Aufgabentypen und die allgemeine Intelligenz g sind untereinander unabhängig. Aufgrund dieser beiden Elemente spricht man auch von der „Zwei-Faktorentheorie“.

3.1.2 Primäre Gruppenfaktortheorie (Thurstone)

Thurstone präsentierte etwas später (1927, bzw. als Buch 1938) eine alternative Ansicht zum Thema. „*Er hielt g für ein statistisches Artefakt“*[184]. Er verstand Intelligenz als *Ensemble spezifischer unabhängiger Fähigkeiten*[185]. Auf Basis seiner empirischen

[181] Vgl. Spearman (1904), S. 201: "*The results indicated that all branches of intellectual activity possess in common one fundamental function, whereas the remaining or specific elements of the activity seem to be wholly different from that in all the others. In adult life no difference between the two sexes was observed.*"

[182] Vgl. Ebd.; Rost (2013), S. 42

[183] Rost (2013)

[184] Funke (2022), S. 93

[185] Rost (2013), S. 50

Untersuchungen identifizierte Thurstone sieben relativ unabhängige Gruppenfaktoren der Intelligenz[186]:

1. *„Faktor **M** (Englisch: memory; Gedächtnis, Merkfähigkeit)*
2. *Faktor **N** (Englisch: number; Rechengewandtheit, Rechengeschwindigkeit)*
3. *Faktor **P** (Englisch: perception; Wahrnehmungsgeschwindigkeit, Auffassungsschnelligkeit)*
4. *Faktor **R** (Englisch: reasoning; schlussfolgerndes Denken, Problemlösen, abstraktes Denken)*
5. *Faktor **S** (Englisch: space; Raumvorstellung, räumliche Vorstellungsfähigkeit)*
6. *Faktor **V** (Englisch: verbal; Sprachbeherrschung, Wortverständnis, sprachliche Gewandtheit)*
7. *Faktor **W** (Englisch: word fluency; Wortflüssigkeit, Geläufigkeit des Wortschatzes“*

Ursprünglich vertrat Thurstone die Auffassung, dass die Primärfaktoren voneinander unabhängig seien, was in Folgeanalysen nicht bestätigt werden konnte und von Thurstone selbst auch im Jahr 1946 korrigiert wurde, womit er schließlich Spearmans ursprüngliche Arbeit bestätigte[187].

3.1.3 Theorie der fluiden und kristallinen Intelligenz (Cattell)

Die von R.B. Cattell und seinem ehemaligen Mitarbeiter J.L. Horn erstmalig in 1941 entwickelte Theorie[188] unterscheidet zwei Typen von Intelligenz: Erstens die fluide Intelligenz als diejenige Intelligenz, die genetisch-neurologisch vordefiniert ist und sich zeigt, wenn man neuartige Probleme löst, die keine oder nur minimale Wissensvoraussetzungen besitzen, und zweitens die kristalline oder kristallisierte Intelligenz, die über Lernprozesse entsteht und sich in der Anwendung erworbenen Wissens in Problemlösungen zeigt. Die beiden Typen entwickeln sich im Laufe eines Menschenlebens höchst unterschiedlich. Die fluide Intelligenz steigt von der Geburt bis in das dritte Lebensjahrzehnt und nimmt dann wieder kontinuierlich an. Die kristalline Intelligenz steigt lebenslang. In der kristallinen Intelligenz zeigen sich kulturelle und subkulturelle Unterschiede. Die Suche nach einem kulturfreien Intelligenztest oder einer kulturunabhängigen Intelligenztheorie war der Auslöser für die Entwicklung dieser Theorie.

---

[186] Die Faktorbeschreibungen stammen aus Rost (2013), S. 51 ff

[187] Thurstone (1946), S. 101: „*There also appears to be a central, 'second order,' factor originally identified by Spearman which influences, with more or less saturation, all the special abilities thus far identified.*“

[188] Cattell (1941), Cattell (1943), Cattell (1963); Anmerkung: Die Zuordnung der Theorie ausschließlich zu Raymond Cattell ist heute durchaus umstritten. Es gibt Evidenz, dass Cattell sich auf frühere Arbeiten und Korrespondenz von Donald O. Hebb bezog; Vgl. Brown (2016)

3.1.4 Structure-of-Intellect Model (SOI) nach Guilford[189]

Wie schon in der Einleitung beschrieben, hat Roy Paul Guilford sein Modell zur Struktur des Intellekts rein theoretisch und als logische Struktur entwickelt und danach empirisch abgesichert. Sein Modell ist dreidimensional und wird oft als Würfel dargestellt. Die drei SOI Gliederungsdimensionen sind[190]:

1. *„Fünf Gegenstandsbereiche (**Inhalte**), Informationsarten, die man unterscheiden kann*
   - ***V**isual: Bildliches*
   - ***A**uditory: Gehörtes*
   - ***S**ymbolic: Symbolisches*
   - *Se**M**antic: Semantisches, Bedeutungen*
   - ***B**ehavioral: Verhaltensbezogenes, Sozialverhalten, Körpersprache*
1. *Fünf **Operationen des Denkens** als zentrale intellektuelle Tätigkeiten*
   - ***C**ognition: Auffassen, Erkennen, Verstehen, Entdecken*
   - ***M**emory: Gedächtnis, Erinnern*
   - ***D**ivergent production: Hervorholen verschiedener alternativer Informationseinheiten aus dem Gedächtnis, z.B. Gegenstände aufzählen*
   - *Co**n**vergent production: Hervorholen einer spezifischen Information aus dem Gedächtnis, z.B. Wort für Kreuzworträtsel finden*
   - ***E**valuation: Bewertung, Feststellung der Richtigkeit, Brauchbarkeit, Güte*
2. *Sechs **Produkte des Denkens**, Arten der Verarbeitung von Information*
   - ***U**nits: Einheiten, kleinstes Element; z.B. Klang eines musikalischen Akkords*
   - ***C**lasses: Klassen; Oberbegriffe, z.B. Wörter mit einer bestimmten Endung oder Berufsgruppen*
   - ***R**elations: Beziehungen, Zusammenhänge, Analogien*
   - ***S**ystems: Systeme, d.h. komplexe miteinander in Verbindung stehende und aufeinander bezogene Relationen; z.B. Melodie oder ein Plan*
   - ***T**ransformations: Transformationen, Umformungen, Veränderungen, Modifikationen*
   - ***I**mplications: Implikationen, Abhängigkeiten, Vorhersagen, Beziehungen; z.B. Ursache-Wirkung, früher-später“*

Dementsprechend kann die Intelligenz als Quader mit 150 Zellen (5 Operationen x 5 Inhalte x 6 Produkte) dargestellt werden[191]. Diese „*150 Faktoren sollten das gesamte intellektuelle Potential des Menschen beschreiben*“[192]. Bei der Anwendung in der

[189] Rost (2013), S. 59 ff

[190] Zitiert aus Ebd.

[191] Vgl. Rost (2013), S. 59, Abb. 2.4

[192] Rost (2013), S.59

Psychologie hat sich dieses Modell nicht nachhaltig bewährt[193]. Für die in dieser Arbeit angestellten Untersuchungen und Überlegungen zu den Fähigkeiten der KI ist es aber fraglos von Nutzen. Die drei Dimensionen sind potentiell geeignet, um die Fähigkeiten der KI mit denjenigen des Menschen innerhalb dieses Rahmens zu vergleichen.

3.1.5 Theorie der Multiplen Intelligenzen (Gardner)

In den letzten Jahrzehnten wurden einige alternative Ansätze der Intelligenztheorie entwickelt, die weitere wichtige Attribute der menschlichen Intelligenz außerhalb der „klassischen" mit IQ-Tests gemessenen Intelligenz ergänzen. Zu erwähnen sind die in diesem Abschnitt dargestellten „Multiplen Intelligenzen" (MI), die „Emotionale Intelligenz" (EI), der Vorläufer der „Sozialen Intelligenz" und auch die im Folgeabschnitt beschriebene „Praktische Intelligenz" (PI). Generell geht es bei all diesen Ansätzen darum, stärker lebensweltliche und alltagspsychologische Aspekte der menschlichen Intelligenz mit zu berücksichtigen. Kritiker wenden ein, dass mit dieser Abkehr von der reinen kognitiven Intelligenz eine Verwässerung eintritt und der Begriff der Intelligenz „*zu einer Leerformel entwertet*"[194] wird.

Howard Gardner stellte in den 1980er Jahren seine Theorie der „Multiplen Intelligenzen" (MI) vor. Zielsetzung war die realitätsnahe Abdeckung des gesamten Umfangs der intellektuellen Fähigkeiten eines Menschen. Konkret sah Gardner in seinem ursprünglichen Konzept von 1982 sieben Ausprägungen von Intelligenzen vor[195]:

> ***„Linguistische Intelligenz**: Sensibilität für sprachliche Phänomene, Sprachlernbegabung, Befähigung zum flexiblen und gewandt-kompetenten Umgang mit gesprochener und geschriebener Sprache, wie man es z.B. bei Dichtern, Schriftstellern, Journalisten, Rechtsanwälten, professionellen Rednern etc. finde.*
>
> ***Logisch-Mathematische Intelligenz**: Fähigkeit zum logisch-deduktiven Denken, zum Entdecken und Verstehen komplexer mathematischer Zusammenhänge und zum naturwissenschaftlichen Arbeiten, so wie man es bei berühmten Mathematikern, Logikern und Naturwissenschaftlern beobachten könne.*
>
> ***Visuell-Räumliche Intelligenz**: Kompetenz zur realen und gedanklichen Orientierung in großen und kleinen Räumen, Imagination räumlicher Gegebenheiten. Diese sei bei Navigatoren, Piloten, Bildhauern, Chirurgen, Architekten und Ingenieuren und auch bei bildenden Künstlern in reichem Ausmaß vorhanden.*
>
> ***Musikalische Intelligenz**: Vor allem produktive, aber auch rezeptive musikalische Befähigung, einschließlich rhythmischer Komponenten. Komponisten, Instrumentalisten und Sänger besäßen diese Fähigkeit.*

[193] Vgl. Rost (2013), S. 61

[194] Rost (2013), S. 111

[195] Liste und Beschreibung der Intelligenzen aus Rost (2013), S. 120 - 121

> ***Körperlich-Kinästhetische Intelligenz**: Befähigung zum koordiniert-flexiblen Umgang mit dem eigenen Körper und die Kompetenz, ihn bzw. einzelne seiner Teile effektiv einzusetzen, z.B. bei Tänzern, Schauspielern, Sportlern, Handwerkern und Chirurgen.*
>
> ***Sozial-Interpersonale Intelligenz**: Fähigkeit, sich in andere Personen, in ihre Gefühle, Empfindungen, Wünsche, Motivationen und Befürchtungen hineinzuversetzen und mit anderen Menschen effektiv zu kooperieren. Ärzte, Lehrer, Pastoren, Politiker, Handelsvertreter und Schauspieler bräuchten diese Fähigkeit, um in ihrem Beruf erfolgreich zu sein.*
>
> ***Sozial-Intrapersonale Intelligenz**: Selbsterkenntnis, d.h. das Verstehen der eigenen Person. Dies umfasse u.a. die Kompetenz, sich selbst und die eigenen Wünsche, Hoffnungen, Befürchtungen, Motivationen und Fähigkeiten etc. realistisch wahrzunehmen und sich dementsprechend adäquat zu verhalten und sein Leben zu planen."*

In seinen späteren Arbeiten (1998 & 1999) ergänzte Gardner noch zwei weitere Intelligenzen:

> ***„Naturalistische Intelligenz**: Die Fähigkeit, Naturphänomene voneinander zu unterscheiden und zu kategorisieren sowie die Gesetze der Natur zu erkennen und gedanklich zu durchdringen, was z.B. für Berufe wie Förster, Landwirt oder Gärtner von entscheidender Relevanz sei"*[196]
>
> ***Existenzielle Intelligenz**, in anderen Werken auch **Lebensintelligenz**[197]* genannt*:*
>
> *„... die Fähigkeit, sich zu den äußersten Grenzen des Kosmos, dem Unendlichen, und dem Infinitesimalen, ins Verhältnis zu setzen, und die verwandte Fähigkeit, sich mit so zentralen existentialen Momenten der conditio humana wie der Bedeutung des Lebens und dem Sinn des Todes, dem endgültigen Schicksal der physischen und psychischen Welten und so ergreifenden Erfahrungen wie der Liebe und dem ungeteilten Aufgehen in einem Kunstwerk auseinanderzusetzen."*[198]

Zwischenzeitlich bestand noch die Idee, eine „*Spirituelle Intelligenz*" und eine „*Moralische Intelligenz*"[199] zu ergänzen, was aber verworfen wurde. Gardner hat dies in seinem Buch „*Intelligenzen*" für die moralische Intelligenz ausführlich erläutert:

> *„Meinem Verständnis nach ist die Hauptkomponente des Moralischen das Gefühl für persönliche Handlungs- und Einsatzbereitschaft, das Bewusstsein, dass man mit Rücksicht auf andere eine unabweisbare Rolle übernimmt und dass das eigene Verhalten in dieser Funktion von der Analyse des gesellschaftlichen Kontextes und der eigenen Willenskraft bestimmt sein muss. Für unser Bild von Gandhi als einem moralischen Menschen ist nicht die Besonderheit seiner Philosophie oder die Beispielhaftigkeit seines Verhaltens ausschlaggebend, sondern ebenso wie bei Mutter Teresa, Nelson Mandela oder Andreij Sacharow, der Wille, eine herausgehobene Rolle im Dienst der Menschheit zu übernehmen. Das Agieren in solchen Schlüsselrollen erfordert zweifellos ein bedeutendes Spektrum*

[196] Rost (2013), S. 121

[197] Gardner (2002), S. 78

[198] Ebd.

[199] Rost (2013), S. 121; Gardner (2002), S. 87f

> *menschlicher Intelligenzen, eingeschlossen die personalen, die sprachliche, die logische und vielleicht auch eine existentiale – es bleibt im wesentlichen* ***Ausdruck der eigenen Persönlichkeit****, genauer, der Persönlichkeit, zu dem man geworden ist, und* ***ist selbst keine Intelligenz****. Moralisches Empfinden wäre danach eine Darstellung der Persönlichkeit, der Individualität, des Wallens, des Charakters und im günstigsten Fall der vollendeten Ausbildung der menschlichen Natur.“*[200]

Die Argumentation erscheint plausibel, allerdings lässt sie die Abwägung zwischen Eigennutz, gesellschaftlichen Gesamtnutzen und dem Interesse der Schöpfung als Ganzem im Sinne des kategorischen Imperativs Kants außer Acht. Es erscheint zweifelhaft, dass die Fähigkeit zur Selbstgesetzgebung außerhalb der Intelligenz und innerhalb der Persönlichkeit zu verorten ist. Das Nachdenken über die „goldene Regel“ und deren Anwendung verdient einen Platz in der Theorie der Intelligenz.

An anderer Stelle sollte hingegen das Argument von Gardner aufgenommen werden, dass die Kombination einer eigenen (autonomen) Willenskraft zusammen mit einer persönlichen Handlungs- und Einsatzbereitschaft bei einer Analyse des gesellschaftlichen Kontextes für moralisches Verhalten erforderlich ist. Aus Sicht des Autors dieser Arbeit ist es ausgeschlossen, dass eine algorithmische KI jemals diese drei Komponenten erfüllen kann.

Auch der Psychologe und Kognitionswissenschaftler Wolfgang Prinz hat in seinen Arbeiten den sozialen Aspekt der menschlichen Intelligenz (er spricht von „Kognition“) herausgearbeitet und als Interaktion zwischen „*geistbegabten Wesen*“ vertieft. Prinz formuliert eine Theorie des „*kollektiven Konstruktivismus*“ (als Kombination aus radikalem Kollektivismus und Konstruktivismus), der zufolge die Menschen ihre Intelligenz im Zusammenspiel mit anderen Menschen und sich selbst entfalten[201]. Er operiert dabei mit den Metaphern des „*inneren*“ und des „*äußeren Spiegels*“:

> *„In erster Linie bieten Spiegel einen Wahrnehmungszugang zu Dingen, die ansonsten unzugänglich sind. Mit Hilfe physikalischer Spiegel können Menschen sehen, was sie tun. Hier ist die Korrelation zwischen Handeln und Sehen perfekt. […]. Wichtiger noch ist die Tatsache, dass Personen durch soziale Spiegelungen sehen und verstehen können, wie andere ihr Handeln sehen und verstehen. […]. In knappen Worten leisten Spiegel also folgendes: dass das Selbst durch die anderen entstehen kann.“*[202]

Bei dem im Zitat beschriebenen Spiegel handelt es sich um den „*äußeren Spiegel*“. Prinz vertritt die Hypothese, dass es beim Menschen und nur beim Menschen zusätzlich einen „*inneren Spiegel*“ gibt:

> *„Mit inneren Spiegeln meine ich spiegelartige Repräsentationsstrukturen, die* [Menschen] *dabei helfen, die Leistung der äußeren Spiegel zu verstehen und zu nutzen. Ebenso wie viele*

[200] Gardner (2002), S. 98; Hervorhebungen DS

[201] Vgl. Prinz (2013)

[202] Prinz (2013), S. 112f

> *andere Verarbeitungsmodule im menschlichen Geist dienen diese Spiegelmodule dazu, die Wahrnehmung mit dem Handeln zu verknüpfen. […] Die Herstellung von Verknüpfungen zwischen Wahrnehmungen und Handeln muss als eine der entscheidendsten und grundlegendsten Leistungen geistiger Architekturen betrachtet werden.“*[203]

Der innere Spiegel geht also über die reine subjektive Sicht der Außenwelt durch das Individuum hinaus. Er beinhaltet die Wahrnehmung und Reflektion der Reaktionen und Handlungen der anderen Individuen in Bezug auf das eigene Handeln.

3.1.6 Praktische Intelligenz (Sternberg)

Eine ähnliche Erweiterung des Intelligenzbegriffs wie Gardner nahm auch R.J. Sternberg vor. Er argumentierte, dass vorherige Intelligenztheorien zu sehr auf „akademische Probleme“ ausgerichtet waren, die oftmals wenig lebensweltliche Relevanz hatten. Seine Praktische Intelligenz (PI) richtete er stärker auf *praktische Probleme* aus. Damit ist dann eher „common sense“ oder der sogenannte „gesunde Menschenverstand“ gemeint.

Tabelle 3: Unterscheidung zwischen akademischen und praktischen Aufgabenstellungen nach Sternberg & Wagner[204]

„akademische Probleme“	***„praktische Probleme“***
• *oft von anderen formuliert*	• *erfordern oft, Probleme selbst zu erkennen / zu formulieren*
• *in der Regel gut definiert*	• *in der Regel unscharf formuliert*
• *eher vollständige Informationsbasis*	• *erfordern oft Informationssuche*
• *oft nur eine einzige Lösung*	• *häufig mehrere Lösungen*
• *häufig nur eine einzige Lösungsmethode*	• *häufig multiple Lösungswege*
• *von Alltagserfahrungen eher abgelöst*	• *in Alltagserfahrungen eingebettet und/oder diese voraussetzend*
• *geringes oder kein intrinsisches Interesse*	• *mehr Motivation und persönliches Engagement*

[203] Prinz (2013), S. 114

[204] Zitiert aus Rost (2013), S. 166

Auf dieser Basis entwickelte er die sogenannte „*Triarchische Intelligenztheorie*“ mit drei breiten Fähigkeitsbereichen der menschlichen Intelligenz[205]:

1. *„Analytische Intelligenz: Diese wird zur Bearbeitung akademischer Problemlöseaufgaben, wie man sie in vielen Intelligenztests findet, benötigt. Hier kommt es beispielsweise auf das Analysieren, Urteilen, Evaluieren, Vergleichen und Entscheiden an, also auf kognitive Strukturen und Prozesse, die bei einer effektiven Informationsverarbeitung eine Rolle spielen (metakognitive Komponenten, Leistungskomponenten und Wissensaneignungskomponenenten; komponentielle Subtheorie)*
2. *Kreativ-kognitive Intelligenz: Sie bezieht sich auf das Entdecken und Umgehen mit neuen Problemstellungen und auf die Entwicklung neuartiger Ansätze und Ideen, ist gewissermaßen ein Experimentieren mit Wissensbestandteilen und automatisierten Informationsverarbeitungskomponenten (experimentelle Subtheorie)*
3. *Praktisch-kognitive Intelligenz: Diese Fähigkeit zielt auf das Identifizieren und Lösen der üblichen Probleme ab, die einem Menschen in seiner spezifischen Umwelt begegnen, also auf alltägliche Anpassungsleistungen. Der Kontext der jeweiligen Umwelt mit den zu lösenden domänenspezifischen Alltagsproblemen ist demnach zentral. Die Umwelt der Ureinwohner Australiens unterscheidet sich beispielsweise erheblich von den Umwelten industrialisierter Gesellschaften (Kontextuelle Subtheorie).“*

Die Dimension der praktischen Intelligenz, die sich aus einer Integration vieler Einzelintelligenzen in einer lebensweltlichen Anwendung ergibt, stellt eine große Herausforderung der KI dar.

3.1.7 Eigenschaften von intelligenten Systemen nach Cruse, Dean & Ritter

Einen disziplinübergreifenden Versuch zur Definition der Intelligenz von Systemen haben der Kybernetiker und theoretische Biologe Holk Cruse, der Neurobiologe Jeffrey Dean und der Neuroinformatiker Helge Ritter unternommen. Sie gingen bei ihren Überlegungen bewusst nicht von der menschlichen Intelligenz aus, die immer auch eine Sprache voraussetzt, sondern haben intelligente Systeme breiter gefasst und generell Lebewesen, Maschinen, Organe und Organgruppen mit einbezogen[206]. Aus ihrer Perspektive „*sollten intelligente Systeme zu nützlichen, effizienten und robusten Verhaltensweisen führen“*. Dafür identifizierten sie die folgenden acht Eigenschaften und Fähigkeiten[207]:

[205] Fähigkeitsbereiche der „Triarchischen Intelligenz“ zitiert aus Rost (2013), S. 167

[206] Vgl. Cruse Dean Ritter (1998), S. 23 ff; Cruse Dean Ritter (1999), S. 102ff

[207] Ebd.

1. ***Autonomie*** = sich selbst das Gesetz bzw. die Regeln geben
   Voraussetzung: ein „Selbst“
2. ***Intentionen*** = eigene Ziele auswählen, Aufmerksamkeit auf bestimmte Objekte in der Umwelt richten und angemessene Verhaltensweisen wählen
3. ***Anpassungsfähigkeit und Lernen aus Erfahrungen***
   Voraussetzung: individuelles Gedächtnis (mehr als nur Datenspeicher)
4. ***Beurteilung des Erfolgs des eigenen Verhaltens***
   Voraussetzung für Lernen aus Erfahrung
5. ***Generalisierung in der Wahrnehmung und Flexibilisierung des eigenen Verhaltens***
   Erkennen trotz Unschärfe und Pragmatismus bei der Wahl der eigenen Mittel
6. ***Kategorienbildung und Abstraktionsvermögen***
7. ***Entscheidung zwischen Alternativen***
   Auch für Moralkonflikte
8. ***Vorhersage von Änderungen in der Umwelt*** auf Basis des eigenen Verhaltens

Es ist allgemein eingängig, dass alle acht Fähigkeiten notwendig und nur in ihrer Gesamtheit hinreichend für einen durchschnittlich intelligenten Menschen sind. Schon das vollständige Fehlen einzelner Komponenten bei einem erwachsenen und gesunden Mitmenschen würde auffallen.

3.2 Kreative Intelligenz

> „*Man muss noch Chaos in sich haben, um einen tanzenden Stern gebären zu können*“ Friedrich Nietzsche, Untertitel zu „Also sprach Zarathustra“[208]

Kreativität bildet innerhalb der Intelligenztheorien fast schon einen eigenständigen Forschungsgegenstand.

Bei der Kreativität handelt es sich um die schöpferischen und gestalterischen Fähigkeiten des Menschen. Runco und Jaeger definieren sie als eine Eigenschaft, die Originelles und Nützliches hervorbringt[209]:

> *“The standard definition is bipartite: Creativity requires both originality and effectiveness*. […] *Originality is undoubtedly required. It is often labeled novelty, but whatever the label, if something is not unusual, novel, or unique, it is commonplace, mundane or conventional. It is nor original, and therefore not creative*. […] *Like originality, effectiveness takes various forms. It may take the form of (and be labeled as) usefulness, fit or appropriateness*.”[210]

In vielen Intelligenztheorien findet die Kreativität ihren Platz, so zum Beispiel als „*flüssige, flexible und ursprüngliche Erzeugung von Konzepten von Lösungen für neuartige Probleme*“[211] und „divergentes Denken“ (divergent production) bei Guilford[212] oder als kreativ-kognitive Intelligenz in der triarchischen Intelligenztheorie von Sternberg.

Guilfords „divergent production” ist die kreative Generierung einer Vielzahl von Antworten oder Möglichkeiten von Antworten für eine gegebene Frage. Kreativität entsteht aus „divergent production“, wenn sie mit „convergent production“ kombiniert wird, also der deduktiven Auswahl der besten Antwort auf die gegebene Frage[213]. In der Psychologie vertritt man die These, dass Individuen entweder stärker in der „divergent production“ oder in der „convergent production“ sind, so dass Kreativität in Teams über die optimierte

---

[208] Nietzsche (2010)

[209] Runco Jaeger (2012), S. 92

[210] Ebd.

[211] Deutscher Ethikrat (2023), S. 91; vgl. auch Guilford (1950), S. 454 zu den Hypothesen zu den Fähigkeiten der kreativen Intelligenz: „*It is suggested that certain kinds of factors will be found, including sensitivity to problems, ideational fluency, flexibility of set, ideational novelty, synthesizing ability, analyzing ability, reorganizing or redefining ability, span of ideational structure, and evaluating ability.*”

[212] Vgl. Guilford (1984)

[213] Vgl. Razumnikowa (2013), S. 546f: “*Divergent thinking is defined as producing a diverse assortment of appropriate responses to an open-ended question or task in which the product is not completely determined by the information. So, divergent thinking concentrates on generating a large number of alternative responses including original, unexpected, or unusual ideas. Thus, divergent thinking is associated with creativity. Convergent thinking involves finding only the single correct answer, conventional to a well-defined problem. Many facts or ideas are examined while convergent thinking for their logical validity or in which a set of rules is followed. Convergent thinking focuses on reaching a problem solution through the recognition and expression of preestablished criteria. Standard intelligence tests are similarly believed to measure convergent thinking.*”

Zusammensetzung der Gruppen maximiert werden kann (siehe auch „*divergent thinking*[214]“).

Einige Wissenschaftler haben sich explizit mit der Kreativität auseinandergesetzt, nicht zwingend nur Vertreter der Psychologie. Die Kognitionswissenschaftlerin Margaret A. Boden und der Mathematiker Marcus du Sautoy unterscheiden zwischen drei Typen von Kreativität[215]:

1. *Explorative Kreativität*
2. *Kombinatorische Kreativität*
3. *Transformative Kreativität*

Die *explorative Kreativität* baut auf bekanntes Wissen auf, nutzt existierende Regeln und findet darauf basierend „*neue Lösungen für bestehende Probleme“*[216]. Beispiele für diese Art von Kreativität sind neue Kombinationen von Schachzügen, Auswege aus Labyrinthen, Bachs Musik, mathematische Beweise und geografische Entdeckungen. Innovationen innerhalb von bestehenden „Denkschulen“ sind explorativ. Boden – so du Sautoy – siedelt etwa 97% der menschlichen Kreativität in dieser Kategorie an.

Kombinatorische Kreativität verknüpft „*zwei komplett unterschiedliche Ideen, um mithilfe von Assoziationen Neues zu erschaffen – so verknüpfte beispielsweise das Smartphone die bestehenden Konzepte des Telefonierens und Fotografierens“*. Durch die Kombination vertrauter und etablierter Ideen und Konzepte entstehen neue Ideen und Anwendungen. Analogien und Metaphern spielen bei dieser Form der Kreativität eine große Rolle. Viele disziplinübergreifende Innovationen fallen in diese Kategorie.

Die *transformative Kreativität* baut nicht auf existierendes Wissen auf, sondern „*bricht mit bestehenden Regeln“*. Dazu zählt die „*verrückte Idee*, *welche die vorherrschenden Denkmuster radikal ändert und zu einem neuen Stil, einer neuen Theorie, also neuen konzeptuellen Räumen führt.“*[217] Beispiele hierfür sind *Descartes‘ Dualismus*, *Einsteins Relativitätstheorie*, *Schönbergs Zwölftonmusik, Picassos Kubismus* und *Gödels Unvollständigkeitstheorem*, aber auch der *Aufklärungsgedanke des 18. Jahrhunderts*[218].

---

[214] “*Divergent thinking is cognition that leads in various directions. Some of these are conventional, and some original. Because some of the resulting ideas are original, divergent thinking represents the potential for creative thinking and problem solving. Originality is not synonymous with creative thinking, but originality is undoubtedly the most commonly recognized facet of creativity. To the degree that tests of divergent thinking are reliable and valid, they can be taken as estimates of the potential for creative thought. Not surprisingly, divergent thinking tests are among the most commonly used in creativity research. Divergent thinking tests are also used in numerous educational programs and in various organizational training packages.”* (Quelle: ScienceDirect, Encyclopedia of Creativity)

[215] Vgl. Jaedtke (2019); du Sautoy (2019), S. 7ff

[216] Jaedtke (2019); sowie die folgenden Beispiele und Zitate

[217] Zipp Vey (2018), S. 31

[218] Vgl. du Sautoy (2019), S 9f;

Die explorative und mit Einschränkungen auch die kombinatorische Kreativität lassen sich mit Algorithmen der künstlichen Intelligenz simulieren, sofern Rechenleistung und Speichermöglichkeiten gegeben sind. Völlig anders sieht es bei der transformativen Kreativität aus. Es ist nicht erkennbar, wie die KI ohne menschliche Mitwirkung transformativ kreativ sein könnte. Gemäß du Sautoy ist transformative Kreativität ohne freien Willen und Bewusstsein nicht denkbar[219]

> „*Our creativity is intimately bound up with our free will, something it would seem impossible to automate. To program free will would be to contradict what free will means.*"

Weiterhin ist sie verbunden mit unserer Sterblichkeit:

> „*Creativity is very tied up with mortality, something very much coded into what it means to be human. Many who seek meaning for their existence but find the stories of the world's religions meaningless hope to leave something behind that will outlast their finite existence – whether it be a painting, a novel, a theorem, or a child.*"

Dieser Ansicht sind auch Jan Sebastian Zipp und Karin Vey. Zusätzlich zum Bewusstsein sind auch Körperlichkeit, Individualität, das Unbewusste und Sinnlichkeit essenziell notwendig für eine transformative Kreativität[220].

Die Möglichkeiten der Schaffung eines künstlichen Bewusstseins werden an späterer Stelle thematisiert.

[219] Vgl. Du Sautoy (2019), S. 282f; sowie die beiden folgenden Blockzitate

[220] Vgl. Zipp Vey (2018), S. 33

3.3 Natürliche Intelligenz vs. Maschinenintelligenz

Zusammenfassend stellt sich die Situation wie in Abbildung 4 dar. Die Intelligenztheorien sind in konzentrischen Ovalen angeordnet. In der Mitte befinden sich die beiden Theorien, die sich am stärksten auf das beschränken, was auch in IQ-Tests gemessen wird, die Generalfaktortheorie von Spearman und die primäre Gruppenfaktortheorie von Thurstone. Die beiden Modelle unterscheiden sich im Wesentlichen dadurch, dass bei Spearman eine viel stärkere Korrelation zwischen den Fachgebieten angenommen wird und dafür die generelle Intelligenz **g** propagiert wird. Thurstone geht von einem Ensemble unabhängiger Fähigkeiten aus.

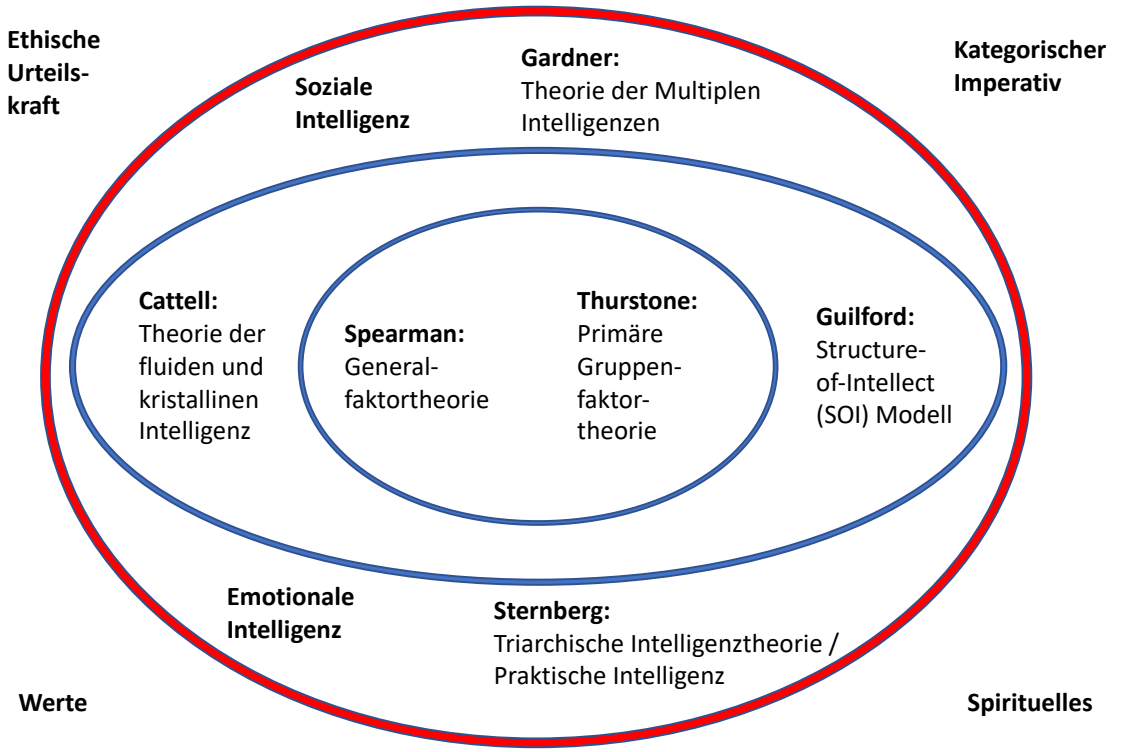

Abbildung 6: Intelligenztheorien

Außerhalb des ersten Ovals sind zwei Theorien dargestellt, die im Falle von Cattell stärker den Lerneffekt und die Erfahrung als Komponenten der kristallinen Intelligenz beinhalten oder auch bei Guilford in einem dreidimensionalen SOI Modell zwischen Inhalten, Operationen und Produkten des Denkens unterscheiden und damit die Intelligenz deutlich feiner strukturieren. Generell bewegt man sich hier noch im klassischen kognitiven Bereich. Außerhalb des zweiten blauen Ovals und innerhalb des roten Ovals befinden sich die alternativen Intelligenztheorien, die bei Gardner eine Reihe von multiplen Intelligenzen beinhalten oder bei Sternberg die praktischen Elemente im Gegensatz zu den akademischen Elementen ergänzen. Beiden Theorien ist gemein, dass sie die „*vielfältigen Dimensionen der menschlichen Lebenswelt*“[221] deutlich besser abbilden als die Theorien im

[221] Deutscher Ethikrat (2023), S. 20

inneren Oval; andererseits haben sich beide dem Vorwurf ausgesetzt, dass sie den Intelligenzbegriff verwässern und empirisch auf tönernen Füßen stehen.

Jenseits aller Theorien befinden sich einige exemplarische Elemente der kognitiven Intelligenz von Menschen, die bisher in keinem Modell angemessen Platz gefunden haben, wie z.B. das moralische Urteilen, die Reflektion über ethische Werte wie die Anwendung des kategorischen Imperativs von Kant oder auch Spirituelles.

Diese Betrachtungen sind bedeutsam für die Diskussion über die Frage, ob die KI nach den anerkannten Intelligenztheorien dem Profil der menschlichen Intelligenz genügen kann. Die Antwort erfolgt wiederum in der obigen Struktur. Es erscheint realistisch, dass schon die heutige KI in einer eigens dafür entwickelten Applikation in allen gängigen IQ-Tests, die i.d.R. Spearman oder Thurstone abbilden, nicht nur mit durchschnittlichen Menschen mithalten kann, sondern – ähnlich wie bei Schach oder Go – alle menschlichen Individuen weit hinter sich lassen kann. Bei der kristallinen Intelligenz nach Cattell und bei vielen der 150 Teilintelligenzen nach Guilford erscheint dies mehr als zweifelhaft. Gänzlich ausgeschlossen erscheint es für das gesamte Portfolio der multiplen Intelligenzen oder auch für die praktische Intelligenz, und noch viel stärker für die Punkte außerhalb der Ovale in der Graphik sowie für den wichtigen Typus der transformativen Kreativität. Grundvoraussetzung für die Abdeckung sämtlicher Fähigkeiten der Intelligenz im Sinne der alternativen Theorien in und außerhalb der äußeren Schale der Darstellung wäre die Existenz eines (künstlichen) Bewusstseins.

Generell konnte in diesem kurzen Exkurs in die Intelligenztheorien der Psychologie gezeigt werden, dass alle Theorien zusammen nicht die gesamte Breite und Tiefe dessen abdecken, was die menschliche Intelligenz umfasst. Insbesondere Kreativität, Moral, Werte und Spiritualität sowie das Zusammenspiel des freien Willens mit Neigungen und Einstellungen liegen außerhalb der etablierten Theorien und damit der Reichweite der Künstlichen Intelligenz. Landgrebe und Smith haben das recht treffend mit einer interessanten Metapher zusammengefasst:

> *"Taken together, all the AI definitions we have looked at try to claim that there can be machine intelligence, but only by using thereby an Ersatzdefinition of what the word intelligence means. As we have seen, however, this strategy will never lead us to the conclusion that there is machine intelligence in any commonly accepted meaning of this term. It is comparable, rather, to defining flying as 'moving in the air', and then jumping up and down and shouting: 'See, I am flying'."*[222]

Es ist nicht erkennbar, wie die Maschinenintelligenz der menschlichen Intelligenz mit all ihren Ausprägungen auch nur nahekommen kann.

[222] Landgrebe Smith (2022), S. 59

4 Leib-Seele-Problem

> *„Das Leib-Seele-Problem ist das ontologische Grundproblem in der Philosophie des Geistes*“
> Thomas Metzinger[223]

Eine geistesphilosophische Diskussion der Künstlichen Intelligenz ist nicht möglich ohne Bezugnahme auf das „Leib-Seele-Problem“, das wiederum zu den ganz großen und ältesten nicht zu Ende geführten Debatten der Philosophie gehört. Die einschlägige Literatur zum Thema ist grenzenlos.

Das „Leib-Seele-Problem“, „Körper-Geist-Problem“ (auf Englisch „mind-body-problem“) oder auch „Subjekt-Objekt-Spalte“ bezieht sich auf den Zusammenhang des Physischen mit dem Mentalen oder Geistigen. Die Frage ergibt sich konkret daraus, dass es uns schwerfällt, die alltäglichen Erfahrungen in uns selbst und unserer jeweiligen Sicht auf die Welt mit den Resultaten naturwissenschaftlicher Methoden und Wirkzusammenhängen in Einklang zu bringen. In der physischen Welt verstehen wir auf uns selbst bezogen die Anatomie unseres Körpers, den Blutkreislauf und unsere Sinnesorgane und in der Welt die Astronomie, die Schwerkraft, Radioaktivität, die Funktionsweise technischer Artefakte wie Autos, Flugzeuge oder Weltraumsatelliten sowie den Impulserhaltungssatz auf dem Billardtisch. In der mentalen Welt gehen wir mit unseren Gedanken, Wünschen, Bedürfnissen, Stimmungen, Empfindungen und Launen als psychische oder seelische Erscheinungen um. Phänomene sind entweder physisch oder geistig. Aber wie passen diese beiden Welten zusammen? Wie korrespondieren sie und wie nehmen sie aufeinander Einfluss? Tun sie es überhaupt?

In diesem Kapitel erfolgt eine Einführung in die Kernfragestellungen und Begrifflichkeiten rund um das Leib-Seele-Problem. Weiterhin wird ein Mittelweg aufgezeigt zwischen den Gegenpolen des akademischen (geistesphilosophischen) Diskurses, der für die weitere Deliberation in dieser Arbeit einen gangbaren Weg darstellt.

Die Problembeschreibung für dieses Kapitel (Leib-Seele-Problem) und das folgende Kapitel (Bewusstsein) fasst das folgende Zitat von Jaegwon Kim zusammen:

> *„Für den zeitgenössischen Physikalisten existieren **zwei Probleme**, die das Leib-Seele-Problem wahrhaft zu einem Weltknoten machen, zu einem schwer zu behandelnden und vielleicht letztlich unauflöslichen Rätsel. Sie beziehen sich auf die mentale Verursachung und auf das Bewusstsein. Das Problem der mentalen Verursachung besteht darin, eine Antwort auf diese Frage zu finden:*
>
> ***Wie kann der Geist seine kausalen Kräfte in einer Welt ausüben, die fundamental physikalischer Natur ist?***
>
> *Das Problem des Bewusstseins besteht darin, eine Antwort auf die folgende Frage zu finden:*

[223] Metzinger (2007), S. 11

> ***Wie kann so etwas wie ein Bewusstsein in einer physikalischen Welt existieren, einer Welt, die letztlich aus nichts anderem besteht als kleinsten Materieteilchen, die über die Raumzeit verteilt sind und sich in Übereinstimmung mit den physikalischen Gesetzen verhalten?***
>
> *Wie sich herausstellt, sind die beiden Probleme miteinander verbunden – die beiden Knoten sind miteinander verflochten, und diese Tatsache macht es umso schwieriger, auch nur einen der beiden zu entwirren."* [Hervorhebungen durch DS][224]

In unserer alltäglichen Erfahrung erscheinen die Übergänge zwischen dem, was in unserm Geist passiert, und den physischen Vorgängen in beide Richtungen fließend und natürlich. Verletzungen und Verbrennungen verursachen Schmerzen und führen dazu, dass wir uns von den schmerzverursachenden Orten entfernen, z.B. die Hand von der Herdplatte nehmen. Aus den Naturwissenschaften und insbesondere aus der Physik wissen wir, dass mentale Vorgänge mit der Physik nichts zu tun haben, dennoch scheinen sie im physischen Bereich – wie im obigen Beispiel – wirksam zu sein. Wenn wir ein Haus bauen oder ein Bild malen, dann entstammt die erste Idee dazu unserer geistigen Vorstellung. Unsere Städte und Museen zeugen von dieser geistig-kreativen Schaffenskraft im wahrsten Sinne des Wortes. Mehr noch: Unsere gesamte Kultur ist Ergebnis der mentalen Schaffenskraft des Menschen. Weiterhin wissen wir ebenfalls aus der Physik, dass die Vorgänge in der Physik vollständig über physikalische Kräfte und Wechselwirkungen sowie Energie- und Impulserhaltungssätze beschrieben werden können. Es gibt (zumindest nach allgemeiner Auffassung, die an späterer Stelle noch widerlegt wird) keinen Platz für sonstige kausale Einflüsse. Die Welt der Physik oder der physischen Phänomene gilt als **kausal geschlossen**.

Peter Bieri hat dies im sogenannten „***Bieri-Trilemma***"[225] in einem System von drei Sätzen zusammengefasst, von deren Wahrheit wir jeweils einzeln betrachtet überzeugt sind, die alle drei zusammen dagegen nicht wahr sein können. Zwei der Sätze schließen den jeweils dritten Satz aus:

> *„1. Mentale Phänomene sind nicht-physische Phänomene*
>
> *2. Mentale Phänomene sind im Bereich physischer Phänomene kausal wirksam*
>
> *3. Der Bereich physischer Phänomene ist kausal geschlossen"*[226]

Satz 1 und Satz 2 zusammen schließen Satz 3 aus. Mentale und physische Phänomene existieren beide nebeneinander und die mentalen Phänomene haben Einfluss auf die physischen. Somit kann die kausale Geschlossenheit nicht stimmen. Dies ist z.B. die Position des interaktionalistischen Dualismus, die im Abschnitt 4.2.2 ausführlicher diskutiert wird.

---

[224] Übersetzte Fassung aus Metzinger (2007), S. 11 – 12; ursprüngliche Quelle: *Physicalism or Something Near Enough* (Princeton und Oxford 2005: S. 7)

[225] Hier zitiert aus Metzinger (2007), S. 14

[226] Ebd.

Satz 1 und Satz 3 zusammen beinhalten auch den Dualismus bei gewahrter kausaler Geschlossenheit des Physischen. Damit kann es keinen kausalen Einfluss des Mentalen auf das Physische geben. In Abschnitt 4.2.2.2 wird dies als paralleler Dualismus in drei Ausprägungen präsentiert: Leibnizscher Parallelismus/Prästabilisierende Harmonie, Okkasionalismus und Epiphänomenalismus.

Satz 2 und Satz 3 zusammen verneinen den Dualismus im ersten Satz und führen zur monistischen Position des Physikalismus in verschiedenen Ausprägungen (nichtreduktiv, reduktiv, eleminativ), u.a. auch als Funktionalismus in Abschnitt 4.3 diskutiert, ebenfalls nicht ohne weitere Probleme, die aus der Aufgabe von Satz 1 resultieren.

Im weiteren Verlauf dieses Kapitels wird zunächst der philosophiehistorische Kontext für die Fragestellung zusammengefasst, dann die Problematik der dualistischen Positionen dargestellt und zuletzt ausführlicher auf die monistischen Positionen eingegangen.

4.1 Ein historischer Zugang – vom antiken Ägypten bis Descartes

In der Menschheitsgeschichte hat man immer wieder über den Begriff „Geist“ und das Phänomen des „Geistigen“ in Abgrenzung zur „Materie“ oder zum „Physischen“ nachgedacht. Bereits in der fünftausend Jahre zurückliegenden Kultur der Ägypter kannte man Konzepte für das „Geistige“ und „Psychische“, *Ach*, *Ba* und *Ka. Ach* steht für die ewigen Daseinskräfte und ergibt sich aus dem Totenkult und wird in Vögeln (Schopfibis) verkörpert. *Ba* entspricht dem, was wir heute unter der Seele verstehen. Damals stellte man sich vor, dass sich Ba im Schlaf und im Tod vom Körper löst. Auch *Ba* wurde als Vogel dargestellt (Falke). *Ka* repräsentiert die Lebens-, Zeugungs- und Geisteskraft des Menschen[227].

Das europäische Denken in Bezug auf Philosophie, Wissenschaft und Logik begann in der griechischen Antike. Bruno Snell spricht in seinem gleichnamigen Buch von der „*Entdeckung des Geistes*“ durch die Griechen:

> *„… die Griechen haben nicht nur mit Hilfe eines schon vorweg gegebenen Denkens nur neue Gegenstände (etwa Wissenschaft und Philosophie) gewonnen und alte Methoden (etwa ein logisches Verfahren) erweitert, sondern haben, was wir Denken nennen, erst geschaffen: der menschliche Geist als tätiger, suchender, forschender Geist ist von ihnen entdeckt; eine neue Selbstauffassung des Menschen liegt dem zugrunde.“*[228]

Die ersten Schriften, die auch das beinhalteten, was wir heute als „Geist“ oder „Seele“ bezeichnen, gehen auf Homer zurück, mutmaßlich aus dem siebten oder achten Jahrhundert v.u.Z. (vor unserer Zeitrechnung). Die Wörter, die Homer in seinen Epen Ilias und Odyssee „*zur Charakterisierung der Sphäre des Psychisch-Geistigen verwendet*“ hat, sind „*Psyché*[229]*, Thymos*[230] *und Nóos*“[231]. Die Bedeutung von „**Psyché**“ bei Homer ist anders als bei den späteren Griechen, die den Begriff „Seele“ nutzten. Er spricht von „Hauch“, „Atem“, „Lebensodem“ oder auch von „Schmetterling“, vermutlich wegen dessen Leichtigkeit und Luftigkeit. „**Thymos**“ verwendet er für das, was Regungen verursacht, was den Menschen in Tätigkeit versetzt, interessanterweise eher für Reaktionen als für Aktionen[232]. Decher spricht vom „*Organ der reagierenden Regung*“[233]. „**Nóos**“ ist das „*Organ, das Vorstellungen aufnimmt*“:

> *„Nóos, das Wort, das im späteren Griechisch ‚Geist‘ meint, gehört zum Verb noein. Und das bedeutet ‚einsehen‘, ‚durchschauen‘, ja weithin lässt es sich bei Homer mit ‚sehen`*

---

[227] Vgl. Decher (2015), S. 17 - 19

[228] Snell (1975), S. 7

[229] Psyche bzw. Psyché, altgr. ψυχή psyché, deutsch Atem', ‚Hauch', ‚Schmetterling', ‚Leben', ‚Seele'

[230] Thymos, altgr. θυμός thymos, deutsch ‚Lebenskraft‘

[231] Decher (2005), S. 21; auf Basis von Bruno Snell: „Die Entdeckung des Geistes“

[232] Vgl. Snell (1975), S. 24

[233] Decher, S. 22

übersetzen. Aber es ist ein Sehen, das nicht nur den rein visuellen Akt bezeichnet, sondern die geistige Wahrnehmung, die mit dem Sehen verbunden ist.“[234]

Bei Homer ist der Mensch noch ein rein reaktives Wesen ohne eigene Willensentscheidungen. Diese bleiben noch vollständig den Göttern vorbehalten:

„… denn es lässt sich wohl denken, dass Seelenorgane Noos und Thymos, die nicht aus sich denken und nicht aus sich regen können, denen überhaupt echte eigene Tätigkeit fremd ist, dem Zauber preisgegeben sind, und dass Menschen, die ihr eigenes Innere so interpretieren, sich selbst als Schauplatz willkürlicher und unheimlicher Gewalten fühlen. Danach mögen wir ahnen, wie vor Homer der Mensch sich und sein Tun gedeutet hat. Die Helden der Ilias aber fühlen sich nicht mehr wüsten Kräften ausgesetzt, sondern ihren olympischen Göttern, und diese sind eine wohlgegliederte und sinnvolle Welt. Immer mehr von dem Wirken dieser Götter nehmen die Griechen, je weiter ihre Selbstauffassung sich entwickelt, gewissermaßen in den menschlichen Geist herein. […] *Die homerische Selbstauffassung des Menschen, wie wir sie in der Sprache Homers begreifen, ist nicht nur primitiv, sondern weist auch weit in die Zukunft – es ist* ***die erste Stufe des europäischen Denkens****.“*[235] [Hervorhebung DS]

In der weiteren griechischen Antike wurden die ägyptischen Konzepte und homerischen Überlegungen unter den Vorsokratikern[236] weiterentwickelt. Aus ‚Nóos‘ wurde ‚**Nous**‘[237].

Platon (427 – 347 v.u.Z.) hat in seinen Dialogen (etwa *Menon, Phaidon, Phaidros, Symposion* und *Politeia*) die Seele bzw. die Psyche als etwas beschrieben, *das unabhängig vom Körper existiert und sich nur während der Spanne eines Lebens mit ihm verbindet, aber, anders als der Körper, nicht sterblich ist*[238]. In der Politeia entwickelt er sein Konzept der Dreiteilung der Seele in Nous (Geist/Vernunft), Thymos (zorniger Drang) und Epithymiai (Begierde)[239], die auch der Dreiteilung des Staates entspricht:

„Und ein gerechter Mann unterscheidet sich in dem Punkt der Gerechtigkeit in nichts vom gerechten Staat, sondern ist ihm ähnlich?“
„Natürlich ähnlich!“
„Der Staat erschien uns dann als gerecht, wenn in ihm drei Arten von Naturen sind, deren jede ihre Aufgabe erfüllt, zudem aber besonnen, tapfer und weise wegen anderer Eigenschaften und Haltungen dieser drei Naturen.“
„Richtig!“
„Der einzelne nun, mein Freund, muß – so fordern wir – dieselben drei Formen in seiner

---

234 Decher (2015), S. 22f, weiterhin bezugnehmend auf Snell

235 Snell (1975), S. 29; dazu detailliert in FN 48, S. 298: „*Daß es wichtig auch für das heutige Philosophieren sei zu wissen, daß die Unterscheidung von Leib und Seele nicht etwas ist, das der Mensch ‚von Natur aus‘ kennt, sondern etwas, das er hat lernen müssen*.“

236 Z.B. Xenophanes (580/577 – 485/480 v. Chr.), Heraklit (544 – 483 v. Chr.), Parmenides (540 – 480 v. Chr.), Demokrit (479 – 370 v. Chr.)

237 Nous aus dem Wörterbuch der philosophischen Begriffe (2013), S. 462: „*nous, gr., spr. Nus (lat. Intellectus), das Vermögen der geistigen Wahrnehmung, svw. Verstand, bei Platon und Aristoteles der edelste und höchste der drei Seelenteile, seit Anaxagoras auch die sinnvoll wirkende, harmonisch ordnende Weltkraft neben dem Weltstoff, der weltordnende Geist. Nous heißt schon bei Homer das Erkenntnisvermögen; von Parmenides und Demokrit wird es der Seele gleichgesetzt.“*

238 Decher (2015), S. 41

239 Decher (2015), S. 42

> *Seele haben und wegen derselben Eigenschaften auch dieselben Namen erhalten wie der Staat."*[240]

Ansgar Beckermann leitet daraus vier Thesen des platonischen Dualismus ab[241]:

1. *„Der Mensch besteht nicht nur aus einem Körper, sondern aus einem Körper und einer Seele, die Seele ist ein immaterielles Wesen.*
2. *Die Seele macht das eigentliche Selbst eines Menschen aus. Sie (und damit der Mensch) ist für ihre Existenz nicht auf den Körper angewiesen.*
3. *Körper und Seele des Menschen sind nur während seines Erdenlebens zusammengespannt, beim Tode löst sich die Seele vom Körper.*
4. *Während der Körper vergänglich ist, ist die Seele unsterblich. (Kann sie den Tod des Körpers überleben)"*

Platons Schüler Aristoteles (384/83 – 323/22 v. Chr.) erarbeitete eine deutlich detailliertere Seelenlehre, die Hellmut Flashar wie folgt zusammenfasst:

> *„Die aristotelische Psychologie als Seelenlehre ist eine ganz eigenständige Konzeption, die weder Vorgänger noch eigentliche Nachfolger hat. Aristoteles stand unter dem Eindruck der ungeheuren Bedeutung, die Platon der Seele in Ethik und Ontologie zugemessen hatte. Damit standen die Probleme der Unsterblichkeit der Seele, des Dualismus von Seele und Körper, die Abwertung des Körpers gegenüber der Seele im Vordergrund.*
>
> *Da musste es wie ein Befreiungsschlag wirken, diese Kluft zu überbrücken, Körper und Seele wieder zusammenzuführen, die Seele als `Entelechie'*[242] *des Körpers zu bezeichnen, die starre Einteilung in Seelenteile zu überwinden. Im Groben unterscheidet auch Aristoteles mit der Gliederung in ernährende, wahrnehmende und denkende Seele wie Platon drei Seelenteile, doch kommt es ihm nicht auf die Trennung von Teilen, sondern auf die Betonung eines großen Bewegungszusammenhanges an, in dem die Seele den Körper zu dem macht, was er eigentlich ist*."[243]

Auch nach Friedhelm Decher sieht Aristoteles den „*Geist als höchste Funktion der Seele"*[244]. Die Unsterblichkeit der Seele findet sich bei Aristoteles, anders als bei Platon, nicht. Im geistesphilosophischen Sinne war Aristoteles der erste Materialist.

Philosophen im weiteren Verlauf der Antike und Spätantike[245] sowie Vertreter des Katholizismus[246] haben Varianten der konzeptionellen Ansätze von Platon und Aristoteles vorgestellt, die hier nicht weiter vertieft werden sollen.

---

[240] Platon, Politaia, Viertes Buch, 434b; oder Platon (1958), S. 226f

[241] Beckermann (2008), S. 12

[242] Siehe auch Glossar

[243] Flashar (2013), S. 316

[244] Decher (2015), S. 44

[245] Z.B. Lukrez, Epikur, Plotin

[246] Z.B. Augustinus, Cusanus, Thomas von Aquin

Für den nächsten größeren, wenn nicht gar den größten Meilenstein in der Entwicklung der Geistesphilosophie ist René Descartes (1596 – 1650) verantwortlich. So schrieb Bertrand Russell in „*Philosophie des Abendlandes*“:

> *„René Descartes gilt im Allgemeinen als der Begründer der modernen Philosophie, und das, wie mir scheint, zu Recht. Er ist der wahrhaft philosophisch veranlagte Denker, der in seiner Weltanschauung von der neuen Physik und Astronomie tief beeindruckt wird. Wenn er auch noch vieles von der Scholastik beibehält, so begnügt er sich doch nicht damit, auf den von seinen Vorgängern geschaffenen Fundamenten weiterzubauen, versucht vielmehr ein vollständiges philosophisches Gebäude neu aufzurichten. Das war seit Aristoteles nicht mehr geschehen und zeugt von dem neuen Selbstvertrauen, das eine Frucht des wissenschaftlichen Fortschritts ist.“*[247]

Descartes verwarf die mehr als zweitausend Jahre Philosophiegeschichte vor ihm, „*so dass einmal im Leben alles vom Grund auf umgeworfen und von den ersten Fundamenten her erneut begonnen werden müsse“*[248]. In seiner ersten Meditation zieht er alle Erkenntnisse und vor allem die Prinzipien der Erkenntnis in Zweifel, an die er bis zu dem Zeitpunkt geglaubt hatte. Insbesondere schloss er alles aus, was nur mittels sinnlicher Wahrnehmung vermittelt wurde:

> „*Aber ich habe entdeckt, dass die Sinne zuweilen täuschen, und Klugheit verlangt, sich niemals blind auf jene zu verlassen, die uns nur einmal betrogen haben*.“[249]

Dabei nimmt er billigend in Kauf, dass mit dieser „Untergrabung des Fundaments“ das ganze Wissenschaftsgebäude zusammenbricht.[250] „*Ein boshafter Genius, ebenso allmächtig wie verschlagen“*, könnte „*all seine Hartnäckigkeit“* daransetzen, ihn „*zu täuschen.“*[251]

In der zweiten „*Meditation*“ identifiziert Descartes einen Punkt, für den der Zweifel aus der ersten Meditation nicht gelten kann, weil er fundamental verschieden ist:

> *„Zweifelsohne bin ich selbst also, wenn er mich täuscht; und er möge mich täuschen, so viel er kann, niemals wird er bewirken, dass ich nichts bin, so lange ich denken werde, dass ich etwas bin; so dass schließlich, nachdem ich es zur Genüge überlegt habe, festgestellt werden muss, dass dieser Grundsatz ‚**Ich bin, Ich existiere**‘, so oft er von mir ausgesprochen oder durch den Geist begriffen wird, notwendig wahr ist.“*[252]

Von diesem Fixpunkt „Ich bin, Ich existiere“ oder „***ego sum, ego existo***“ geht er dann aus und kommt zur folgenden Schlussfolgerung:

---

[247] Russell (1945), S. 567

[248] Descartes (2009), S. 19

[249] Descartes (2009), S. 20

[250] Ebd.: *„… ich will unverzüglich auf die Prinzipien selbst losgehen, auf die sich alles stützte, das ich einst geglaubt habe; denn wenn die Fundamente untergraben sind, fällt alles, was auf ihnen errichtet ist, von selbst zusammen.*“

[251] Descartes (2009), S. 24

[252] Descartes (2009), S. 28

> *„Was aber bin ich demnach? Ein denkendes Ding. Was ist das? Nun – ein denkendes, einsehendes, behauptendes, bestreitendes, wollendes, nicht wollendes und auch etwas sich vorstellendes und sinnlich wahrnehmendes Ding.“*[253]

Diesen Gedanken entwickelte er in seiner „Abhandlung von der Methode“ weiter zu dem berühmten „**Ich denke, also bin ich**“[254] oder „***cogito, ergo sum***“.

Das „*denkende Wesen*“, das wahrnimmt, zweifelt, einsieht, bejaht, verneint, will, empfindet, sich erinnert oder sich etwas vorstellt, nennt er auf lateinisch ***res cogitans***. Darin enthalten sind also alle mentalen Eigenschaften des Menschen: Denken, Geist und Bewusstsein[255]. Die mentale Seite grenzt er bewusst ab von der physischen, körperlichen Seite, der ***res extensa***, auf Deutsch „ausgedehnte Substanz“. Der Körper ist materiell, ausgedehnt und teilbar. In seinen Schriften bezeichnet er den menschlichen Leib als „Automat“ oder auch „Maschine“. Aus seiner Sicht wird der Geist „*nicht von allen Teilen des Körpers affiziert, sondern lediglich durch das Gehirn*“[256].

Friedhelm Decher fasst Descartes‘ Philosophie als ein „*Konzept eines Geistes in einer Maschine*“ (unsterblicher Geist in sterblicher Maschine) zusammen:

> *„Einerseits betont sie* [Descartes‘ Philosophie; Anmerkung DS], *der Mensch sei eine Einheit, sei ein Ganzes aus Geist und Körper. Andererseits gibt sie zu bedenken, Geist und Körper seien, da der Körper materiell und teilbar, der Geist hingegen inmateriell und unteilbar sei, nicht nur ‚verschiedenartig‘, sondern ‚in gewisser Weise gegensätzlich‘. Mit anderen Worten: Geist und Körper,* ***res cogitans*** *und* ***res extensa****, besitzen jeweils unterschiedliche Eigenschaften: Der Geist beziehungsweise die (vernünftige) Seele ist wesentlich eine denkende, nicht ausgedehnte, nicht teilbare, immaterielle, unsterbliche Substanz[257]. Der Körper dagegen ist ausgedehnt, teilbar, materiell, sterblich, ist eine Maschine.“*[258]

Weiterhin geht Descartes davon aus, dass der immaterielle Geist mit dem materiellen Körper interagiert, d.h. der Geist übt Einfluss auf den Körper aus und umgekehrt. Genau mit dieser Annahme stößt er schon zu Lebzeiten auf Widerspruch und Unverständnis. Wie soll diese Interaktion funktionieren, etwa konkret im menschlichen Gehirn?

---

253 Descartes (2009), S. 32

254 Descartes (1637), S. 64: „*je pense, donc je suis*“; und auf deutsch Descartes (1637), S. 65: „Ich denke, also bin ich“; die lateinische Fassung („ego cogito ergo sum“) erschien in einer späteren Schrift (1644, Die Prinzipien der Philosophie)

255 Vgl. Gabriel (2018), S. 261: „*Descartes versteht unter Denken (Lateinisch cogitare) übrigens auch sinnliche Vollzüge. Für Descartes sind Empfinden (sentire) und Vorstellen (imaginari) benso wie Wollen (velle) Denkvorgänge. Er reduziert das Denken gerade nicht auf das intelligere, also auf die Ausübung rationaler Berechnungsvorgänge*“

256 Descartes (2009), S. 93

257 Decher bezieht sich hier auf Descartes‘ Meditationen aus der „Übersicht über die sechs folgenden Meditationen“ (Descartes (2009), S. 13f

258 Decher (2015), S. 96

Dieses **Wie** (modus operandi) der Wechselwirkung ist nach Godehard Brüntrup *„das eigentliche empirische Kernproblem des Dualismus"*[259]. In unserer menschlichen Lebenswelt erfahren wir ständig, wie physische Beeinträchtigungen (Verletzungen) Schmerzen verursachen und wie wir umgekehrt Einfluss auf die physische Welt nehmen, wenn wir unsere Entscheidungen umsetzen oder unsere Bedürfnisse befriedigen wollen, zum Beispiel beim Aufheben eines Apfels, um ihn zu essen. Aber wie funktioniert dieser Übergang zwischen den beiden so verschiedenen Substanzen? Auch Descartes hat für diese Frage keine nachvollziehbare Antwort gefunden. Er entwickelte die Vorstellung, dass die Zirbeldrüse[260] die Rolle des „Transformators" zwischen Geist und Körper wahrnehme – eine Annahme, die schon damals kaum ernst genommen wurde.

Derzeit gibt es kaum noch Wissenschaftler, die die von Descartes entwickelte Theorie des Substanzdualismus unterstützen. Trotzdem hat kein anderer Philosoph die Entwicklung der Geistesphilosophie in den zurückliegenden 450 Jahren stärker geprägt. Der kartesianische Dualismus hatte im 17. Jahrhundert eine über die Philosophie hinausgehende Bedeutung, da er das Materielle klar vom Geistlichen abgrenzte und damit auch die Naturwissenschaften von der Religion[261]. In anderen Worten: Es zeichnete sich eine vorläufige Auflösung der Konflikte zwischen Glauben und Vernunft bzw. Naturwissenschaftlern und Theologen ab. So schrieb John Searle:

> *„Der Geist wurde als unsterbliche Seele betrachtet und war kein angemessenes Thema für naturwissenschaftliche Untersuchungen. Körper konnten von Wissenschaften wie der Biologie, der Physik und der Astronomie erforscht werden. Die Philosophie, so glaubte Descartes übrigens, könne sowohl den Geist als auch den Körper untersuchen*."[262]

In den folgenden Abschnitten soll es einerseits um die Argumente für und wider den Dualismus mit all seinen Spielarten und andererseits um die verschiedenen monistischen Ansätze gehen.

---

[259] Brüntrup (2018), S. 49

[260] „*Die Zirbeldrüse ist eine pinienzapfenähnliche, 8-14 mm lange Drüse, die beim Menschen am Mittelhirn liegt. Sie hat direkt weder etwas mit der neuronalen Verarbeitung von Sinnesreizen noch mit der Steuerung von Körperbewegungen zu tun; vielmehr dient sie der Produktion des Hormons Melatonin. Dieses Hormon wirkt jedoch – wie andere Hormone – indirekt auf die gesamte neuronale Informationsverarbeitung*." (Quelle: Beckermann (2008), S. 44)

[261] Vgl. Searle (2006), S. 21

[262] Searle (2006), S. 21

4.2 Probleme und offene Fragen zum Substanzdualismus

4.2.1 Begriffliches Raster und weitere Dualismen

Aufbauend auf dem Geist-Körper Dualismus nach Descartes zählt Brüntrup eine Reihe weiterer Dualismen auf, die damit zusammenhängen[263]:

- ***Subjektiv-objektiv***
  Geistige Phänomenen und Empfindungen kann immer nur der betroffene Mensch beschreiben, subjektiv, aus seiner Ich-Perspektive oder „Erste-Person-Perspektive". Was er beschreibt, ist beobachterrelativ. Meine Zahnschmerzen kann nur ich selbst beschreiben. Physische Phänomene werden objektiv, für jedermann nachvollziehbar beschrieben. Es ist egal, wer sie beschreibt, sie sind beobachterunabhängig.
- ***Privat-öffentlich***
  Dieser Punkt hängt eng mit dem vorherigen zusammen. Mentale Erfahrungen werden aus der Innenperspektive der Menschen wahrgenommen und sind für Dritte nicht ersichtlich. Man kann Gedanken nicht lesen. Anders ist dies bei physischen Ereignissen und Phänomenen. Sie sind für jedermann ersichtlich und öffentlich.
- ***Unkorrigierbar-korrigierbar***
  Die menschlichen Erkenntnisse über die physische Welt können fehlerhaft sein, wie zum Beispiel die gefühlte Geschwindigkeit des eigenen Autos, die Lufttemperatur oder das Gewicht einer Melone in der Hand. Diese Wahrnehmungen können allerdings über bessere Messmethoden oder den Austausch mit anderen Menschen korrigiert werden. Bei psychischen Phänomenen ist die Situation anders. Wenn man Schmerzen hat, kann man nicht davon überzeugt werden, sie nicht zu haben. Mentale Phänomene sind also nicht korrigierbar.
- ***Temporal-spatio temporal*** (zeitlich-zeitlich&räumlich)
  Wenn man Phänomene der Quantenmechanik außer Acht lässt, haben physische Gegenstände eine eindeutige Position in den drei Dimensionen des Ortes und der vierten Dimension der Zeit. In der mentalen Welt kann man den Gedanken zwar eine zeitliche Positionierung und Reihung zuordnen, aber keine räumliche Positionierung oder Größenordnung. Ein Gedanke hat keinen Ort und keine Ausdehnung.
- ***Intentional-nichtintentional***
  Im weiteren Fortgang dieser Arbeit wird noch ausführlich auf die Intentionalität eingegangen. An dieser Stelle soll nur knapp herausgestellt werden, dass die Intentionalität bedeutet, dass unsere Gedanken gerichtet sind auf Objekte in

[263] Auflistung nach Brüntrup (2018), S. 16f

unserm Innern oder auch außerhalb oder gar auf Fantasieobjekte. Dies beinhaltet auch unsere Überzeugungen über diverse Sachverhalte, die aus anderen Überzeugungen folgen können. Intentionale „*Zustände lassen sich beispielhaft durch eine Analyse unserer Überzeugungen darstellen. Sie lassen sich in die Form bringen: ‚X glaubt, dass P'*“[264]. So kommt Rationalität zustande. Nach derzeitigem Forschungsstand gibt es dafür kein Äquivalent in der physischen Welt. *„Wenn wir einer Person Rationalität zuschreiben, dann deshalb, weil ihre Überzeugungen logisch stimmig sind und sie diese Überzeugungen zur Ursache ihres Handelns werden lässt. Ein Gegenstand des physischen Bereichs hingegen ist weder rational noch irrational, auch handelt er nicht.“*[265]

- ***Frei-determiniert***
  Der physische Bereich ist vollständig durch die Naturgesetze determiniert, wie zum Beispiel die Energie- und Impulserhaltungssätze[266]. Der physische Zustand zum Zeitpunkt t2 ist abhängig vom physischen Zustand zum vorherigen Zeitpunkt t1 und von den Naturgesetzen. Die Geschwindigkeit eines fallenden Gegenstandes ist abhängig von der vorherigen Geschwindigkeit und der Erdbeschleunigung. Kompliziert wird dies bei Bewegungen, die wir mit unserm Körper ausführen, wie zum Beispiel das Heben des Armes, um einen Apfel aufzuheben. Dies ist ein Ergebnis einer mentalen Regung oder Willensbekundung, die frei ist und nicht determiniert ist, wie wir meinen.[267]

4.2.2 Das Problem der Interaktion von Geist und Körper

Der ursprünglich von Platon entwickelte und von René Descartes noch einmal neu begründete Substanzdualismus lässt sich in drei Aussagen zusammenfassen[268]:

1. *„Der menschliche Körper ist eine res extensa; er ist ausgedehnt, teilbar, materiell.*
2. *Der menschliche Geist ist eine res cogitans; er ist nicht ausgedehnt, nicht teilbar, immateriell.*
3. *Der immaterielle Geist und der materielle Körper interagieren: Der Geist wirkt auf die Maschine ein und die Maschine auf den Geist.“*[269]

[264] Brüntrup (2018), S. 18

[265] Brüntrup (2018), S. 19

[266] Auch hier sollte die Quantenmechanik zunächst ausgeklammert werden

[267] Die grundsätzlichen Überlegungen zur kausalen Geschlossenheit der physischen Welt, zur psychophysischen Interaktion und zum mentalen Determinismus bzw. freien Willen sollen an dieser Stelle nicht weiter vertieft werden. Die Meinungen zur Determinierung mentaler Zustände sind sehr breit gestreut.

[268] Vgl. Decher (2009), S. 98

[269] Decher (2009), S. 98

Die weitere philosophische Diskussion noch zu Lebzeiten von Descartes und seither hat sich zunächst auf den dritten Punkt fokussiert, also auf die Frage, ob und wie Körper und Geist interagieren.

4.2.2.1 Dualistische Ansätze

Vier grundsätzliche Lösungsansätze sind von verschiedenen Philosophen über die Jahrhunderte erarbeitet worden[270]:

1. ***Der interaktionalistische Dualismus***
   Physische Zustände verursachen mentale Zustände und mentale Zustände verursachen physische Zustände, „*Geist und Körper"* beeinflussen sich „*gegenseitig kausal, obwohl sie Substanzen verschiedener Art sind"*[271].
   Vertreter: René Descartes, John Eccles
2. ***Parallelismus***
   Physische und mentale Zustände beeinflussen sich nicht gegenseitig über Kausalbeziehungen, sondern über eine „prästabilisierende Harmonie". Göttliche Fügung sorgt dafür, dass Zustände in Körper und Geist einander entsprechen. Die Analogie ist ein Uhrmacher, der zwei Uhren synchronisiert und damit dafür sorgt, dass sie beide dieselbe Zeit anzeigen, ohne dass zwischen Ihnen ein Zusammenhang bestünde.
   Vertreter: Gottfried Wilhelm Leibniz (1646 – 1716), Baruch Spinoza (1632 – 1677)
3. ***Okkasionalismus***
   Der systematische Zusammenhang zwischen Geist und Körper wird hier über den direkten Eingriff sichergestellt. Bei bestimmten Zuständen im Körper werden die entsprechenden Zustände im Geist hervorgebracht und umgekehrt. Dies geschieht dann direkt ohne Umweg über eine „prästabilisierende Harmonie".
   Vertreter: Arnold Geulincx (1624 – 1669), Nicolas Malebranche (1638 – 1715)
4. ***Epiphänomenalismus***
   Zustände im Geist sind quasi redundante zusätzliche Phänomene der Zustände im Körper, sogenannte Epiphänomene. Es gibt also eine Kausalität von den physischen Zuständen zu den mentalen Zuständen, allerdings NICHT umgekehrt.
   Vertreter: Thomas Huxley (1825-1895), Ernst Haeckel (1834-1919)

---

[270] Vgl. Beckermann (2008), S. 42f, die Zusammenfassungen der Theoriebeschreibungen wurden dort größtenteils übernommen; Vgl. ebenfalls: https://plato.stanford.edu/entries/dualism/: "*If mind and body are different realms, in the way required by either property or substance dualism, then there arises the question of how they are related. Common sense tells us that they interact: thoughts and feelings are at least sometimes caused by bodily events and at least sometimes themselves give rise to bodily responses. I shall now consider briefly the problems for* ***interactionism****, and its main rivals,* ***epiphenomenalism*** *and* ***parallelism****."*; Hervorhebungen durch DS

[271] Beckermann (2008), S. 38

Der interaktionalistische Dualismus wird bis heute immer wieder aufgegriffen. Allerdings bestehen einige empirische und theoretische Ungereimtheiten. Einerseits konnte die Wirkung des Geistes auf das Gehirn empirisch nicht nachgewiesen werden. Auch ist das Einwirken des Geistes mit den Impuls- und Energieerhaltungssätzen der Physik nicht zu vereinbaren (kausale Geschlossenheit der physischen Welt) . Letztlich gibt es keinerlei überzeugende Erklärungen zu den Mechanismen der Wechselwirkung.

Der Parallelismus und auch der Okkasionalismus haben sich als „Krücken" zur Überbrückung der Lücke zwischen Körper und Seele erwiesen und werden beide heute nicht mehr unterstützt.

Der Epiphänomenalismus wird in einigen philosophischen Kreisen bis heute ernsthaft diskutiert. Insbesondere war man lange der Meinung, dass es dafür auch empirische Evidenz gebe. Andererseits bestehen auch auch hier erhebliche theoretische Probleme. Wenn das Geistige nur eine Begleiterscheinung des Physischen ist und selbst keinen Einfluss auf den Gang der Dinge hat, dann würde die Welt ohne den geistigen Teil genauso aussehen wie mit ihm. Dies wiederum erscheint aus lebensweltlicher Sicht kaum plausibel.

Insgesamt unterstützen in der Gegenwart nur noch sehr wenige Philosophen den Substanzdualismus. Im nächsten Abschnitt wenden wir uns dem Gegenpart zu, dem Monismus.

4.2.2.2 Die kausale Geschlossenheit des Physischen

Neben dieser Frage nach dem „Wie" gab es immer wieder eine Diskussion zur „kausalen Geschlossenheit des physischen Bereiches". Alle Theorien, die eine Wirkung des Mentalen auf das Physische unterstellen, brechen mit diesem physikalischen Prinzip.

Die heutige Physik konstatiert vier *fundamentale Wechselwirkungen*[272],

> „*... die durch Felder entstehen: die schwache, die starke, die elektromagnetische und die Gravitation. Jede Wechselwirkung hat ihre Träger. Die schwache Wechselwirkung hat alle Elementarteilchen als Träger, die starke Wechselwirkung hat die sogenannten Hadronen*[273] *als Träger, die elektromagnetische Wechselwirkung wird von elektrischen Ladungen getragen, die Gravitation hat schließlich Massen als Träger.*"[274]

Nun behauptet die Physik, dass man mit diesen vier Wechselwirkungen im Kontext weiterer physikalischer Grundgesetze wie dem Energie- und dem Impulserhaltungssatz alle physikalischen Phänomene beschreiben und erklären kann. Da besteht kein Spielraum für sonstige Eingriffe und Interventionen.

Zusammengefasst wird dies mit dem Prinzip des *methodologischen Physikalismus*:

---

[272] Vgl. Brüntrup (2018), S. 51

[273] **Hadronen** sind zum Teil Neutronen und Protonen, also Elementarteilchen, die wiederum aus Quarks zusammengesetzt sind.

[274] Brüntrup (2018), S. 51

> *„Eine Kausalerklärung eines physischen Ereignisses p_1 gilt dann und nur dann als gelungen, wenn sie nur physische Ereignisse p_2, p_3, …, p_n identifiziert, die p_1 verursacht haben.“*[275]

Es ist aber unsere Alltagserfahrung, dass wir mit unseren mental verursachten Körperbewegungen auf der Makroebene in die physische Welt eingreifen. Damit „*wäre die Idee von verlässlichen physikalischen Gesetzen außer Kraft gesetzt“*. Gleiches gilt für den obigen Energieerhaltungssatz: Wenn dem physischen Gesamtsystem unserer Welt Energie aus der geistigen Welt zugeführt oder entnommen wird, gilt der Energieerhaltungssatz nicht mehr, es sei denn, die geistige/mentale Welt ist Teil der physischen Welt. Dies führt zu den monistischen Theorien, die in einem späteren Abschnitt thematisiert werden.

Auf alle Fälle gilt folgender von Brüntrup formulierte Satz:

> *„Wenn* [der interaktionistische Dualist] *überhaupt eine Chance haben will, seine Theorie einigermaßen plausibel zu machen, dann muss er zeigen, dass die physische Welt auf solche Weise für eine Einflussnahme des Mentalen offen ist, dass dadurch der innere gesetzesmäßige Zusammenhang der physischen Welt (z.B. Energieerhaltung) nicht gefährdet ist.“*[276]

Philosophen der letzten hundert Jahre (zum Beispiel John Eccles) erkennen diese Offenheit für die Einflussnahme des Mentalen in der physischen Welt in der Quantenmechanik, die immer wieder als „Silver Bullet“ oder Allheilmittel für das Unerklärliche herhalten muss.

Ausgerechnet die Physikerin (und Philosophin) Brigitte Falkenburg schränkt in ihrem Buch zum *„Mythos Determinismus“*[277] die Forderung nach der kausalen Geschlossenheit als zu erfüllender Randbedingung aus der Perspektive drastisch ein. Sie begründet dies ebenfalls mit der Quantenmechanik und mit der Thermodynamik, beides Prozesse mit stochastischen Grundlagen, die eine Reversibilität, welche ein Merkmal der geschlossenen Kausalität wäre, nicht zulassen:

> „[Die] *Physik kann den* […] *Kausalitätsbegriff nicht eindeutig präzisieren. Die kausalen Prozesse der Physik sind <u>entweder</u> deterministisch, reversibel und zeitsymmetrisch (Mechanik, Elektrodynamik, Signal-Ausbreitung nach Einstein). <u>Oder</u> aber sie sind Zeit asymmetrisch, irreversibel und indeterministisch (Thermodynamik; quantenmechanischer Messprozess).“*[278]

Sie führt aus, dass es einerseits „*reversible dynamische*“ und andererseits „*irreversible statistische Naturgesetze*“[279] gibt, die in beide Richtungen nicht aufeinander reduzierbar sind, oder „*höchstens mit beträchtlichem metaphysischem Aufwand*“, wie sie ergänzt. Damit gilt (Falkenburg spricht den Leser an):

---

[275] Brüntrup (2018), S. 51; sowie das folgende Zitat

[276] Brüntrup (2018), S. 55

[277] Falkenburg (2012), S. 399

[278] Ebd.; Hervorhebungen durch die Autorin

[279] Falkenburg (2012), S. 401

> *„Alle kausalen Prozesse der Physik, Chemie und Biologie verlaufen partiell indeterministisch, soweit die Thermodynamik im Spiel ist. Und da dies auch für das neuronale Gesetz in Ihrem Kopf gilt, sind Sie nicht strikt determiniert."*

Damit löst sich bei ihr auch das Bieri-Trilemma „*in Luft auf*"[280]*:*

> *„All dies spricht dafür, die These von der kausalen Geschlossenheit der physischen Welt als irreführenden Restbestand der frühneuzeitlichen Metaphysik aufzugeben – und die beiden anderen Thesen beizubehalten, solange sie nicht empirisch widerlegt sind."*

Bei den beiden anderen Thesen handelt es sich erstens um die radikale Verschiedenheit von mentalen und physischen Phänomenen und zweitens um die These der mentalen Wirksamkeit. Die Frage nach dem „Wie" dieser Wirksamkeit bleibt bestehen.

[280] Falkenburg (2012), S. 411

4.3 Monistische Ansätze

Die Gegenposition zum Dualismus von Leib und Seele bildet der Monismus, der eine Identität von mentalen und physischen Phänomenen postuliert. Deswegen spricht man bei den monistischen Ansätzen auch von der Identitätstheorie:

> *„Die These der klassischen Identitätstheorie lautet, dass jede mentale Eigenschaft mit einer physischen Eigenschaft identisch ist.“*[281]

Schon Thomas Hobbes (1588 – 1679) hat die Vorstellung vertreten, dass „*geistige Zustände Gehirnzustände sind“*[282].

Metzinger sieht vier Stärken der Identitätstheorie[283]:

- *„Erstens ist sie empirisch plausibel: Ihre zentrale Prämisse des Vorliegens einer engen und durchgängigen Korrelation zwischen Gehirn- und Bewusstseinsvorgängen ist mittlerweile eine bestens dokumentierte wissenschaftliche Tatsache, an der vernünftigerweise nicht mehr gezweifelt werden kann.*
- *Zweitens zeichnet sich die Identitätsthese als Lösungsvorschlag für das philosophische Leib-Seele-Problem durch maximale ontologische Sparsamkeit aus. Sie ist parsimonisch*[284]*, denn sie hat empirisch exakt denselben Gehalt wie die Korrelationsthese (die auch die meisten Dualisten akzeptieren), ist aber deutlich einfacher als diese.* [...]
- [Drittens besteht ihre Stärke darin], *dass sie das Leib-Seele-Problem nicht löst, sondern auflöst: Die Identitätstheorie beantwortet das Problem nicht, indem sie etwa auf eine dubiose Form psychophysischer Kausalität verweist, sondern sie rettet die kausale Wirklichkeit mentaler Zustände einfach dadurch, das sie sie in den Bereich des Physischen eingliedert und die Vermutung nahelegt, dass es sich bei der Verursachung von Handlungen durch geistige Zustände im Grunde um längst bekannte und wahrscheinlich neurobiologisch zu erklärende Formen der Kausalität handelt.*
- [Viertens weist sie] *unter wissenschaftstheoretischen Aspekten eine große Kohärenz mit bereits existierenden und gut bewährten naturwissenschaftlichen Theorien auf. Sie kann an Modelle der innertheorethischen Reduktion angepasst werden und verträgt sich deshalb gut mit Schichtenmodellen der Wirklichkeit.“*

---

[281] Metzinger (2007), S. 92

[282] Ravencroft (2008), S. 76

[283] Metzinger (2007), S. 107f; Hervorhebungen durch DS

[284] Ergänzung von Metzinger an gleicher Stelle: „*Das Prinzip der **Parsimonität** besagt: Eine Theorie, die zur Erklärung ihres Zielgegenstandes weniger ontologische Entitäten oder strukturelle Annahmen benötigt, ist einer zweiten und auf denselben Gegenstandsbereich bezogenen Theorie dann vorzuziehen, wenn sie ansonsten gegenüber dem konkurrierenden Modell gegenüber keinen Nachteil hat.*“; Hervorhebung durch DS

Die Darstellung der monistischen Ansätze in dieser Arbeit orientiert sich sehr stark an der Struktur von Godehard Brüntrup. Er benennt jenseits des bereits diskutierten kartesischen Dualismus drei mögliche monistische Hauptpositionen zum Leib-Seele-Problem[285]:

1. *„Es gibt mentale Entitäten. Sie gehören nicht einem vom Bereich physischer Entitäten unabhängigen Bereich an. Sie sind abhängig von ihnen zugrundeliegenden physischen Entitäten, ohne jedoch vollständig auf diese reduziert zu sein.*
2. *Es gibt mentale Entitäten. Sie gehören nicht einem vom Bereich physischer Entitäten unabhängigen Bereich an. Sie sind abhängig von ihnen zugrundeliegenden physischen Entitäten und können vollständig auf diese reduziert werden.* [Unterschied zum vorherigen hervorgehoben]
3. *Es gibt keine mentalen Entitäten.“*

Diese drei Positionen sollen in den folgenden Abschnitten vorgestellt und diskutiert werden.

4.3.1 Nichtreduktiver Physikalismus – Mentale Eigenschaften in der physischen Welt

Diese Hauptposition geht explizit von der Existenz mentaler Eigenschaften auf einer oberen Makroebene in einem Schichtenmodell aus. Auf der unteren detaillierteren Mikroebene existieren nur physische Eigenschaften. Die mentalen Erscheinungen auf der Makroebene werden durch physische Eigenschaften auf der unteren Mikroebene determiniert, jedoch können sie umgekehrt nicht eindeutig auf sie zurückgeführt werden. Mentale Eigenschaften können trotz der kausalen Geschlossenheit des physischen Bereichs auf diesen Einfluss nehmen.

Für den nichtreduktiven Physikalismus wurden zwei Theorien entwickelt: Zum einen die Emergenztheorie, nach der mentale Phänomene „neu entstehen“ (englisch „emerge“) und zum anderen die Supervenienztheorie („supervene“ = „noch dazu kommen“) nach Donald Davidson.

4.3.1.1 Emergenztheorie

Nach Beckermann vertraten die britischen Emergentisten[286]

> „*die These, dass in der Evolution zwar nur natürliche Faktoren – also keine übernatürlichen Kräfte oder Entitäten – wirksam sind, dass aber trotzdem immer wieder genuin Neues entsteht – z.B. Eigenschaften von komplexen Gegenständen, die sich nicht auf die*

[285] Brüntrup (2018), S. 23; Brüntrup hat vier Hauptpositionen definiert, von denen die erste den klassischen Dualismus abdeckt: „*Es gibt mentale Entitäten. Sie gehören einem vom Bereich physischer Entitäten unabhängigen Bereich an.*“

[286] Conwy Lloyd Morgan (1852 – 1936), Samuel Alexander (1859 – 1938) und Charles Dunbar Broad (1852 – 1936)

> *Eigenschaften der Teile dieser Gegenstände zurückführen lassen. Menschliches Bewusstsein ist in den Augen der britischen Emergentisten ein emergentes Phänomen*."[287]

Nach Brüntrup sind emergente Eigenschaften „*nichtreduzierbar*, *unvorhersagbar*" und „*neuartig*"[288]:

> *„Es gibt keine allgemeine Theorie, aus der sich das Entstehen der neuartigen Phänomene hätte vorhersagen lassen (ante factum). Erst nachdem sie einmal aufgetreten sind (post factum), lassen sich die Bedingungen ihres Entstehens angeben."*

Er fasst zusammen:

> *„Die psychophysische Emergenztheorie besagt, dass mentale Eigenschaften emergente Eigenschaften sind. Sie sind real, mikrodeterminiert durch die physische Ebene, irreduzibel, unvorhersagbar, neuartig und kausal wirksam."*[289]

In diesem Zitat kombiniert er ein ontologisches (z.B. mit der Aussage zur Nichtreduzierbarkeit und Neuartigkeit) und ein erkenntnistheoretisches (z.B. mit der Aussage zur Unvorhersagbarkeit oder zur kausalen Wirksamkeit) Verständnis der Emergenz. Dazu etwas detaillierter bei Thomas Metzinger, der die Emergenztheorie nicht vollumfänglich dem Monismus zurechnet, sondern eher einem Eigenschaftsdualismus:

> „*Im Kontext des Leib-Seele Problems sind jedoch drei generelle Kernthesen mit dem Begriff der ‚Emergenz' verknüpft. Die erste These ist der ontologische Physikalismus: Die Gesamtheit der konkreten Realität erschöpft sich in den von der Physik postulierten Elementarteilchen und in Aggregaten dieser Elementarteilchen. Die zweite These ist die Emergenz von Makroeigenschaften: Ab einer gewissen Ebene struktureller Komplexität entstehen aus Mengen von Mikroeigenschaften genuin neue, emergente Makro-Eigenschaften. Drittens sind diese Makroeigenschaften aber irreduzibel, denn sie sind real und kausal wirksam: Die Eigenschaften des Ganzen wirken wieder auf Eigenschaften der Teile zurück (‚abwärtsgerichtete Kausalität'. Makro-Eigenschaften sind nicht auf Mikro-Eigenschaften reduzierbar, insbesondere sind sie nicht aus der Kenntnis solcher Eigenschaften heraus zu prognostizieren. Trotzdem existieren nomologische Korrelationen, die den Geist mit dem Körper verbinden*."[290]

Das Problem der oben bereits geschilderte Unvereinbarkeit der kausalen Geschlossenheit des Physischen mit der kausalen Wirksamkeit der mentalen Eigenschaften über die „Abwärts-Verursachung" kann mit der plausiblen Argumentation von Brigitte Falkenburg (siehe Abschnitt 4.2.2.1) gelöst werden. Dennoch verbleiben Fragen nach dem „Wie" der Abwärts-Verursachung.

---

[287] Beckermann (2008), S. 116

[288] Brüntrup (2018), S. 74 & 75

[289] Brüntrup (2018), S. 76

[290] Metzinger (2007), S. 275

4.3.1.2 Supervenienztheorie

Der Begriff „Supervenienz“[291] wurde ursprünglich synonym mit „Emergenz“ verwendet. Seit der zweiten Hälfte des letzten Jahrhunderts hat er eine eigenständige Positionierung, die leicht von dem der „Emergenz“ abweicht.

Thomas Metzinger erklärt die Theorie über die asymmetrische Relation zweier Eigenschaftsmengen:

> *„Die Supervenienz-Theorie formuliert ihren Lösungsvorschlag für das Leib-Seele-Problem als einen neuen Typ von psychophysischer Relation. Diese Relation besteht zwischen Familien beziehungsweise Mengen von Eigenschaften. Es gibt eine Eigenschaftsmenge A (die sogenannte Basismenge, in unserem Falle typischerweise gebildet durch bestimmte funktionale Eigenschaften des menschlichen Gehirns) und eine Eigenschaftsmenge B (die Supervenienzmenge, in unserem Falle die mentalen Eigenschaften eines Organismus), wobei die These dann in der Behauptung besteht, dass B supervenient auf A ist. Das bedeutet, dass zwei Elemente x und y des Gegenstandsbereichs, die alle Eigenschaften in der Basisfamilie A teilen, notwendigerweise auch alle Eigenschaften in der Supervenienz-Familie B teilen. Wenn x und y in Bezug auf A ununterscheidbar sind, müssen sie es notwendigerweise auch in Bezug auf B sein.“*[292]

„*Umgekehrt gilt dies hingegen nicht“*. Es kann also zwei unterschiedliche Elemente a und b der Supervenienz Familie B geben, mit identischen Eigenschaften auf der Ebene, die jeweils unterschiedliche Eigenschaften in der Basisfamilie A besitzen. Der gleiche mentale Zustand kann mit unterschiedlichen physischen Zuständen korrespondieren. „*Eine Reduktion des Mentalen auf das Physische ist damit nicht möglich*“. Es gibt also eine „*asymmetrische Abhängigkeit des Mentalen vom Physischen“*.

Die Vorteile dieser Theorie liegen einerseits darin, dass sie es ermöglicht, „*dualistische Intuitionen*“ zu wahren, und gleichzeitig „*dem Primat des Physikalischen Rechnung trägt“*. Trotzdem weist sie auch einige Schwächen auf. Erstens bietet sie keine Lösung für die „*abwärtsgerichtete mentale Verursachung“*. Zweitens erfasst sie nicht die „*Individuengebundenheit des Mentalen*“, also die Subjektivität und die Erste-Person-Perspektive. Vor allem erklärt sie nicht das „Wie“ der „*Verbindung zwischen den beiden Eigenschaftsmengen*“.

4.3.1.3 Das Problem der abwärtsgerichteten Verursachung

Aufbauend auf der Diskussion der kausalen Geschlossenheit des Physischen aus Abschnitt 4.2.2.1 soll hier das Problem der abwärtsgerichteten Verursachung bei monistischen Theorien vertieft werden.

---

[291] Zur Definition des Begriffes Supervenienz sei auf den entsprechenden Eintrag im Glossar verwiesen.

[292] Dieses und die folgenden Zitate: Metzinger (2007), S. 245 - 247

Jaegwon Kim hat dazu das sogenannte „*Exklusionsargument*“[293] formuliert. Es lautet wie folgt:

- Eine emergente Eigenschaft M verursacht eine andere emergente Eigenschaft M*. Diese kann man als Verursachung auf derselben Ebene bezeichnen, nämlich der mentalen Ebene.
- Nun muss M* physikalische Basiseigenschaften besitzen, aus denen sie emergiert; diese nennen wir P*.
- Umgekehrt gilt: P* selbst garantiert, dass M* emergiert. Dies geschieht unabhängig von dem, was M* auf der mentalen Ebene vorausgegangen ist, also auch wenn vorher kein M existiert hätte.
- Nun existiert aber M, und M hat M* mit der dazugehörigen Basiseigenschaft P* verursacht, also hat M P* verursacht.
- M weist aber auch eine dazugehörige Basiseigenschaft auf, die wir P nennen.
- P emergiert M und M verursacht M* und M* instantiiert P*; entlang dieser Kette verursacht P also P*.
- In der physischen Welt kann P aber auch direkt P* verursachen. Die obige Kette ist dafür nicht erforderlich.
- Fazit: *Die Abwärtsverursachung überdeterminiert die geschlossene physische Welt.*

Kim schreibt dazu:

> *„Emergente Entitäten sollen einen charakteristischen, unverwechselbaren und neuartigen kausalen Beitrag leisten. Wenn es jedoch in allen Fällen abwärtsgerichteter Verursachung eine systematische kausale Überdetermination gibt, dann können emergente Eigenschaften ihr kausales Versprechen nicht einlösen. Alles, was sie in kausaler Hinsicht beitragen können, kann und wird auch durch eine physikalische Ursache geleistet. Dieses Ergebnis* [das in gleicher Weise auch für die Supervenienz gilt, Anmerkung DS] *droht, wenn es nicht erfolgreich widerlegt wird, eine der zentralen Thesen des Emergentismus zu ruinieren. Wenn die abwärtsgerichtete Verursachung wegfällt, dann fällt auch der Emergentismus weg.“*[294]

Kim argumentiert zwar aus einer monistischen Position, fällt allerdings in dualistische Denkmuster zurück. Wie wir später sehen werden: Weder John Searle mit seinem biologischen Naturalismus noch Mario Bunge mit seinem emergentistischen psychoneuralen Monismus problematisieren eine mögliche Überdeterminierung der physischen Welt.

[293] Vgl. Kim (2007), S. 314 – 318; Struktur und Logik des folgenden Arguments wurden von dort übernommen

[294] Kim (2007), S. 316

4.3.2 Reduktiver Physikalismus – Zurückführung des Mentalen auf das Physische

Mit Hilfe der Reduktion, also der Reduzierung oder Rückführung des Mentalen auf das Physische garantiert man die kausale Relevanz des Mentalen. Mentale Ereignisse sind physische Ereignisse. Der Dualismus existiert nicht.

Die zentrale Theorie des reduktiven Physikalismus ist der Funktionalismus, in seiner extremen Ausprägung der Computer-Funktionalismus. Geprägt wurde der Begriff des Funktionalismus vom amerikanischen Philosophen Hilary Putnam. Weiterhin wurde er sehr stark unterstützt von Jerry Fodor, Daniel Dennett und David Lewis. Physische Zustände werden mit funktionalen Zuständen identifiziert; dies sind wiederum Systemzustände, die eine „*kausale Rolle"* spielen[295]. Man muss sich das System als eine Einheit vorstellen, die „Inputs" und „Outputs" hat und von einem aktuellen Zustand in einen neuen Zustand übergeht. Bei den Inputs kann es sich um Verletzungen, Verbrennungen, Stöße oder Schnitte handeln. Bei den Outputs kann es sich um Reaktionen handeln, wie „*Schreie, Stöhnen, Jammern, Erbleichen beim Anblick der Wunde*"[296] und den Wunsch nach Maßnahmen zur Reduzierung der Schmerzen handeln. Sämtliche Funktionen, die den gleichen Output bei gegebenem Input herbeiführen sind beliebig austauschbar oder multipel realisierbar[297]. Damit geht in diesem Modell die Einzigartigkeit des menschlichen Gehirns bzw. der Biologie verloren. Das Ganze wird zum „Computerfunktionalismus" mit der Vorstellung, dass das Gehirn wie eine Computerhardware zu verstehen ist, und die funktionalen Zustände (die mentale Zustände sind) der Software entsprechen.

> „*Mittels eines so verstandenen Computerfunktionalismus wird der psychophysische Zusammenhang des menschlichen Verhaltens nicht einfach auf ein Ineinandergreifen von physikalisch-chemischen Zuständen, wie es die Identitätstheorie unterstellt, zurückgeführt, sondern als ein Zusammenspiel einer materiellen Struktur – nämlich des Gehirns – mit einer funktionalen Ordnung begriffen*."[298]

Der gedankliche Schritt ist nicht weit entfernt von der Annahme, dass man ähnlich wie der Portierung von Software von einem Computer zum nächsten und auf unterschiedliche Betriebssysteme auch den menschlichen Geist multipel realisieren kann. Max Tegmark

[295] Vgl. Decher (2015), S. 245

[296] Decher (2015), S. 245f

[297] Vgl. https://plato.stanford.edu/entries/multiple-realizability/: „*In the philosophy of mind, the multiple realizability thesis contends that a single mental kind (property, state, event) can be realized by many distinct physical kinds.*"

[298] Decher (2015), S. 248

spricht von „*Substratunabhängigkeit*“[299] und denkt über die Möglichkeiten eines „Uploads“[300] nach, also des Hochladens des menschlichen Geistes auf eine Computerplattform.

Der Bonner Philosoph Markus Gabriel warnt vor dem Funktionalismus als einer „*Form von Religion*, [die aus seiner Sicht] *durch keine empirischen Belege jemals bewiesen noch widerlegt werden kann*“[301]. Gabriel ist mit dieser ablehnenden Haltung nicht allein; der prominenteste Kritiker ist der amerikanische Philosoph John Searle, dessen Position im Abschnitt 4.4 dargestellt wird.

Aus Sicht des Autors dieser Arbeit überzeugt das Argument der nicht erkennbaren Reduktion mentaler Phänomene (etwa Subjektivität, Intentionalität und qualitatives Erleben) auf die physische Ebene. Hierzu sei auf die in der folgenden Tabelle zusammengefassten Gedankenmodelle von Thomas Nagel verwiesen („*What is it like to be a bat?“),* John Searle („*Chinese Room*“) und Frank Cameron Jackson („*Mary’s room*“).

[299] Tegmark (2017), S. 102: „*Die Tatsache, dass genau dieselbe Berechnung auf jedem beliebigen universellen Computer durchgeführt werden kann, heißt, dass Rechnen substratunabhängig ist, und zwar auf dieselbe Weise, wie es die Informationen sind: Es kann unabhängig von seinem Substrat sein Eigenleben annehmen.* „

[300] Tegmark (2017), S. 249f

[301] Gabriel (2018), S. 121

Tabelle 4: Gedankenmodelle zur Widerlegung des reduktiven Physikalismus

Gedankenmodell	***„Wie ist es, eine Fledermaus zu sein?“***[302]	***„Das chinesische Zimmer“***[303]	**„Mary's Zimmer“**[304]
Autor	**Thomas Nagel**	**John Searle**	**Frank Cameron Jackson**
Eckpunkte	Am Beispiel der Fledermaus erläutert Thomas Nagel, dass nur Fledermäuse wissen, wie es ist, Fledermaus zu sein. Subjektive Tatsachen kann nur erfassen, wer selbst die Erfahrungsperspektive einnehmen kann.	John Searle entwickelte das Gedankenexperiment mit einem menschlichen Individuum ohne Kenntnis der chinesischen Sprache in einem geschlossenen Zimmer, dem man Karten mit Fragen auf chinesisch durch den Türspalt zuschiebt und die er mit Handbüchern, die Regeln beinhalten, beantwortet. Der Mensch würde wie eine Turing Maschine arbeiten und den Turing Test bestehen, ohne die Sprache imf insbesondere deren Semantikzu beherrschen.	In dem Gedankenexperiment wird eine Wissenschaftlerin, die in ihrem ganzen Leben noch nie andere Farben als Schwarz, Weiß oder Grau gesehen hat und alles, was man über Physik und Biologie lernen kann, verinnerlicht hat, erstmalig mit der Farbe Rot konfrontiert. Obwohl sie über das gesamte physikalische Wissen verfügt, lernt sie mit dieser Erfahrung etwas dazu.
Schlussfolgerung	Subjektive mentale Zustände lassen sich nicht auf objektive physikalische Zustände reduzieren. Wir werden nie wissen, wie es ist, ein radikal anderes Wesen zu sein.	Searle widerlegt mit diesem Gedankenexperiment den Computer-Funktionalismus, weil er schlussfolgert, dass bestenfalls die Syntax beherrscht, aber niemals die Semantik.	Jackson leitet daraus das sogenannte Wissensargument ab: der Physikalismus ist widerlegt, weil physikalische Erklärungen und Beschreibungen mentaler Zustände unvollständig seien und immer unvololständig beliben würden.

[302] Nagel (1974)

[303] Searle (1980), S. 417ff

[304] Jackson (1982), S. 127ff

4.3.3 Eleminativer Physikalismus – Zweifel an der Realität des Mentalen

Die dritte Hauptposition des Monismus, der eleminative Physikalismus oder auch eleminative Materialismus, beschreibt am besten das folgende Zitat eines ihrer prominenten Vertreter, Paul M. Churchland:

> *„Eleminativer Materialismus ist durch die These charakterisiert, dass unsere Alltagskonzeption psychologischer Phänomene eine radikal falsche Theorie ist; eine Theorie, die so fundamentale Defekte aufweist, dass sowohl ihre Prinzipien als auch ihre Ontologie irgendwann schließlich durch eine entwickelte Neurowissenschaft ersetzt werden, statt problemlos auf sie reduziert zu werden. Unser gegenseitiges Verstehen und selbst unsere Introspektion könnten dann innerhalb des begrifflichen Rahmens einer voll entwickelten Neurowissenschaft rekonstituiert werden; einer Theorie, von der wir erwarten dürfen, dass sie bei weitem stärker ist als die Alltagspsychologie, die sie verdrängt, und zudem besser in den allgemeinen Kanon der physikalischen Wissenschaften integriert.“*[305]

Die von der „*Alltagskonzeption psychologischer Phänomene*“ oder auch kürzer „Alltagspsychologie (abgekürzt ATP)“[306] postulierten „*Entitäten*“ (Meinungen, Wünschen, Vorstellungen und Gedanken) existieren den Eleminativisten zufolge gar nicht. Es handele sich um eine falsche Theorie, um einen wissenschaftstheoretischen Irrtum oder eine „*überholte empirische Theorie“*[307].

Brüntrup fasst drei Gründe der Eleminativisten „*für die Überflüssigkeit und Inadäquatheit der ATP* zusammen“[308]:

- Erstens hat die ATP in vielen „*zentralen Gebieten keine oder nicht befriedigende Erklärungen“* hervorgebracht. Beispiele: Theorie des Schlafes, psychische Erkrankungen und Umgang mit Verletzungen des Gehirns. „*Die ATP ist eine lückenhafte Theorie von geringer Erklärungskraft“*.
- Zweitens ist die Begrifflichkeit der ATP komplett verschieden von derjenigen der Neurowissenschaften. Eine Reduktion im Sinne des reduktiven Physikalismus erscheint völlig ausgeschlossen. Die ATP lässt sich „*nicht in den wachsenden Korpus naturwissenschaftlichen Wissens integrieren“*.
- Der dritte Grund ist ein induktiver. Die ATP ist eine „*primitive und sehr alte Theorie“*, die ein ähnliches Schicksal erleiden wird, wie „*einstmals akzeptierte Theorien“* in anderen Disziplinen. Beispiele: „*Theorien über die Verbrennung, die Körperbewegung, den Aufbau des Universums, die Natur des Lebens wurden ersetzt, weil sich viele der in ihnen postulierten Entitäten (Phlogiston, Impetus, Äther) als nichtexistent erwiesen“*.

---

[305] Zitiert aus der Übersetzung von Churchlands Originaltext „*Eleminativer Materialismus und propositionale Einstellungen*“ aus Metzinger (2007), S. 189

[306] Brüntrup (2018), S. 128

[307] Brüntrup (2018), S. 130

[308] Brüntrup (2018), S. 132f; bezieht sich auf alle Zitate in der Auflistung der drei Gründe

Die radikale Hauptposition und deren Begründung haben heftige Diskussionen und eine Vielzahl von Gegenargumenten provoziert. Dem stärksten Gegenargument zufolge liegen keinerlei Hinweise darauf vor, dass die Neurowissenschaften mit ihren physischen Beschreibungen und Modellen auch nur annähernd Antworten auf die vielen Fragen zur Natur des Bewusstseins, zur Naturalisierung der Intentionalität und zu den qualitativen Phänomenen produzieren:

> *„Der eleminative Physikalismus verspricht die Lösung des Leib-Seele-Problems durch Fortschritt in den Naturwissenschaften. Aber es scheint nicht möglich zu sein, dieses Versprechen einzulösen*."[309]

Auch Metzinger sieht das wesentliche Problem darin, dass die Neurowissenschaften über keinerlei „*Nachfolgebegriffe"*[310] für die Terminologie der ATP verfügen, die den Anforderungen in Bezug auf empirische Plausibilität und erkenntnistheoretische Überzeugungskraft genügen und zusätzlich „*intuitiv einleuchtend"* sind. Trotzdem sieht er einen Wert in der Diskussion über den eleminativen Physikalismus an sich. Auch wenn die extreme Position von Churchland überzogen ist, erscheint es durchaus angemessen, den einen oder anderen mentalistischen Begriff zu überdenken und an den neurowissenschaftlichen Fortschritt anzupassen.

John Searle, der erklärte Gegner des reduktiven Physikalismus und insbesondere des Computer-Funktionalismus, aber auch des cartesischen Dualismus, hat sich ebenfalls sehr kritisch mit dem eleminativen Physikalismus auseinandergesetzt. In seiner Analyse kommt er zu dem Ergebnis, dass letzten Endes Reduktionisten und Eleminativisten ihre Positionen mit unterschiedlichem Vokabular erklären, aber das Gleiche meinen[311]:

> „*Die frühen Materialisten* [die reduktiven Physikalisten, Anmerkung DS] *wollten zeigen, dass mentale Zustände als solche nicht existierten, indem sie zeigten, dass sie sich in Typ-Typ-Reduktionen auf Entitäten der Neurobiologie reduzieren ließen. Die späteren eleminativen Materialisten wollten zeigen, dass die Entitäten der Alltagspsychologie überhaupt nicht existieren, weil sie sich in Typ-Typ-Reduktionen auf Entitäten der Neurobiologie reduzieren lassen. Keines der beiden Argumente taugt etwas, aber sie legen nahe, dass diese Leute mit aller Kraft zu zeigen versuchen, dass unsere gewöhnlichen Alltagsbegriffe des Mentalen sich auf nichts in der wirklichen Welt beziehen und dass sie bereit sind, für diese Konklusion jedes Argument, das ihnen in den Sinn kommt, vorzubringen."*[312]

Searle selbst entwickelte eine eigene Theorie zum Leib-Seele-Problem, die im Abschnitt 4.5 vorgestellt wird.

---

[309] Brüntrup (2018), S. 138

[310] Metzinger (2007), S. 185

[311] Vgl. Searle (2006), S. 91

[312] Searle (2006), S. 91

4.4 Kants Auflösung des Leib-Seele-Problems

> „Der logische Paralogismus besteht in der Falschheit eines Vernunftschlusses der Form nach, sein Inhalt mag übrigens sein, welcher er wolle. Ein transzendentaler Paralogismus aber hat einen transzendentalen Grund: der Form nach falsch zu schließen. Auf solche Weise wird ein dergleichen Fehlschluß in der Natur der Menschenvernunft seinen Grund haben, und eine unvermeidliche, obzwar nicht unauflösliche Illusion bei sich führen“
>
> Immanuel Kant, Kritik der reinen Vernunft (2. Auflage 1787), AA III, 262

Immanuel Kant war nicht der Auffassung, dass das Leib-Seele-Problem unlösbar sei, sondern eher, dass es inexistent sei[313]. Der in diesem Abschnitt zitierte Philosoph Michael Wolff hat sich ausführlich mit Kants Sicht auf das Thema auseinandergesetzt.

Kant hat im Disput zwischen den Verfechtern des cartesischen Dualismus und den Vertretern materialistischer oder nichtmaterialistischer Positionen des Monismus keine Partei bezogen. Dazu schreibt Wolff:

> *„Nach seiner Ansicht besteht das Falsche an diesen Fragen* […] *darin, dass sie auf einem unzulässigen Gebrauch nicht-empirischer Begriffe beruhen, die gemäß seiner systematisch vorgenommenen Begriffseinteilung als Kategorien zu bezeichnen sind.* […]
>
> *... für Kant* [sind] *Wörter wie „Ding“, „Ursache“, „Wirkung“ usw. Kategorien oder doch wenigstens Ausdrücke, die als gleichbedeutend mit Kategorien im terminologischen Sinne gebraucht werden können, so dass ihr Gebrauch in der Philosophie des Geistes zu falsch gestellten Fragen verleitet.“*[314]

Nach Hegel, so berichtet Wolff, „*hat Kant maßgeblich dazu beigetragen, das Leib-Seele-Problem zum Verschwinden zu bringen und Fragen wie die nach der Immaterialität der Seele durchschaubar zu machen als Fragen, die weder einfach mit Ja noch einfach mit Nein beantwortet werden können“*[315].

Wolff zitiert für sein Argument Passagen aus dem Paralogismus-Kapitel der „*Kritik der reinen Vernunft*“ von 1781 und 1787. In seiner „*Kritik an der Rationalen Psychologie der Cartesianischen und nach-Cartesianischen Metaphysik* […] *bezeichnet er das Leib-Seele-Problem als die Aufgabe, die „Gemeinschaft“ (das commercium) von Seele und Leib zu erklären“*[316]. Mit der Gemeinschaft meint er die „*wechselseitige ursächliche Abhängigkeit seelischer und körperlicher Vorgänge“.* Problematisch ist – wie wir und vor allem schon Descartes vorher schon festgestellt haben – die „*Ungleichartigkeit seelischer und körperlicher Vorgänge“*. Körperliche Vorgänge finden im Raum statt und sind auch dort eindeutig zuzuordnen. Alle körperlichen Wechselwirkungen finden im Raume statt. Seelische Vorgänge hingegen „*bestehen aus Veränderungen seelischer Zustände“*, wie

313 Vgl. Wolff (2013)

314 Wolff (2013), S. 50; dieses und folgende Zitate

315 Ebd.

316 Wolff (2013), S. 51; dieses und folgende Zitate

z.B. „*Empfindungen, Vorstellungen, Gedanken und Erinnerungen"*. Sie stehen in keinem räumlichen Zusammenhang mit den körperlichen Vorgängen und werden auch nicht räumlich bestimmt oder festgelegt. Auf der Zeitachse sind sie jedoch festgelegt.

Anders als seine Vordenker führt Kant die Ungleichartigkeit des Seelischen und des Körperlichen, die Descartes als „*res extensa*" und „*res cogitans*" voneinander abgrenzte, *auf die* ***Unterscheidbarkeit von innerem und äußerem Sinn*** *zurück*[317]*:*

> *„Der innere und der äußere Sinn machen nach Kant zusammengenommen eine der beiden Quellen empirischer Erkenntnis aus, nämlich das, was er Anschauungsvermögen nennt. Anschauungen sind unmittelbare Vorstellungen von etwas Einzelnem, und genau dadurch unterscheiden sie sich von Vorstellungen des Verstandes, der, als zweite Quelle empirischer Erkenntnis, nur dadurch einen Beitrag zur Erfahrung leisten kann, dass seine Vorstellungen allgemein sind und sich mittelbar auf Anschauungen beziehen. Kant unterscheidet zwischen innerer und äußerer Anschauung und dementsprechend zwischen innerem und äußerem Sinn."*[318]

Der innere Sinn beinhaltet Zustände und „*Zustandsveränderungen unseres Vorstellungsvermögens"*. Kant nennt dies „*Gemüth*". Der äußere Sinn kann sich auf Objekte beziehen. Nur er kann sich auf räumliche Verhältnisse und den Raum beziehen. Zwischen den Zuständen des Gemüts herrschen nur zeitliche Zusammenhänge.

> *„Also kann alles, was nur im inneren Sinn angeschaut wird, niemals räumlich, sondern nur zeitlich bestimmt sein. Andererseits kann alle, was nur im äußeren Sinn angeschaut wird, immer nur räumlich bestimmt sein"*[319]*.*

In Konsequenz ergibt sich die unterschiedliche Wahrnehmung seelischer und körperlicher Vorgänge und Zustände aus der unterschiedlichen Art, wie diese im inneren und äußeren Sinn gegeben sind. Daraus folgt nach Kant auch eine andere Fragestellung:

> *„Nun ist die Frage nicht mehr von der Gemeinschaft der Seele mit anderen bekannten und fremdartigen Substanzen ausser uns sondern blos von der Verknüpfung der Vorstellungen des inneren Sinnes mit den Modificationen unserer äußeren Sinnlichkeit , und diese unter einander nach beständigen Gesetzen verknüpft sein mögen, so daß sie in einer Erfahrung zusammenhängen."*[320]

Auch die Frage nach der Immaterialität der Seele weist er entschieden zurück. Die *Seele als „Ding an sich"*[321] kann *„kein Gegenstand von Anschauung sein"*. Deswegen kann sie auch nicht gleichwertig sein mit etwas, was nur „*Gegenstand von Anschauung*" ist:

> *„Wäre Materie ein Ding an sich selbst, so würde sie als ein zusammengesetztes Wesen von der Seele als einem einfachen sich ganz und gar unterscheiden. Nun ist sie aber blos äußere Erscheinung, deren Substratum durch gar keine anzugebende Prädicate erkannt wird;*

317 Vgl. Wolff (2013), S. 52

318 Wolff (2013), S. 52; sowie die folgenden Zitate

319 Wolff (2013), S. 53

320 Kant (1781), Kritik der reinen Vernunft, AA IV 241; Kant (1787b), S. 386 (A 385, 386)

321 Wolff (2013), S. 65

> *mithin kann ich von diesem wohl annehmen, daß es an sich einfach sei, ob es zwar in der Art, wie es unsere Sinne afficirt, in uns die Anschauung des Ausgedehnten und mithin Zusammengesetzten hervorbringt, und daß also der Substanz, der in Ansehung unseres äußeren Sinnes Ausdehnung zukommt, an sich selbst Gedanken beiwohnen, die durch ihren eigenen inneren Sinn mit Bewußtsein vorgestellt werden können. Auf solche Weise würde eben dasselbe, was in einer Beziehung körperlich heißt, in einer andern zugleich ein denkend Wesen sein, dessen Gedanken wir zwar nicht, aber doch die Zeichen derselben in der Erscheinung anschauen können.* ***Dadurch würde der Ausdruck wegfallen, daß nur Seelen*** *(als besondere Arten von Substanzen)* ***denken; es würde vielmehr wie gewöhnlich heißen, daß Menschen denken****, d. i. eben dasselbe, was als äußere Erscheinung ausgedehnt ist, innerlich (an sich selbst) ein Subject sei, was nicht zusammengesetzt, sondern einfach ist und denkt."*[322] [Hervorhebung DS]

Die (von den Dualisten, insbesondere Descartes) angenommene Gleichartigkeit der Substanzen Seele und Materie bezeichnet Kant als „*unschicklich*":

> *„Aber ohne dergleichen Hypothesen zu erlauben, kann man allgemein bemerken, daß, wenn ich unter Seele ein denkend Wesen an sich selbst verstehe, die* ***Frage an sich schon unschicklich sei: ob sie nämlich mit der Materie*** *(die gar kein Ding an sich selbst, sondern nur eine Art Vorstellungen in uns ist)* ***von gleicher Art sei****, oder nicht; denn das versteht sich schon von selbst, daß ein Ding an sich selbst von anderer Natur sei, als die Bestimmungen, die blos seinen Zustand ausmachen."*[323] [Hervorhebung DS]

Die Konsequenz ist eindeutig:

> *„Für Kant ist damit das Leib-Seele-Problem kein unlösbares metaphysisches Problem, wie manche meinen. Vielmehr existiert es für ihn nicht."*[324]

Er bleibt aber eine Antwort auf die Frage schuldig, wie der innere Sinn zustande kommt. Es kann jedoch davon ausgegangen werden, dass er jegliche Vorstellung einer Naturalisierung des inneren Sinnes durch technische Artefakte ablehnen würde.

[322] Kant (1781), Kritik der reinen Vernunft, AA IV 226; Kant (1787b), S. 369 (Werkausgabe Bd. IV)

[323] Ebd.

[324] Wolff (2013), S. 66-67

4.5 Der biologische Naturalismus von John Searle

In seinem Buch „*Geist. Eine Einführung*“ entwickelte John Searle eine naturalistische Lösung des traditionellen Körper-Geist-Problems, „*die den biologischen Charakter mentaler Zustände betont und sowohl den Materialismus als auch den Dualismus vermeidet“*[325]. Für ihn „*entsteht das Leib-Seele-Problem aus der vermeintlichen Unvereinbarkeit eines naiven Mentalismus, für den mentale Phänomene wirklich und irreduzibel sind, mit einem naiven Physikalismus*“[326]. Aus Searles Sicht sind beide Perspektiven wahr, „*denn mentale Phänomene seien 1. durch zentralnervöse Vorgänge verursacht und 2. in diesen realisiert*“.

Er beschreibt seinen „*biologischen Naturalismus*“ entlang von vier Thesen[327]:

- In der ersten These stellt er noch einmal heraus, dass „*Bewusstseinszustände mit ihrer subjektiven, Erste-Person-Ontologie* […] *wirkliche Phänomene in der wirklichen Welt“* seien, die sich einerseits weder als *Täuschung* oder wissenschaftliche Irrung *eleminativ reduzieren* lassen und andererseits auch nicht auf eine *neurobiologische Basis*. Mit einer derartigen ontologischen Reduktion würde die Erste-Person-Perspektive verloren gehen und in eine Dritte-Person-Perspektive überführt werden.
- Die zweite These besagt zweierlei: Einerseits werden die Bewusstseinszustände „*vollständig von neurobiologischen Gehirnprozessen der niedrigeren Ebene verursacht“* und andererseits sind sie damit ebenfalls vollständig „*kausal reduzierbar auf neurobiologische Prozesse“*. Die Betonung liegt auf der kausalen Reduktion, im Gegensatz zur ontologischen Reduktion.
- In seiner dritten These behauptet Searle, dass die Bewusstseinszustände „*im Gehirn als Eigenschaften des Gehirnsystems realisiert“* seien und dort „*auf einer höheren Ebene als der Ebene der Neuronen und Synapsen*“ existieren würden. Dazu schreibt er: „*Einzelne Neuronen haben kein Bewusstsein, aber Teile des aus Neuronen zusammengesetzten Gehirnsystems haben Bewusstsein“*. Er illustriert dies mit den Systemeigenschaften von Wasser als „flüssig“ und „nass“. Wasser besteht aus H_2O-Molekülen, die einzeln nicht flüssig oder nass sind.
- Seine vierte These lautet: „*Weil Bewusstseinszustände wirkliche Merkmale der wirklichen Welt sind, haben sie kausale Funktionen*.“ Der bewusste Durst bewirkt, dass man Wasser trinkt.

Das Innovative von Searles Ansatz gegenüber allen vorherigen und teilweise in diesem Kapitel beschriebenen Diskussionen liegt **erstens** darin, dass er die strikte Abgrenzung

[325] Searle (2006), S. 123f

[326] Schäfer (1994), S. 18; sowie das folgende Zitat

[327] Searle (2006), S. 123f

von mentalen Zuständen gegenüber physischen Zuständen der Dualisten nicht übernimmt und das Bewusstsein als eine biologische Eigenschaft des Systems Gehirn versteht.

Zweitens präzisiert er den Begriff der Reduzierung und stellt heraus, dass zwischen einer ontologischen und einer kausalen Reduktion zu unterscheiden ist:

> *„Bewusstsein lässt sich kausal reduzieren, aber nicht ontologisch, denn dann würde der eigentliche Sinn, überhaupt einen Begriff von Bewusstsein zu haben, verlorengehen.“*[328]

Bei einer ontologischen Reduktion würde die „Erste-Person-Ontologie“ in eine „Dritte-Person-Ontologie“ überführt, was weder empirisch noch theoretisch funktioniert.

Drittens erweitert er das Verständnis von Kausalbeziehungen dergestalt, dass es einerseits kausale Prozesse zwischen Mikrophänomenen geben kann und andererseits auch Makroeigenschaften von Mikrophänomenen beeinflusst werden können und umgekehrt:

> *„Searle scheint allgemein für jeden Gegenstand ein Kontinuum möglicher Beschreibungsebenen anzunehmen, innerhalb dessen zwischen den Phänomenen der jeweils verschiedenen Ebenen die Verursachungs- und Realisierungsbeziehung darin besteht, dass das Phänomen der höheren Beschreibungseben durch das Phänomen der niedrigeren verursacht und darin realisiert ist.“*[329]

Viertens postuliert er, einhergehend mit den drei vorherigen Punkten, auch keine strikte Identitätsforderung zwischen dem Bewusstsein oder mentalen Zuständen und den dazugehörigen neurobiologischen Prozessen.

Für Searle stellen sich dann die folgenden Fragen:

> *„Wie passen qualitative, subjektive und intentionale Phänomene in die physische Welt? Was genau sind die physischen Eigenschaften der Welt, in die sie hineinpassen müssen?“*[330]

Dies sind genau einige der Dualismen, die schon im Abschnitt 4.2.1 aufgeführt wurden. Im linken Teil der folgenden Abbildung werden die mentalen Eigenschaften und die physischen Eigenschaften gemäß einer modernen Auslegung von René Descartes gegenübergestellt.

In Searles biologischem Naturalismus werden die ersten drei Eigenschaften der mentalen Seite mit den letzten drei Eigenschaften auf der physischen Seite widerspruchsfrei kombiniert. Das bedeutet: Die Eigenschaften Subjektivität, Qualitativität und Intentionalität werden mit der räumlichen Lokalisierbarkeit, der physischen (oder neurobiologischen) Erklärbarkeit und der kausalen Wirksamkeit und Geschlossenheit kombiniert. Die letzten drei Eigenschaften auf der Liste des Mentalen (nicht räumlich lokalisierbar, nicht durch physische Prozesse erklärbar, kausale Unwirksamkeit) sind nach Searle „*einfach*

[328] Searle (2006), S. 130

[329] Schäfer (1994), S. 23

[330] Searle (2006), S. 127

falsch"[331]. Die ersten drei Eigenschaften auf der Liste des Physischen „*sind einfach keine notwendigen Bedingungen, um Teil des physischen Universums zu sein"*.

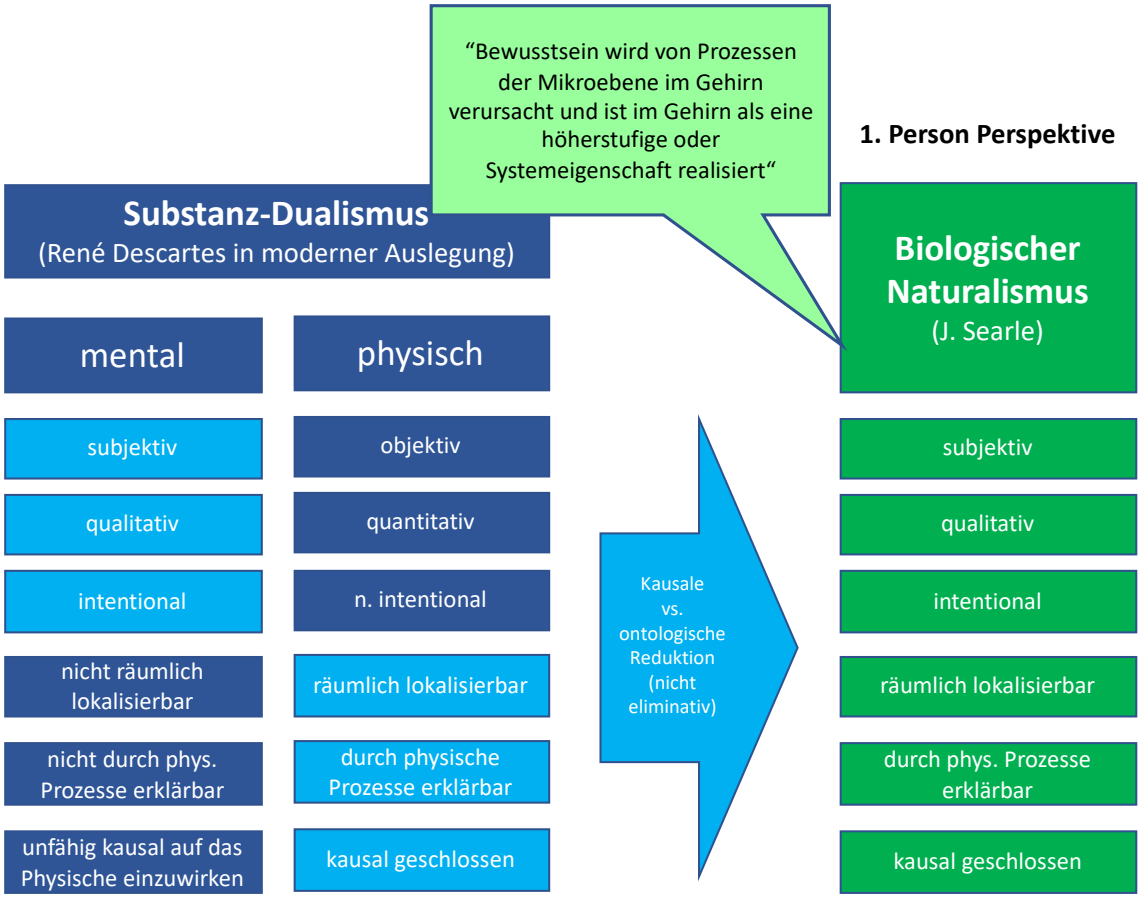

Abbildung 7: Substanzdualismus und biologischer Naturalismus[332]

Mit dieser Theorie wendet Searle sich gegen den Dualismus und auch gegen den Monismus in der Form des Physikalismus (reduktiv, nicht reduktiv bzw. eleminativ) oder Funktionalismus. Er vermeidet die unnötige Verdoppelung der Wirklichkeit, die zu den beschriebenen Schwierigkeiten führt, und die Probleme der Identitätstheorien. Für ihn gibt es nur eine Wirklichkeit, und das ist diejenige der Naturwissenschaften. Hier hat er Platz geschaffen für Geist, Bewusstsein, Intentionalität, qualitative Phänomene und auch die „Ich-Perspektive".

Ungeklärt bleibt die Frage nach den physischen Wirkmechanismen und der Verbindung der physischen Eigenschaften der unteren Ebene und den mentalen Eigenschaften der oberen Ebene.

[331] Searle (2006), S. 128

[332] Skizze auf Basis von Searle (2006), S. 127

4.6 Mario Bunge: Emergentistischer Materialismus und Konsequenzen für die künstliche Intelligenz

An dieser Stelle soll ein Philosoph zu Wort kommen, der nicht zwingend für seine Arbeiten zur Philosophie des Geistes bekannt war, sondern eher als Universalist galt und mit seinen Beiträgen ein deutlich breiteres Feld in der Philosophie abdeckte: Mario Bunge. In seinem Lebenswerk fand die Geistesphilosophie allerdings ihren Platz, vor allem in seinem Buch „*Das Leib-Seele-Problem, Ein psychobiologischer Versuch*“ von 1984 (Originalausgabe 1980) und „Matter and Mind“[333] aus dem Jahre 2010. Er verfolgte einen Ansatz des Realismus und verankerte seine Schlussfolgerungen dabei stets in einem umfassenden wissenschaftlichen Fundament („*strives to formulate a comprehensive outlook based on scientific principles*“, siehe Biografie). Insofern ist seine Sicht auf die Geistesphilosophie insgesamt und speziell in Bezug auf die Positionierung der Künstlichen Intelligenz von Interesse.

Mario Bunge hat knapp zwanzig Jahre vor Searle seinen „*emergentistischen psychoneuralen Monismus*“ als seine Antwort auf das Leib-Seele-Problem publiziert[334]. Er fasst diesen in drei Thesen zusammen:

> *„I. Alle psychischen Zustände, Vorgänge und Prozesse sind Zustände, Vorgänge und Prozesse in Gehirnen der höheren Wirbeltiere.*
>
> *II. Diese Zustände, Vorgänge und Prozesse sind gegenüber solchen der zellulären Komponenten des Gehirns als emergent zu betrachten.*
>
> *III. Die sogenannten psychophysischen (bzw. psychosomatischen) Beziehungen sind Wechselwirkungen zwischen unterschiedlichen Teilsystemen des Gehirns oder zwischen einigen von ihnen und anderen Teilen des Organismus.“*[335]

Nach Bunge ist das Mentale etwas, *„was gegenüber dem Physikalischen emergent ist, nur ist es nicht erlaubt, ersteres zu verdinglichen“*[336].

Im Kern entspricht diese Theorie derjenigen von John Searle, allerdings vermeidet Bunge die Thematisierung der kausalen Wirkungskraft der psychischen Zustände und der Subjekt-Objekt-Spalte. Es war ihm wichtig, die Psychologie – anders als viele zeitgenössische Geistesphilosophen – als Naturwissenschaft zu verstehen:

> „***psychische Zustände bilden eine Teilmenge*** *(wenn auch eine sehr wichtige)* ***der Gesamtheit der Gehirnzustände****, die ihrerseits wiederum eine Teilmenge des Zustandsraums des gesamten Lebewesens bilden.*“[337] Hervorhebung MB

[333] Bunge (2010)

[334] Bunge (1977); Bunge und Llinás (1978); Bunge (1980) bzw. (Bunge (1984)

[335] Bunge (1984), S. 32

[336] Bunge (1984), S. 271

[337] Bunge (1984), S. 33

Ähnlich wie Searle muss auch Bunge eingestehen, dass dies eine empirisch nicht bewiesene Hypothese ist, allerdings eine plausible.

Seine Begründung in sieben Punkten lautet wie folgt[338]:

1. Vermeidung einer „*geheimnisvollen ‚Geistsubstanz'*" und bessere Verträglichkeit mit „*wissenschaftlicher Methodik*"
2. „*frei von Verschwommenheit*" zwischen physischen und mentalen Phänomenen
3. Bessere Einbingung in das „*Schema der Allgemeinbegriffe Zustand und Prozesse*"; klare Abgrenzung vom Dualismus
4. Förderung des „*Zusammenwirken von Psychologie und den übrigen Wissenschaften*"
5. „*Einklang mit Entwicklungsphysiologie und Neurophysiologie*", d.h. kongruente Reifung von Gehirn und Verhalten
6. Einfügung in das „*Konzept der Evolutionsbiologie*"
7. Verständnis des Gehirns als „*vielschichtiges System*"

Bernulf Kanitscheider bestätigte dies recht überzeugend in seinem Vorwort zu Bunges Buch:

> „*Bunge prägte selbst den Namen ‚**emergentistischer Materialismus**' für seine Auffassung, eine Bezeichnung, die auf zweierlei hinweist. Erstens auf die Tatsache, dass man der Autonomie des Mentalen Rechnung tragen kann, ohne eine rätselhafte Substanz einzuführen, deren gesetzesartige Wechselwirkung mit der physikalischen, chemischen und biologischen Ebene stets im Dunkeln bleiben müsste, und zweitens auf eine schichtartige Verfassung der Realität, in der Seelisch-Geistiges als Aktivität des zentralen Nervensystems die komplexestes Ebene darstellt und sich als vorläufig letzter Schritt der Evolution in der Morphogenese erweist.*"[339] [Hervorhebung DS]

Bunge grenzt sich klar gegenüber dem psychophysikalischen Dualismus ab, allerdings auch gegenüber jedem „*törichten Reduktionismus*"[340]:

> [Der emergentistische Materialismus] „*wirkt* [...] *als Beschützer für die Freiheit und die Schöpferkraft des Menschen, der kein programmierbarer Apparat ist, aber auch keine nach Belieben abrichtbare Taube. Schließlich ist der Mensch das einzige von Haus aus kreative Geschöpf, das einzige, das nach langem Bemühen vielleicht am Ende eine ‚Wissenschaft des Geistes' zu schaffen vermag. Nur der Mensch ist imstande, sein eigenes Leben zu gestalten, sei es im Licht seiner Einsicht oder im Dunstkreis seiner Vorurteile, sei es zum Guten oder zum Bösen.*"[341]

Zusätzlich wird seine klare Ablehnung der beiden physikalistischen Theorien des Computerfunktionalismus (programmierbarer Apparat) und des Behaviorismus (abrichtbare

[338] Vgl. Bunge (1984), S. 35

[339] Kanitscheider in Bunge (1984), Seite XI

[340] Bunge (1984), S. 273

[341] Ebd.

Taube) deutlich[342]. Ebenso wie Searle hat auch Bunge keine Antwort auf die Frage nach dem „Wie“ der mentalen Verursachung, die dann auch die Grundlage für einen Nachbau des menschlichen Geistes sein könnte. Die Physikerin (!) Brigitte Falkenburg wendet sich zu dieser Frage vom Verständnis des Systems Gehirn-und-Geist als physisches System ab, zugunsten des Systems Körper-und-Geist als Person:

> *„Das Gehirn lässt sich natürlich als ein komplexes physisches System verstehen (das menschliche Gehirn gilt als das komplexeste System im ganzen Universum); doch das System Gehirn-und-Geist lässt sich nicht als ein komplexes physisches System verstehen, sondern höchstens das System Körper-und-Geist als intelligentes Lebewesen bzw. als Person.“*[343]

Die grundsätzliche Motivation, sich in dieser Arbeit ausführlich mit dem Leib-Seele-Problem zu beschäftigen, liegt darin, eine Antwort auf die Frage nach der Verbindung zwischen physischer und geistiger Welt zu finden. Dabei geht es um die Suche nach einer Brücke für zukünftige Arbeiten über die KI. Abstrakt ist die Aufgabe beschrieben mit der mehr oder weniger vorsichtig artikulierten Hypothese der Substratunabhängigkeit des menschlichen Bewusstseins und den daraus resultierenden Möglichkeiten eines „Uploads“ des menschlichen Geistes auf eine Halbleiterplattform.

Aus den für diese Arbeit untersuchten dualistischen und monistischen Theorien zum Leib-Seele-Problem lässt sich in der Tat kein Ansatz für einen Computerfunktionalismus ableiten, der alle Aspekte des menschlichen Bewusstseins abdeckt.

Das mit obiger Grundannahme verbundene Missverständnis bringt Mario Bunge in seinem folgenden Zitat sehr gut zum Ausdruck:

> “*... computationalism involves a confusion between science and technology, particularly artificial intelligence (AI). The ultimate goal of AI is not to explain cognition in terms of natural laws – the job of cognitive psychology – but to design efficient, fast, reliable, and inexpensive machines and programs likely to successfully mimic or replace certain cognitive processes. The resulting artifact cannot be a good guide to psychologists because they study animals, which are products of blind, opportunistic, and tortuous evolution, not intelligent design – whence such animals are likely to operate in complex, slow, unreliable, and expensive ways.”*[344]

Bunge kontrastiert die per Design erzeugten Computer mit den durch die Evolution entstandenen menschlichen (und tierischen) Gehirnen. Erstere können einige kognitive

---

[342] Anmerkung: Im Epilog seines Buches (Bunge (1984), S. 274f) gibt er Donald O. Hebb, dem im ersten Kapitel dieser Arbeit bereits zitierten Begründer der Hebbschen Regel, Gelegenheit, den Behaviorismus zu verteidigen und gleichzeitig den Dualismus zu bestätigen: „*Der dualistische interaktionalistische* [Ansatz zur Lösung des Leib-Seele-Problems, Ergänzung DS] *kann uns hier , wie Professor Bunge dartut, nicht in die richtige Richtung führen, er schließt das wesentliche Problem aus der Sphäre wissenschaftlicher Betrachtungsweise aus. Aber Befürworter wie Abstinenten einer Verwendung mathematischer Methoden können sich gegenseitig beim Bau des monistischen Denksystems unterstützen, und selbst der Black Box-Theoretiker* [also der Behaviorist, Anmerkung DS] *ist in der Lage, wertvolle Beiträge zu leisten.“*

[343] Falkenburg (2012), S. 413

[344] Bunge Ardila (1987), S. 110

Prozesse nachahmen oder gar ersetzen, eignen sich ansonsten aber nicht als Gegenstand der Psychologie.

In "*Matter and Mind*" setzt sich Mario Bunge auch ausführlich mit dem Computer-Funktionalismus kritisch auseinander:[345]

1. *Seine Kritik zielt zunächst auf die inflationäre Verwendung des Begriffs „computation", der für die unterschiedlichsten Prozesse verwendet wird, die nichts mit algorithmisch gesteuerten Turing-Maschinen zu tun haben.*
2. *Anders als Computer sind Menschen und Tiere nicht Produkte eines „Designs", sondern Ergebnisse der Evolution und Erfahrung. Damit verliert das immer wieder gerne verwendete Bild von Hardware und Software in Bezug auf das menschliche Gehirn und den menschlichen Geist seinen Sinn. Die menschliche „Software" ergibt sich aus Lernprozessen und ist Resultat einer Entwicklung in Verbindung mit der Evolution. Nur das verstehende Gehirn ist das Ergebnis der Evolution, nicht aber der Verstand. Die Rolle der biologischen Evolution wird bei der Computer-funktionalistischen Beschreibung des menschlichen Geistes völlig vernachlässigt.*
3. *Computerprogramme oder Algorithmen sind keine lebendigen Dinge oder Prozesse, sondern Artefakte. Wie Maschinen, werden sie designt und repariert, bestellt und verkauft, verloren und gestohlen. Daher gehorchen Algorithmen bestimmten Gesetzen und technischen Normen, von denen manche klug und angemessen sind und andere weniger. Somit befinden sich Algorithmen in einem kontinuierlichen Upgrade-Prozess.*
4. *Wiederholte Abläufe des menschlichen Verhaltens funktionieren wie Algorithmen und genügen der Definition von „Computation". Alle anderen mentalen Prozesse wie Lieben, Fürchten, Hassen, Raten, Erfinden und Kritisieren sind nicht-algorithmisch. Insbesondere gibt es keine bekannten Regeln für das Entstehen guter Ideen. Selbst bei der erstgenannten „Computation" bei Routineprozessen verstehen wir die neuralen Abläufe nicht.*
5. *Die Begriffe „Information" und „Computation" werden inflationär bei der Beschreibung der Abläufe im menschlichen Gehirn und Geist verwendet, oftmals ohne oder mit fehlerhafter Bezugnahme auf die korrespondierenden technischen Konzepte, wie z.B. der Kommunikationstheorie nach Shannon. Damit verharrt der Diskurs auf einem intuitiven oder metaphorischen Level.*
6. *Da der Computer-Funktionalismus sich ausschließlich an rationalen (und damit algorithmischen) Prozessen orientiert, unterbricht er die starken Verbindungen zwischen Kognition auf der einen Seite und Motivation, Emotion und Gemeinschaftlich-Sozialem auf der anderen Seite. Daher kann er weder Neugier noch*

[345] Vgl. Bunge (2010), S. 233 - 234, sinngemäße Übersetzung DS

Lernen noch die Tatsache erklären, dass soziale Rahmenbedingungen Lernen fördern oder unterdrücken können.

7. *Im Gegensatz zu Computern, die darauf programmiert sind zu gehorchen, verfügen Menschen über Fähigkeiten und Möglichkeiten zur Innovation sowie zum Ungehorsam und Betrug. Insbesondere können sie Regelwerke in Form von moralischen Prinzipien oder technischen und legalen Normen entwerfen, diskutieren, kritisieren und umsetzen oder verwerfen. Obwohl diese Regeln auf Basis von Wissen entwickelt wurden, haben sie keinen epistemischen Charakter. Motivation und Inkraftsetzung dieser Normen beruhen auf sozialen Emotionen wie Empathie, Sympathie, Mitleid, Scham, Stolz, Vertrauen und Misstrauen, was alles jenseits der Möglichkeiten von Maschinen liegt. In Kürze: Computer, anders als Menschen, haben kein Gefühl für Moral und sind auch zu einem moralischen Urteil nicht in der Lage.*
8. *Der Computer-Funktionalismus ignoriert die Tatsache, dass menschliche Gehirne, im Gegensatz zu Computern, sozial sensibilisiert sind und über Interaktion, Anpassung, Kooperation und Konflikt lernen.*
9. *Menschliche Gehirne verfügen über Intentionalität und Intensionalität, zwei Begriffe, die im weiteren Verlauf dieser Arbeit noch vertieft werden. Spezifisch geht es hier um das Selbst-Bewusstsein, die Intensionalität erster Ordnung, und das Nachdenken über das, was jemand über eine andere Person denkt, also die Intensionalität zweiter, dritter und n-ter Ordnung.* Man beachte den Unterschied zwischen Intension und Intention. Intension bezieht sich auf Bedeutung und Intention auf Absicht. Diese Begriffe werden bei Bunge durchaus etwas anders verstanden als bei Brentano und Husserl.
10. *Bunge schließt seine Kritik in zehn Punkten mit der Feststellung ab, dass der Computer-Funktionalismus gegen das erste Prinzip der Technikphilosophie verstoße: Damit meint er die These, dass hergestellte Dinge, anders als vorgefundene Dinge, Ideen verkörpern. Hergestellte Dinge repräsentieren eine andere ontologische Ebene als vorgefundene Dinge wie unser menschliches Gehirn. Daraus ergibt sich aus seiner Sicht ein Kategorienfehler.*

Auf Basis dieses „Rundumschlags“ gegen den Computer-Funktionalismus und die weit verbreitete Euphorie über die KI kommt Bunge zu einem unmissverständlichen Schlusswort, das hier im Original zitiert wird:[346]

> *“Computer worshippers believe that the future of psychology belongs to AI. This is like saying that the future of human autonomy and physiology belongs to robotics. Since the goal of AI and robotics is to imitate us in some regards, they can advance only insofar as they learn about humans from the sciences of man. In general, to imitate anything, start by learning about the genuine article.*

[346] Bunge (2010), S. 237

> *Much of the same holds for the current vogue of the "its from bits" recipe in popular physics: it is wrong to try and reduce the natural to the artificial, and in particular to attempt to base physics upon engineering, rather than the other way around, and this for two reasons. First, because machines and engineering are artificial items built by brains, which belong in both nature and culture. Second, because the basic sets of digital computer science are denumerable, whereas reality is continuous in most respects, which is why continuous functions and differential equations have been so successful in physics and engineering.*
>
> *In sum, computers are useful as long as they are regarded as aids to brains, not as substitutes for them."*

Bunge war für seinen Realismus bekannt und galt auch als Philosoph, der noch den Anspruch erhob, mit seinen Arbeiten das gesamte Spektrum der Philosophie abzudecken. Sein Urteil zur Künstlichen Intelligenz ist klar: Computer werden immer wertvolle Werkzeuge für den menschlichen Verstand oder das menschliche Gehirn sein, dieses aber niemals ersetzen.

5 Bewusstsein

> *„Das Problem des Bewusstseins bildet heute – vielleicht zusammen mit der Frage nach der Entstehung unseres Universums – die äußerste Grenze des menschlichen Strebens nach Erkenntnis.“*[347] Thomas Metzinger

Diese Beschreibung der Grenze unseres Erkenntnisvermögens findet sich bereits 1872 bei Emil du Bois-Reymond in seinem berühmten Vortrag „*Über die Grenzen des Naturerkennens*“ (Rechtschreibung und Wortwahl wie im Original, Anmerkung DS):

> „*Allein es tritt nunmehr. An irgend einem Punkte der Entwickelung des Lebens auf Erden, den wir nicht kennen und auf den es hier nicht ankommt, etwas Neues, bis dahin Unerhörtes auf, etwas wiederum, gleich dem Wesen von Materie und Kraft, Unbegreifliches. Der in negativ unendlicher Zeit angesponnene Faden des Verständnisses zerreißt, und unser Naturerkennen gelangt in eine Kluft, über die kein Steg, kein Fittig trägt: wir stehen an der Grenze unseres Witzes. Dies neue Unbegreifliche ist das Bewusstsein. Ich werde jetzt, wie ich glaube in sehr zwingender Weise, dartun, dass nicht allein bei dem heutigen Stand unserer Kenntnis des Bewusstsein aus seinen materiellen Bedingungen nicht erklärbar ist, was wohl jeder zugibt, sondern dass es auch der Natur der Dinge nach aus diesen Bedingungen nie erklärbar sein wird.“*[348]

Für lange Zeit war das Nachdenken über unseren Geist und unser Bewusstsein eher eine akademische Übung für Philosophen und Psychologen mit insgesamt begrenztem praktischem Nutzen. Mit der Entwicklung der KI hat sich dies drastisch verändert. Signifikante Entwicklungsschritte hin zu einer starken KI mit einer menschenähnlichen Intelligenz mit allen in den jüngeren Intelligenztheorien beschriebenen Merkmalen scheinen auch ein „Künstliches Bewusstsein“ zu erfordern. Weiterhin wird dies immer wieder gerne mit der Vorstellung kombiniert, dass ja auch der menschliche Geist determiniert sei. Dazu schreibt Bettina Walde:

> *„Wenn der menschliche Geist tatsächlich vollständig durch neuronale Aktivität determiniert ist oder zumindest auf ihrer Grundlage existiert, dann liegt es ja auch nahe anzunehmen, daß es irgendwann möglich sein wird, ihn auf künstlicher Basis nachzubilden. Auf diese Weise würde dann nicht nur die herausragende Stellung des Menschen unter anderen Lebewesen verloren gehen, sondern er bekäme sogar noch Konkurrenz durch ein ihm ebenbürtiges künstliches Wesen.“*[349]

In diesem Kapitel geht es zunächst (Abschnitt 5.1) um den Versuch einer begrifflichen Einordnung und Definition, dann um eine Konkretisierung des erkenntnistheoretischen Problems, zwei wichtige Vertiefungen zu den Teilfragestellungen der Intentionalität (Abschnitt 5.2) und des phänomenalen Bewusstseins (Abschnitt 5.3) und die Darstellung eines philosophischen Modells des Bewusstseins, das versucht, eine Brücke zu schlagen

[347] Metzinger (1995), S. 15

[348] Du Bois-Reymond (1872), S. 16f

[349] Walde (2002), S. 15

zwischen den empirischen Wissenschaften und der Philosophie (Abschnitt 5.4). In einem Abstecher in die empirische Neurowissenschaft soll der Stand der Forschung zur Entstehung des Bewusstseins dargestellt werden (Abschnitt 5.5 und Abschnitt 5.6) mit einem kurzen Exkurs zur Willensfreiheit und zum Determinismus (Abschnitt 5.5.2 und 5.5.3). Abschließend wird über die Möglichkeiten der Schaffung eines „künstlichen Bewusstseins" reflektiert (Abschnitt 5.7).

Ähnlich wie beim Begriff der „Intelligenz" fehlt es auch bei dem des „Bewusstseins" an einer klaren und allgemein anerkannten Definition. So schreibt David Chalmers:

> *"Conscious experience is at once the most familiar thing in the world and the most mysterious. There is nothing we know about more directly than consciousness, but it is far from clear how to reconcile it with everything else we know. Why does it exist? What does it do? How could it possibly arise from lumpy gray matter? We know consciousness far more intimately than we know the rest of the world, but we understand the rest of the world far better than we understand consciousness."*[350]

Für das Bewusstsein lassen sich keine notwendigen und hinreichenden Bedingungen finden. Angrenzende und überlappende Begriffe, wie z.B. Intentionalität, Gewahrsein oder Geist, werden nicht scharf genug voneinander abgegrenzt. Nicht zu verwechseln ist das Bewusstsein mit Selbstbewusstsein, Gewissen oder Verstand.

Es gibt eine immer wieder zitierte Definition von John Searle, der man leicht zustimmen kann, die aber auch nur bedingt weiterhilft:

> *„Was ich unter ‚Bewusstsein' verstehe, lässt sich am besten mit Hilfe von Beispielen veranschaulichen. Wenn ich aus einem traumlosen Schlaf erwache, dann gelange ich in einen Zustand des Bewusstseins – ein Zustand, der so lange anhält, wie ich wach bin. Wenn ich einschlafe oder eine Vollnarkose bekomme oder sterbe, dann hören meine Bewusstseinszustände auf. Wenn ich im Schlaf träume, dann bin ich in einem Bewusstseinszustand, obwohl Traumformen von Bewusstsein im Allgemeinen viel weniger intensiv und lebhaft sind als gewöhnliches Wach-Bewusstsein."*[351]

5.1 Konkrete Eigenschaften des bewussten Erlebens

Metzinger hat in einem umfassenden Werk die wichtigsten phänomenologischen Merkmale des Bewusstseins herausgearbeitet. Unser Bewusstsein ist bestimmt von unseren Sinnesempfindungen, Emotionen und Gedanken, die wie selbstverständlich und vor jeder Begriffsbildung einfach da sind und unser Erleben der Welt bestimmen. „*In der Philosophie des Geistes wird dieses pure Erleben als der phänomenologische Gehalt unserer*

[350] Chalmers (1996), S. 3

[351] Searle (1993-2), S. 102

mentalen Zustände bezeichnet."[352] Deswegen spricht man auch vom „phänomenalen Bewusstsein":

> *„Es geht darum, dass manche mentalen Zustände nicht nur einen Wissens- oder Informationsgehalt besitzen, sondern, dass sie sich auf eine bestimmte Weise anfühlen"*[353]

Thomas Nagel hat dieses „sich auf eine bestimmte Weise anfühlen" in seinem oben (Kapitel 4.3.2) bereits erwähnten Aufsatz „***Wie ist es, eine Fledermaus zu sein***"[354] sehr gut beschrieben, nämlich anhand des subjektiven Erlebens von Fledermäusen. Wir können uns unmöglich vorstellen, „*wie es ist, eine Fledermaus zu sein*". Wir können dies nicht aus unserem eigenen Bewusstsein extrapolieren. Die Extrapolation wird immer unvollständig sein und kann niemals sinnliche Eindrücke beinhalten, für die wir noch nicht einmal ein Organ besitzen.

Metzinger nähert sich dem Phänomen „Bewusstsein" über vier Merkmale[355]: Transparenz, Perspektivität, Präsenz und Einheit. Diese vier Merkmale werden in den folgenden Abschnitten erläutert.

A) Die Transparenz phänomenaler Zustände

Die Transparenz phänomenaler Zustände bedeutet, dass wir sie unmittelbar, nach unserm Eindruck unverfälscht und direkt erleben. Wenn wir die Augen öffnen, sehen wir Blumen im Garten, Autos auf der Straße, den Mond am Himmel oder den Freund vor uns und nicht etwa Signale auf der Netzhaut unseres Auges oder „*die Gehirnzustände, die an der Erzeugung dieser Perzepte beteiligt sind"*[356].

Die phänomenalen Zustände bringen uns „*in einen direkten Kontakt mit der Welt* und sind uns damit *unendlich nah"*[357]. Diese „*unendliche Nähe*" hat aber auch etwas von „*unendlich Fernem*", weil wir nicht sehen und wissen, wie die Daten- und Informationsstrukturen in unserm Gehirn die scheinbar unmittelbaren, unverfälschten und direkten Erlebniseindrücke schaffen. Daraus ergibt sich eine grundsätzliche Unsicherheit, die sehr an das Zweifeln von Descartes erinnert:

> *„Woher wissen wir überhaupt, dass die einfachen Tatsachen des Bewusstseins wirklich Tatsachen sind?"*[358]

B) Die Perspektivität phänomenaler Zustände

---

[352] Metzinger (1995), S. 22

[353] Metzinger (1995), S. 22

[354] Nagel (1974)

[355] Metzinger (1995), S. 25ff; Auflistung der Merkmale und Titel der drei Folgeabschnitte zitiert aus der Originalquelle; Überlegungen zur Einheit des Bewusstseins, S. 46: *„Die globale Einheit des Bewusstseins in diesem Sinne einer konkret erlebten Ganzheitsqualität höchster Stufe scheint die allgemeinste phänomenologische Eigenschaft des bewussten Erlebens überhaupt zu sein."*

[356] Prinz (2013), S. 375f

[357] Metzinger (1995), S. 25

[358] Metzinger (1995), S. 27ff

Bei der Perspektivität geht es auch um die Subjektivität des phänomenalen Bewusstseins. Wir selbst stehen im Mittelpunkt unseres „Bewusstseinsraumes“, wir nehmen alles aus der subjektiven „Innenperspektive“ wahr. Diese Perspektive bezeichnet man auch als „Ich-Perspektive“. Das ungelöste Rätsel ist hier die fehlende Brücke zwischen subjektiver Innenwelt und objektiver Außenwelt:

> *„Kann das aperspektivistische Weltbild der Wissenschaft überhaupt dem phänomenalen Gehalt Rechnung tragen, der an die vielen individuellen Bewusstseinsperspektiven geknüpft ist? Wird Bewusstsein unter dem Blick der Wissenschaft nicht automatisch von etwas ganz Nahem zu etwas ganz Fernem, von etwas unbezweifelbar Realem zu einer Illusion?“*[359]

Für Searle ist dies die allergrößte epistemische Herausforderung: Wie kann uns eine epistemisch objektive Beschreibung der ontologisch subjektiven Realität unseres Bewusstseins gelingen?[360]

Bettina Walde konstatiert eine „*offenkundige Vermischung der Epistemologie und der Ontologie von Qualia oder Erlebnissen auf Seiten des erlebenden Subjekts*“[361]:

> *„Es scheint so, als mache es aus der Perspektive eines erlebenden Subjekts keinen Sinn, zwischen der Epistemologie und der Ontologie von bewussten Erlebnissen zu differenzieren. Diese Vermischung von Epistemologie und Ontologie ist ein wesentlicher Bestandteil der Probleme bei der Erklärung von bewusstem Erleben oder Qualia.“*

Es ergibt sich eine „*deutliche Asymmetrie zwischen der Zugangsweise zu den eigenen bewussten mentalen Zuständen oder Erlebnissen und den bewussten mentalen Zuständen anderer Objekte“*[362]. Dies ist die Lücke zwischen der „*Perspektive der ersten“* und der „*dritten Person“*. Weiterhin lässt sich „*die Perspektive, in der wir alle unsere Erlebnisse haben, naturwissenschaftlich nicht fassen*“[363].

C) Die Präsenz phänomenaler Zustände

Das dritte Merkmal hat mit dem zeitlichen Aspekt des Erlebens zu tun. Das phänomenale Bewusstsein präsentiert uns die Welt im Hier und Jetzt. Diese Gegenwart in unserem Bewusstsein bildet zu jedem Zeitpunkt die Achse zwischen der Erinnerung an das, was vorher war, und die Erwartung (Hoffnungen, Wünsche, Pläne und Absichten) an das, was künftig sein wird. Dieses Zeitbewusstsein ist subjektiv. Auch hier liegt das Problem in der Brücke zwischen dem subjektiven Zeiterleben und der objektiven Weltzeit (Kalender, Uhrzeit).

---

[359] Metzinger (1995), S. 31

[360] Searle (1993-1), S. 9: „*What I am arguing here is that we can have an epistemically objective science of a domain that is ontologically subjective*“

[361] Walde (2002), S. 29; Hervorhebung der Autorin; sowie folgendes Blockzitat

[362] Ebd. und folgendes Zitat

[363] Walde (2002), S. 30

Brigitte Falkenburg schreibt dazu:

> *„Aus unserer subjektiven Perspektive ist nur die Gegenwart wirklich. Vergangenheit und Zukunft, Erinnerungen und Hoffnungen existieren nur in unserer Vorstellung. Dagegen betrachten wir den objektiven Zeitablauf als unabhängig vom subjektiven Erleben.* […] *Die Grenze zwischen Innen und Außen erleben wir räumlich, nicht zeitlich. Aus der Innenperspektive erleben wir uns als räumliche Körper in einer räumlichen Umgebung, aber nicht als Jetzt-Bewusstsein in einer zeitlichen Umgebung. Vergangenheit und Zukunft sind nicht da, obwohl wir den Zeitablauf als wirklich erleben und durch die Zeitmessung objektivieren können.“*[364]

D) Die Einheit des Bewusstseins

Das Bewusstsein präsentiert sich uns als integriertes Bild, das alle Sinneseindrücke und Gedanken synthetisiert. Der sich daraus ergebende „*Holismus phänomenaler Zustände“* ist eine „*höherstufige Eigenschaft“* des „*phänomenalen Modells der Wirklichkeit“*[365]:

> *„Die globale Einheit des Bewusstseins in diesem Sinne einer konkret erlebten Ganzheitsqualität höchster Stufe scheint die allgemeinste phänomenologische Eigenschaft des bewussten Erlebens überhaupt zu sein. Deshalb ist sie auf begrifflicher Ebene nur sehr schwer zu erfassen.“*

Searle arbeitet zwei wichtige Aspekte der Einheit des Bewusstseins heraus:

> *“First, at any given instant all of our experiences are unified in one single conscious field. Second, the organization of our consciousness extends over more than just simple instants. So, for example, if I begin a sentence, I have to maintain in some sense at least an iconic memory of the beginning of the sentence so that I know what I am saying by the time I get to the end of the sentence.”*[366]

Es erscheint sehr zweifelhaft, wie diese vier Eigenschaften Transparenz, Perspektivität, Präsenz und Einheit je künstlich realisiert werden könnten, insbesondere da wir die Lücken zwischen der Innen- und Außenperspektive, zwischen der Innen- und Außenwelt und dem subjektiven Zeitempfinden und der objektiven Zeit weder philosophisch noch erkenntnistheoretisch – wie wir sehen werden - schließen können.

---

[364] Falkenburg (2012), S. 213; sie zitiert weiterhin Augustinus (354–430), der in den „Confessiones“ schrieb: „*Was ist also die Zeit? Solange mich niemand fragt, weiß ich es; wenn ich es einem auf seine Frage hin erklären will, weiß ich es nicht. Dennoch sage ich zuversichtlich: Ich weiß, wenn nichts verginge, gäbe es keine vergangene Zeit, wenn nichts hinzukäme, gäbe es keine zukünftige Zeit, wenn nichts wäre, gäbe es keine gegenwärtige Zeit.“* Quelle: Augustinus Confessiones Bekenntnisse, 11. Buch, Kapitel XIV; Augustinus (2009), S. 587

[365] Metzinger (1995), S. 46; sowie folgendes Blockzitat

[366] Searle (1993-1), S. 9

5.2 Intentionalität

> *„Die Erkenntnis oder reine ‚Vorstellung (représentation)' ist nur eine der möglichen Formen meines Bewusstseins ‚von' diesem Baum; ich kann ihn auch lieben, fürchten, und diese Überschreitung des Bewusstseins durch sich selbst, die man ‚**Intentionalität**' nennt, findet sich in der Furcht, dem Hass und der Liebe wieder.*"[367]
>
> Jean-Paul Sartre

Intentionalität ist begrifflich und definitorisch noch deutlich schwieriger einzugrenzen als das Bewusstsein. Schon zur Beziehung zwischen Intentionalität und Bewusstsein finden sich unter den Philosophen gegensätzliche Positionen. Einige gehen davon aus, dass Intentionalität und Bewusstsein zwei unterschiedliche Perspektiven auf den gleichen Gegenstand seien, andere sehen eine teilweise Überlappung und wieder andere betrachten die Intentionalität als einen Teil des Bewusstseins. Auch die vorliegende Arbeit folgt der Betrachtungsweise, wonach es im Bewusstsein auch Nicht-Intentionales (wie z.B. Qualia, siehe nächster Abschnitt 5.3) gibt.

„Die kleine Routledge Enzyklopädie der Philosophie" weist zum Begriff „Intentionalität" folgenden Eintrag auf:

> ***„Intentionalität**. Als I. bezeichnet man die Fähigkeit des Geistes, sich selbst auf Dinge zu richten. Geistige Zustände wie Gedanken, Überzeugungen, Glauben, Wünsche, Hoffnungen etc. weisen I. in dem Sinne auf, dass sie immer auf etwas gerichtet sind: wenn man hofft, glaubt oder wünscht, dann muss man etwas hoffen, glauben oder wünschen. Hoffnung, Glaube, Wunsch und andere geistige Zustände, die sich auf etwas richten, sind als intentionale Zustände bekannt. Die I. in diesem Sinne ist nur am Rande mit der ursprünglichen Vorstellung von einer Absicht oder dem Beabsichtigen von etwas verbunden. Die Absicht etwas zu tun, ist ein intentionaler Zustand, weil man keine Absicht hegen kann, ohne etwas zu beabsichtigen. **Die Absichten sind aber eine von vielen Gruppen oder Arten intentionaler Zustände.** Die Terminologie der I. leitet sich von der scholastischen Philosophie des Mittelalters ab und wurde 1874 von Brentano zu neuem Leben erweckt. Brentano charakterisiert die I. als die Gerichtetheit des Geistes auf einen Gegenstand und betonte, dass dieser Gegenstand nicht notwendig existieren muss. Ferner behauptet er, dass es genau die I. der geistigen Phänomene sei, die sie von den physischen Phänomenen unterscheiden."*[368] [Hervorhebung DS]

Beim erwähnten Franz Brentano findet sich in seiner „*Psychologie vom empirischen Standpunkt*" aus dem Jahr 1874 folgendes Zitat, das in kaum einer Publikation zur Intentionalität fehlt:

> *„Jedes psychische Phänomen ist durch das charakterisiert, was die Scholastiker des Mittelalters die intentionale (auch wohl mentale) Inexistenz eines Gegenstandes genannt haben, und was wir, obwohl mit nicht ganz unzweideutigen Ausdrücken, die Beziehung auf einen Inhalt, die Richtung auf ein Objekt (worunter hier eine Realität zu verstehen ist), oder*

[367] Sartre (1997), S. 36

[368] „Die kleine Routledge Enzyklopädie der Philosophie" (2007), Band 2, S. 208

> *die immanente Gegenständlichkeit nennen würden. Jedes enthält etwas als Objekt in sich, obwohl nicht jedes in gleicher Weise. In der Vorstellung ist etwas vorgestellt, in dem Urteile ist etwas anerkannt oder verworfen, in der Liebe geliebt, in dem Hasse gehasst, in dem Begehren begehrt usw.* ***Diese intentionale Inexistenz ist den psychischen Phänomenen ausschließlich eigentümlich. Kein physisches Phänomen zeigt etwas Ähnliches****.* [Hervorhebung DS] *Und somit können wir die psychischen Phänomene definieren, sie seien solche Phänomene, welche intentional einen Gegenstand in sich enthalten.“*[369]

Intentionalität bedeutet für Brentano also, dass wir unsere Gedanken auf Inhalte beziehen und auf Gegenstände richten. Diese Objekte können real existieren oder auch fiktiv sein. Diese „*Gerichtetheit*“[370] ist ein zentrales Merkmal der Intentionalität. Weiterhin grenzt Brentano psychische Phänomene von physischen Phänomenen ab. Gemäß seiner Definition sind die psychischen Phänomene bewusste Erlebnisse, wie etwa bildliche Vorstellungen, Hoffnungen, Wünsche, Zuneigungen oder Ablehnungen von etwas. Als physische Phänomene gelten für ihn Töne, Kälte, Hitze oder Helligkeit. Sowohl physische als auch psychische Phänomene sind Erscheinungen. Die physischen Phänomene können trügerisch sein, die psychischen Phänomene sind genau so, wie sie wahrgenommen werden. Intentionalität ist nach seiner Definition eine Gemeinsamkeit der psychischen Phänomene. Damit grenzt er das Geistige vom Physischen über die Intentionalität klar ab.

Der Begriff der „intentionalen Inexistenz“ ist bei Brentano nicht eindeutig. Einerseits könnte damit gemeint sein, dass „*der intentionale Gegenstand* […] *im Akt enthalten ist*“, und andererseits, „*dass der Akt auf ihn* [den intentionalen Gegenstand] *gerichtet sein kann, selbst wenn er nicht existieren sollte*“[371]. Gianfranco Soldati argumentiert in seinem Aufsatz „*Intentionale Existenz und Bewusstsein*“ für Letzteres:

> *„Die Idee, dass Erscheinungen, als intentionale Gegenstände psychischer Akte, nicht im Geist enthalten sein müssen, hatte uns dazu veranlasst, intentionale Inexistenz anders zu verstehen. Nicht als Existenz im Geiste, sondern als Existenz, die für das Bestehen des Aktes* ***nicht erforderlich*** *ist*.“[372] [Hervorhebung DS]

Dan Zahavi interpretiert Brentano eher im Sinne der ersten Bedeutung:

> „*... Brentano* […] *spricht von der intentionalen ‚(In-) Existenz‘ des Objekts im Bewusstsein, wobei ‚Inexistenz‘ als ‚Existenz-In‘ oder ‚innere Existenz‘ verstanden werden soll. Das Objekt des Bewusstseins ist* ***immanent im psychischen Akt enthalten****, und darum wird der*

---

[369] Zitiert aus Metzinger (2010), S. 13; dort Verweis auf die Originalquelle: Brentano, Franz: *Psychologie vom empirischen Standpunkt*. Buch 2,1: §5 (Hamburg 1971 [1874]: 124f)

[370] Vgl. ebd.; Stanford Encyclopedia of Philosophy: *Consciousness and Intentionality*. https://plato.stanford.edu/entries/consciousness-intentionality/: “*Intentionality, on the other hand, has to do with the directedness, aboutness, or reference of mental states—the fact that, for example, you think of or about something. Intentionality includes, and is sometimes seen as equivalent to, what is called “mental representation*”

[371] Soldati (2016), S. 83f, §3

[372] Soldati (2016), S. 83f, §5

> *existentielle Modus dieses Objekts, sein ontologischer Status, intentional genannt.* [Hervorhebung DS]“[373]

Mutmaßlich hat Brentano beide Lesarten gemeint: den intentionalen Gegenstand, der im Akt enthalten ist und den Bezug auf real nicht-existierende Entitäten.

> „*Das Grundproblem der ‚intentionalen Inexistenz‘ besteht darin, dass die Intentionalitätsbeziehung oft als Beziehung zwischen etwas Existierendem und etwas Nicht-Existierendem gedacht werden muss und dass sie deshalb keine physikalische oder natürliche Beziehung sein kann.“*[374]

So offenbart sich laut Daniel Dennett die „*intentionalistische These*“ als „*eine nicht zu überwindende Lücke zwischen dem Geistigen und dem Physischen*“[375].

Diese Frage soll im nächsten Abschnitt thematisiert werden.

5.2.1 Husserl und die Naturalisierung der Intentionalität

Edmund Husserl, der Schüler Brentanos und Begründer der philosophischen Strömung der Phänomenologie , hat sich eingehend mit dem Verhältnis von Bewusstsein und Intentionalität auseinandergesetzt. Auch für ihn waren beide Begriffe mehrdeutig. Er unterschied „*drei miteinander verflochtene Bedeutungen*“[376]:

1. *„Bewusstsein als der gesamte reelle phänomenologische Bestand des empirischen Ich, als Verwebung der psychischen Erlebnisse in der Einheit des Erlebnisstroms.*
2. *Bewusstsein als inneres Gewahrwerden von eigenen psychischen Erlebnissen.*
3. *Bewusstsein als zusammenfassende Bezeichnung für jederlei ‚psychische Akte‘ oder ‚intentionale Erlebnisse‘. “*[377]

Die erste Bedeutung entspricht auch der Definition mit den vier Merkmalen (Transparenz, Perspektivität, Präsenz, Einheit) von Thomas Metzinger aus Kapitel 5.1. Es ist „*die Bedeutung, die wir aufrufen, wenn wir zum Beispiel vom Bewusstseinsstrom sprechen “*[378].

Die zweite Bedeutung ist eher intransitiv (auf kein Objekt gerichtet) und bezieht sich auf das „*Problem des Selbstbewusstseins*“. Wir können von einer Erfahrung sprechen, von der wir sagen, „*sie sei uns innerlich gegeben und daher bewusst*“.

Bei der dritten Bedeutung geht es um das Bewusstsein „*in einem transitiven Sinn*“, nämlich als Summe der auf Objekte zielenden Gedanken des Subjekts.

---

[373] Zahavi (2008), S. 141

[374] Metzinger (2010), S. 345

[375] Daniel Dennett zitiert in Crane (2007), S. 45

[376] Zahavi (2008), S. 139

[377] Ebd.; Originalzitat: Hua XIX/1, 356 (Husserl, Logische Untersuchungen. Zweiter Band, I. Teil)

[378] Dieses und die folgenden Zitate: Zahavi (2008), S. 139

Zusammengefasst unterscheidet Husserl, so Zahavi, erstens „*die Einheit des Bewusstseinstroms*“[379], zweitens „*das innere Bewusstsein oder Selbstbewusstsein*“ und drittens „*Intentionalität*“.

In der analytischen Philosophie und den Kognitionswissenschaften finden sich nach Zahavi drei Anwendungen der Intentionalität[380]:

1. In der Sprachphilosophie bei der Analyse von Sätzen, die *psychologische Phänomene* beschreiben
2. In der Philosophie des Geistes insbesondere bei Physikalisten (Quine, Dennett, Fodor, Churchland) als Gegenstand, den es über eine Reduktion zu naturalisieren gilt[381]
3. Ebenfalls in der Philosophie des Geistes bei Philosophen (Searle, Strawson, Crane), die eine Reduktion und Naturalisierung für ausgeschlossen halten, aber dennoch die Intentionalität und die Erste-Person-Perspektive in die Untersuchungen des Bewusstseins einbeziehen

Husserl hat sich an Diskussionen zum Leib-Seele-Problem wenig oder nur indirekt beteiligt. Gegenstand seiner Überlegungen war stets das Verhältnis zwischen Geist und Welt und nicht dasjenige von Geist und Gehirn[382]. Bezüglich der *Intentionalität* interessierte er sich primär für deren Darstellung „*aus der Erste-Person-Perspektive, d.h. vom Standpunkt des Subjekts*“. Dazu schlussfolgert Zahavi:

> *„Daher würde sich Husserl offensichtlich nicht an einer Naturalisierung von Intentionalität beteiligen, wenn man darunter den Versuch versteht, Intentionalität unter Berufung auf nicht-intentionale Mechanismen und Prozesse zu erklären.“*[383]

Bezüglich der KI sind die Überlegungen zum zweiten Anwendungsbereich von Interesse. Bei der Naturalisierung der Intentionalität geht es im Kern um folgende Frage: Kann Brentanos Behauptung, dass nur psychische Phänomene Intentionalität in sich tragen und niemals physische Phänomene, widerlegt werden?

Zahavi verweist auf zwei klassische Ansätze zur Naturalisierung der Intentionalität[384], die aber beide nicht belastbar sind, nämlich über Ähnlichkeit oder Kausalität.

---

379 Dieses und die folgenden Zitate: Zahavi (2008), S. 140

380 Vgl. Zahavi (2008), S. 142

381 Interessant ist dazu der folgende Satz bei Zahavi (2008), S. 142: „*Also entweder Intentionalität wird naturalisiert, indem man intentionale Zustände auf Verhalten, Neurophysiologie und letztlich Physik reduziert; oder man argumentiert, dass eine solche Reduktion nicht möglich sei, und schließt draus, das intentionale Vokabular sei leeres Gerede und sollte aus unserem wissenschaftlichen Diskurs eliminiert werden.*“

382 Vgl. Zahavi (2008), S. 143

383 Zahavi (2008), S. 142

384 Vgl. Zahavi (2008), S. 143f

Die Kernfrage lautet: Wie könnte eine reduktive Erklärung für die Gerichtetheit von psychischen Akten aussehen? Der Ansatz über die Ähnlichkeit könnte so aussehen, dass ein Bild oder ein Bitmuster ein Objekt referenziert. Genau dies wird aber von Zahavi unter Verweis auf Husserl zurückgewiesen:

> *„Jedes Objekt ähnelt sich selbst, es repräsentiert aber nicht sich selbst. Ferner, während die Ähnlichkeit eine symmetrische Relation ist, so ist es die Repräsentation nicht, d.h. auch wenn die dänische Königin ihrem Portrait ähnlich sehen mag, sie ist keine Repräsentation davon."*[385]

Ein Zeichen oder Bild repräsentiert nicht aus sich selbst heraus ohne weitere Erläuterung oder „*kognitive Apprehension*" ein Objekt. Genau dies stellt Searle in seiner Kritik des Computer-Funktionalismus in seinem „Chinese Room" Gedankenmodell als die fehlende Semantik heraus. Die „*Repräsentationstheorie der Wahrnehmung*" muss „*zurückgewiesen*" werden, da sie voraussetzt, „*was sie zu erklären versucht*"; in Husserls Worten:

> *„Das Gemälde ist nur Bild für ein bildkonstituierendes Bewusstsein, das nämlich einem primären und wahrnehmungsmäßig ihm erscheinenden Objekt durch seine (hier also in einer Wahrnehmung fundierte) imaginative Apperzeption erst die ‚Geltung' oder ‚Bedeutung' eines Bildes verleiht. Setzt danach die Auffassung als Bild schon ein dem Bewusstsein intentional gegebenes Objekt voraus, so würde es offenbar auf einen unendlichen Regress führen, dieses selbst und immer wieder durch ein Bild konstituiert sein zu lassen, also hinsichtlich einer schlichten Wahrnehmung ernstlich von einem ihr einwohnenden ‚Wahrnehmungsbild' zu sprechen, mittelst dessen sie sich auf die ‚Sache selbst' beziehe."*[386]

Dieser Weg der Naturalisierung der Intentionalität über die Ähnlichkeit (zum Beispiel über die Darstellung in einem repräsentierenden Bild) funktioniert also nicht.

Ein weiterer Kandidat für die Naturalisierung der Repräsentation ist die Kausalität. Rauch repräsentiert Feuer und rote Flecken die Masern[387]. Es besteht eine asymmetrische Kausalbeziehung zwischen dem Repräsentierten (Feuer, Masern) und der Repräsentation (Rauch, rote Flecken). Nun ist diese Repräsentation eher die Ausnahme und gilt maximal für real existierende Objekte. Wenn wir an ein Einhorn oder den Osterhasen denken, existiert keine Kausalkette, die zur Naturalisierung dienlich wäre. Zahavi dazu:

> *„Die Tatsache, dass es möglich ist, Objekte zu intendieren, die nicht existieren, scheint ein schlagendes Argument gegen eine Theorie, die beansprucht, ein Objekt müsse kausalen Einfluss auf mich haben, wenn ich seiner bewusst sein soll."*[388]

---

[385] Zahavi (2008), S. 143

[386] Zahavi (2008), S. 144; Originalzitat: Husserl (1900), S. 423 (Hua XIX, I S. 437)

[387] Ebd.

[388] Zahavi (2008), S.145

Damit funktioniert auch die zweite Option einer Naturalisierung der Intentionalität nicht. Die ontologische Frage der „*physikalische(n) Instantiierung intentionaler Zustände*“[389] bleibt ungelöst:

- Es gibt „*keinen physikalischen Zustand mit Intentionalität*“[390].
- Es „*scheint sogar ausgeschlossen zu sein, dass physikalische Zustände Intentionalität haben können*“.

Trotzdem sieht Wolfgang Barz einen Ausweg, der darin besteht, die Krux der Intentionalität nicht als ein Problem der Nichterklärbarkeit der physikalischen Etablierung intentionaler mentaler Zustände zu verstehen, sondern als eines der logischen Form intensionaler Sätze[391]. Er zitiert dafür Fred Dretske:

> „[T]*he intentionality of our cognitive states has its source in the intentionality of informational structures.* […] *And this intentionality derives, in turn, from the non-extensionality of statements describing nomic dependencies.* […] *It seems, then, that the intentionality associated with our cognitive states can be viewed as a manifestation of an underlying network of nomic regularities*.“[392]

Damit wird das ursprüngliche Problem der Naturalisierung der Intentionalität nicht gelöst, aber als ontologisches Problem für nichtexistent erklärt:

> „*Die Quelle der Intentionalität unserer mentalen Zustände liegt darin, dass sich Ausdrücke in den ‚dass‘-Nebensätzen intensionaler Konstruktionen nicht auf die von ihnen normalerweise bezeichneten Gegenstände, sondern auf deren intensionale Gegenstücke beziehen.“*[393]

Dretske‘s informationstheoretischer Lösungsansatz postuliert also, dass intentionale Zustände von Menschen **informationstragende Zustände** sind[394]:

> „*In essence, the information-theoretic proposal is that device S carries information about instantiations of property G if and only if S’s being F is nomically correlated with instantiations of G. If S would not be F unless property G were instantiated, then S’s being F carries information about, or as Dretske likes to say, indicates G-ness. A fingerprint carries information about the identity of the human being whose finger was imprinted. Spots on a human face carry information about a disease. The height of the column of mercury in a*

---

389 Barz (2006), S. 198

390 Dieses und das folgende Zitat: Barz (2006), S. 189

391 Vgl. Barz (2006), S. 198

392 Ebd.; Originalzitat: Dretske (1980), S. 287; etwas ausführlicher auf S. 281: „*To know, perceive, or remember is to know, perceive, or remember something. Subtleties aside, this something may be either a thing or a fact. We remember a party, see a game, and know a person; but we also remember that the party was a bore, see that the game has started, and know that Hilda is a grouch. It may be, as some have argued, that we cannot know, remember, or perceive a thing without knowing, remembering, or perceiving some fact about that thing.* […] *I am concerned with knowing, seeing, and remembering that your dog is lame, not with knowing, seeing and remembering your lame dog. I shall call such states cognitive states. The belief that your dog is lame is not, on this characterization, a cognitive state*.“

393 Barz (2006), S. 198

394 Vgl. auch https://www.staatslexikon-online.de/Lexikon/Intentionalität

> *thermometer carries information about the temperature. A gas-gauge on the dashboard of a car carries information about the amount of fuel in the car tank. The position of a needle in a galvanometer carries information about the flow of electric current. A compass carries information about the location of the North pole. In all such cases, a property of a physical device nomically covaries with some physical property instantiated in its environment.*"[395]

Damit wäre ein Weg beschrieben, wie in der Künstlichen Intelligenz über Logik, Mathematik und Sprache Intentionalität im ersten Ansatz nachgebaut werden könnte. Trotzdem wäre immer noch nicht geklärt, wie die eigentlichen mentalen Zustände des „jemanden lieben oder hassen" oder „eine bestimmte Überzeugung hegen" naturalisiert werden können. Die syntaktischen Strukturen und die Informationsstrukturen können nachgebaut werden, die Semantik und Bedeutung indes nicht. Nur der Mensch integriert sein Wissen über die Erde als Globus mit Nord- und Südpol, sein Verständnis des Magnetfeldes sowieder Funktionsweise des Kompasses und die Deutung einer Richtung der Kompassnadel.

Dieter Sturma, der „*Intentionalität* [...] *als das Grundphänomen menschlicher Intelligenz*"[396] begreift, bestreitet die Gleichwertigkeit einer „Als-ob-Intentionalität" mit der menschlichen Intentionalität:

> „*Intentionalität ist mentale Repräsentation, Als-ob-Intentionalität blendet dagegen eine dichte und insofern phänomengerechte Bestimmung vom Ansatz her aus und verschreibt sich gezielt deskriptiver Oberflächlichkeit. Im Fall von KI und Robotik wird damit jener Indifferentialismus verstärkt, der zum Verkennen der grundsätzlichen Unterschiede zwischen Mensch und Maschine und problematischen Reaktionsformen aus Maschinen führen kann.*"[397]

In der Künstlichen Intelligenz wird immer nur eine „Als-ob-Intentionalität" realisierbar sein, die sich mit dem betrachteten Gegenstand nur objektiv auseinandersetzt, niemals aber aus einer echten subjektiven Perspektive.

[395] Stanford Encyclopedia of Philosophy: Jacob (2023), Abschnitt 9

[396] Sturma (2003), S. 45

[397] Sturma (2003), S. 48

5.3 Qualia und das phänomenale Bewusstsein

> *"We cannot form to ourselves a just idea of the taste of a pineapple, without having actually tested it.*"
>
> David Hume, "A Treatise of Human Nature", 1739[398]

Ohne die Vokabel „Qualia" zu verwenden (oder zu kennen), beschrieb David Hume bereits vor mehr als 280 Jahren das zentrale Problem des phänomenalen Bewusstseins, das anders als das im vorherigen Abschnitt diskutierte intentionale Bewusstsein nicht auf existierende oder nicht-existierende Objekte gerichtet ist.

Schon vor Hume konnte sich der deutsche Philosoph Leibniz das Zustandekommen einer Perzeption, wie er das phänomenale Empfinden nannte, nicht erklären:

> *„Man muß ferner notwendig zugestehen, daß die **Perception** und was von ihr abhängt, aus **mechanischen** Gründen, d.h. aus Gestalt und Bewegung, nicht erklärbar ist. Denkt man sich etwa eine Maschine, deren Einrichtung so beschaffen wäre, daß sie zu denken, zu empfinden und zu perzipieren vermöchte. So kann man sie sich unter Beibehaltung derselben Verhältnisse vergrößert denken, so daß man in sie wie in eine Mühle hineintreten könnte. Untersucht man alsdann ihr Inneres, so wird man nichts als Stücke finden, die einander stoßen, niemals aber Etwas, woraus man eine Perzeption erklären könnte."*[399]
> [Hervorhebungen DS]

Qualia ist der Plural von Quale und bezeichnet die ungerichteten Sinneseindrücke, wie Wärme, Schmerz, Licht, Gerüche und Süße. Sie sind quasi die „*Atome des Bewusstseins*" oder auch „*phänomenale Eigenschaften erster Ordnung*"[400].

Das dominierende philosophische Problem der Qualia liegt in ihrer Subjektivität und damit verbunden in der „*epistemischen Asymmetrie*"[401]. Dies soll am folgenden Beispiel erläutert werden, dem Duft einer Rose. Wenn wir an einer Rose riechen, entwickeln wir einen ganz spezifischen Eindruck vom Duft dieser Pflanze. Bei der späteren Konfrontation mit diesem Duft im Dunkeln erkennen wir ihn wieder und können klar sagen, dass wir Rosenduft wahrgenommen haben. Dies ist ein Eindruck, den wir über lange Zeiträume nicht vergessen und manchmal noch nach Jahrzehnten wieder aktivieren, wie der Geruch bestimmter Speisen aus der Kindheit. Den Geruch der Rosen und anderer Stoffe können wir in unserer Sprache nicht hinreichend und voll verständlich beschreiben. Das Gleiche gilt für andere Sinneseindrücke, wie Schmerzen, die Röte des Sonnenuntergangs oder der Klang eines Musikinstruments. Diese Unmöglichkeit der Objektivierung von Qualia bildet die epistemische Asymmetrie. Wir können sie nicht in Worte fassen. Sie stellen zunächst einmal ein sprachliches Problem dar. Aber nicht nur das: Eine

[398] Hume (1739), S. 5

[399] Zitiert aus Siebert (1998), S. 13; Originalzitat: Leibniz (1925), S.605f

[400] Metzinger (2006), S. 57

[401] Metzinger (1995), S. 40

vollständige Reduktion von Qualia auf die physische und subjektive Ebene ist so wenig möglich wie umgekehrt die Erklärung des Rosenduftes aus der chemischen Zusammensetzung der beteiligten Stoffe und deren Interaktion mit unserem Nervensystem.

Auch der bereits zitierte Emil du Bois-Reymond bringt diese Asymmetrie in seinem berühmten Vortrag „*Über die Grenzen des Naturerkennens*“ von 1872 klar zum Ausdruck:

> *„Welche denkbare Verbindung besteht zwischen bestimmten Bewegungen bestimmter Atome in meinem Gehirn einerseits, andererseits den für mich ursprünglichen, nicht weiter definierbaren, nicht wegzuleugnenden Tatsachen ‚Ich fühle Schmerz, fühle Lust, ich schmecke Süßes, rieche Rosenduft, höre Orgelton, sehe Rot,‘ und der ebenso unmittelbar daraus fließenden Gewissheit: ‚Also bin ich‘?“*[402]

Dies ist ein Verweis auf Descartes‘ Cogito, im Ergebnis konsistent mit dessen Vorgehensweise der philosophischen Argumentation aus der Ich-Perspektive. Die Beschreibung des Bewusstseins aus der inneren Wahrnehmung war das Ziel der Phänomenologen mit durchaus interessanten theoretischen Schlussfolgerungen bei Brentano, Husserl und Sartre. Die Brücke zu den naturalistischen Objektivsten, die sich dem Problem des Bewusstseins von außen nähern wollten, konnte aber niemals geschlossen werden.

Wichtig ist auch, darauf hinzuweisen, dass Qualia meist nicht isolierte Sinneseindrücke sind, sondern verschiedene Sinneseindrücke verbinden. Farbe, Geruch und Geschmack eines Rotweins aus Bordeaux machen die holistische und homogene Perzeption dieses Weines aus. Charles Sanders Peirce beschreibt dies treffend:

> *“The quale consciousness is not confined to simple sensations. There is a peculiar quale to purple, though it be only a mixture of red and blue. There is a distinctive quale to every combination of sensations so far as it is really synthesized – a distinctive quale to every work of art – a distinctive quale to this moment as it is to me – a distinctive quale to every day and every week – a peculiar quale to my whole personal consciousness. I appeal to your introspection to bear me out of this.”*[403]

Daniel Dennett weist den Qualia *vier Eigenschaften zweiter Stufe* zu[404]:

- Sie sind „***unaussprechlich“***. Wir können anderen nicht mitteilen, „*wie sich das, was man gerade sieht, schmeckt, riecht oder anfühlt“*.
- Dies liegt daran, dass sie „***intrinsische“*** Eigenschaften sind, die man nicht weiter analysieren oder zerlegen kann. In dem Sinne sind sie auch in sich homogen.
- Sie sind „***privat“*** aus Sicht der wahrnehmenden Person und damit auch subjektiv. Ein Vergleich der Wahrnehmung zwischen zwei Personen ist nicht möglich.
- Abschließend sind Qualia „***dem Bewusstsein unmittelbar und direkt zugänglich***“.

[402] Du Bois-Reymond (1872), S. 29

[403] Peirce (1934), S. 150

[404] Vgl. Dennett (2006), S. 210 f; gilt auch für die vier folgenden Eigenschaften

Dennett erläutert dies am Beispiel der Wahrnehmung von Farben, die wir nicht weiter beschreiben können, die intrinsisch im Auge und Gehirn des Betrachters sind, aus seiner Sicht privat und in seinem Bewusstsein direkt präsent. Hierzu zitiert er in seinem Buch „*Brainchildren*“[405] Ornstein und Thompson[406]:

> *“‘Color‘ as such does not exist in the world; it exists only in the eye and brain of the beholder. Objects reflect many different wavelengths of light, but these light waves themselves have no color.”*

In einem anderen Aufsatz[407] bemüht er dazu Einstein:

> *„Ich glaube, es war Einstein, der sagte, die Wissenschaft könne uns keine Auskunft darüber geben, wie Suppe schmeckt. Könnte ein so weiser Mann sich geirrt haben?“*

Thomas Nagel hat das phänomenale Bewusstsein mit der Formel „Wie ist es x zu sein“ oder „Wie ist es y zu erleben“ zusammengefasst, ganz spezifisch in dem schon mehrfach erwähnten Aufsatz „*Wie ist es, eine Fledermaus zu sein*“[408]. Nur die Fledermaus weiß, wie es ist, eine Fledermaus zu sein; nur wer schon einmal einen Fallschirmsprung unternommen hat, kann wissen, wie sich so ein Abenteuer anfühlt, und nur wenn man schon einmal auf einem Weihnachtsmarkt war, weiß man, wie gebrannte Mandeln und Glühwein riechen. Wie bereits oben dargelegt, sind Qualia subjektiv und privat und – nach der Argumentation von Peirce – holistisch und in sich geschlossen.

Mit einigen Gedankenexperimenten – insbesondere Nagel (1974) und Jackson (1982) – konnte überzeugend argumentiert werden, wie Siebert zusammenfasst, „*daß die Beziehung, die jemand aus der Perspektive der ersten Person singular zu seinen Qualia unterhält, eine Form von Wissen darstellt, die sich aus einer rein physikalistischen oder funktionalistischen Position nicht darstellen lässt“.*[409]

Diese Erkenntnislücke zwischen dem, was mit der Physik darstellbar ist, und was wir selbst erleben (unsere Innenperspektive), nennt man die Erklärungslücke, die für uns Menschen nicht zu schließen ist, auch von allen Neurobiologen, Geistesphilosophen und KI-Forschern dieser Welt nicht.

---

405 Dennett (1998), S. 142

406 Ornstein R., Thompson R.F. (1984): *The Amazing Brain*

407 Dennett (2006), S. 213

408 Nagel (1974)

409 Siebert (1998), S. 15

5.4 Zwei Arten des Bewusstseins nach Block und Burge

Die beiden Philosophen Ned Block und Tyler Burge haben in ihren aufeinander aufbauenden Arbeiten[410] versucht, eine Brücke zwischen Philosophie und Neurobiologie zu schlagen, und damit auch empirische Argumente und Betrachtungsweisen in das philosophische Modell des Bewusstseins eingebracht. Ein wesentliches Ergebnis dieser Arbeiten bildet die Unterscheidung zwischen zwei Arten des Bewusstseins, dem phänomenalen Bewusstsein (P-Bewusstsein) und dem Zugriffsbewusstsein (Z-Bewusstsein). Diese Unterscheidung korreliert in Teilbereichen mit derjenigen zwischen Qualia und Intentionalität, nimmt jedoch stärker Bezug auf empirische Erfahrungen und schafft die Basis für eine Quasi-Architektur des gesamten Bewusstseins einschließlich Selbstbewusstsein und Unterbewusstsein. Vor allem wird das Zusammenspiel dieser beiden Arten beschrieben. Im Folgenden werden die konzeptionellen Eckpunkte dargestellt und Schlussfolgerungen auch für die Folgediskussion des Künstlichen Bewusstseins und der Künstlichen Intelligenz abgeleitet.

Die Argumentation für die (modellhafte) Zweiteilung des Bewusstseins baut auf einer Reihe von empirischen Experimenten und Beobachtungen auf. Die zwei wichtigsten sollen hier kurz zusammengefasst werden.

Die erste empirische Beobachtung stützt sich auf Untersuchungen an Patienten, die aufgrund von Verletzungen des visuellen Kortex im Gehirn ihre Sehfähigkeit teilweise oder komplett eingebüßt haben, obwohl ihre Augen intakt blieben. Beim Kortex handelt es sich gleichsam um das „*Rechenzentrum unseres Gesichtssinns*“[411]. Dort werden die Signale aus den Augen zu einem Bild zusammengefügt. Diese Patienten hatten also „*blinde Regionen*“ oder „blinde Flecken“ in ihrem Gesichtsfeld. In Experimenten präsentierte man ihnen optische Reize und Signale innerhalb dieser Flecken, zum Beispiel verschiedenfarbige Blitze, unterschiedliche Symbole, horizontale oder vertikale Streifen oder kleinere Gegenstände. Erwartungsgemäß verneinten die Patienten, dass sie überhaupt irgendetwas sahen. Wenn man sie bat, die Farbe eines Lichtblitzes oder die Ausrichtung von Streifen oder gezeigten Gegenständen zu erraten, lagen sie zu einem hohen Prozentsatz der Fälle richtig. Wenn man sie aufforderte, einen Gegenstand innerhalb des blinden Flecks zu ergreifen, formten sie ihre Hand in angemessener Weise. Zum Beispiel würde man ein liegendes Wasserglas anders ergreifen als ein stehendes. Die gleiche Beobachtung hat man auch bei Probanden gemacht, deren Kortex über eine transkranielle Magnetstimulation[412] außer Kraft gesetzt wurde. Aus diesen Experimenten läßt sich

[410] Block, Ned: *Eine Verwirrung über eine Funktion des Bewusstseins* in Metzinger (1995), S. 523 – 581 und Burge, Tyler: *Zwei Arten von Bewusstsein* in Metzinger (1995), S. 583 - 594

[411] Vgl. https://medlexi.de/Rindenblindheit

[412] „*Die transkranielle Magnetstimulation (TMS) ist eine in den Neurowissenschaften inzwischen weit verbreitete Methode zur Untersuchung neurophysiologischer Prozesse sowie des Zusammenhangs*

schlussfolgern, dass die für das Bewusstsein unsichtbaren Reize offensichtlich unbewusst wahrgenommen wurden. Aus dem Unbewussten stehen sie allerdings nicht für ein rationales Nachdenken zur Verfügung und somit auch nicht für eine Handlungskontrolle oder Sprachsteuerung. Daraus lässt sich eine direkte Verbindung zwischen phänomenalem Bewusstsein und Rationalität ableiten. Ohne phänomenales Bewusstsein keine Rationalität.

Die zweite empirische Beobachtung bezieht sich auf Epileptiker, die während einer Tätigkeit wie zum Beispiel Autofahren, Klavierspielen oder Spazierengehen einen Anfall erleiden und die Aktivität unbeirrt fortsetzen und nicht – wie man vielleicht erwarten würde – abrupt abbrechen oder verändern[413]. Aus diesem Phänomen schließt Block, der dafür Searle zitiert, dass Bewusstsein offensichtlich Verhaltensveränderungen, also Flexibilität und Kreativität befördert. Das Ausführen einer beschlossenen Handlung geschieht quasi automatisch im Unterbewusstsein.

Das zweiteilige Bewusstseinsmodell von Ned Block, das später von Tyler Burge aufgegriffen wurde, besteht aus dem *phänomenalen Bewusstsein (kurz: P-Bewusstsein)* und dem *(rationalen) Zugriffsbewusstsein (kurz: Z-Bewusstsein)*[414].

P-Bewusstsein manifestiert sich im Erleben. Dies beinhaltet Sehen, Hören, Riechen, Schmecken und Fühlen also Empfindungen und Wahrnehmungen. P-bewusste Zustände können einen (sekundären) intentionalen Gehalt haben, müssen dies aber nicht. Ein Beispiel wäre die Wahrnehmung der Quelle eines Geräusches oder eines Lichtstrahles. P-bewusste Zustände sind niemals kognitiv im Sinne von Denken und Argumentieren oder funktional, also auf Handlungsabläufe und Algorithmen bezogen, und selten intentional (Repräsentation bezogen oder gerichtet auf etwas).

Davon klar abgegrenzt ist das Z-Bewusstsein, bei Tyler Burge das „rationale Zugriffsbewusstsein“. *Z-bewusste Zustände*[415] sind *ungebunden und frei*, also nicht davon abhängig, dass ein phänomenaler Zustand anhält und damit als Prämisse für rationale Überlegungen zur Verfügung steht, so zum Beispiel für *rationale Handlungskontrolle* und *rationale Sprachkontrolle*.

Block arbeitet drei Unterschiede zwischen P- und Z-Bewusstsein heraus[416]:

- P-bewusste Zustände sind primär *phänomenal* und Z-bewusste Zustände sind primär *repräsentational.*

zwischen fokaler Gehirnaktivität und Verhalten. Von wissenschaftlichem Interesse ist vor allem die kurzfristige Störung einer kleinen Hirnregion, um deren physiologische Funktion zu untersuchen. So kann man mit der Magnetstimulation über dem motorischen Kortex Muskelzuckungen auslösen, über der Sehrinde kann man Phosphene, aber auch Skotome erzeugen.“ Quelle: https://www.thieme-connect.com/products/ejournals/abstract/10.1055/s-2005-866866

413 Block (1995), S. 525

414 Vgl. Block (1995), S. 531

415 Vgl. Block (1995), S. 535

416 Vgl. Block (1995), S. 537f

- Das P-Bewusstsein beinhaltet das Erlebnis im Sinne von „Wie ist es, in diesem Zustand zu sein“, in der Regel ist es nicht transitiv. Das Z-Bewusstsein ist immer transitiv: Bewusstsein-von-etwas.
- Paradigmatische P-bewusste Zustände sind Empfindungen und paradigmatische Z-bewusste Zustände sind propositionale Einstellungen wie Gedanken, Überzeugungen und Wünsche.

Damit ist das intransitive Bewusstsein das P-Bewusstsein und das transitive Bewusstsein (B. von etwas) das Z-Bewusstsein. Zusammenfassend drängt sich folgende Unterscheidung auf:

- P-Bewusstsein: privat, subjektiv, intrinsisch, nicht-korrigierbar, nicht kommunizierbar
- Z-Bewusstsein: objektiv, extrinsisch, korrigierbar und kommunizierbar

Im Zusammenspiel zwischen den beiden Arten von Bewusstsein führt der übliche Fluss vom P-Bewusstsein in das Z-Bewusstsein. P-bewusste Zustände stehen im Z-Bewusstsein zur Verfügung für die weitere Verarbeitung in Gedanken, die Handlungskontrolle oder die Artikulation in Sprache. Auch gibt es eine Rückwirkung vom rationalen Z-Bewusstsein in das P-Bewusstsein, und zwar dergestalt, dass unsere Rationalität das Erleben beeinflusst. Glücksgefühle bei einer Achterbahnfahrt stellen sich nur deshalb ein, weil die Rationalität einem sagt, dass die Technik sicher ist und nichts passieren kann. Ohne diesen Einfluss des Z-Bewusstseins auf das P-Bewusstsein würden wir eine derartige Fahrt anders erleben. Das Gleiche in umgekehrter Richtung gilt für die Wahrnehmung der Rückenflosse eines Haies, die Angst auslöst.

Für die übergreifende Themenstellung ergibt sich die wichtige Frage, ob beim Menschen P-Bewusstsein ohne Z-Bewusstsein denkbar ist. Block sieht dies erfüllt, wenn wir die generelle Wahrnehmung von Störgeräuschen (Presslufthammer von Bauarbeitern auf der Straße vor dem eigenen Haus) aus unserem Z-Bewusstsein „*ausblenden*“[417], was letzten Endes ein P-Bewusstsein ohne Aufmerksamkeit ist.

Auch bestehen große Vorbehalte gegenüber der umgekehrten Konstellation: Z-Bewusstsein ohne P-Bewusstsein. Block konstruiert dies über eine sogenannte Superblindsichtigkeit für Teilbereiche der Wahrnehmung. Dies entspricht dem Herumlaufen in einem stockfinsteren Raum ohne jegliches Restlicht und der Konstruktion der Wahrnehmung der Möbelstücke in dem Raum. Tyler Burge ist bei dem Thema sehr kategorisch. Auch wenn das Z-Bewusstsein in Teilbereichen das P-Bewusstsein ersetzen kann, ist ein Z-Bewusstsein ohne jegliches P-Bewusstsein nicht vorstellbar[418]:

[417] Block (1995), S. 543f

[418] Burge (1995), S. 586

> *„Ein phänomenaler Zombie hat kein Bewusstsein – ganz gleich wie wirkungsvoll und rational sein Verhalten, seine Verbalisierungen und seine kognitiven Leistungen auch sein mögen.“*

Für Burge ist das phänomenale Bewusstsein eine notwendige Voraussetzung des Bewusstseins insgesamt[419].

Genau da setzt die Problematik der Naturalisierung des Bewusstseins an. Das Z-Bewusstsein lässt sich in Bezug auf Handlungs- und Sprachkontrolle „nachbauen“, hingegen nicht in Bezug auf die propositionalen Einstellungen (Gedanken, Wünsche und Überzeugungen). Vor allem kann das P-Bewusstsein nicht mit der Sensorik eines Roboters gleichgesetzt werden. Ein Roboter mit hochauflösender Stereokamera, Positionsmelder und Thermometer hat kein „Erleben“ und damit auch kein Gefühl davon, wie es ist, Roboter zu sein. Damit verfügt er auch über kein P-Bewusstsein und somit auch insgesamt über kein Bewusstsein.

[419] Vgl. Burge (1995), S. 587

5.5 Neurobiologie: Wie Bewusstsein entsteht

> *"Science is the greatest of communal cultural achievements and is among the highest achievements of human consciousness. Nevertheless, it must be said that, however grand, the scientific view derives from other cultural ingredients and does not compel them. Science is only a partial experience of consciousness that, once born and developed in human culture, has a potentially endless sweep in subjective personal experience, in art, and in the creation of myths. What spins up out of consciousness, language, and culture is necessarily full of novelty. From this point of view, whatever we establish scientifically to be true, there is much in our experience that is of our own making and much of it the most precious part of our life. That is so, however, should not be allowed to obscure our knowledge of the structural conditions of the world's being, and of our being, and of how we know them, all of which can come reliably only from scientific inquiry itself."*[420]
>
> Gerald Edelman

Zahlreiche philosophische Erörterungen zum Bewusstsein nehmen Bezug auf empirische Studien in der Psychologie oder Neurobiologie. Schon Descartes bezog sich in seinen Aufsätzen auf pathologische Untersuchungen und entwickelte so seine These, wonach die Zirbeldrüse das verbindende Organ zwischen Körper und Geist sein könnte. Mittlerweile ist die Neurobiologie und -medizin deutlich vorangeschritten, insbesondere in den letzten fünf Jahrzehnten, vor allem durch die Arbeiten von Nobelpreisträgern wie Charles Scott Sherrington, Gerald Edelman und Francis Crick.

In der Tat ist die Hirnforschung gemäß einem Manifest von elf führenden Wissenschaftlern der Disziplin aus dem Jahre 2004 noch immer weit davon entfernt zu verstehen, wie das Gehirn konkret funktioniert und wie insbesondere Bewusstsein zustande kommt. Im gleichen Manifest wurde jedoch große Zuversicht zum Ausdruck gebracht, dass man innerhalb von zwei bis drei Jahrzehnten „*widerspruchsfrei Geist, Bewusstsein, Gefühle, Willensakte und Handlungsfreiheit als natürliche Vorgänge ansehen werde, denn sie beruhen auf biologischen Prozessen*"[421]. Dahinter verbirgt sich einerseits die Hoffnung auf einen signifikanten weiteren Erkenntnisgewinn und andererseits eine Hypothese über dessen Inhalt. Das Manifest ist im zweiten Anhang dieser Arbeit vollständig wiedergegeben.

Parallel hat sich die geistesphilosophische Diskussion zum Körper-Geist Problem und zum Rätsel des Bewusstseins auch mit den oben zitierten Philosophen weiterentwickelt. Es ist so etwas wie ein Kampf um die Deutungshoheit in Bezug auf den menschlichen Geist entstanden.

Die Neurobiologie verfolgt die gleiche Kernfrage des Leib-Seele Problems bzw. des Gehirn-Bewusstsein Problems, die auch schon die Philosophen der letzten Jahrhunderte

[420] Edelman (1989), S. 270f

[421] Manifest der Hirnforscher: siehe Anhang 2

beschäftigte: Wie kommt Bewusstsein im menschlichen Gehirn zustande? Darauf bauen etliche Fragen auf, die denjenigen der Philosophen sehr ähneln[422]: Erstens geht es um den kausalen Status des Bewusstseins. Inwieweit ist das Bewusstsein *ein reines Epiphänomen* ohne Einfluss auf die Materie oder kann es physische Ereignisse anstoßen? Zweitens geht es um die Frage nach dem „neuralen Wirkmechanismus", der subjektive Bewusstseinszustände (Qualia) impliziert. In dieser Arbeit soll das Modell von Gerald Edelman vorgestellt werden, das die Mechanismen der Evolutionstheorie von Charles Darwin auch im neuronalen Bereich anwendet.

In einem weiteren für diese Arbeit interessanten Fragenkomplex geht es um den Determinismus und den freien Willen des Menschen. Was sagt uns die Hirnforschung dazu und was bedeutet das für die Beurteilung von Determinismus und Willensfreiheit der Künstlichen Intelligenz?

5.5.1 Neuraler Darwinismus

> *"The fundamental subject underlying all theories of brain function is morphologic evolution."*[423]
>
> Gerald Edelman

Seit den frühen 1980er Jahren arbeitete Medizin-Nobelpreisträger Gerald Edelman an seiner „*biologischen Theorie des Bewusstseins*" (Untertitel eines seiner Bücher)[424], in der er die Selektionsmechanismen nach Darwin als „*Theorie der Selektion neuronaler Gruppen*" anwendet. Das Innovative in seinem Ansatz liegt darin, dass er Bewusstsein nicht als Funktion versteht, die in einer bestimmten Region des Gehirns erbracht wird, sondern als Prozess, der quer über eine Vielzahl von Regionen des Gehirns in der Interaktion zwischen Organismus und Umwelt abläuft. Sämtliche instruktionistische Modelle, die davon ausgehen, dass das Gehirn wie ein Computer, wie eine Turing-Maschine funktioniert, weist er klar zurück[425]:

> *„Eine Hirntheorie muss den Facettenreichtum, die Vielfalt und die Bandbreite bewussten Erlebens erfassen, mit den Tatsachen der Evolution und der Entwicklung von Individuen vereinbar sein und ein Grundprinzip beschreiben. Das Grundprinzip soll in den wesentlichen Wirkmechanismen, mit denen das Gehirn Information und neuartige Erfahrungen verarbeitet, nachweisbar sein."*

Hinweise dafür, dass im Gehirn Programme und Algorithmen abgearbeitet werden, hat er in seinen empirischen Studien nicht feststellen können. Allerdings verweist er auf viele

[422] Vgl. Edelman (2004), S. 17
[423] Edelman (1989), S. 273
[424] Edelman (1989)
[425] Edelman (2004), S. 44

Anzeichen, die in Richtung einer „*auf dem Populationsdenken beruhende Theorie*“[426] weisen. Er kommt zu dem Ergebnis, dass „*jedes Gehirn* […] *ein außerordentlich hohes Maß an individueller Variation*“ aufweist:

> *„Das gilt für Strukturen und Funktionen auf allen Ebenen. Individuen unterscheiden sich in genetischen Festlegungen, epigenetischen Entwicklungssequenzen, körperlichen Reaktionsmustern und den Erfahrungen, die sie in ihrer jeweiligen Umwelt machen. Die Folge ist, dass chemische Vorgänge in Neuronen, Netzwerkstrukturen, Synapsenstärken, zeitliche Charakteristika, gespeicherte Erinnerungen und von Bewertungssystemen regulierte Motivationsmuster eine ungeheure Variationsbreite aufweisen. Menschen unterscheiden sich deshalb in Inhalt und Stil ihres Bewusstseinsstroms beträchtlich voneinander*.“

Edelman kann sich diese Vielfalt nur als Ergebnis eines Selektionsprozesses nach der von Darwin erfundenen populationsdynamischen Denkweise vorstellen. Synapsen-Populationen, die einen Überlebensvorteil in spezifischen neuartigen Situationen bewiesen haben und damit eine gelungene Anpassung an die Umwelt darstellen, definieren mit größerer Wahrscheinlichkeit das zukünftige Verhalten. Diesen Selektionsprozess „*im Verlauf eines einzelnen Lebens*“ bezeichnet Edelman als „*somatische Selektion*“[427]. In früheren Arbeiten hat Edelman genau diese somatische Selektion im Immunsystem nachweisen können. Die ursprüngliche „*instruktivistische*“ *Hypothese*, „*dass sich Antikörper um das* ‚feindliche‘ *Antigen herumlegen und so dessen Form entsprechend abnehmen und erhalten*“ hat sich als falsch herausgestellt. Bestätigt werden konnte die somatische Selektion nach folgendem Ablauf: Ein riesiges „*Repertoire an unterschiedlichen Antikörpern*“ wird mit einem „*Fremdartigem konfrontiert*“. Danach erfolgten eine Selektion und eine Vermehrung der Zellen, deren Antikörper einigermaßen gut passen, selbst wenn das Antigen „*in der Erdgeschichte nie zuvor aufgetaucht war*“.

Die Anwendung der somatischen Selektion im menschlichen Gehirn unterscheidet sich signifikant vom Computer-Konnektionismus als Modell des Gehirns und des Bewusstseins. Computermodelle lassen keine Mehrdeutigkeiten zu und arbeiten Eingangssignale aus der Außenwelt in logisch deduktiven Schritten ab:

> „Das Gehirn ist kein Computer und die Welt kein Magnetband, von dem das Gehirn Informationen abtastet.“[428]

Edelman belegt dies mit einer Reihe von empirischen Beobachtungen und Experimenten, die an dieser Stelle nicht weiter vertieft werden.

---

[426] Dieses und das folgende Blockzitat: Edelman (2004), S. 44

[427] Dieses und folgende Zitate: Edelman Tononi (2000), S. 113f

[428] Edelman (2004), S. 49; Edelman schrieb dies in einer Zeit, in der das Speichermedium für Computer Magnetplatten und Magnetbänder waren. Seine grundsätzliche Schlussfolgerung gilt auch für zeitgenössische Speichermedien (Festplatten, Cloud-Speicher).

Die Edelman'sche ***Theorie der Selektion neuronaler Gruppen*** *(TNGS, Theory of neuronal group selection)* beinhaltet *drei Hauptaussagen*:[429]

1. ***„Entwicklungsselektion**. Im Verlauf der frühen Individualentwicklung von Vertretern einer Art wird die Ausprägung der ursprünglichen Hirnanatomie zunächst einmal, soviel steht fest, durch Gene und Vererbung eingeschränkt. Doch bereits von einem sehr frühen embryonalen Stadium an wird die weitere Etablierung von Kontakten auf synaptischer Ebene in hohem Maße durch somatische Selektion geleistet. Mit fortschreitender Entwicklung senden Neurone beispielsweise Myriaden verzweigter Fortsätze in die unterschiedlichsten Richtungen aus. Diese Verästelungen garantieren eine außerordentlich hohe Variabilität im Hinblick auf die möglichen Verknüpfungsmuster bei diesem speziellen Wesen und schaffen ein riesiges und vielfältiges Repertoire an neuronalen Schaltkreisen. Im weiteren Entwicklungsverlauf haben die Neurone die Möglichkeit, diese individuellen Muster an elektrischer Aktivität zu verstärken oder wieder abzuschwächen: Neurone, die gleichseitig feuern, werden verkabelt. Die Folge davon ist, dass die Neurone innerhalb einer Gruppe enger miteinander verknüpft sind als Neuronen aus verschiedenen Gruppen.*
2. ***Erfahrungsselektion**. Überlappt wird diese frühe Phase durch einen Prozess der synaptischen Selektion innerhalb der Repertoires einzelner neuronaler Gruppen, der auf Verhaltenserfahrungen zurückzuführen ist und das ganze Leben hindurch stattfindet. Beispielsweise ist bekannt, dass ‚Karten', die sich im Gehirn auf der Grundlage der taktilen Signale aus den Fingern gebildet haben, in ihren Umrissen verändert werden, sobald mehr oder weniger Finger verwendet werden. Diese Veränderungen ergeben sich, weil gewisse Synapsen innerhalb einer Gruppe lokal gekoppelter Neurone beziehungsweise zwischen mehreren Gruppen ohne sonstige anatomische Veränderungen gestärkt, andere hingegen abgeschwächt werden. Gelenkt und mit gewissen Randbedingungen versehen wird dieser Selektionsprozess durch Hirnsignale, die sich aus der Aktivität diffus projizierender Bewertungssysteme ergeben, eine Einschränkung, die durch den erfolgreichen Output kontinuierlich modifiziert wird.*
3. **Wiedereinspeisende Kartierung**[430] ***(Reentry)**. Zur Korrelation zwischen selektiven Ereignissen innerhalb verschiedener Karten des Gehirns kommt es durch den dynamischen Prozess des Reentrys. Reentrante Wechselwirkungen machen es einem Tier* [oder Menschen] *mit einem hochvariablen, einzigartigen Nervensystem möglich, auch ohne die Hilfe eines Homunkulus oder eines Computerprogramms eine nicht mit Begriffen belegte Welt in Objekte und Ereignisse zu unterteilen.* [Der] *Prozess des Reentry* [führt] *zur Synchronisation der Aktivität neuronaler Gruppen in verschiedenen Hirnkarten und verbindet diese zu Schaltkreisen, die einen zeitlich kohärenten Output entstehen lassen. Das Phänomen des Reentry bildet somit den zentralen Mechanismus zur räumlichen und zeitlichen Koordination verschiedener sensorischer und motorischer Ereignisse."*

[429] Edelman Tononi (2000), S. 116f

[430] Abweichende Übersetzung aus Siebert (1999), S. 95

Der wichtigste, kontroverseste und am wenigsten bewiesene Teil von Edelmans Theorie liegt in der dritten Hauptaussage. Es ist allgemein bekannt und bewiesen, dass die Rezeptoren der Sinnesorgane in jeweils spezifischen Gebieten des Kortex Erregungsmuster erzeugen, die man als „mentale Karten" („mental maps") bezeichnet. So gibt es allein für das visuelle System eine Vielzahl dieser Karten, „*in denen z.B. Farbe, Bewegung, Orientierung im Raum, relative Position auf der Retina etc. verarbeitet werden"*[431]. Kontrovers und nicht bewiesen ist der Mechanismus des Reentry, also der Rückkopplung der Karten untereinander, so dass Karten aktiviert werden, obwohl kein spezifischer Reiz für sie vorliegt. Faszinierend ist diese Idee, weil mit dieser „*weiträumigen Synchronisation der Aktivität zahlreicher Gruppen von aktiven Neuronen, die über verschiedene, funktional spezialisierte Gehirnareale verteilt sind*, auch *die Integration von Wahrnehmungs- und Bewegungsabläufen"*[432] erklärt werden kann. Der Prozess ermöglicht damit die Lösung des sogenannten Bindungsproblems, der Frage nach der sensorischen Integration verschiedenster Sinneseindrücke in einer einheitlichen Wahrnehmung des Bewusstseins. Damit ist Reentry bzw. die wiedereinspeisende Kartierung „*das bestimmende Prinzip in der raumzeitlichen Koordination selektionaler Netzwerke"*[433] und nicht Logik.

Edelman stellt noch ein weiteres Prinzip heraus, das er „*Degeneriertheit*" nennt. Damit meint er die Eigenschaft selektiver Systeme, mehrere Komponenten vorzusehen, die den gleichen Output produzieren. Ingenieure würden dies „Redundancy" nennen. Aus Edelmans Sicht ist dies die Grundvoraussetzung dafür, dass der Selektionsprozess des Reentry nicht in eine Sackgasse läuft. Konkret geht es z.B. um den Umgang mit lokal begrenzten Läsionen im Gehirn. Auch ist dies eine wichtige Erklärung für die Kontingenz in unserer Wahrnehmung und unserem Denken.

Nun zurück zum Bewusstsein: Edelman unterscheidet „*zwischen einem primären Bewusstsein und einem Bewusstsein höherer Ordnung"*[434]. Primäres Bewusstsein ist derjenige „*Zustand mentalen Gewahrseins, bei dem ein Individuum in der Gegenwart mentale Bilder von Dingen in der Welt aufbaut"*. Über ein derartiges Bewusstsein verfügen Menschen und auch Tiere, die über ein ähnlich organisiertes Gehirn verfügen. Ein Wesen mit einem „*Bewusstsein höherer Ordnung"* dagegen ist sich dessen bewusst, „*dass es Bewusstsein hat"* und kann „*als denkendes Subjekt die eigenen Handlungen und Gefühlsregungen in den Blick nehmen"*. Dies beinhaltet semantische und sprachliche Fähigkeiten.

Diese Zweiteilung erinnert ein wenig an die Unterscheidung zwischen phänomenalem Bewusstsein (P-Bewusstsein) und Zugriffsbewusstsein (Z-Bewusstsein) nach Block und Burge. Auch Edelman kommt zu dem Ergebnis, dass ein Bewusstsein höherer Ordnung

431 Siebert (1999), S. 98

432 Edelman Tononi (2000), S. 72

433 Edelman (2004), S. 50

434 Dieses Zitat und die folgenden: Edelman (2004), S. 22

zwar für gewisse Zeiträume und zu bestimmten Inhalten ohne primäres Bewusstsein auskommt, aber nicht dauerhaft.

In den Kategorien des Leib-Seele-Problems vertritt Edelman eine Position, die dem biologischen Naturalismus von John Searle und dem Epiphänomenalismus entspricht:

> *„Und wie steht es mit dem Geist, der das Denken hervorbringt? Die Antwort lautet, dass dieser gleichermaßen materiell und mit einem Bedeutungsinhalt versehen ist. Für den Geist als Netzwerk von Beziehungen gibt es eine materielle Basis: Das Wirken Ihres Gehirns und all seiner Mechanismen von ganz unten bis nach ganz oben, von den Atomen bis zum Verhalten, resultiert in einem Geist, der sich mit Bedeutungen befassen kann. Während dieser Geist solche immateriellen Beziehungen schafft, die er selbst und andere Geister zu erkennen vermögen, wurzelt er dennoch gleichzeitig ganz und gar in den physikalischen Prozessen, die seinem eigenen Wirken, dem anderer Geister und all jenen Ereignissen, die Teil einer Kommunikation sind, zugrunde liegen. Es gibt keine zwei vollständig voneinander getrennten Domänen der Materie und des Geistes und keine Basis für einen Dualismus. Doch offenbar gibt es eine durch die physikalische Ordnung von Gehirn, Körper und sozialer Welt geschaffene Sphäre, durch die Bedeutung geschaffen wird.* […] *Es sind die unglaublich komplexen materiellen Strukturen des Nervensystems und des Körpers, aus denen dynamische mentale Prozesse und Bedeutung hervorgehen.“*[435]

Aus Sicht der Funktionalisten und insbesondere der Computer-Funktionalisten könnte dies eine unterstützende Sichtweise sein, vor allem wenn man die Idee der Substratneutralität verfolgt, wie es viele der KI-Enthusiasten tun. Edelman hat sich dazu aber schon in seinem ersten Werk sehr skeptisch geäußert:

> *“The second point has to do with whether artifacts designed to have primary consciousness are necessarily confined to carbon chemistry and, more specifically, to biochemistry (the organic chemical or chauvinist position). The provisional answer is that, while we cannot completely dismiss a particular material basis for consciousness in the liberal fashion of functionalism, it is probable that there will be severe but not unique constraints on the design of any artifact that is supposed to acquire conscious behavior. Such constraints are likely to exist because there is every indication that an intricate, stochastically variant anatomy and synaptic chemistry underlie brain function and because consciousness is definitely a process based on immensely intricate and unusual morphology. (This conclusion is based on our evolutionary assumption).“*[436]

Edelman vermutet („*provisional answer*“), dass die Kohlenstoffchemie in dem auf der Erde vorherrschenden und für das Leben erforderlichen Temperaturspektrum (um die 300 Kelvin) Grundvoraussetzung für das hier beschriebene Bewusstsein ist. Beweisen kann er dies hingegen nicht.

Ein weiteres für die Computer-Funktionalisten problematisches Faktum ist allerdings klar bewiesen: Unser Geist und unser Bewusstsein entstehen nicht in einer durch Algorithmen

[435] Edelman Tononi (2000), S. 300f

[436] Edelman (1989), S. 32f

gesteuerten Turing-Maschine. Eine universale Turing-Maschine kann logische Operationen nach einem Programm abarbeiten. Die Leistungsfähigkeit und vielseitige Verwendung derartiger Computer hat zu der Vermutung geführt, dass auch unser Gehirn wie eine Turing-Maschine funktionieren könnte. Letztlich war dies ein Rückschluss von der vermeintlich ähnlich leistungsfähigen Kopie auf das Original. Edelman und Tononi legen klar und umfassend dar, dass dies definitiv so nicht funktionieren kann:

> *„Jedes Gehirn ist so angelegt, dass seine Verkabelung und seine Dynamik auf der Ebene seiner synaptischen Verbindungen von kolossaler Variabilität ist. Es ist ein selektionales System und deshalb ist jedes Gehirn einzigartig. Diese Einzigartigkeit und die damit verbundene mangelnde Vorhersehbarkeit kann bei der Ausführung gewisser Operationen des Gehirns bedeutsam werden und muss im Zusammenhang mit jeder einzelnen Hirnfunktion in Betracht gezogen werden. Hinzu kommt, dass die Gehirnfunktion degeneriert ist: Bei der Auseinandersetzung mit einem zufälligen Ereignis können unterschiedlich gestaltete (nicht isomorphe) Gehirnstrukturen auf mehreren Ebenen der Konstruktion und Operation zum selben Output bzw. zur selben Funktion führen.“*[437]

Deutlicher kann man die Zurückweisung eines menschlichen Determinismus nicht formulieren. Gerade die durch die von der Evolution geschaffene Variabilität und Degeneration (bei Edelman: Degeneriertheit) verursachte Kontingenz ist die Basis für den freien Willen des Menschen:

> *“Selective systems* […] *breach the deadlock imposed by bottom-up determinism. Just as the order of nucleotides in a given gene depends on a higher order of phenotypes in the evolutionary past, so a selectionistic brain theory allows alterations of particular synaptic populations that then become stable.* […] *The molecular constraints are highly determined, but the particular historical and differential amplifications acting on a hereditary or autocorrelated system resulted in various unpredictable but definite events and structures. The building up of a system of higher-order consciousness capable of rehearsal and planning based on value-laden memory and goals adds one more level of selection capable of acting on particular synaptic subpopulations. Moreover, this selection is plastic: while it depends upon a number of internal and external events, in a degenerate system such as that postulated by the TNGS,* ***multiple choices always remain****.* [Hervorhebung DS]. *The self based on consciousness needs no homunculus and can nonetheless exercise a voluntary choice within a microdeterministic frame*.”[438]

Mit einem starren Programm von Instruktionen könnte ein biologisches Wesen nicht auf die Vielzahl und die Variabilität der nicht eindeutigen Inputs aus der Außenwelt reagieren. Die *Welt „mag zwar physikalischen Gesetzen gehorchen, aber sie verhält sich nicht wie der Lochstreifen eines Computers“*[439].

Die Konsequenz ist für die Wissenschaftstheorie und Epistemologie durchaus weitreichen:

---

[437] Edelman Tononi (2000), S. 292

[438] Edelman (1989), 261

[439] Ebd.

> *„Wenn aber das Gehirn in der Evolution auf eine solche Weise entstanden ist und diese Evolution die biologische Basis für die endgültige Entstehung und Verfeinerung logischer Systeme innerhalb der menschlichen Kultur bildet, dann können wir freilich zu dem Schluss kommen, dass – im generativen Sinne –* ***Selektion mächtiger ist als Logik*** [Hervorhebung DS]. *Selektion – natürliche und somatische – ließ Sprache und Metapher entstehen, und Selektion, nicht Logik liegt der Mustererkennung und dem Denken in Metaphern zugrunde*."[440]

Edelman schließt daraus weiterhin, dass unsere Fähigkeit der Mustererkennung derjenigen des Beweises von Lehrsätzen und Behauptungen mit Logik übersteigt, was jeder bestätigen kann, der schon einmal mit Beweisen in der schulischen Mathematik gekämpft hat. Mit unserem Bewusstsein können wir neue Axiome hervorbringen, wozu ein Computer nicht imstande ist. Abduktion liegt uns eher als Deduktion und Induktion. Außerdem sind wir fähig – und damit kommen wir zurück zu Gödel und seinen Unvollständigkeitssätzen – wahre Axiome zu erfinden, die zum Teil mit Logik nicht beweisbar sind.

5.5.2 Willensfreiheit und Determinismus im Lichte der Hirnforschung

Die Frage der Existenz des freien Willens als Bedingung für Autonomie und Menschenwürde wird an späterer Stelle noch thematisiert. Hier soll es um die empirische Sicht auf die Willensfreiheit und damit einhergehend den Determinismus gehen. Insbesondere im deutschsprachigen Raum entwickelte sich in den vergangenen zwei Jahrzehnten eine heftige Auseinandersetzung rund um die beiden Neurowissenschaftler Gerhard Roth und Wolf Singer sowie den Psychologen und Kognitionswissenschaftler Wolfgang Prinz und deren Positionen zum menschlichen Determinismus, der Willensfreiheut und der Rolle des menschlichen Gehirns.

In ihren Veröffentlichungen haben Roth und Singer drei zentrale Thesen postuliert[441]:

1. *„Die Welt ist deterministisch und also ist auch der Mensch in ihr determiniert*
2. *Nicht wir, sondern unsere Gehirne entscheiden*
3. *Die Willensfreiheit ist eine Illusion"*

Alle drei Wissenschaftler und Deterministen nehmen in ihren Publikationen Bezug auf das nach Benjamin Libet benannte Libet-Experiment[442]. Bei dem Anfang der 1980er Jahre veröffentlichten Laborexperiment wurde den Probanden eine einfache Aufgabe gestellt: Sie sollten zu einem selbst gewählten Zeitpunkt die Hand beugen und sich anhand einer umlaufenden Uhr merken, wann sie sich dazu entschieden haben. Gleichzeitig wurde mittels Elektroenzephalografie (EEG) der Zeitpunkt beginnender Hirnaktivität festgestellt. Libet und sein Team verglichen dann den Zeitpunkt der Entscheidung des Probanden

[440] Edelman Tononi (2000), S. 293

[441] Rott (2009), S. 119f

[442] Vgl. auch: https://www.spektrum.de/news/die-wiederentdeckung-des-willens/1341194

(bzw. des von ihm wahrgenommenen Zeitpunkts) und dem Zeitpunkt der beginnenden Hirnaktivität. Erwartet wurde, dass Probanden zuerst entscheiden, dann die Hirnaktivität einsetzt und schließlich die Bewegung der Hand stattfindet. Das Ergebnis fiel bei Libet und auch bei anderen Teams, die das Experiment wiederholten, jedoch anders aus: Im Durchschnitt lag die bewusste Entscheidung (gemäß der angegebenen Uhrzeit) 200 msec vor der Handlung. Die Hirnaktivität, auch Bereitschaftspotential genannt, begann jedoch schon 500 Millisekunden vor der Handlung, also 300 Millisekunden[443] vor der bewussten Entscheidung. Diese Erkenntnis schlug in der wissenschaftlichen Gemeinschaft wie eine Bombe ein und führte zu hochkontroversen Diskussionen, die bis zum heutigen Tag andauern. Die Deterministen, zu denen auch Prinz, Singer und Roth gehören, sahen darin einen (weiteren) Beweis dafür, dass es den freien Willen nicht gibt. So antwortete Wolfgang Prinz in einem Interview auf die Frage, ob die Libet-Experimente ein Hinweis für die Determinierung unseres Gehirns seien:

> *„Ja. Aber um festzustellen, dass wir determiniert sind, bräuchten wir die Libet Experimente nicht. Die Idee eines freien menschlichen Willens ist mit wissenschaftlichen Überlegungen prinzipiell nicht zu vereinbaren. Wissenschaft geht davon aus, dass alles was geschieht, seine Ursachen hat und dass man diese Ursachen finden kann. Für mich ist unverständlich, dass jemand, der empirische Wissenschaft betreibt, glauben kann, dass freies, also nichtdeterminiertes Handeln denkbar ist.“*[444]

Diese Interpretation und ihre Schlussfolgerung lassen keine Spielräume. Libet selbst hat seine Messergebnisse um einiges flexibler ausgelegt:

> *„Was bedeutet das? Erstens wird der Prozess, der zu einer Willenshandlung führt, vom Gehirn <u>unbewusst</u> eingeleitet, und zwar deutlich vor dem Erscheinen des bewussten Handlungswillens. Das bedeutet, dass der freie Wille, wenn es ihn gibt, eine Willenshandlung nicht einleiten würde.*“[445]

An einer anderen Textstelle, in einem imaginären Interview oder Gespräch mit René Descartes (RD), wird Libet noch präziser:

> *„RD: Gibt es dann noch eine Möglichkeit, wie der freie Wille eine Rolle bei der Willenshandlung spielen könnte?*
>
> *BL: Ja. Die bewusste Absicht erscheint etwa 150 msec vor der motorischen Bewegung. Das lässt genügend Zeit dafür, dass die Bewusstseinsfunktion in diesen Prozess eingreift. Der Prozess kann ein Auslöser sein, der ermöglicht, dass der Willensprozess vollendet wird: dafür gibt es jedoch keine direkten Belege. Es gibt jedoch Belege dafür, dass der bewusste Wille den Prozess stoppen oder unterdrücken kann, so dass es nicht zu einer Handlung kommt. In einem solchen Fall könnte der freie Wille das Ergebnis steuern. Das passt zu*

[443] Je nach Publikation waren es mitunter mehr als 300 msec, z.B. auch 350 msec.

[444] Prinz (2004), S. 22

[445] Libet (2005), S. 175; Hervorhebung durch den Autor des Ursprungszitats

> *unserem Gefühl, dass wir Kontrolle über uns selbst haben, etwas, das die ethischen Systeme von uns verlangen.“*[446]

Er stellt sich so etwas wie ein „bewusstes Veto“ („*‘Veto-Fähigkeit‘ des bewussten Willens*[447]“) als seine Version des freien Willens vor. Libet geht davon aus, dass es direkt „*vor dem Beginn der Körperbewegung*“ ein „*Zeitfenster von ca. 100 Millisekunden*“ gebe, währenddessen die Bewegung noch gestoppt werden könnte. Dies könne, so Libet, dann kein Ergebnis einer unbewussten Aktivierung oder Deaktivierung sein. „*Dafür sei nicht genug Zeit*“.

Wie oben schon angesprochen, hat die Publikation zu heftigen Diskussionen zum freien Willen und Determinismus geführt. Einerseits sehen die anfangs genannten Deterministen darin einen Beweis der Widerlegung des freien Willens. Andererseits wird von einigen Philosophen, z.B. Geert Keil, die Eignung des Experiments insgesamt angezweifelt[448].

Für das Bewusstsein hat Libet eine eigene Theorie entwickelt, diejenige des bewussten mentalen Feldes (BMF). Darunter versteht er eine emergente Eigenschaft des Gehirns als „*Vermittler zwischen den physischen Aktivitäten der Nervenzellen und dem Auftauchen von subjektivem Erleben“*[449]. In diesem Feld sieht er weiterhin eine „*kausale Fähigkeit, bestimmte neuronale Funktionen zu beeinflussen und zu verändern*“. Libet machte weitreichende Vorschläge, wie ein derartiges Feld empirisch nachgewiesen werden könnte, die bisher aber noch nicht realisiert wurden.

Einer von Wolf Singers Aufsätzen trägt einen provokativ deterministischen Titel: „*Verschaltungen legen uns fest: Wir sollten aufhören, von Freiheit zu sprechen*“[450]:

> *„Die in der lebensweltlichen Praxis gängige Unterscheidung von gänzlich unfreien, etwas freieren und ganz freien Entscheidungen erscheint in Kenntnis der zugrundeliegenden neuronalen Prozesse problematisch. Unterschiedlich sind lediglich die Herkunft der Variablen und die Art ihrer Verhandlung: Genetische Faktoren, frühe Prägungen, soziale Lernvorgänge und aktuelle Auslöser, zu denen auch Befehle, Wünsche und Argumente anderer zählen, wirken stets untrennbar zusammen und legen das Ergebnis fest, gleich, ob sich Entscheidungen mehr unbewussten oder bewussten Motiven verdanken. Sie bestimmen gemeinsam die dynamischen Zustände der ‚entscheidenden Nervennetze‘.“*[451]

Solch einer Position (der entscheidenden nervennetze) würden vermutlich Edelman und Libet zustimmen. Hier könnte man hineininterpretieren, dass der freie Wille im Kontext

---

[446] Libet (2005), S. 246

[447] Dieses und die folgenden Zitate: Keil (2007), S. 168

[448] Vgl. Keil (2007), S. 168f

[449] Libet (2005), S. 212

[450] Singer (2004)

[451] Singer (2004), S. 62

der eigenen Herkunft und Erziehung bei Berücksichtigung relevanter Neigungen entscheidet. Dass Singer aber genau dieses so nicht versteht, wird an anderer Stelle deutlich:

> *„Wir erfahren unsere Gedanken und unseren Willen als frei, als jedweden neuronalen Prozessen vorgängig.* […] *Wir erfahren uns als wertende, mit Intentionalität ausgestattete Wesen, die sich selbst und anderen Verantwortung zuschreiben für das, was sie tun, und wir empfinden uns in der Lage, mit unserem Gewissen in Zwiegespräche einzutreten, mit unseren kategorischen Imperativen zu argumentieren, unsere Stimmungen zu beherrschen und uns über diese Handlungsdeterminanten hinwegzusetzen.* […] *Bei alldem begleitet uns das Gefühl, dass wir es sind, die diese Prozesse kontrollieren. Dies aber ist mit den deterministischen Gesetzen, die in der dinglichen Welt herrschen, nicht kompatibel.“*[452]

Auch für Prinz stimmt es „*mit Sicherheit nicht“*, „*dass wir die freien autonomen Subjekte, für die wir uns halten, auch tatsächlich sind.“*[453] Bei Roth heißt es dann unmissverständlich: „***Es gibt keine Willensfreiheit*** (im starken Sinne)“[454].

Roth, Prinz und Singer folgern aus ihren Erkenntnissen einen signifikanten Anpassungsbedarf im Strafrecht, weil das „Anders-handeln-können“ in ihrer Sicht systematisch eingeschränkt ist.

5.5.3 Mythos Determinismus und der epistemische Libertarismus

Einige Philosophen, wie z.B. der im vorherigen Abschnitt zitierte Geert Keil, haben sich nicht nur umfassend mit dem Experiment von Libet und dessen Eignung für die Diskussion um Willensfreiheit und Determinismus, sondern auch kritisch mit den Deterministen unter den Hirnforschern auseinandergesetzt. Sie konstatieren begriffliche Verwechselungen (Kausalprinzip vs. Determinismusprinzip, diachrone und synchrone Determination[455], Tun und Herbeiführen), Fehldeutungen (Anderskönnen, physische Realisierung als freiheitsgefährdend) und Äquivokationen (Drang, Wille, Absicht und Entschluss)[456] .

Interessant und überzeugend ist die Argumentation der bereits im Zusammenhang mit der kausalen Geschlossenheit zitierten Physikerin und Philosophin Brigitte Falkenburg. In dem erwähnten Buch „*Mythos Determinismus – Wieviel erklärt uns die Hirnforschung?*“ fasst sie zum einen die philosophische Position zum Thema zusammen und steigt zum anderen sehr tief in die Befunde der Hirnforschung ein und argumentiert ergänzend auch aus der Perspektive der Physik gegen den neuronalen Determinismus und für die menschliche Willensfreiheit.

---

[452] Singer (2204), S. 36; Hervorhebungen DS

[453] Prinz (2004), S. 25

[454] Roth (2004), S. 81; Hervorhebung DS

[455] Synchron ist auf unterschiedlichen ontologischen Ebenen: mental parallel zu physisch; diachron ist auf der gleichen Ebene im zeitlichen Ablauf; z.B. bei Determinismus beim Billard

[456] Vgl. Keil (2007), S. 185

Falkenburg bestreitet nicht, dass die menschlichen kognitiven Fähigkeiten biologisch bedingt sind. Zweifelsohne ist „*„Bedingtheit' etwas ganz anderes als ‚vollständige Determination'"*[457]:

> *„Die Freiheit des Menschen realisiert sich immer in bestimmten Schranken. Diese Schranken sind vielfältiger Natur; es gibt soziale Zwänge, Erziehungseinflüsse, die Muttersprache, kulturelle Wurzeln, genetische Dispositionen, Nahrung, klimatische Bedingungen, körperliche Beeinträchtigungen und vieles mehr als Randbedingungen für unser Leben."*[458]

In ihrer Kritik am Determinismus sind folgende Argumente zentral:

1. Das neuronale Geschehen kann „*nicht strikt deterministisch"* verlaufen, da es sich dabei um „*thermodynamische Prozesse handelt, die stochastisch, irreversibel und nicht-linear*" ablaufen[459]. Es gibt zwei wichtige Beispiele für Prozesse, die diese drei Eigenschaften aufweisen: die Quantenphysik und die Thermodynamik. Damit einher geht dann auch die an anderer Stelle bereits dargestellte Widerlegung der kausalen Geschlossenheit der Natur (siehe Abschnitt 4.2.2.2).
2. Die bisherigen Erkenntnisse der Neurowissenschaften sind extrem begrenzt und beruhen im Wesentlichen auf Analogieschlüssen mit dem „*Informationsbegriff als Brücke"* und der Modellierung der „*Kortex als Computer"*, was erhebliche Grenzen hat[460]. Dies ist eine Überzeugung, die schon bei Edelman klar herausgestellt wurde.
3. „*Mentale und physische Phänomene sind inkommensurabel.*" Weder in die eine noch in die andere Richtung können die Phänomene reduziert werden. Sie unterstellt denjenigen, die es trotzdem versuchen, mereologische[461] Fehlschlüsse.
4. Allerdings bestehen kausale wechselseitige Beziehungen zwischen Geist und Gehirn, also von oben nach unten und umgekehrt.

Damit unterstützt sie die in Abschnitt 4.3.1.1 dargestellte Emergenztheorie sowie den biologischen Naturalismus von John Searle (Kapitel 4.5) und den emergentistischen Materialismus von Mario Bunge (Kapitel 4.6). Die kausale Geschlossenheit der Natur galt immer als das gewichtigste Argument gegen das Konzept der Emergenz. Dieses Argument hat Falkenburg überzeugend widerlegt. Das Bewusstsein „*steht in einer nicht-*

---

457 Falkenburg (2012), S. 388

458 Ebd.; an anderer Stelle (S. 414) schreibt sie: „*Die menschliche Freiheit liegt nicht in unbegrenzten Möglichkeiten. Sie liegt in der Fähigkeit, unter den gegebenen Bedingungen, Einschränkungen und Grenzen im Rahmen der Möglichkeiten zu handeln.*"

459 Vgl. Falkenburg (2012), S. 409

460 Dieses und folgende Zitate: Falkenburg (2012), S. 410f

461 „*Mereologie: Teilgebiet der Ontologie; Lehre von Teil und Ganzen";* Quelle: Metzler Lexikon der Philosophie

reduktiven Einheit mit unserm Gehirn, und die menschliche Freiheit in einer nicht-reduktiven Einheit zur Natur".[462]

Die Philosophin Bettina Walde hat in ihrem Freiheitsmodell des „*epistemischen Libertarismus*"[463] „*drei Bedingungen der Willensfreiheit*" formuliert:

1. „*Die Relevanzbedingung: Bewusste, mentale Zustände finden Eingang in den Bereich des Physikalischen. In der Ausformulierung als Monismusbedingung:*
   Alle mentalen Zustände (Entscheidungen, Handlungsabsichten) sind Zustände, die sich auch physikalisch beschreiben lassen - dies ist vereinbar mit ihrer epistemischen Irreduzibilität"[464]
2. *„Die Bedingung der geeigneten Determination: Freie Willensentscheidungen sind nicht nicht-determiniert, sondern auf bestimmte Weise determiniert."*[465]
3. *„Die epistemische Offenheit der Zukunft: Freie Entscheidungen sind solche, von denen die Person glaubt, dass sie auch eine andere Entscheidung hätte treffen können.*
   Voraussetzung ist der epistemische Indeterminismus aus der Perspektive der Person."

Besonders bedeutsam ist der dritte Punkt. Menschen, die davon überzeugt sind, eine Willens- und Handlungsfreiheit zu besitzen, verfügen gerade deswegen (wegen dieser Überzeugung) darüber. Der Umkehrschluss gilt ebenfalls: Die Willensfreiheit „*verschwindet, wenn Personen glauben, dass sie keine Willensfreiheit haben*"[466].

Das ist der epistemische Libertarismus: Das Wissen von der Freiheit determiniert ihre Existenz.

---

462 Falkenburg (2012), S. 414

463 Walde (2006), S. 190f

464 Walde (2006), S. 194

465 Walde (2006), S. 196f

466 Walde (2006), S. 206; Walde schreibt dazu weiterhin: „*Willensfreiheit im hier entwickelten Sinne lebt von der epistemischen Offenheit der Zukunft und der Überzeugung der Personen, zwischen Alternativen entscheiden zu können – auch wenn diese Entscheidungen mitbestimmt sind von den Wünschen und Überzeugungen der Personen.*"

5.6 Sonstige Theorien zur Erklärung des Bewusstseins aus Sicht der empirischen Wissenschaften

Eine vollständige Darstellung der in der Wissenschaft diskutierten Theorien zum menschlichen Bewusstsein ist im Rahmen dieser Arbeit nicht realistisch, könnte wahrscheinlich Thema einer eigenständigen Dissertation sein. Anil Seth und Tim Bayne haben in ihrem Paper zu den „Theories of Consciousness“[467] eine Auswahl von 22 Bewusstseinstheorien allein aus der Neurobiologie aufgelistet. Das Spektrum der Erklärungsansätze für das Zustandekommen des Bewusstseins ist breit. Dies kann als Fingerzeig dafür gesehen werden, wie weit die Wissenschaft immer noch von ihrem Ziel entfernt ist.

In diesem Abschnitt sollen drei Theorien vorgestellt werden, die interessante Diskussionsansätze aus der Wissenschaft darstellen. Sie haben entweder hypothetischen Charakter mit unzureichenden empirischen Bestätigungen oder sind komplett theoretischer Natur. Für die Überlegungen zu den Grenzen der Künstlichen Intelligenz und spezifisch den Möglichkeiten des Künstlichen Bewusstseins lohnt sich dennoch eine Beschäftigung mit diesen Ansätzen.

5.6.1 Neuronales Korrelat des Bewusstseins

Die Suche nach dem neuronalen Korrelat des Bewusstseins (englisch: Neuronal Correlate of Consciousness; NCC) beschäftigt die Hirnforschung mittlerweile seit mehr als dreißig Jahren, mit immer noch sehr begrenzten empirischen Ergebnissen.

Die Association of Scientific Study of Consciousness (ASSC) definiert das neuronale Korrelat des Bewusstseins wie folgt:

> “A neural correlate of consciousness is a *`specific system in the brain whose activity correlates directly with states of conscious experience'*”.[468]

Chalmers entwickelte dieses weiter zu:

> *“An NCC is a minimal neural system N such that there is a mapping from states of N to states of consciousness, where a given state of N is sufficient* […] *for the corresponding state of consciousness.”*[469]

Oder in den Worten von Christof Koch:

> *“Whenever information is represented in the NCC you are conscious of it. The goal is to discover the minimal set of neuronal events and mechanisms jointly sufficient for a specific conscious percept.”*[470]

---

[467] Seth Bayne (2022), siehe Anhang 3

[468] Zitiert aus Chalmers (2000), S. 18 (Seitenzahl trianguliert, da Originalaufsatz nicht verfügbar)

[469] Chalmers (2000), S. 32 (Seitenzahl trianguliert, da Originalaufsatz nicht verfügbar)

[470] C. Koch (2004), S. 16

Wenn man das System N mit dem gesamten Gehirn identifiziert, wäre dies schnell bewiesen, doch dann bräuchte man N gar nicht. Die Grundhypothese lautet also, dass nicht das ganze Gehirn mit dem Bewusstsein direkt korreliert, sondern nur bestimmte Bereiche oder bestimmte Teilprozesse. Die Ideen für diese Bereiche oder Teilprozesse sind vielfältig und sämtlich ohne abschließende empirische Beweise. Francis Krick und Christof Koch vermuteten in frühen Aufsätzen, dass es sich beim NCC um einen speziellen Typ von Neuronen handeln könnte, die im ganzen Gehirn verteilt seien. In späteren Aufsätzen machten sie die *40-Hertz Oszillation* im Klein- und Zwischenhirn dafür verantwortlich, eine Hypothese, der sich auch Wolf Singer anschloss[471].

Christof Koch hat ein ganzes Buch der Suche nach dem neuronalen Korrelat des Bewusstseins gewidmet, allerdings ohne Resultat und nur mit (zum Teil begründeten) Vermutungen[472]. Sein Schlusswort ist aufschlussreich und klingt optimistisch:

> *"Francis* [Crick] *and I aim to explain all aspects of the first-person perspective on consciousness in terms of the activity of identified nerve cells, their interconnectivities, and the dynamics of coalitions of neurons. This is a bit like playing three-dimensional chess: You must keep simultaneous track of the phenomenology of consciousness, the behavior of the organism, and the underlying neuronal events. It won't be easy, but then no truly worthwhile task ever is. We live at a unique point in the history of science. The technology to discover and characterize how the subjective mind emerges out of the objective brain is within reach. The next years will prove decisive."*[473]

Auch 2023 (19 Jahre später) liegt noch immer keine finale empirisch bestätigte Theorie für das neuronale Korrelat des Bewusstseins vor.

5.6.2 Integrierte Informationstheorie des Bewusstseins

Bei der Integrierten Informationstheorie des Bewusstseins (engl.: Integrated Information Theory of Consciousness) handelt es sich um den Versuch der substratunabhängigen theoretischen Beschreibung des Bewusstseins bei voller Berücksichtigung der zentralen Erkenntnisse aus Neurowissenschaften und Philosophie. Die Anforderungen an die Theorie lauten folgendermaßen:

---

471 Vgl. Barinaga (1990), S. 856: „*Last year Singer's research team at the Max Planck Institute for Brain Research in Frankfurt, West Germany, published a dramatic finding. While recording electrical signals from widely spaced neurons in the brains of cats, they found that the neurons tend to fire synchronous electrical impulses when responding to electrical stimuli that appear to come from the same object.* […] *But here was clear evidence of correlated firing, revealed by Singer's instruments as an oscillating wave of synchronous electrical activity with a frequency of roughly 40 hertz. Singer and some others think the osciallation may provide the answer to a fundamental puzzle: how do distant neurons responding to a single visual object pool their information to create a coherent image?"*

472 C. Koch (2004)

473 C. Koch (2004), S. 314

- Konsistenz mit der Beschreibung des Bewusstseins aus philosophischer Sicht in der kartesischen Tradition, insbesondere in Bezug auf Einheit, Direktheit und Unmittelbarkeit
- Konsistenz mit den Ergebnissen der neurowissenschaftlichen Forschung der letzten Jahrzehnte
- Unabhängigkeit von jeglichen „Substratmaterialien“ oder anderen möglichen „physischen Voraussetzungen“ für Bewusstsein

In der dritten Anforderung schwingt auch die Erwartung mit, dass das menschliche Bewusstsein substratunabhängig sein könnte und damit auch die vielbeschworene „Upload-Möglichkeit“ bestehen könnte. Hingegen ergeben sich aus der Theorie andere Hürden einer Übertragung des menschlichen Bewusstseins auf andere Medien, auf die später noch eingegangen wird.

Ursprünglich entwickelt wurde IIT von Giulio Tononi im Jahr 2004.[474] Auch der im vorherigen Abschnitt zu NCC zitierte Neurowissenschaftler Christof Koch griff das Konzept auf und publizierte zu dem Thema.[475] Als Motivation für das Nachdenken über die Integrierte Informationstheorie des Bewusstseins zitiert Koch den Philosophen Colin McGinn: „*Wie wird das Wasser des Gehirns in den Wein des Erlebens verwandelt?*“[476] Und Koch weiter:

> *„Die integrierte Informationstheorie (IIT)* […] *setzt* […] *beim Erleben an und fragt, wie Materie organisiert sein muss, damit daraus ein Geist erwachsen kann. Ist jede Art von Materie geeignet? Bergen komplexe Systeme von Materie eher ein Erleben als weniger komplexe? Was genau ist mit „komplex“ eigentlich gemeint? Ist die Tendenz in der organischen Chemie stärker als bei dotierten Halbleitern? Oder bei evolvierten Lebewesen stärker als bei konstruierten Artefakten?“*[477]

Für ihn handelt es sich bei IIT um eine „*Grundlagentheorie“*, welche die „*Ontologie (die Erforschung des Wesens des Lebendigen) mit der Phänomenologie (der Erforschung dessen, wie die Dinge erscheinen) mit der Biologie und Physik verbindet“.*

5.6.2.1 Phänomenologische Axiome

Den Einstieg in die Theorie bildet das „Cogito ergo sum“ von Déscartes. Weder kann man die Existenz des eigenen Denkens noch die Existenz des eigenen Bewusstseins leugnen. Darauf basierend formuliert Koch die fünf (phänomenologischen) Axiome der IIT, die widerspruchsfrei und konsistent sind:

> *„jedes Erlebnis existiert für sich, ist strukturiert, informativ, integriert und definitiv“*[478]

[474] Vgl. Tononi (2004)
[475] Zuletzt: C. Koch (2020)
[476] C. Koch (2020), S. 69
[477] C. Koch (2020), S. 72; sowie folgende Zitate
[478] C. Koch (2020), S. 73

Etwas ausführlicher beschrieben, handelt es sich bei diesen Axiomen um die Dimensionen der Erfahrung:

1. ***Intrinsische Existenz***
   „Consciousness is real and undeniable; moreover, a subject's consciousness has this reality intrinsically; it exists from its own perspective."[479]
   Die Bewusstseinserfahrung existiert in dem Subjekt, das etwas bewusst wahrnimmt und nirgendwo sonst. Weder an der Örtlichkeit noch an der Existenz kann gezweifelt werden. Nach Markus Gabriel besteht die intrinsische Existenz „*darin, dass etwas von sich weiß, dass es existiert*"[480]. Das erinnert sehr stark an Descartes' Cogito ergo sum.
2. ***Zusammensetzung***
   „Consciousness has composition. In other words, each experience has structure. Color and shape, for example, structure visual experience. Such structure allows for various distinctions."
   Auch die grundsätzliche Zusammensetzung der phänomenologischen Erfahrung ist unbezweifelbar. Bei Kant waren dies die sog. „Kategorien". Jeder Bewusstseinszustand hat eine Komposition und Struktur.
3. ***Information***
   „Third is the axiom of information: the way an experience distinguishes from other possible experiences. An experience specifies; it is specific to certain things, distinct from others."[481]
   Beim „Wie" der Erfahrung handelt es sich zweifellos um einen der meistdiskutierten Aspekte der phänomenologischen Erfahrung. Wie Koch betont, unterscheidet sich subjektive intrinsische Information grundsätzlich von der extrinsischen objektiven Information.
4. ***Integration***
   „Fourth, consciousness has the characteristic of integration. The elements of an experience are interdependent. For example, the particular colors and shapes that structure a visual conscious state are experienced together. As we read these words, we experience the font-shape and letter-color inseparably. We do not have isolated experiences of each and then add them together. This integration means that consciousness is irreducible to separate elements. Consciousness is unified."
   Bei jedem Erleben werden Sinneseindrücke zu einem integrierten Bild zusammengeführt. Beim Genuss eines Gerichts in einem schönen Restaurant

[479] Dieses und die folgenden vier Zitate: „Integrated Information Theory of Consciousness". In: *Internet Encyclopedia of Philosophy. Iep.utm.edu/int-info/ (18.12.2020)*

[480] Gabriel (2018), S. 214

[481] Schreibfehler des Originaltexts korrigiert

unterscheiden wir nicht zwischen den Gerüchen, den Geschmackserlebnissen, den optischen Eindrücken und den gehörten Tönen. Alles nehmen wir integriert wahr.

5. ***Ausgrenzung (Exklusion)***
   „Fifth, consciousness has the property of exclusion. Every experience has borders. Precisely because consciousness specifies certain things, it excludes others. Consciousness also flows at a particular speed."
   Hier sind gleich mehrere Merkmale des Bewusstseins enthalten: die Privatheit des Erlebens, die Ausgrenzung eines „alternativen Erlebens" und der fließende Übergang im Erlebnisstrom.

5.6.2.2 Ontologische Postulate

Beim Übergang zu ontologischen Postulaten reflektiert Koch in der gleichen Struktur darüber, welche Mechanismen die Erfüllung der Axiome überhaupt ermöglichen. Wie er und seine Mitverfechter argumentieren, muss es innerhalb des Bewusstseinsbildes eine spatio-temporale Ursache-Wirkungs-Kraft (räumlich, zeitlich und sensorisch) geben.

Koch erläutert diesen Gedanken entlang der Struktur der fünf Axiome als ***fünf Postulate***:

1. ***Intrinsische Existenz***
   Für Koch ist die kausale Kraft der intrinsischen Existenz Grundvoraussetzung ihrer selbst:
   „*Das* […] *Postulat der intrinsischen Existenz behauptet, dass jede Anordnung, jeder Satz physikalischer Elemente, um intrinsisch zu existieren, eine Reihe von* ***Unterschieden, die einen Unterschied für die Anordnung selbst machen***[482] *darlegen muss."*[483]
   Mit dieser kausalen Kraft des Bewusstseins meint er nicht zwingend die Kraft des Bewusstseins auf die äußere Welt (obwohl das nicht ausgeschlossen ist), sondern die Kraft auf das Bewusstsein selbst. Die von den unterschiedlichen Sinnesorganen aufgenommenen Eindrücke haben kausale Kraft auf das Gesamtbild. Das Bewusstseinsbild der Vergangenheit beeinflusst dasjenige der Gegenwart und hat auch Einfluss auf dasjenige der Zukunft.
2. ***Zusammensetzung***
   Das Postulat der Zusammensetzung besagt, dass sich das Erleben aus der Kombination der Inputs aller Sensoren, auch aus der Vergangenheit zusammensetzt. Er illustriert dies mit einem einfachen Schaltkreis, bei dem die Bewertungen der einzelnen Elemente (Mechanismen erster Ordnung), Kombinationen von jeweils

[482] Hervorhebung des zitierten Autors
[483] C. Koch (2020), S. 78

zwei Elementen (Mechanismen zweiter Ordnung) und des gesamten Schaltkreises (Mechanismen dritter bis n-ter Ordnung) im Zusammenspiel erforderlich sind.[484]

3. ***Information***
   Wie schon im vorherigen Abschnitt beschrieben, setzt Information voraus, dass sie einen Unterschied macht und eine diesbezügliche kausale Kraft nach innen und/oder außen wirken lässt.
4. ***Integration***
   Koch erläutert die kausale Kraft, die aus der Integration des Bewusstseinsbildes erwächst, mit einer Metapher:
   "*Angenommen, einige Bürger wollen eine Initiative gründen, um den Bau einer Autobahn durch ihr Wohngebiet zu verhindern. Wenn sie sich nie treffen, nie miteinander austauschen und ihre Aktionen nicht koordinieren, dann existiert ihre Gruppe hinsichtlich ihrer kausalen Kraft in der Lokalpolitik nicht.*"[485]
   Nur aus dem integrierten Gesamtbild des Bewusstseins erwächst die kausale Kraft. Tononi veranschaulicht dies in seinen Aufsätzen gern mit einer anderen Metapher: Die einzelnen Pixel in einem Bild eines angreifenden Tigers in einer Digitalkamera haben keinerlei kausale Kraft. Erst durch die Integration auf verschiedenen Ebenen und im Vergleich zu vorherigen Bildern kann daraus eine kausale Kraft entstehen und das vierte Postulat erfüllt werden.
5. ***Exklusion***
   Das fünfte Postulat bringt die Reduzierung der möglichen Erlebnisse auf das Minimum zu Ausdruck. Generell ist unser Erlebnisraum durch die Vielzahl der einzelnen phänomenologischen Inputs hoffnungslos überdeterminiert. Auch hier dient eine Metapher der Verdeutlichung: Eine an zwei Punkten befestigte Fahrradkette könnte unendlich viele Formen einnehmen, sie nimmt aber nur eine an. Das ist diejenige, bei der die potentielle Energie minimiert ist.
   Die philosophischere Referenz ist Ockhams Rasiermesser: „*Ursachen dürfen nie über das Notwendige hinaus vermehrt werden.*"[486]

5.6.2.3 Bewusstsein als maximal irreduzible Ursache-Wirkung-Struktur

Gemäß dieser Theorie bilden in einem bewusstseinserfüllten System alle Elemente mit möglichst vielen anderen Elementen eine Ursache-Wirkung-Struktur, die maximal irreduzibel ist. Durch die Zahl der Elemente mit einer noch größeren Anzahl von Verbindungen ergibt sich eine „*schwindelerregende Komplexität*"[487]. Es gilt:

[484] Vgl. C. Koch (2020), S. 82
[485] C. Koch (2020), S. 84
[486] C. Koch (2020), S. 85
[487] Ebd.; sowie folgendes Blockzitat

> „*Jedes Erleben ist identisch mit der maximal irreduziblen Ursache-Wirkung-Struktur, die in diesem Zustand mit dem System assoziiert ist.*“

Koch vergleicht diese maximal irreduzible Struktur mit einem Kristall:

> *„Der Kristall ist das System, von innen gesehen. Er ist die Stimme im Kopf, das Licht im Innern des Schädels. Er ist alles, was wir jemals über die Welt erfahren werden. Er ist unsere einzige Realität. Er ist die Quintessenz der Erfahrung. Der Traum des Lotusessers, die Gewahrsamkeit des meditierenden Mönchs und die Agonie des Krebspatienten fühlen sich so an, wie sie es tun, weil die jeweiligen Kristalle in einem milliardendimensionalen Raum so geformt sind – eine wahrhaft überirdische Vision.“*[488]

Das Bewusstsein ergibt sich gemäß dieser Theorie nicht aus einem Algorithmus, es wird vielmehr *von* einer *„auf sich selbst wirkenden kausale*[n] *Kraft geschaffen“*[489]. Er schließt explizit aus, dass diese kausale Kraft simuliert werden kann. Damit wendet er sich entschieden gegen den Computer-Funktionalismus und die Annahme, dass das Gehirn wie eine Turing-Maschine arbeite und daher durch eine solche kopiert werden könne. Eine Simulation kann Abläufe nachvollziehen. Die Realität zu ersetzen vermag sie nicht.

Koch räumt ein, dass nur eine einzige denkbare technische Möglichkeit bestehe, eine Maschine mit Bewusstsein zu erschaffen. Dabei handelt es sich um eine sogenannte „*neuromorphe elektronische Hardware*“. Das wäre allerdings explizit KEINE Turing-Maschine. Eine derartige Maschine wäre „*gemäß den Designprinzipien des Gehirns konstruiert*“, als Hardware und nicht als Software. Die einzelnen den Neuronen entsprechenden „*Logikgatter bekämen Input von Zehntausenden anderer Gatter“* und hätten „*Output-Verknüpfungen zu Zehntausenden anderer Gatter“.* Dies wäre grundsätzlich anders als in den heute existierenden künstlichen neuronalen Netzwerken, in denen die Logikeinheiten Input von einer „*Handvoll“* anderer Logikeinheiten verarbeiten und ihren Output an eine ähnlich überschaubare Anzahl weiterer Logikeinheiten übermitteln. Neuromorphe Computer könnten, so argumentiert Christof Koch, eine ähnliche „*intrinsische Ursache-Wirkungs-Kraft entwickeln“* wie das menschliche Gehirn und damit über Bewusstsein verfügen.

Dies wäre der physische Nachbau des menschlichen Gehirns auf anderem Substrat. Dazu sollte man sich daran erinnern, dass das menschliche Gehirn über 100 Milliarden (10^11) Neuronen und 100 Billionen Synapsen (10^14) verfügt.

Die IIT ist in ihrer Logik plausibel und schlüssig, konnte bisher aber weder theoretisch noch empirisch bewiesen werden. Sie besticht in ihrer Überzeugungskraft mit der Annahme der intrinsischen Ursache-Wirkungs-Kraft und den damit direkt verbundenen Schlussfolgerungen. Der Beweis der fünf Postulate, die auf der Basis von intuitiv wahren Axiomen entwickelt wurden, und der empirische Nachweis einer

[488] C. Koch (2020), S. 86

[489] C. Koch (2020), S. 145f (einschließlich der folgenden Zitate)

bewusstseinsschaffenden Ursache-Wirkungs-Kraft wären zwar wesentliche Schritte hin zum Verständnis der Wirkzusammenhänge des Bewusstseins und der Möglichkeiten zur Schaffung eines künstlichen Bewusstseins, aber bis dato bleibt diese Theorie eine Hypothese.

Nicht beantwortet wird in der Theorie, wie Subjektivität eines phänomenalen Bewusstseins entsteht in einer derartigen Struktur entsteht. Ein Bewusstsein ohne Subjektivität ist kein Bewusstsein.

5.6.3 Roger Penrose: Physik des Bewusstseins

An mehreren Stellen dieser Arbeit, insbesondere im Kontext des Gödel'schen Unvollständigkeitssatzes (siehe Abschnitt 2.2.1.1.) und der Church-Turing-These (siehe Abschnitt 2.2.1.3) wurde festgestellt, dass kognitive Prozesse des Menschen, vor allem „*mathematische Erkenntnisprozesse nicht-algorithmisch ablaufen"*, wohingegen nach der Überzeugung vieler Naturwissenschaftler „*Makroprozesse der klassischen Biochemie algorithmisch sind"*[490]. Diese Sicht wird nicht von allen Naturwissenschaftlern geteilt, darunter Brigitte Falkenburg, die in ihrem Buch zum Determinismus des Gehirns (siehe Abschnitt 6.4) die Thermodynamik als Erklärung für nicht-algorithmische (und nicht-deterministische) Prozesse heranzieht.

Der Physiker und jüngst mit dem Nobelpreis[491] ausgezeichnete Roger Penrose hat in den frühen 1990er Jahren eine Theorie zur Physik des Bewusstseins entwickelt, welche die Quantenmechanik für die nicht-algorithmischen Prozesse im Gehirn verantwortlich macht[492].

Seine Argumentation, die er auf mehr als 500 Seiten darlegt, lautet kurz zusammengefasst wie folgt und beginnt mit Gödel:

1. Der Gödel'sche Unvollständigkeitssatz besagt, „*dass kein formales System vernünftiger mathematischer Beweisregeln auch im Prinzip jemals ausreichen kann, um alle wahren Aussagen der üblichen Arithmetik zu beweisen* […] . *Aber aus dem Satz folgt darüber hinaus, daß sich Verständnis und Einsicht nicht auf ein System von Rechenregeln reduzieren lassen"*[493]. Mit Rechenregeln lassen sich Kreativität, Verständnis, Einsicht und menschliche Intuition also nicht nachvollziehen.
2. Das menschliche Denken ist „*stimmig, ohne algorithmisch zu sein"*[494]. Damit ist es auch nicht „computational", also kann es nicht durch einen Computer mit

[490] Crush Churchland (1995), S. 223

[491] 2020 für Forschung über Schwarze Löcher

[492] Penrose (1994a)

[493] Penrose (1994a), S. 82; dieses Zitat und die folgenden

[494] Crush Churchland (1995), S. 224

einem Algorithmus ersetzt werden. Penrose dazu: „*Kein erkennbarer, rechnerisch noch so gut abgesicherter Mechanismus kann korrektes menschliches mathematisches Schließen komplett umfassen*“[495]. Dies war auch die zentrale Erkenntnis von Edelman (siehe Abschnitt 6.2).

3. „*Bewusstsein* [kann] *sich nur dann einstellen, wenn im Gehirn gewisse nicht-rechnerische Vorgänge ablaufen.“*[496] Penrose stützt das dazugehörige Argument auf das mathematische Verstehen und dabei insbesondere auf das Verständnis der Eigenschaften natürlicher Zahlen und des Begriffes „Zahl“. Beides lässt sich nicht mit rechnerischen Methoden abbilden. Ein Computer kann mit natürlichen Zahlen operieren, sie addieren, multiplizieren und beliebig kombinieren. Er hat jedoch kein Verständnis vom Begriff der Zahl. Auch diese Schlussfolgerung ist konsistent den Gegnern des Computer-Funktionalismus, wie z.B. John Searle und Mario Bunge.
4. Daraus schließt Penrose: „*Bewusstes menschliches Denken beruht zumindest manchmal, vielleicht sogar immer, auf Prinzipien, die unser heutiges physikalisches Verständnis überschreiten, auch wenn sie nicht prinzipiell jedes (z.B. zukünftige) wissenschaftliche, physikalische Verständnis überschreiten.*“[497]. Dies ist letzten Endes der mathematisch-physikalische Ausdruck des Leib-Seele-Problems.
5. In einem gewagten weiteren Schritt postuliert es, dass es „*zukünftige physikalische Theorien“* geben müsse, die „*nicht-algorithmische, physikalische Prozesse“* beinhalten.
6. Der aus seiner Sicht einzige Kandidat dafür ist die Quantenmechanik und spezifisch der Kollaps der Wellenfunktion (als Teil der Theorie der Quantengravitation).
7. In einem abschließenden Argumentationsschritt vermutet er quantenmechanische Prozesse auf subneuronaler Ebene innerhalb der Nervenzellen, spezifisch in den sogenannten Mikrotubili[498]: „*Mikrotubili stellen ein Fenster für nicht-algorithmische Prozesse der menschlichen Kognition dar, da sie sowohl mit quantenmechanischen Prozessen als auch mit Bewusstseinsprozessen etwas zu tun haben.*“ Und

---

[495] Penrose (1994a), S. 224

[496] Penrose (1994a), S. 272

[497] Dieses und die folgenden Zitate: Crush Churchland (1995), S. 225

[498] „***Mikrotubuli*** *[von *mikro- , latein. tubuli = kleine Röhren], Cytotubuli, Gruppe von Filamenten, die am Aufbau des Cytoskeletts (Zellskelett), der Geißeln und Cilien und bei der Ausbildung der Spindelapparate bei der Zellteilung (Cytokinese) beteiligt sind. Mikrotubuli sind röhrenförmige Strukturen mit einem Gesamt-Durchmesser von 25 nm (Durchmesser innen: 15 nm); sie können bis zu 100 µm lang werden.*“, Quelle: Lexikon der Biologie, Online: https://www.spektrum.de/lexikon/biologie/mikrotubuli/43021

> weiter: „*Quantengravitation oder etwas sehr Ähnliches, muss – durch Mikrotubili vermittelt – eine zentrale Rolle für Bewusstsein und Kognition spielen*.“[499]

Insgesamt liest sich das Buch wie der verzweifelte Versuch, die von Gödel aufgezeigte Lücke mithilfe von Mechanismen der Quantenmechanik und der Quantengravitation zu schließen. Die Einwände verschiedener Philosophen und Physiker waren vernichtend. Rick Crush und Patricia Smith Churchland, die in der obigen Zusammenfassung mehrfach zitiert wurden, haben einige überzeugende Argumente gegen die Theorie von Penrose herausgearbeitet. Hier sind die gewichtigsten[500]:

- *Sie bestreiten, dass nicht-algorithmische Prozesse für das Verständnis des menschlichen Bewusstseins notwendig oder hinreichend sind.*
- *Die Charakterisierung der Quantengravitation als nicht-algorithmischen Prozess halten sie für spekulativ.*
- *Und für noch spekulativer halten sie die Vermutung, dass es Quantenphänomene in den Mikrotubili geben könnte.*

Das Schlusswort ist gleichermaßen verheerend:

> *„Nichts von dem, was wir hier gesagt haben, zeigt die Falschheit der Quanten-Bewusstsein-Verbindung. Unsere Meinung ist lediglich, dass für diese Verbindung genauso wenig spricht, wie für Abermillionen anderer Pfeifenträume und Luftblasen der wissenschaftlichen Phantasie.“*[501]

Die Quantentheorie und die Möglichkeit von Quantencomputern werden immer wieder von namhaften Wissenschaftlern ins Gespräch gebracht, um einerseits weitere Ansätze für das Verständnis des Zustandekommens des menschlichen Bewusstseins zu finden und andererseits Möglichkeiten der Schaffung von künstlichem Bewusstsein im Gesamtkontext der KI zu diskutieren. Tatsächlich stecken die Forschungsansätze ähnlich fest wie diejenigen zur Findung des neuronalen Korrelats des Bewusstseins.

[499] Crush Churchland (1995), S. 226

[500] Crush Churchland (1995), S. 246

[501] Crush Churchland (1995), S. 246

5.7 Schlussgedanken zum Bewusstsein

> *„Es ist nämlich ganz gewiß, daß wir die organisierten Wesen und deren innere Möglichkeit nach bloß mechanischen Prinzipien der Natur nicht einmal zureichend kennen lernen, viel weniger uns erklären können; und zwar so gewiß, daß man dreist sagen kann, es ist für Menschen ungereimt, auch nur einen solchen Anschlag zu fassen, oder zu hoffen , daß noch etwa dereinst ein Newton aufstehen könne, der auch nur die Erzeugung eines Grashalms nach Naturgesetzen, die keine Absicht geordnet hat, begreiflich machen werde: sondern man muß diese Einsicht den Menschen schlechterdings absprechen.“*[502]
>
> Immanuel Kant, Kritik der Urteilskraft, § 75, 1790

„*Factum brutum*“[503] nennt Christian Siebert seine Schlussfolgerung am Ende seiner umfangreichen Dissertation über Qualia, nämlich die Erkenntnis, dass wir als Menschen die Erklärungslücke zwischen dem, was wir mit der Physik und den benachbarten Disziplinen und mit unserer subjektiven Innenperspektive nicht schließen können, auch nicht mit der Philosophie des Geistes.

Wir können das Phänomen des menschlichen Bewusstseins beschreiben und Modelle der inneren Strukturen und der Wirkzusammenhänge entwickeln. Einige dieser Modelle sind für die Psychologie, Medizin oder auch für die KI-Forschung hilfreich. Trotzdem sind wir nach wie vor weit von einem vollständigen Modell unseres Bewusstseins entfernt. Damit sind auch alle Ideen des Nachbaus oder der unabhängigen Schaffung eines künstlichen Bewusstseins illusorisch.

Wie Siebert in seiner Arbeit mutmaßt, sind wir vermutlich kategorisch nicht in der Lage, unser eigenes Bewusstsein zu verstehen. Aufgrund dieser definitiven Grenze unserer Erkenntnis scheint es ausgeschlossen, dass es jemals gelingen wird, ein künstliches Bewusstsein zu entwickeln.

Siebert schreibt dazu in den Schlussgedanken seiner Dissertation:

> *„Nur, weil wir über eine* […] *Innenperspektive verfügen, sind wir in semantisch gehaltvollen Zuständen; nur weil wir letztere haben, sind wir in der Lage, uns mit dem interpretierend auseinanderzusetzen, was uns in der Welt begegnet. Uns braucht dabei nicht weiter zu verwundern, daß unser Blick auf uns selbst schon methodologisch durch uns selbst verstellt ist. Als Wesen, die einen Blickwinkel auf die Welt einnehmen können, die ihre Umwelt überhaupt erst als Welt verstehen können, haben wir verschiedene Praktiken ausgebildet, um mit dem umzugehen, was uns begegnet. Das macht unsere Lebensform aus, hinter der wir nicht zurücktreten können und deren Teil letztlich auch die Naturwissenschaften sind. Daher führt der Versuch, unsere Lebensform ihrerseits in ihrer Gesamtheit naturwissenschaftlich zu erklären, notwendig in Paradoxa.* […] *Wenn wir so tun, als seien wir uns völlig durchsichtig, verlieren wir uns aus den Augen.“*[504]

---

[502] Kant (1790), S. 352; B338/A334; Kritik der Urtheilskraft, AA V 400

[503] Siebert (1998), S. 191

[504] Siebert (1998), S. 193f

Siebert hat das erkenntnistheoretische Grundproblem der Bewusstseinsforschung sehr gut beschrieben. Unsere eigene Intelligenz und unser eigenes Bewusstsein werden niemals für das Verständnis desselbigen ausreichen. Dafür bräuchte es vermutlich die nächsthöhere Kategorie von Intelligenz und Bewusstsein.

Damit liegt er sehr nahe bei Immanuel Kant von 1790 (siehe Eingangszitat dieses Abschnitts), dass es ganz gewiss sei, „*daß wir die organisierten Wesen und deren Möglichkeit nach bloß mechanischen Prinzipien der Natur nicht einmal zureichend kennen lernen, viel weniger uns erklären können*."[505]

Weil dies aber so ist, werden wir auch niemals in der Lage sein, eine künstliche Intelligenz zu entwickeln, die unserer eigenen Intelligenz nahekommt und unser eigenes Bewusstsein dupliziert oder repliziert.

[505] Kant (1790), S. 352; B338/A334; Kritik der Urtheilskraft, AA V 400

6 Fazit 1: Grenzen der KI aus der deskriptiven Betrachtung

In diesem Kapitel werden die bisherigen Erkenntnisse zusammengefasst. Darauf folgen eine fokussierte Diskussion eingangs vorgestellter Irrtümer und eine Überleitung zum weiteren Diskurs.

6.1 Zusammenfassung der epistemischen und ontologischen Basis

Die Künstliche Intelligenz hat ihren Ursprung in den 1930er bis 1950er Jahren. In den Anfangsphasen gab es vier wichtige Meilensteine, die bis heute für die Entwicklung und Wahrnehmung der KI prägend waren:

- Die Arbeiten von McCulloch, Pitts und Hebb an Modellen für künstliche Neuronen
- Turings Überlegungen zum Denken von Maschinen, das den Maßstab für den Vergleich der maschinellen und natürlichen Intelligenz setzen sollte
- Gödels Unvollständigkeitssätze, welche die grundsätzlichen Grenzen der Technologie aufzeigen sollten
- Die Dartmouth Konferenz von 1956, an der einige der späteren Vordenker der KI teilnahmen und bei der der Begriff der „Künstlichen Intelligenz“ erstmalig prominent positioniert wurde

In den folgenden Jahrzehnten wechselten sich immer wieder Phasen der relativen Euphorie (KI-Sommer) und der Ernüchterung (KI-Winter) ab. Der große Sprung in die breitere Anwendung der Technologie hat erst in den ersten beiden Jahrzehnten dieses Jahrtausends stattgefunden.

Dazu haben einige technologische Innovationen zum Durchbruch der Technologie beigetragen: z.B. die Umsetzung der Prädikatenlogik, das probalistisches Schließen mit Bayes-Netzwerken und Expertensysteme. Der eigentlich wichtige Fortschritt wurde allerdings mit den Künstlichen Neuronalen Netzen (KNN) erreicht, die Machine Learning und Deep Learning in Kombination mit Big Data ermöglichten.

Das maschinelle Lernen entwickelte sich vom überwachten Lernen („supervised Learning“ zum unüberwachten Lernen („unsupervised Learning“) und verstärkenden Lernen („reinforcement learning“). Dies bewirkte große Fortschritte in der Text- und Bildverarbeitung, insbesondere in der Gesichts- und Schrifterkennung. Auch viele Anwendungen in der Wissenschaft konnten darauf basierend umgesetzt werden. Begünstigend waren die drastische Steigerung der Rechnerleistungen und die Verfügbarkeit von immer größeren

Datenmengen, so dass mit Big-Data-Analysen neue Erkenntnisse gewonnen werden konnten.

In jüngster Vergangenheit konnten beeindruckende Fortschritte im Bereich der generativen KI (generative AI) erzielt werden. Auf Basis der statistischen Analyse großer Mengen von Rohdaten (Texten, Bildern, Videos) sind derartige Applikationen in der Lage, neue Texte, Bilder oder gar Videos zu generieren, die von menschlichen Artefakten und Schöpfungen kaum zu unterscheiden sind. Besondere Aufmerksamkeit erregte das sogenannte „Large Language Model" ChatGPT von OpenAI, das in kurzer Zeit vielen Millionen Nutzern gestattete, mit der Technologie zu experimentieren, und zunehmend auch im Alltag von Wirtschaft und Wissenschaft genutzt wird. Trotz der beeindruckenden Qualität der erstellten Texte ist auch hier klar zu konstatieren, dass die Algorithmen der KI auf Basis von statistischen Analysen von vorgegebenen Texten den wahrscheinlich besten Output ausgibt. Von einem Verständnis der Bedeutung der Texte kann keinesfalls ausgegangen werden[506].

Für die weitere philosophische Betrachtung ist es wichtig festzuhalten, dass die Künstliche Intelligenz bis zum heutigen Tag in Turing Maschinen streng deterministisch abläuft. Nicht in allen Fällen sind die deterministischen Abläufe im Nachhinein nachzuvollziehen, da sich interne Parameter im Zeitverlauf und erfahrungsbasiert verändern können. Die Technologie wird nach dem Kybernetiker Heinz von Foerster nichttrivial[507]: synthetisch determiniert, allerdings analytisch nicht determiniert, vergangenheitsabhängig und nicht prognostizierbar. Daraus ergibt sich ein Black-Box-Problem insofern, dass eine Zuweisung von Verantwortung, z.B. bei Funktionsstörungen oder Unfällen und Schadensfällen, nicht oder nur sehr eingeschränkt möglich ist.

Die Fantasievorstellung von der Entwicklung einer KI, die vollkommen die menschliche Intelligenz abbildet, also einer „Artificial General Intelligence", bleibt bestehen und wird von einigen Experten weiterhin für sehr realistisch gehalten. Dies ist jedoch höchst umstritten und zweifelhaft.

Im Rahmen einer Diskussion der Künstlichen Intelligenz lohnt sich eine Beschäftigung mit dem Original, der menschlichen natürlichen Intelligenz. Die Psychologie hat sich insbesondere im letzten Jahrhundert eingehend damit befasst. Eine vollständige, in sich geschlossene und allgemein als endgültig anerkannte Theorie der menschlichen Intelligenz liegt indes nicht vor.

Das Spektrum der bestehenden Theorien zeigt allerdings die breit gefächerte Vielfalt der von der menschlichen natürlichen Intelligenz abgedeckten Bereiche, die weit über die

[506] Vgl. Bender Koller (2020), S. 5185

[507] Vgl. Kaminski Gelhard (2014), S. 66

kognitiven Kernbereiche hinausgehen. Hier sind insbesondere transformative Kreativität, moralische Urteilskraft und Ethik, Werte und Spiritualität zu nennen.

Es existiert keine Theorie, die das alles abdeckt, und somit besteht auch keine Erwartung, dass eine KI dies jemals duplizieren könnte. Die Geistesphilosophie hat die KI von Beginn an begleitet. Zwei Fragen sind dabei von höchster Bedeutung, die Jaegwon Kim sehr treffend zusammengefasst hat:

- *„Wie kann der Geist seine kausalen Kräfte in einer Welt ausüben, die fundamental physikalischer Natur ist?*
- *Wie kann so etwas wie ein Bewusstsein in einer physikalischen Welt existieren, einer Welt, die letztlich aus nichts anderem besteht als kleinsten Materieteilchen, die über die Raumzeit verteilt sind und sich in Übereinstimmung mit den physikalischen Gesetzen verhalten?“*[508]

Die erste Frage wird in der Darstellung des Leib-Seele-Problems thematisiert und die zweite in den Debatten zum Bewusstsein des Menschen.

Seit frühester Zeit beschäftigen sich Philosophen mit dem menschlichen Geist. Dreh- und Angelpunkt der geistesphilosophischen Diskurse zum Leib-Seele-Problem ist der von René Descartes begründete Dualismus in Abgrenzung zu einer monistischen Identitätstheorie.

Die verschiedenen Denkansätze im Dualismus und im Monismus sind herausgearbeitet. Der Dualismus wird klar zurückgewiesen und ein Monismus präferiert, der sehr stark die Argumentation von Immanuel Kant, John Searle und Mario Bunge aufgreift, wobei die beiden letztgenannten Philosophen auch eine klare Perspektive des Realismus zur Künstlichen Intelligenz mit einbringen.

Die vier Grundmodelle des Substanzdualismus – interaktionalistischer Dualismus, Parallelismus, Okkasionalismus und Epiphänomenalismus – können mangels empirischer Evidenz zurückgewiesen werden. Entweder wäre die Welt überbestimmt oder es wären zusätzliche externe Entitäten wie z.B. göttliche Kräfte erforderlich.

Die große Hürde für viele monistische Überlegungen zum Thema stellte die Forderung nach der kausalen Geschlossenheit des Physischen dar. Mit einer sehr überzeugenden Argumentation begründet Brigitte Falkenberg den Verzicht auf diese Randbedingung.

Wenn man die Existenz mentaler Entitäten als gegeben ansieht, bestehen zwei Typen von Physikalismus, der nichtreduktive und der reduktive Physikalismus. Nichtreduktiv wäre die Emergenztheorie oder die Supervenienztheorie. Das Problem bei beiden Theorien ist die abwärtsgerichtete Verursachung, die zu einer Überdeterminierung der geschlossenen physischen Welt führt.

[508] Übersetzung aus Metzinger (2007), S. 11f; ursprüngliche Quelle: *Physicalism or Something Near Enough* (Princeton und Oxford 2005: S. 7)

Der reduktive Physikalismus ist die Zurückführung des Mentalen auf das Physische. Eine wichtige Ausprägung des reduktiven Physikalismus ist der Funktionalismus, insbesondere der Computer-Funktionalismus, bei dem mentale Zustände als funktionale Zustände in der Software ausgeführt werden. Verschiedene Gedankenmodelle widerlegen ihn.

Immanuel Kant hat sich nicht mit der Lösung des Leib-Seele-Problems beschäftigt, sondern es insgesamt als nicht existent betrachtet. Er führt die Ungleichartigkeit des Seelischen und des Körperlichen auf die Unterscheidbarkeit von innerem und äußerem Sinn zurück. Allerdings kann auch er die Frage nicht beantworten, wie der innere Sinn zustande kommt.

Überzeugend ist der biologische Naturalismus von John Searle, der alle wichtigen Randbedingungen aus der dualistischen Betrachtungsweise miteinander in Einklang bringt. Mit dieser Theorie wendet sich Searle gegen den Dualismus und auch gegen den Monismus in der Form des Physikalismus (reduktiv, nicht reduktiv oder eleminativ) oder Funktionalismus. Er vermeidet die unnötige Verdoppelung der Wirklichkeit, die zu den beschriebenen Schwierigkeiten führt, und die Probleme der Identitätstheorien. Für ihn existiert nur eine Wirklichkeit, und zwar diejenige der Naturwissenschaften. Hier schuf er Platz für die Begriffe Geist, Bewusstsein, Intentionalität, qualitative Phänomene und Ich-Perspektive.

Auch die Behandlung der Frage nach dem menschlichen Bewusstsein beginnt Searle mit einer philosophischen Charakterisierung und Strukturierung; dann beschäftigt er sich mit der neurobiologischen Perspektive auf das Thema und diskutiert schließlich (Denk-)Ansätze zur Naturalisierung des Bewusstseins außerhalb des menschlichen Gehirns, um sie letztlich allesamt zu verwerfen.

Der etwas früher entwickelte emergentistische psychoneurale Monismus von Mario Bunge zielt in die gleiche Richtung. Auch er schließt die (törichte) Reduktion (Verdinglichung) des Mentalen aus. Für Bunge gehören Computer zu den hergestellten Dingen; sie bilden ontologisch eine andere Kategorie als die vorgefundenen Dinge, zu denen das menschliche Gehirn gehört.

Nach Thomas Metzinger bildet das „*Problem des Bewusstseins zusammen mit der Frage nach der Entstehung des Universums die äußerste Grenze menschlichen Strebens nach Erkenntnis“*[509]. Vier Merkmale – so Metzinger weiter – beschreiben es: Transparenz, Perspektivität, Präsenz und Einheit. Mit der Transparenz ist das direkte und unverfälschte Erleben phänomenaler Zustände gemeint. Die Perspektivität bezieht sich auf die subjektive Innensicht. Präsenz bedeutet die zeitliche Wahrnehmung der Welt in der Gegenwart. Und bei der Einheit handelt es sich um die Integration aller Sinneseindrücke zu einem Bild.

[509] Metzinger (1995), S. 15

Das Bewusstsein hat zwei wichtige Bestandteile: Intentionalität und Qualia. Bei der Intentionalität handelt es sich um psychische Akte, die auf reale oder ideale Ziele gerichtet sind. In anderen Worten handelt es sich hierbei um Gedanken, die sich sowohl auf real existierende Objekte als auch auf nichtexistierende Entitäten richten können. Die Intentionalität lässt sich nicht naturalisieren. Man kann sie zwar – wie von Fred Dretske angeregt – informationstheoretisch verstehen oder interpretieren, allerdings bleiben offene Fragen bezüglich der Semantik und Bedeutung von intentionalen Akten.

Qualia beschreiben ungerichtete Sinneseindrücke, wie Wärme, Schmerz, Licht, Gerüche und Süße. Sie sind quasi die *„Atome des Bewusstseins“* oder auch *„phänomenale Eigenschaften erster Ordnung“*[510]. Nach Daniel Dennett besitzen sie vier Eigenschaften zweiter Stufe: sie sind *„unaussprechlich“*, *„intrinsisch“*, *„privat* und *„dem Bewusstsein unmittelbar und direkt zugänglich“*[511].

Die beiden Philosophen Ned Block und Tyler Burge entwickelten eine Theorie, mit der sie eine Brücke zwischen den empirischen Neurowissenschaften und der Geistesphilosophie schlagen. Sie unterscheiden zwischen zwei Arten des Bewusstseins, dem phänomenalen Bewusstsein (P-Bewusstsein) und dem Zugriffsbewusstsein (Z-Bewusstsein). Mit ihrer Theorie bilden sie auch die sonstigen angesprochenen Dualismen ab: subjektiv (Z) – objektiv (P); intrinsisch (P) – extrinsisch (Z); nicht-nicht korrigierbar (P) – korrigierbar (Z) usw. Eine Naturalisierung erscheint nur für einen Teil des Zugriffsbewusstseins realistisch.

In den Neurowissenschaften besteht schon seit langer Zeit die Ambition, das Bewusstsein im menschlichen Gehirn zu lokalisieren und dessen Funktionsweise zu verstehen. In der Tat ist man trotz großer Anstrengungen und beachtlicher Forschungsbudgets immer noch weit von den selbstgesteckten Erkenntniszielen entfernt.

Es wurden einige vielversprechende Theorien formuliert, die allerdings auch (noch) nicht empirisch bestätigt werden konnten. Plausibel erscheint die Theorie der Selektion neuronaler Gruppen von Gerald Edelman, die sich sehr stark auch an den Mechanismen der Evolution orientiert. Eine Reihe interessanter Implikationen ergibt sich aus dieser Theorie, insbesondere diejenige, dass Edelman einen Nachbau des menschlichen Gehirns auf einem anderen Substrat für ausgeschlossen hält und auch klar verneint, dass das menschliche Gehirn eine durch Algorithmen gesteuerte Turing-Maschine sei.

Eine Gruppe von Neurowissenschaftlern und Psychologen um Gerhard Roth, Wolf Singer und Wolfgang Prinz haben mit ihren Thesen zum menschlichen Determinismus und der Illusion der Willensfreiheit Aufsehen erregt. Überzeugend ist hingegen die sorgfältig

[510] Metzinger (2006), S. 57

[511] Vgl. Dennett (2006), S. 210f

entwickelte Gegenposition der Physikerin und Philosophin Brigitte Falkenburg. Ihre Schlussfolgerung soll an dieser Stelle noch einmal wiederholt werden:

> **Das Bewusstsein „*steht in einer nicht-reduktiven Einheit mit unserm Gehirn, und die menschliche Freiheit in einer nicht-reduktiven Einheit zur Natur*“.**[512]

Dies bestätigt auch Bettina Walde mit ihrem Freiheitsmodell des epistemischen Libertarismus. Das Wissen um unsere Willensfreiheit determiniert deren Existenz.

Weitere Erklärungsansätze für das Bewusstsein aus den Neurowissenschaften (Neuronales Korrelat des Bewusstseins), der Informationstheorie (Integrierte Informationstheorie des Bewusstseins) und der Physik (Physik des Bewusstseins) wurden im Rahmen dieser Arbeit gesichtet und zusammengefasst. Sie weisen jeweils interessante Komponenten auf, sind aber empirisch nicht bewiesen.

Nach dieser recht umfassenden Erhebung der theoretischen und empirischen Arbeiten in der Geistesphilosophie, der Neurobiologie und anderen Disziplinen lässt sich festhalten, dass es keine bewiesene Theorie des Bewusstseins gibt, jedoch durchaus eine Vielzahl von Indizien, die den auf der Turing-Maschine beruhenden Computerfunktionalismus als Erklärungsansatz für das menschliche Bewusstsein ausschließen.

Es zeichnet sich ab, dass wir vermutlich gar nicht in der Lage sind bzw. sein werden, unser eigenes Bewusstsein zu verstehen. Wir können nicht gleichzeitig eine subjektive Innenperspektive innerhalb unseres Bewusstseins und eine objektive Außenperspektive auf uns selbst einnehmen. Dazu schreibt Siebert:

> ***„Wenn wir so tun, als seien wir uns völlig durchsichtig, verlieren wir uns aus den Augen.“***[513]

Ein vollständiges Verständnis unserer selbst wird niemals möglich sein, und damit wird die von uns entwickelte KI uns zwar immer näherkommen, uns jedoch niemals erreichen.

[512] Falkenburg (2012), S. 414

[513] Siebert (1998), S. 194

6.2 Irrtümer

Kapitel 2 bis 6 zielen darauf ab, die für die Belange dieser Arbeit erforderliche epistemische Basis und die Historie der relevanten Disziplinen darzustellen. Dabei sollten insbesondere die sieben zentralen Irrtümer zur KI, die in der Einleitung eingeführt wurden, widerlegt werden:

Argument:

Verschiedene Irrtümer zur KI vernebeln den Blick auf diese Technologie und deren Potenziale, Schwächen und Risiken, sie postulieren folgende Annahmen:

I. Die Wirkmechanismen der KI ermöglichten eine Intelligenz, die der natürlichen Intelligenz des Menschen vergleichbar oder überlegen sei.
II. Die KI sei aus sich selbst heraus kreativ und könne eigenständig wissenserweiternd wirken.
III. In der KI lasse sich eine Willensfreiheit darstellen; Roboter und Agenten der KI verfügten über einen „freien Willen".
IV. Einige einflussreiche Vertreter der Hirnforschung behaupten weiterhin, es gebe den freien Willen des Menschen nicht, der Determinismus sei wahr.
V. Der Computer-Funktionalismus sei die plausible Lösung des Leib-Seele-Problems.
VI. Das menschliche Bewusstsein mit seinen wichtigen phänomenalen und intentionalen Komponenten sei „naturalisierbar".
VII. Die Hirnforschung besitze ein klares und detailliertes Verständnis vom Zustandekommendes menschlichen Bewusstseins und liefere konkrete Ansätze für den Nachbau des menschlichen Gehirns auf einem anderen Substrat. Das menschliche Gehirn sei ein Computer und funktioniere wie ein solcher.

In den folgenden Abschnitten werden diese „Irrtümer" anhand der erworbenen Erkenntnisse widerlegt.

Irrtum I:

Die Wirkmechanismen der KI ermöglichten eine Intelligenz, die der natürlichen Intelligenz des Menschen vergleichbar oder überlegen sei.

Überzeugende Argumente gegen die Annahme einer der natürlichen Intelligenz ebenbürtigen oder gar überlegenen Maschinenintelligenz liegen seit den frühen Entwicklungsphasen dieser Technologie vor, nämlich die Gödelschen Unvollständigkeitssätze. Sie definieren die Grenze zwischen dem, was in sich geschlossen und beweisbar ist, und dem, was nur mit der menschlichen Intelligenz und Kreativität geschaffen werden kann, zwar gilt und wahr ist, aber nicht bewiesen und abgeleitet werden kann. Dies gilt für die Mathematik und alle Wissenschaften, einschließlich der Philosophie. Letzteres soll in den folgenden Überlegungen noch einmal ausgeführt werden.

Wilhelm Vossenkuhl hat sich ausführlich mit dem Begriff der „Geltung“ als normativem Konzept beschäftigt, indem er zwischen zwei Typen von Tatsachen unterscheidet, die als solche wirksam und anerkannt sind: die *„unabgeleitete Geltung“* und die *„abgeleitete Geltung*“[514]:

> *„Ich spreche deswegen von ‚unabgeleiteter Geltung‘ im Unterschied zu allem, was daraus argumentativ und in Verfahren abgeleitet und begründet werden kann. „Unabgeleitet“ bedeutet, dass es keine weiteren allgemeineren Grundlagen für den Anspruch gibt, dass etwas gilt. Alle theoretischen und praktischen Prinzipien gelten unabgeleitet.“*[515]

Was unabgeleitet ist und bedingungslos gilt, kann auch nicht bewiesen oder begründet werden. Beispiele für derartige bedingungslose und unabgeleitete Tatsachen sind etwa Kodizes wie „Die Würde des Menschen ist unantastbar“, „Du sollst nicht töten“ oder das Widerspruchsprinzip. Solche unabgeleiteten und bindungslosen Geltungen können zwar Eingangsgrößen für Algorithmen sein, aber niemals durch Algorithmen und damit durch Maschinen formuliert werden. Es ist durch Gödel bewiesen, dass mathematische Wahrheiten existieren, die aus einem definierten Axiomenraum heraus nicht bewiesen, aber eben auch nicht widerlegt werden können. Dies gilt auch für normative Wahrheiten. Nur die bewusstseinsgestützte menschliche Intelligenz oder – präziser – die menschliche Urteilskraft kann dies leisten.

Irrtum II:

Die KI sei aus sich selbst heraus kreativ und könne eigenständig wissenserweiternd wirken.

Drei Typen von Kreativität lassen sich unterscheiden: explorative, kombinatorische und transformative Kreativität. Mit den Algorithmen der KI lassen sich die explorative Kreativität, zum Beispiel in Schachcomputern, und in Ansätzen auch die kombinatorische Kreativität simulieren. Umfänglich wissenserweiternd ist nur die transformative Kreativität. Mit ihr entsteht bewusst jenseits des etablierten Wissens „aus dem Nichts heraus“ neues Wissen. Transformative Kreativität wird aus Sicht der Wissenschaft mit Algorithmen der KI nicht simulierbar sein, denn dann müsste die zu findende Innovation schon vom Programmierer im Algorithmus vorgesehen sein. Echte Kreativität erfordert den freien Willen. Ihn zu programmieren wäre ein Widerspruch in sich.[516]

Die drei logischen Schlussformen Deduktion, Induktion und Abduktion können mit Hilfe der Algorithmen der KI simuliert werden. Jedoch ist der Beitrag zur Wissenserweiterung dabei nur eingeschränkt. Bei der Deduktion entsteht kein neues Wissen, bei der Abduktion ergeben sich lediglich Hypothesen, die mit den Methoden der Induktion falsifiziert werden können. Echte Wissenserweiterung findet nicht statt. Trotzdem kann die KI mit

[514] Vossenkuhl (2021), S. 22f

[515] Ebd.

[516] Vgl. Du Sautoy (2019), S. 282

all ihren Methoden und Werkzeugen den Menschen bei der Wissenserweiterung unterstützen. Auch zukünftig sind Probleme, Gedanken(-gänge), Theorien und von Menschen entwickelte Experimente unverzichtbar in der Epistemologie[517]

Irrtum III:

In der KI lasse sich eine Willensfreiheit darstellen; Roboter und Agenten der KI verfügten über einen „freien Willen".

Die Künstliche Intelligenz besitzt keinen freien Willen. Sie ist synthetisch vollständig determiniert, nicht immer analytisch determiniert und damit auch oftmals nicht prognostizierbar oder vergangenheitsunabhängig. Für die KI gibt es aber an keiner Stelle Verzweigungen, an denen sie frei wählen könnte, entweder links oder rechts abzubiegen; jede Kontingenz wird mit den Methoden der Wahrscheinlichkeitsrechnung überwunden: sie verfügt über keinen freien Willen.

Irrtum IV:

Einige einflussreiche Vertreter der Hirnforschung behaupten weiterhin, es gebe den freien Willen des Menschen nicht, der Determinismus sei wahr.

Einige einflussreiche Philosophen, Hirnforscher und Psychologen vertreten die Hypothese, das menschliche Gehirn sei in ähnlicher Weise determiniert wie klassische Computer und der freie Willen des Menschen ein Mythos.

Gerald Edelman und Brigitte Falkenburg haben diese Betrachtungsweise mit empirischen und theoretischen Erkenntnissen der Physik, Biologie und Neurowissenschaften bei einem Rückgriff auf Argumente der Philosophie eindeutig widerlegt. Die zentrale Schlussfolgerung lautet, dass Physik, Biologie, Genetik und Psychologie dem Menschen „Leitplanken" auferlegen (insbesondere die Sterblichkeit), hingegen in nahezu allen Fällen genügend Spielraum für den freien Willen lassen.

Falkenburg dazu:

> „*Die Freiheit des Menschen realisiert sich immer in bestimmten Schranken. Diese Schranken sind vielfältiger Natur, es gibt soziale Zwänge, Erziehungseinflüsse. Die Muttersprache, kulturelle Wurzeln, genetische Dispositionen, Nahrung, klimatische Bedingungen, körperliche Beeinträchtigungen und vieles mehr als Randbedingungen für unser Leben.*"[518]

---

[517] Vgl. Kitchin(2014); Frické (2015): *"The ability to cheaply and easily gather large amounts of data does have advantages: Sample sizes can be larger, testing of theories can be better, there can be continuous assessments, and so on. But data-driven science, the 'fourth paradigm', is a chimera. Science needs problems, thoughts, theories, and designed experiments. If anything, science needs more theories and less data."*

[518] Falkenburg (2012), S. 388

Es steht uns Menschen frei, dass wir uns von den genannten Randbedingungen innerlich distanzieren, sie herausfordern und gestalten und unsere Ziele nicht wegen der Randbedingungen, sondern trotz derselben erreichen.

Irrtum V:

Der Computer-Funktionalismus sei die plausible Lösung des Leib-Seele-Problems.

Verschiedene Gedankenexperimente und theoretische Argumentationen gegen den Funktionalismus im Allgemeinen und den Computer-Funktionalismus im Besonderen mit Beiträgen von Searle[519], Jackson[520], Nagel[521], Block[522] und Bunge[523] sind nachgezeichnet und diskutiert worden. In ihrer Gesamtheit ist die Zurückweisung des Computer-Funktionalismus plausibel.

Brigitte Falkenburg, die an anderer Stelle bereits wiederholt zur kausalen Geschlossenheit und zum menschlichen Determinismus zitiert wurde, kam in ihrer Argumentation auch zu einer klaren Schlussfolgerung zum Computer-Modell des menschlichen Gehirns:

> *„Bei aller prima facie-Ähnlichkeit der Kortex-Architektur und der Architektur künstlicher neuronaler Netze dürfen Sie eines nicht vergessen: Auch an der Atomvorstellung von Newton und seinen Zeitgenossen war einiges dran. Verglichen mit der Wirklichkeit der Quantenprozesse lag sie aber ziemlich daneben. Dies tat sie jedoch vermutlich in ganz anderer Weise als das Computer-Modell die Wirklichkeit von Gehirn und Bewusstsein verfehlen dürfte.*
>
> *Beim Analogieschluss von den makroskopischen Körpern mikroskopischen Bestandteile funktioniert die Teile-Ganzes-Beziehung bestens, bis hinab zum Quark Modell. Beim parallelen Analogieschluss vom Gehirn auf den Computer und von der Computer-Information zurück auf das Bewusstsein funktioniert die Teile-Ganzes-Beziehung gerade nicht. Das Bewusstsein hat weder mit der Computer-Information noch mit dem Feuern der Neurone irgendeine Eigenschaft gemeinsam, die sich nur ansatzweise messen, quantifizieren und im Sinne einer Teile-Ganzes-Beziehung deuten ließe. Stattdessen verleiten die üblichen top-down- und bottom-up-Verfahren hier zu mereologischen und kausalen Fehlschlüssen.“*[524]

Der Computer-Funktionalismus löst das Leib-Seele-Problem nicht.

[519] Vgl. Searle (1980), S. 417ff
[520] Vgl. Jackson (1982), S. 127ff
[521] Vgl. Nagel (1974)
[522] Vgl. Block (1995), S. 523ff
[523] Vgl. Bunge (2010), S. 227ff
[524] Falkenburg (2012), S. 412

Irrtum VI:

Das menschliche Bewusstsein mit seinen wichtigen phänomenalen und intentionalen Komponenten sei „naturalisierbar".

Es liegen keine auch nur halbwegs erfolgversprechenden Konzepte zur Schaffung eines „künstlichen Bewusstseins" vor. Alle Theorien der Naturalisierung von Intentionalität und phänomenalem Bewusstsein (Qualia) sind aus Sicht der Philosophie und der Neurowissenschaften erfolglos. Der detaillierteste Ansatz – der einer „Integrierten Informationstheorie des Bewusstseins" (Abschnitt 6.5.2) – kommt explizit zu dem Ergebnis, dass das Bewusstsein nicht digitalisiert werden kann[525].

Irrtum VII:

Die Hirnforschung besitze ein klares und differenziertes Verständnis der Entstehung des menschlichen Bewusstseins und liefere konkrete Ansätze für den Nachbau des menschlichen Gehirns auf einem anderen Substrat. Das menschliche Gehirn sei ein Computer und funktioniere wie ein solcher.

Die Hirnforschung erzielte signifikante Erkenntnisse über die Wirkzusammenhänge auf der Ebene der Neuronen und Synapsen. Ähnlich große Fortschritte sehen wir beim Verständnis des Gehirns auf der „Behavior"-Ebene. Allerdings ist man in Bezug auf die Ebene dazwischen, die erklären helfen könnte, wie Bewusstsein zustande kommt, über Absichtserklärungen nicht oder kaum hinausgekommen.

Brigitte Falkenburg hält in ihrem Buch zum „*Mythos Determinismus*" fest:

> „*Die kognitive Neurowissenschaft kann nach ihrem derzeitigen Stand keinen neuronalen Mechanismus vorweisen – egal ob strikt deterministisch oder nicht –, der erklären könnte, wie das Bewusstsein aus dem neuronalen Geschehen hervorgeht.*"[526]

Überzeugende Antworten auf die großen Fragen zum Zustandekommen des menschlichen Bewusstseins liegen weiterhin nicht vor. Erklärungsansätze aus der Quantenmechanik und die Suche nach dem neuronalen Korrelat des Bewusstseins können als gescheitert betrachtet werden. Auch informationstheoretische Überlegungen haben allenfalls zu interessanten Hypothesen geführt, die bisher empirisch nicht bestätigt werden konnten.

[525] Vgl. C. Koch (2020), S. 145f

[526] Falkenburg (2012), S. 379

6.3 Standort und Überleitung

Eine solide Basis für weitere normative Überlegungen zur Künstlichen Intelligenz bilden die bis zu diesem Punkt erarbeiteten Erkenntnisse zur technischen Entwicklung der Künstlichen Intelligenz, die Perspektiven der Psychologie zur menschlichen Intelligenz, die zentralen Diskussionen der Philosophie des Geistes zum Leib-Seele-Problem und zur Bewusstseinstheorie sowie der aktuelle Forschungsstand der Neurowissenschaften.

Davon ausgehend soll hier ein Perspektivwechsel erfolgen: Der Fokus verlagert sich dabei von einer deskriptiven Beschreibung des Gegenstands KI im Vergleich zur menschlichen Intelligenz und den diversen Theorien der Geistesphilosophie und Neurowissenschaften hin zu der Frage, was diese Technologie mit all ihren Stärken und Schwächen für uns Menschen bedeutet. Dies soll aus vier Einzelperspektiven betrachtet werden, die sich zum Teil überlappen:

1. In welchem Umfang ist die KI im philosophischen Sinne autonom und was bedeutet das für unseren Umgang mit ihr? Inwieweit wird unsere Freiheit durch die zunehmende Delegation an die KI eingeschränkt? (Kapitel 7)
2. Was bedeutet die Einführung der KI für die Zuweisung von Verantwortung? Kann die KI Verantwortung übernehmen? Welche Probleme entstehen möglicherweise kurz-, mittel- und langfristig durch die Verlagerung von Zuständigkeiten? (Kapitel 8)
3. Was bedeutet die KI für das Subjekt, die Person bzw. das Individuum? Wird der Mensch zum Objekt? (Kapitel 9)
4. Abschließend erfolgt eine Vertiefung der Implikationen für die Menschenwürde? Worin liegt der Kern der Einzigartigkeit der Würde des Menschen gegenüber allen anderen Kategorien, Gegenständen und Arten. (Kapitel 10)

Darauf basierend soll dann über die Implikationen für die zentrale These dieser Arbeit reflektiert werden, wonach der Mensch mit zunehmendem Einsatz der KI erneut in eine selbstverschuldete Unmündigkeit gerät und damit wichtige Errungenschaften der Aufklärung einbüßt. (Kapitel 11)

Die Ergebnisse des zweiten Teiles dieser Dissertation werden dann in einem 2. Fazit, der philosophischen Gesamtbeurteilung zusammengefasst. (Kapitel 12)

Daran anschließend erfolgt eine Anwendung der normativen Überlegungen in zwei Fallstudien: KI in der Pflege und KI beim Einsatz in militärischen Waffen. (Kapitel 13)

Abschließend werden die Erkenntnisse dieser Arbeit zusammengefasst in einem Ausblick Implikationen und mögliche neue Forschungsfragen angerissen. (Kapitel 14)

7 Ist die KI autonom oder heteronom?

Viele Systeme der Künstlichen Intelligenz werden in der Fachliteratur und in den Medien als autonom beschrieben[527]. Bei fahrerlosen Kraftfahrzeugen spricht man von „autonomem Fahren“, das Militär plant oder verwendet „autonome Waffen“ und in der Fabrikfertigung werden „autonome Werkstätten“ oder „autonome Produktionszellen“ eingesetzt. Diese und andere Fälle suggerieren eine Vergleichbarkeit mit der Autonomie des Menschen.

In diesem Kapitel soll als Eingangsargument für die folgenden Kapitel die These überprüft werden, dass KI-Systeme im philosophischen Sinn über keinerlei Autonomie verfügen und verfügen können und damit die Verwendung dieses menschlichen Attributs irreführend und im Sinne der weiteren Diskussionen zur Verantwortlichkeit, Individualität und Menschenwürde gefährlich ist.

Dieter Sturma hat die Problematik treffend als eine „*Reihe von semantischen Transformationen*“[528] beschrieben, deren Durchlaufen den „Begriff der Autonomie“ „*von seiner ursprünglichen Bedeutung beträchtlich*“ abrückt. Aus seiner Sicht gehen die Transformationen weit über „*metaphorische Umdeutungen“* hinaus; sie erschweren die angemessene Erfassung von Sachverhalten in der KI und Robotik und ziehen „*jenseits konkreter Anwendungsfälle weitergehende Folgen im sozialen Raum nach sich*“[529].

7.1 Der Begriff der Autonomie nach Immanuel Kant

> „*Die Autonomie des Willens ist das alleinige Prinzip aller moralischen Gesetze und der ihnen gemäßen Pflichten: alle Heteronomie der Willkür gründet dagegen nicht allein gar keine Verbindlichkeit, sondern ist vielmehr dem Prinzip derselben und der Sittlichkeit entgegen. In der Unabhängigkeit nämlich von aller Materie des Gesetzes (nämlich einem begehrten Objekte) und zugleich doch Bestimmung der Willkür durch die bloße allgemeine gesetzgebende Form, deren eine Maxime fähig sein muss, besteht das alleinige Prinzip der Sittlichkeit. Jene Unabhängigkeit ist Freiheit im negativen, diese eigene Gesetzgebung aber der reinen und als solche praktischen Vernunft ist Freiheit im positiven Verstande. Also drückt das moralische Gesetz nichts anders aus, als die Autonomie der reinen praktischen Vernunft, d.i. der Freiheit, und diese ist selbst die formale Bedingung aller Maximen, unter der sie allein mit dem obersten praktischen Gesetze zusammenstimmen können. Wenn daher die Materie des Wollens, welche nichts anders als das Objekt einer Begierde sein kann, die mit dem Gesetz verbunden wird, in das praktische Gesetz als Bedingung der Möglichkeit desselben hineinkommt, so wird daraus Heteronomie der Willkür, nämlich Abhängigkeit vom Naturgesetze, irgendeinem Antriebe oder Neigung zu*

[527] Anmerkung: Das für die KI zuständige Geschäftsfeld der Fraunhofer-Gesellschaft heisst „Künstliche Intelligenz und Autonome Systeme“, vgl. https://www.iosb.fraunhofer.de/de/geschaeftsfelder/kuenstliche-intelligenz-autonome-systeme.html

[528] Sturma (2003), S. 38

[529] Sturma (2003), S. 39

> *folgen, und der Wille gibt sich nicht selbst das Gesetz, sondern nur die Vorschrift zur Verfolgung pathologischer Gesetze; die Maxime aber, die auf solche Weise niemals die allgemeingesetzgebende Form in sich enthalten kann, stiftet auf diese Weise nicht allein keine Verbindlichkeit, sondern ist selbst dem Prinzip einer reinen praktischen Vernunft, hiermit also auch der sittlichen Gesinnung entgegen, wenn gleich die Handlung, die daraus entspringt gesetzmäßig sein sollte.*"
> Immanuel Kant, 1781, Von den Grundsätzen der reinen praktischen Vernunft, §8 Lehrsatz IV[530]

An diversen Stellen in seinem Werk hat sich Kant mit dem Begriff der Autonomie als einem der zentralen Eckpfeiler seiner von Konfession und Religion unabhängigen Ethik beschäftigt. In dem obigen Zitat sind drei der vier wichtigsten philosophischen Funktionen der Autonomie gemäß dem Kant-Lexikon (Willaschek et al.) für Kants Verhältnisse sehr klar und verständlich dargelegt[531]: Autonomie als Selbst-Gesetzgebung, Autonomie als oberstes Prinzip der Moral und Autonomie als Freiheit. Nicht enthalten ist die vierte Funktion nach Willaschek et al.: Autonomie als Grund der Würde, eine Funktion, die in einem späteren Kapitel dieser Arbeit von großer Bedeutung sein wird.

„Für Kant ist Autonomie eine Form von Gesetzgebung"[532] (die erste Funktion). Das beinhaltet einerseits die inhaltliche Formulierung eines Gesetzes und andererseits seine verbindliche Geltung und Autorität. Es ist aber eine „*Selbst-Gesetzgebung*". Dieses „Selbst" kann nach Willaschek et al. drei verschiedene Bedeutungen haben: Es kann ein „*empirisches Selbst*" sein, d.h. ein konkretes Gesetz, das ein Mensch aus seiner 1. Person Perspektive für sich selbst formuliert, wie z.B. täglich Sport zu treiben oder keinen Alkohol zu trinken. Die zweite Möglichkeit ist ein „*transzendentales Selbst*" im Sinne von „eine Person wie ich müsste eigentlich …". Dieses Selbst ist unabhängig von empirischen Bedingungen. Die dritte Möglichkeit geht noch einen Schritt weiter: eine „*eigene Gesetzgebung*"[533], a priori und komplett unabhängig von jeglicher Empirie und konkreter Existenz.

Zur zweiten Funktion: Kant legt viel Wert auf die Autonomie als das „*alleinige Prinzip aller moralischen Gesetze*"[534]. „*Dahinter steckt die These, dass nur eine eigene Gesetzgebung der reinen Vernunft einen kategorischen Imperativ und damit unbedingte Verbindlichkeit gewährleisten könne.*"[535] Der Gegenspieler der Autonomie ist die Heteronomie, d.h. die Abhängigkeit von einem fremden Gesetz.[536] Diese kann keine moralische Verbindlichkeit ergeben, da „*sie zur Motivation auf eine kontingente und relative*

[530] Kant, Kritik der praktischen Vernunft, AA V 33; Kant (1788), S. 144 (A 59)
[531] Vgl. Willaschek et al. (2017), S. 45ff
[532] Willaschek et al. (2017), S. 46
[533] Kant, Grundlegung der Metaphysik der Sitten, AA IV 450
[534] Siehe oben und Kant, Kritik der praktischen Vernunft, AA V 33; Kant (1788), S. 144 (A 59)
[535] Willaschek et al. (2017), S. 46
[536] Willaschek et al. (2017), S. 226

Neigung angewiesen wäre und somit keine unbedingte und allgemeine Forderung ergeben“[537] kann.

Freiheit ist die dritte philosophische Funktion der Autonomie. Der Mensch wird nicht nur von seinen Neigungen getrieben, sondern besitzt die Freiheit, „*sich nach einem Gesetz der Vernunft selbst zu bestimmen*“. Die Unabhängigkeit von eigenen Neigungen und externen Einflüssen bezeichnet er als negative Freiheit und die Möglichkeit der eigenen Gesetzgebung nach dem kategorischen Imperativ als positive Freiheit.

Gemäß der vierten Funktion nach Willaschek et al. ist die Autonomie als Grund der Würde anzusehen. „*Die Fähigkeit, allgemein gesetzgebend zu sein*“[538], grenzt den Menschen vom Rest der Natur ab. In Kants Worten: „*Autonomie ist also der Grund der Würde der menschlichen und jeder vernünftigen Natur*“[539]. Nur der Mensch ist nicht bloß ein Spielball natürlicher Neigungen oder anderer Kräfte der Natur. Allein der Mensch kann Subjekt sein.

[537] Willaschek et al. (2017), S. 47

[538] Willaschek et al. (2017), S. 48

[539] Kant, Grundlegung der Metaphysik der Sitten, AA IV 436; Kant (1785b), S.69 (BA 79)

7.2 Die „semantischen Transformationen“ der Autonomie

Dieter Sturma hat sich ausführlich mit dem Autonomiebegriff, dessen Historie und insbesondere mit seiner Anwendung auf die Künstliche Intelligenz beschäftigt[540]. Aus seiner Sicht begegnen wir bei der derzeitig üblichen Verwendung des Begriffs der Künstlichen Intelligenz einigen „*konstruktiven Verstellungen von Sachverhalten“*, die „*Klarstellungen“* und teilweise auch „*Revisionen“*[541] erforderlich machen.

Den Autonomiebegriff kannte man bereits in der Antike, doch Sturma setzt bei Kant an. Autonomie mit „*bloßer Handlungsfreiheit, Selbstorganisation oder Selbststeuerung*“ gleichzusetzen, „*liefe auf gravierende Unterbestimmungen hinaus*“.

Wie schon im vorigen Kapitel dargestellt, besteht die Funktion der Autonomie darin, „*Gesetze für Maximen – worunter Lebensregeln zu verstehen sind – festzulegen*“. Kant führt an dieser Stelle noch den Begriff der „Person“ ein. Selbstbewusste Wesen, die „*für Gründe empfänglich*“ sind, von ihren Neigungen abgrenzen und somit „*verallgemeinern, differenzieren und handeln können*“, sind Personen.

Kant (wie auch Rousseau), so Sturma, stellt „*dem naturwissenschaftlichen Raum der Ursachen*“ die Welt der selbstgesetzgebenden Personen gegenüber:

> *„Jedes Objekt in Raum und Zeit wirke nach Gesetzen. Nur eine Person habe das Vermögen, nach der ‚Vorstellung der Gesetze, d.i. nach Prinzipien, zu handeln‘*[542]*. Autonomie vollzieht sich damit nicht allein durch die zwangsläufigen Auswirkungen von Gesetzen. Eine autonome Handlung kommt vielmehr erst dadurch zustande, dass eine Person ein vorgestelltes Gesetz zur Anwendung bringt. Die vorgestellten Gesetze der Autonomie weisen einen gänzlich anderen Status auf als die Gesetze der sogenannten exakten Wissenschaften. In ihrer formalen Struktur und Anwendbarkeit sind sie zudem ungleich komplizierter als diese und verdienen insofern keineswegs weniger Beachtung. Auf diesen Sachverhalt spielt Kant mit seiner berühmten Formulierung vom bestirnten Himmel über mir und dem moralischen Gesetz in mir an.“*[543]

Auf dieser Basis erarbeitet Sturma noch weitere Merkmale autonomer Personen, die nicht zwingend der Spezies „homo sapiens sapiens“ angehören müssen. Die wichtigsten sind hier aufgelistet[544]:

---

[540] Sturma (2003)

[541] Sturma (2003), S. 38f (und folgende Zitate)

[542] Zitat im Zitat: Kant (1785a), S. 412; Kant (1785b), S. 41 (BA 37); ausführlich: „*Ein jedes Ding der Natur wirkt nach Gesetzen. Nur ein vernünftiges Wesen hat das Vermögen, nach der Vorstellung der Gesetze, d. i. nach Principien, zu handeln, oder einen Willen. Da zur Ableitung der Handlungen von Gesetzen Vernunft erfordert wird, so ist der Wille nichts anders als praktische Vernunft. Wenn die Vernunft den Willen unausbleiblich bestimmt, so sind die Handlungen eines solchen Wesens, die als objectiv nothwendig erkannt werden, auch subjectiv nothwendig, d. i. der Wille ist ein Vermögen, nur dasjenige zu wählen, was die Vernunft unabhängig von der Neigung als praktisch nothwendig, d. i. als gut, erkennt.“*

[543] Sturma (2003), S. 40

[544] Sturma (2003), S. 40f

1. Sie verfügen über „*Fähigkeiten und Eigenschaften wie Intelligenz, Emotivität, Selbstbewusstsein, Wille, Selbstverständnis, Intentionalität, Sprache, Handlungsfreiheit, Rationalität, Zuschreibungen und Anerkennungen*“.
2. Dazu gehört eine Vorstellung von der „*zeitlichen Ausdehnung in Vergangenheit, Gegenwart und Zukunft“* und ein „*Verständnis der fremden und eigenen Innerlichkeit*“ sowie die dafür erforderliche „*Ausdrucksfähigkeit*“.
3. Personen sind in der Lage, soziale Fähigkeiten zu entwickeln, die eine Einbindung in eine „*epistemische, ethische*“ und kulturelle Gemeinschaft erlauben.
4. Sie bewegen sich in einen „*logischen Raum der Gründe*“, d.h. Handlungen von Personen sind von Gründen geleitet. Diese sind mehr als nur Syntax und Semantik: „*Eine Aussage im Raum der Gründe ist daher mehr als eine informationsartige Veränderung. Sie liegt erst dann vor, wenn Meinungen oder Annahmen in den Raum des Gebens und Entgegennehmens von Gründen gestellt werden*.“

Daher kann der Begriff der Autonomie nur auf eine Lebensform angewandt werden, die sich dem „Raum der Gründe“ öffnet[545].

Bettina Walde führt dies in ihren Betrachtungen über „*Gründe als Ursachen?*“[546] etwas detaillierter aus:

> *„Offenbar sind Gründe keine im weitesten Sinne physikalischen Entitäten (die Sorte von Entitäten, die es in einer naturalistischen, monistischen Ontologie gibt). Dennoch werden Gründe oder allgemein Abwägungsprozesse immer dann interessant und von Personen in Erwägung gezogen, wenn bisherige Erfahrungen und bisheriges Wissen es zunächst nicht erlauben, für eine aktuelle und konkrete Entscheidungssituation eine angemessene Antwort zu finden. In solchen Fällen entfalten (nicht-physikalische) Gründe so etwas wie kausale Wirksamkeit, und möglicher Weise lassen Gründe sich auch dazu heranziehen, zu erklären, weshalb freie Willensentscheidungen im unbedingten Sinne nicht mit Zufallsereignissen gleich zu setzen sein müssen.“*[547]

Die KI folgt einem Algorithmus, wird aber nicht von Gründen geleitet. Über den Algorithmus ist sie „*prinzipiell vorhersagbar*“[548] und damit, so Sturma, „*uneigentlich autonom*“. Eine von Gründen geleitete Person ist dagegen „*prinzipiell nicht vorhersagbar*“ und damit „*eigentlich autonom*“:

> *„Hat eine Person mit spezifischen Dispositionen und Eigenschaften unter bestimmten Bedingungen verschiedene begründbare Handlungsoptionen zur Verfügung, dann ist prinzipiell – d.h. jenseits von bloßer Plausibilität – niemals vorhersagbar, für welche Option sie sich entscheiden wird.“*[549]

---

[545] Sturma (2003), S. 49

[546] Walde (2006), S. 50f

[547] Walde (2006), S. 51

[548] Sturma (2003), S. 50, sowie folgende Zitate

[549] Sturma (2003), S. 51, sowie weitere Zitate

Beim Elfmeterschießen im Fußballspiel kann niemand vorhersagen, ob ein Schütze den Ball in die linke oder rechte Ecke des Tores schießen wird, selbst dann, wenn er bisher alle seine Schüsse in eine bestimmte Ecke gesteuert hat[550].

Die Summe aller Begründungszusammenhänge der Personen schafft eine „*künstliche Ordnung von Rationalität und Moralität*“[551]:

> *„Der notorische Übergang von Naturgeschichte in Kulturgeschichte, der für die menschliche Lebensform kennzeichnend ist, besteht letztlich in nichts anderem als in der Initiierung und Realisierung von eigenen Regeln, Gesetzen und Institutionen*.“

Dieser Argumentation folgend ist der Begriff der Autonomie ein Alleinstellungsmerkmal der menschlichen Lebensform. Interessant ist die klare Abgrenzung von Naturgeschichte und Kulturgeschichte. Die Kulturentwicklung geht klar über das Befolgen von Naturgesetzen hinaus. Nur der von Gründen geleitete und sich selbst das Gesetz gebende Mensch ist zur Kulturentwicklung in der Lage.

Die Möglichkeit der Selbstgesetzgebung von Maschinen auf Basis von Gründen (und nicht nur Algorithmen) ist zurzeit nicht erkennbar. Sturma warnt allerdings davor, dies überhaupt in Betracht zu ziehen. Falls es je KI-Systeme oder Roboter geben sollte, die im Raum der Gründe gemäß der hier vorgestellten Bedeutung selbstständig agieren und „*über Selbstbewusstsein und mentale Repräsentation verfügen können, müssten wir sie in unsere ethische Gemeinschaft aufnehmen*“, was eine Reihe unlösbarer Fragen nach sich ziehen würde. Daher sind KI und Roboter bis auf weiteres „*schlicht Maschinen und müssen ausschließlich als Mittel für menschliche Zwecke behandelt werden*“[552]:

> „*Die Politik der Autonomie sollte mit Entschiedenheit eine parallele Entwicklung von Humanität und Robotik als Verlängerung und Unterstützung personalen Handelns vorantreiben. Das hieße auch, den Traum vom künstlichen Menschen zugunsten konkreter Projekte der Erweiterung des Spielraums menschlicher Personen aufzugeben.*“[553]

Dies ist eine Vorgabe, auf die im weiteren Verlauf dieser Arbeit noch einzugehen ist. Nur Menschen können autonom sein. Jegliches Delegieren von Autonomie an Wesen oder Artefakte, die nicht autonom sein können, unterminiert die eigene Autonomie.

[550] Vgl. Sturma (2003), S. 51

[551] Sturma (2003), S. 51

[552] Sturma (2003), S. 53

[553] Ebd.

7.3 Autonomie und Urteilskraft

Wie oben in den Zitaten von Willaschek dargelegt wurde, spielt die Autonomie bei Kant eine zentrale Rolle in der Ethik: Autonomie als Selbst-Gesetzgebung, als oberstes Prinzip der Moral, als Freiheit und als Grund der Würde. Sie ist somit *oberstes Princip der Sittlichkeit*[554]:

> *„Autonomie des Willens ist die Beschaffenheit des Willens, dadurch derselbe ihm selbst (unabhängig von aller Beschaffenheit der Gegenstände des Wollens) ein Gesetz ist. Das Princip der Autonomie ist also: nicht anders zu wählen als so, daß die Maximen seiner Wahl in demselben Wollen zugleich als allgemeines Gesetz mit Begriffen seien. Daß diese praktische Regel ein Imperativ sei, d. i. der Wille jedes vernünftigen Wesens an sie als Bedingung nothwendig gebunden sei, kann durch bloße Zergliederung der in ihm vorkommenden Begriffe nicht bewiesen werden, weil es ein synthetischer Satz ist; man müßte über die Erkenntniß der Objecte und zu einer Kritik des Subjects, d. i. der reinen praktischen Vernunft, hinausgehen, denn völlig a priori muß dieser synthetische Satz, der apodiktisch gebietet, erkannt werden können, dieses Geschäft aber gehört nicht in gegenwärtigen Abschnitt.“*[555]

Die Autonomie stößt also beim kategorischen Imperativ an ihre Grenze. Etwas anders verhält es sich bei der Urteilskraft. Zunächst einmal zur Definition der Urteilskraft bei Kant in der Kritik der Urteilskraft:

> *„Urtheilskraft überhaupt ist das Vermögen, das Besondere als enthalten unter dem Allgemeinen zu denken. Ist das Allgemeine (die Regel, das Princip, das Gesetz) gegeben, so ist die Urtheilskraft, welche das Besondere darunter subsumirt, (auch wenn sie als transscendentale Urtheilskraft a priori die Bedingungen angiebt, welchen gemäß allein unter jenem Allgemeinen subsumirt werden kann) bestimmend. Ist aber nur das Besondere gegeben, wozu sie das Allgemeine finden soll, so ist die Urtheilskraft bloß reflectirend.“*[556]

Die Urteilskraft dient einerseits der Zuordnung des Besonderen zum Allgemeinen, also einer bekannten Regel, einem etablierten Prinzip oder einem Gesetz. An anderer Stelle (In der Kritik der praktischen Vernunft) bezeichnet Kant dies auch als praktische Urteilskraft[557]. Andererseits ist sie „reflektierend“, wenn es das Allgemeine (noch) nicht gibt. Bei Willaschek in der Langfassung des Kant-Lexikons findet sich dazu[558]:

> *„Die reflektierende Urteilskraft, deren Konzeption erst in der KU voll entwickelt wird, gibt sich im Unterschied zur bestimmenden Urteilskraft das Prinzip ihrer Reflexion selbst bzw.*

---

554 Zitiert bei Willaschek (2015), S. 200

555 Kant, Grundlegung der Metaphysik der Sitten, AA IV 440; Kant (1785b), S. 74 (BA 87)

556 Kant, Kritik der Urteilskraft, AA V 179; Kant (1790), S. 87

557 Kant, Kritik der praktischen Vernunft, AA V 67; Kant (1788), S. 186 (A 120): „*Die Begriffe des Guten und Bösen bestimmen dem Willen zuerst ein Object. Sie stehen selbst aber unter einer praktischen Regel der Vernunft, welche, wenn sie reine Vernunft ist, den Willen a priori in Ansehung seines Gegenstandes bestimmt. Ob nun eine uns in der Sinnlichkeit mögliche Handlung der Fall sei, der unter der Regel stehe, oder nicht, dazu gehört* ***praktische Urtheilskraft****, wodurch dasjenige, was in der Regel allgemein (in abstracto) gesagt wurde, auf eine Handlung in concreto angewandt wird“.* [Hervorhebung DS]

558 Willaschek (2015), S. 2443

> *autonom und hat die Aufgabe, ‚von dem Besonderen in der Natur zum Allgemeinen aufzusteigen'"*[559].

Die praktische (oder auch bestimmende) Urteilskraft ist „*subsumiert heteronom"* unter vorgegebenen allgemeinen Regeln. Die reflektierende Urteilskraft erteilt „*sich selbst das (transzendentale) Prinzip ihrer Reflexion"*[560] und ist damit autonom.

> Diese Autonomie der „*reflektierenden Urteilskraft unterscheidet sich einerseits von der Autonomie des Verstandes, der vermittels des einen Verstandesbegriffe oder Kategorien gesetzgebend für die Natur (im allgemeinen) ist, wie auch von der Autonomie der praktischen Vernunft, die im Blick auf den Gedanken eines Endzweckes des menschlichen Daseins gesetzgebend für das Begehrungsvermögen ist*"[561].

Zusätzlich zur praktischen und reflektierenden Urteilskraft nennt Kant noch die teleologische Urteilskraft[562] in der Naturphilosophie sowie die ästhetische Urteilskraft:

> *„Wir haben ein Vermögen der bloß ästhetischen Urtheilskraft, ohne Begriffe über Formen zu urtheilen und an der bloßen Beurtheilung derselben ein Wohlgefallen zu finden, welches wir zugleich jedermann zur Regel machen, ohne daß dieses Urtheil sich auf einem Interesse gründet, noch ein solches hervorbringt."*[563]

Einige Philosophen sehen für diese „erweiterte Autonomie" die Notwendigkeit einer begrifflichen Abgrenzung und sprechen von der Heautonomie, so der Theologe Ernst Feil:

> *„Im Unterschied zur Autonomie des Verstandes oder der Vernunft ist die Autonomie der Urteilskraft „bloß subjectiv, für das Urtheil aus Gefühl gültig", wobei das Urteil nur dann Allgemeingültigkeit beanspruchen kann, wenn es auf Prinzipien a priori und eben nicht auf Erfahrung gegründet ist; wegen dieser Subjekt-Bezogenheit müßte nach Kant diese Gesetzgebung eigentlich `Heautonomie´ genannt werden."*[564]

Und an anderer Stelle:

> *„Aufgrund der eben vorgelegten Hinweise lässt sich jedoch schon soviel sagen, dass sich auch in der ‚Kritik der Urteilskraft' ein spezifischer Gebrauch von Autonomie bestätigt hat: Autonomie meint wie schon in den ethischen Schriften Unabhängigkeit von Empirie und Erfahrung; durch diese Unabhängigkeit der (ästhetischen) Urteilskraft vermögen ästhetische Urteile Notwendigkeit und Allgemeingültigkeit zu erlangen, selbst wenn sie subjektbezogen sind. Um diesem Sachverhalt Rechnung zu tragen, wäre es nach Kant besser, statt von „Autonomie" bezüglich der Urteilskraft von ‚Heautonomie' zu sprechen. Jedenfalls ist Vorsicht geboten, uneingeschränkt von ‚Autonomie der Urteilskraft' zu sprechen."*[565]

---

559 Willaschek zitiert hier Kant, Kritik der Urteilskraft, AA V 385; Kant (1790), S. 88 (A XXV, B XXVII)

560 Willaschek (2015), S. 203

561 Willaschek (2015), S. 205

562 Willaschek (2015), S. 2456: „*Teleologische Urteilskraft oder die teleologisch gebrauchte Urteilskraft gibt `die Bedingungen bestimmt an, unter denen etwas (z.B. ein organisirter Körper) nach der Idee eines Zweckes der Natur zu beurtheilen sei´. Als reflektierende Urteilskraft entwirft sie die Hypothese der Organisiertheit der Natur."* Willaschek zitiert hier Kant, Kritik der Urteilskraft, AA V 194

563 Kant, Kritik der Urteilskraft, AA V 300; Kant (1790), S. 233 (A 167, B 169)

564 Feil (1982), S. 418

565 Feil (1982), S. 420-421

Diesen Gedanken hat Wilhelm Vossenkuhl in seinem Buch zur Geltung[566] weiterentwickelt. In Anlehnung an Kants *Konzept der Willenskraft* versteht er die „*Willensbildung als freies Spiel von Willen und Vernunft"*. Darin enthalten sind einige stillschweigende Annahmen, wie z.B. diejenige, dass „*der Wille gut ist*, die *Vernunft kognitiv anspruchsvoll* und die *Willensbildung vorurteilsfrei, intersubjektiv und kohärent"*. Damit kann „*das Resultat der Willensbildung immer nur exemplarisch, für bestimmte Fälle und nicht wie der Kategorische Imperative universal wirksam sein"*.

Die Kombination von praktischer, reflektierender telelogischer und ästhetischer Urteilskraft in einem freien Spiel von Willen und Vernunft ermöglicht den Umgang mit Zielkonflikten, die Navigation zwischen widersprüchlichen Geltungen bei gleichzeitiger „*Anerkennung von Prinzipien, die für die soziale, politische und rechtliche Praxis unverzichtbar sind"*[567]. Es besteht eine innere Korrespondenz zwischen der Urteilskraft und der Welt jenseits des ursprünglichen Apriorismus[568] Kants.

Die Betrachtungen in diesem Abschnitt kommen im Wesentlichen zu dem Ergebnis, dass Urteilskraft im kantschen Verständnis, in einer Weiterentwicklung aber auch in einem zeitgenössischen Verständnis, mehr Spielraum braucht und hat, als es die Ausrichtung auf die Selbstgesetzgebung im Sinne des kategorischen Imperativs auf den ersten Blick zulässt. Gefordert ist die Kombination aus Verstand, Vernunft, gutem Willen, praktischer Urteilskraft, reflektierender, telelogischer und ästhetischer Urteilskraft.

Das im ersten Abschnitt vorgestellte engere Verständnis des Autonomiebegriffs weicht bereits weit von den Möglichkeiten der Künstlichen Intelligenz ab. Reflektierende und ästhetische Urteilskraft oder gar das von Vossenkuhl thematisierte freie Spiel zwischen (gutem) Willen und Vernunft im Sinne des Gemeinsinns liegen außerhalb jeder technischen Realisierbarkeit.

---

566 Vossenkuhl (2021), S. 320 f

567 Vossenkuhl (2021), S. 332

568 *„**Apriorismus**, zusammenfassende Bezeichnung der philosophischen Lehren, nach denen es von der Erfahrung unabhängige Erkenntnis gibt."*, Quelle: Wörterbuch der philosophischen Begriffe (2013), S. 59

7.4 Zusammenfassung: Klärung eines Missverständnisses

Bei der Beschreibung der KI ist die Verwendung des Begriffs der Autonomie irreführend und *unterbestimmt*[569]. Die heute bekannten KI-Systeme werden als „autonom" bezeichnet, weil sie mit Hilfe ihrer einprogrammierten Algorithmen die ihnen vorgegebenen Ziele selbstständig und ohne steuernden Eingriff von außen erreichen. Wie wir gesehen haben, sind viele Systeme sowohl synthetisch als auch analytisch determiniert. Selbst lernende Maschinen sind zwar nicht analytisch determiniert, jedoch immer synthetisch[570]. Drei grundsätzliche und unverzichtbare Prämissen der menschlichen Autonomie – oder, wie Sturma schreibt, der Autonomie von Personen – werden von der KI nicht erfüllt.

Erstens sind KI-Systeme nicht in der Lage, ihre eigenen Ziele unabgeleitet zu definieren. Unabgeleitete Ziele sind sowohl deskriptiv als auch präskriptiv. Sie kombinieren empirische und vorgegebene normative Tatsachen. Dies soll an einem Beispiel erläutert werden, das Julian Nida-Rümelin zur Erläuterung von inferenziellen Beziehungen zwischen empirischen und normativen Tatsachen verwendet[571]. Das Gebot, einem Kind kein scharfes Messer zu geben, basiert auf der empirischen Tatsache, dass man sich mit scharfen Messern verletzen kann, und der normativen Tatsache, dass ein Selbstverletzungsrisiko von Kindern nicht eingegangen werden darf. Ein KI-System kann sich die empirische Tatsache erschließen, die normative Tatsache aber nur dann, wenn sie direkt vorgegeben ist oder aus einer allgemeineren Vorgabe abgeleitet werden kann: Menschen dürfen sich nicht selbst verletzen können bzw. dürfen keiner Gefahr ausgesetzt werden. Autonome Personen können sich beides erschließen.

Damit kommen wir zur zweiten Prämisse. Autonome Personen sind zur allgemeinen Selbstgesetzgebung auf Basis des kategorischen Imperativs fähig, d.h. sie können bei Berücksichtigung der Forderung, andere Menschen immer auch als Zweck und niemals nur als Mittel zu behandeln, Handlungsmaximen entwickeln. Systeme der KI vermögen das nicht.

Die dritte Prämisse besteht darin, dass sich autonome Wesen von Gründen affizieren lassen, die auch, aber nicht nur, einer Logik oder einem Ursache-Wirkung-Mechanismen folgen. Die Gründe von autonomen Wesen sind daher nicht streng deterministisch und logisch ableitbar. Dies gilt insbesondere für die Ableitung von normativen Begründungen aus empirischen Tatsachen. Nida-Rümelin spricht von der „*naturalistischen Unterbestimmtheit unserer Handlungs- und Urteilsgründe*"[572]. Die oben angesprochene Grenze zwischen Natur und Kultur kann nur der autonome Mensch überschreiten, eine Künstliche Intelligenz keinesfalls. Sie kann allerdings kulturelle Schöpfungen des Menschen

[569] Vgl. Sturma (2003), S. 38f

[570] Siehe auch Kapitel 2.4

[571] Vgl. Nida-Rümelin (2020), S. 335

[572] Nida-Rümelin (2005), S. 35

analysieren, sortieren und auf Muster untersuchen und darauf basierend nach Statistik und Stochastik neue Artefakte generieren. Dabei bleibt sie aber Instrument des Menschen und wird nicht zum Kulturschaffenden.

Aufgrund dieser dreifachen Lücke (Zieldefinition, Selbstgesetzgebung, Gründe) können Systeme der KI zwar in hohem Maß automatisiert und selbstständig operierend sein, aber niemals autonom im philosophischen Sinn. Sie werden immer heteronom sein.

Dies gilt umso stärker für die zuletzt dargestellte autonome Urteilskraft, die zwar in ihrer praktischen Auslegung das Anwenden existierender Regeln, Gesetze und Strukturen zulässt, aber als reflektierende Urteilskraft eine Autonomie erfordert, die das heteronome System der Künstlichen Intelligenz niemals aufweisen kann, womit es auch niemals zu einem freien Spiel von Willen und Vernunft imstande sein wird.

8 Kann die Künstliche Intelligenz Verantwortung übernehmen?

Mit der voranschreitenden Entwicklung und Nutzung der Künstlichen Intelligenz zeichnen sich zwei Arten von Verantwortungslücken ab. Einerseits ergeben sich durch die Nutzung von Systemen der KI wachsende Bereiche, bei denen die Zuweisung einer echten Verantwortung von Menschen an Menschen nicht oder nur sehr eingeschränkt möglich ist. Verstärkt wird dies insbesondere durch selbstlernende Systeme, also Systeme, die ihre eigenen Algorithmus auf Basis von Erfahrungen im technischen Sinne anpassen und dafür keine Gründe im humanistischen Sinne angeben können. Konkret wird man dies schon sehr bald beim sogenannten „autonomen Fahren" sehen. Es wird Unfälle mit Sach- und Personenschäden geben, deren Verlauf aus Sicht von Beteiligten und Zeugen irrational ist und die auch mit technischen und juristischen Methoden nicht nachvollzogen werden können. Aus zivilrechtlicher Sicht werden Sachschäden über Produkthaftpflichtversicherungen abgewickelt. Dies deckt aber nur einen kleinen Teil dessen ab, was wir üblicherweise unter Verantwortung verstehen.

Die zweite Verantwortungslücke ergibt sich durch die mit der KI geschaffenen irreversibel veränderten Lebenswelt für den Menschen mit massiven Einschränkungen der menschlichen Freiheit. Diese Einschränkungen stellen sich erst mittel- bis langfristig durch die sukzessive Einführung der neuen Technologien ein. Auch dies kann am Beispiel des autonomen Fahrens erläutert werden. Wenn sich 80, 90 oder 95 Prozent der Fahrzeuge autonom bewegen, wird man irgendwann feststellen, dass der größte und folgenreichste Störfaktor im Straßenverkehr die von Menschen gesteuerten manuellen Fahrzeuge sind. Spätestens dann wird es Diskussionen zum Verbot des „manuellen Fahrens" geben. Der Mensch gewinnt Komfort und Sicherheit, verliert aber auch Freiheit. Der von BMW 1965 geprägte Werbeslogan „*Freude am Fahren*"[573] verliert seinen Sinn und Gegenstand. Dies gilt potenziell für viele andere Bereiche unserer Lebenswelt, nicht zuletzt auch unsere Arbeitswelt, in der nicht nur unangenehme und schwere Arbeiten durch Roboter und KI-Systeme erledigt werden. Insofern ergibt sich eine zusätzliche Verantwortungslücke auf der Zeitachse bezüglich der Lebenswelt, die wir unseren Kindern und Enkeln hinterlassen.

[573] https://www.bmw.com/de/automotive-life/die-geschichte-des-bmw-slogan.html

8.1 Herkunft und Entwicklung des Begriffs der Verantwortung

Der Verantwortungsbegriff hat erst im 20. Jahrhundert Bedeutung in der Moralphilosophie erlangt[574]. Dem Begriff der „Verantwortung" oder äquivalenten Konzepten wurden in der klassischen Ethik von Aristoteles bis Kant keine herausragende Bedeutung beigemessen. Erst in der zweiten Hälfte des 19. Jahrhunderts schrieb z.B. John Stuart Mill in seinem Essay „*On Liberty*" über „*responsibility*" und „*moral responsibility*". Etwas später beschäftigte sich auch Friedrich Nietzsche mit dem Begriff in seiner „*Genealogie der Moral*". Max Weber führte am Ende des 1. Weltkrieges die Unterscheidung von „Gesinnungsethik und Verantwortungsethik"[575] ein und charakterisierte mit dem Begriff „Verantwortung" „*einen ganzen Typus ethischer Theorien"*[576]. Im weiteren Verlauf des 20. Jahrhunderts wurde der Begriff dann von vielen weiteren Philosophen aufgegriffen bis zu Hans Jonas, der gegen Ende des Jahrhunderts eine spezifische Ethik rund um das „**Prinzip Verantwortung**" entwickelte.

Das heutige Verständnis des Verantwortungsbegriffs hat sich, wie an späterer Stelle insbesondere beim „Prinzip Verantwortung" nach Hans Jonas noch vertieft werden wird, sehr stark von den Ursprüngen im späten 19. Jahrhundert entfernt. Kurt Bayertz stellt „*fünf Charakteristika"* des aktuellen Verantwortungsverständnisses heraus[577].

1. *„Neuartige Gegenstände der Verantwortung treten in den Vordergrund: Haftung für technische Unfälle, Zuständigkeit für anspruchsvolle Aufgaben innerhalb einer Organisation oder die Sorge für die Erhaltung der Natur.* […] *Die Verantwortung wächst über den Rahmen unmittelbarer Beziehungen zwischen Individuen hinaus und wird mehr und mehr zur* ***Sicherung öffentlicher Güter*** *mobilisiert.* […] *Jeder Bürger kann für die Folgen seines Handelns von seinen Mitbürgern zur Rede gestellt und zur Verantwortung gezogen werden."*
2. Früher stellte sich die Frage nach der Verantwortung „*immer nur ex post"*. In der Moderne orientiert sich der Verantwortungsbegriff zusätzlich in die Zukunft. Es geht zunehmend um die Verminderung oder Vermeidung von Risiken. Aus dem

[574] Vgl. Bayertz (1995), S. 3; einschließlich der folgenden Überlegungen

[575] Vgl. Weber (1919b), S.70: „*Wir müssen uns klar machen, dass alles ethisch orientierte Handeln unter zwei voneinander grundverschiedenen, unaustragbar gegensätzlichen Maximen stehen kann: es kann ‚gesinnungsethisch' oder ‚verantwortungsethisch' orientiert sein. Nicht dass Gesinnungsethik mit Verantwortungslosigkeit und Verantwortungsethik mit Gesinnungslosigkeit identifiziert wäre. Davon ist natürlich keine Rede. Aber es ist ein abgrundtiefer Gegensatz, ob man unter der gesinnungsethischen Maxime handelt – religiös geredet -: ‚der Christ tut recht und stellt den Erfolg Gott anheim', oder unter der verantwortungsethischen: dass man für die (voraussehbaren) Folgen seines Handelns aufzukommen hat."*

[576] Ebd.

[577] Alle folgenden Zitate: Bayertz (1995), S. 43 f; Hervorhebungen DS

klassischen rein *„retrospektiven Verantwortungsbegriff“* wird einer, der **zunehmend *„prospektiv“*** ist.

3. Die Umkehr oder Ausweitung der *„Zeitrichtung* ist *keinesfalls neutral gegenüber dem Inhalt der Verantwortung. Retrospektive Verantwortung bezieht sich auf negativ bewertete Folgen*, […] die *prospektive **Verantwortung auf positiv bewertete Zustände“***.
4. *„Im klassischen, d.h. retrospektiven Sinne verantwortlich gemacht wurde man in aller Regel für die Folgen seines Handelns“*. In der Moderne *„treten **Unterlassungen gleichberechtigt neben Handlungen“***.
5. *„Die Verantwortung für einen Schaden kann unter bestimmten Bedingungen **von den inneren Bedingungen** (insbesondere: der Absicht) **abgekoppelt** werden.* […] *Max Webers Unterscheidung zwischen einer Gesinnungs- und einer Verantwortungsethik bringt dies unmissverständlich auf den Punkt: aus verantwortungsethischer Perspektive zählen bei der Bewertung einer Handlung nicht die guten oder schlechten Absichten des Akteurs, sondern ausschließlich ihre tatsächlichen Folgen“.*

8.2 Vorverständnis: Relata von Verantwortung

Die Vokabeln „Verantwortung“ als Substantiv, „verantworten“ als Verb und „verantwortlich“ als Adjektiv werden alltagssprachlich höchst unterschiedlich verwendet[578]. Dies soll anhand von drei Beispielsätzen veranschaulicht werden[579]:

1. Der Vorstandsvorsitzende (CEO, Chief Executive Officer) trägt vor dem Aufsichtsrat die Verantwortung dafür, den Marketingchef entlassen zu haben.
2. Nicht nur der CEO, sondern der gesamte Vorstand verantwortet das Wachstum des Unternehmens, die Erlöse, die Kosten und die Ergebnissituation des Unternehmens.
3. Das Marktumfeld und die COVID-Pandemie sind für den Einbruch der Vertriebszahlen verantwortlich.

Gemäß dem ersten Satze legt der CEO gegenüber dem vorgesetzten Organ, dem Aufsichtsrat, Rechenschaft über die Gründe ab, die ihn dazu bewogen haben, den Marketingchef zu entlassen. Es handelt sich bei der Verantwortung um *„eine Form des Rechenschaftsablegens“*[580]. Der zweite Satz bringt zum Ausdruck, dass nicht nur er, sondern auch seine Vorstandskollegen zusammen mit ihm für die Leistungsdaten des Unternehmens verantwortlich sind. Hier wird der Begriff der Verantwortung im Sinne einer *„Pflichten- und Aufgabenzuschreibung“* verwendet. Der dritte Satz beschäftigt sich mit

[578] Vgl. Buddeberg (2011), S. 11

[579] In Abwandlung von Buddeberg (2011), S. 12; grundsätzliche Struktur der Beispiele wie bei Buddeberg

[580] Buddeberg (2011), S. 12

dem Marktumfeld und der Pandemie als Ursache der schlechten Vertriebszahlen. Es geht also um die Verantwortung als „*Verursachung“*.

Eva Buddeberg extrahiert vier Relata[581], welche die Relationen des Verantwortungsbegriffes beschreiben:

- *Wer trägt Verantwortung? -> Das **Subjekt** der Verantwortung*
- *Wofür wird Verantwortung übernommen? -> Das **Objekt** der Verantwortung*
- *Vor wem wird Rechenschaft abgelegt? -> Die **Instanz** der Verantwortung*
- *Warum ist das Subjekt verantwortlich? -> Der **normative** Bezugsrahmen*

Im Folgenden sollen die vier Relata vertiefend betrachtet werden, insbesondere in Bezug auf die Künstliche Intelligenz.

8.2.1 Subjekt der Verantwortung

In den drei im vorherigen Abschnitt aufgelisteten Sätzen sind die Subjekte der Verantwortung jeweils eindeutig bestimmbar. Im ersten Satz ist es der Vorstandsvorsitzende, der gleichzeitig Träger der Verantwortung gegenüber dem Aufsichtsrat ist und Ausübender der Handlung, nämlich der Entlassung des Marketingschefs. Diese „*Korrelation von Verantwortungssubjekt und Handlungssubjekt*“[582] ist bedeutsam: Der CEO trägt die Verantwortung für sein Handeln. Er rechtfertigt sein Handeln mit Gründen gegenüber dem Aufsichtsrat. Für die Rechtfertigung mit Gründen braucht er Kommunikationsfähigkeiten. Ohne kommunizierte Begründung ist auch keine Rechtfertigung und damit keine Verantwortung (im Sinne des ersten Satzes) möglich. Beim zweiten Satz handelt es sich beim Subjekt der Verantwortung und der Handlung um mehrere Personen. Eine eindeutige Zuschreibung der Verantwortlichkeit auf einzelne Individuen wird nicht angeboten, ergibt sich aber möglicherweise aus dem Kontext: Der Produktionschef ist verantwortlich für die Produktionszahlen, der Marketingchef für die Vertriebszahlen und der Finanzchef für die Ergebniszahlen. Bei den Handlungen kann es sich um vollzogene oder noch zu vollziehende Handlungen halten. Beim dritten Beispielsatz ist das Subjekt der Verantwortung die kausale Ursache. Von einer echten Handlung, für die es Alternativen gebe, kann in der Deterministik zwischen Ursache und Wirkung nicht ausgegangen werden. Auch kann das Subjekt keine Gründe für seine Verursachung angeben. Das dritte Beispiel

[581] In der Literatur finden sich auch höherstellige Relationen, wie z.B. die sechsstellige Relation bei Lenk und Maring (1993), S. 229, ebenfalls zitiert bei Bayertz (1995), S. 15:
- jemand: Verantwortungssubjektträger (Personen, Korporationen) ist
- für: etwas (Handlungen, Handlungsfolgen, Zustände, Aufgaben, usw.)
- gegenüber: einem Adressaten
- vor: einer (Sanktions-, Urteils-) Instanz
- in Bezug auf: ein (präskriptives, normatives) Kriterium
- im Rahmen eines: (Verantwortungs-, Handlungsbereiches) verantwortlich

[582] Buddeberg (2011), S. 13

weicht also grundsätzlich von Beispiel 1 und 2 ab. Für die philosophische Diskussion der Verantwortung im Kontext dieser Arbeit ist es nur bedingt bedeutsam.

Anhand eines Beispiels aus der Anwendung der KI soll die Problematik der Übernahme von Verantwortung erläutert werden. Inwieweit ist ein selbststeuerndes Fahrzeug („Autonomes Vehikel"), das Unfälle verursachen kann, Subjekt im Sinne von Beispielsatz 1 oder Beispielsatz 3? Für Beispielsatz 1 könnte sprechen, dass das Fahrzeug die Entscheidungsbäume seines Algorithmus für die Verursachung des Unfallereignisses nutzen würde. Nun handelt es sich dabei explizit nicht um Gründe, die Willens- und Handlungsfreiheit voraussetzen, und nicht nur die reine Deterministik des Algorithmus. Nida-Rümelin spricht von der „*naturalistischen Unterbestimmtheit unserer Überzeugungen, Handlungen und Gefühle"*[583]. Daher entspricht die Verantwortungszuweisung eher Beispiel 3. Alternativ könnte man die Ursache des Unfalls bei Programmierfehlern suchen und den Programmierer dafür verantwortlich machen, was bei großen Programmierteams potenziell zu Problemen führt. Viel relevanter ist indes der Umstand, dass viele KI-Systeme selbstlernend sind und ihre Algorithmen selbstständig anpassen oder deren Entstehung von der immer wieder aktualisierten Mustererkennung in großen Datensätzen abhängig ist. Das wäre ein rein deterministischer Ablauf nach Regeln, die nur zum Teil menschengemacht sind. Genau dadurch ergibt sich eine Lücke: Die Programmierer sind nicht verantwortlich, weil sie die Ergebnisse des selbstlernenden Algorithmus nicht vorhersagen oder vorherbestimmen können. Die KI kann nicht moralisch, ethisch oder juristisch verantwortlich sein, weil sie über keine Willens- oder Handlungsfreiheit verfügt. Es besteht also eine Verantwortungslücke.

8.2.2 Objekt der Verantwortung

Beim Objekt der Verantwortung geht es um die Frage: „*Wofür ist jemand verantwortlich oder was hat jemand zu verantworten?*"[584] Im ersten oben ausgeführten Beispielsatz trägt der CEO die Verantwortung für die Entscheidung, den Marketingchef zu entlassen. Er ist verantwortlich für die vollzogene Handlung. Der Aufsichtsrat interessiert sich nicht nur für die Tatsache, dass er die Entlassung vollzogen hat, sondern auch für dessen Begründung, einen möglichen Anlass, ein Fehlverhalten oder neue Informationen zu seiner Qualifikation und Eignung für das Amt. Zusätzlich interessiert er sich für die Motive und Ziele des Vorstandsvorsitzenden, also die Einbettung der Handlung in einen übergeordneten Zweck. In die Verantwortung schließt der Aufsichtsrat auch die möglichen Implikationen und Konsequenzen ein, ein mögliches arbeitsgerichtliches Verfahren, die Neubesetzung der Stelle oder die Verunsicherung der Mitarbeiter. Einige dieser

[583] Nida-Rümelin (2011), S. 15f

[584] Buddeberg (2011), S.23

Konsequenzen mag der CEO im Blick haben, andere berücksichtigt er nicht und viele sind ihm vielleicht nicht bekannt, da noch einer größeren Unsicherheit unterworfen.

Beim zweiten Beispielsatz ist die Festlegung des Objekts wesentlich breiter und auch viel weniger präzise. Es geht um Wachstum, Erlöse, Kosten und Erträge des Unternehmens. Dies beinhaltet alle Faktoren, die diese vier Kennzahlen beeinflussen, retrospektiv und prospektiv. Es handelt sich also um eine Beschreibung des Handlungs- und Aufgabenfeldes und des Verantwortungsbereiches. Im prospektiven zukunftsorientierten Teil findet „*in der normativen Dimension des Verantwortungsbegriffs"* eine Vorzeichenvertauschung statt: „*man ist nicht mehr für negative Folgen verantwortlich, sondern für positive Zustände"*[585]. In dem konkreten Beispiel handelt es sich bei den positiven Zuständen um Produkte, die Wachstum ermöglichen, neue Prozesse, die Kostenreduzierungen ermöglichen, oder um Investitionen, die das Ergebnis verbessern. Dieser Teil der Verantwortung wirkt über die Laufzeit der Arbeitsverträge der Vorstände hinaus. Oftmals werden noch nach mehreren Jahren Boni reduziert oder sind zurückzuzahlen. Beispiel: Ein Familienvater, der für das Wohl seiner Kinder verantwortlich ist, muss Rechenschaft ablegen, wenn den Schutzbefohlenen aufgrund einer Vernachlässigung der Aufsichtspflicht etwas passiert ist (retrospektiv). Er ist aber auch für deren zukünftiges Wohlergehen zuständig. Er ist verantwortlich für die „*positiven Zustände"* (prospektiv), wie zum Beispiel die Gesundheit, Erziehung und Bildung der Kinder.

Auch ohne eine explizit definierte Rolle (Vorstand, Familienvater) ist der Mensch als menschliches Subjekt für menschliche Objekte, also für andere Personen verantwortlich. Dazu ein weiteres Beispiel: Eine Gruppe von vier zufällig zusammen gekommenen Individuen unternimmt einen Segeltörn mit einem Bootsführer. Das Schiff gerät bei Sturm in Seenot. Der Skipper geht über Bord und ist verschwunden. Von diesem Moment an ist jeder der vier Bootsinsassen für das eigene Überleben verantwortlich, aber auch für das Überleben der drei anderen, zumindest würde das in den meisten dem Autor bekannten sittlichen Gemeinschaften so gelten.[586]

Auch aus dieser Betrachtung lässt sich verneinen, dass eine Künstliche Intelligenz als nichtmenschliches Subjekt Verantwortung für Menschen als Objekte im Hier und Jetzt oder in der Zukunft übernehmen kann.

8.2.3 Instanz der Verantwortung

Nun zur dritten Frage: Vor wem, vor welcher Instanz muss sich das Subjekt retrospektiv oder prospektiv rechtfertigen? Bei den Beispielen ist die Instanz direkt benannt oder ergibt sich aus dem Kontext. Beim ersten Beispiel ist der Aufsichtsrat die Instanz. Er

[585] Buddeberg (2011), S. 25f; sie zitiert hier Bayertz, „Geschichte der Herkunft von Verantwortung", S. 32

[586] Ähnlich strukturierte Beispiele bei Buddeberg (2011), Bayertz (1995)

erteilt dem Vorstand die Verantwortung für alle seine (in diesem Fall) Personalentscheidungen und erwartet retrospektiv eine Rechtfertigung konkreter Handlungen. Beim zweiten Beispiel erschließt sich dies aus dem Kontext. Die Instanz für die Vorstände ist der Vorstandsvorsitzende im ersten Schritt und der Aufsichtsrat im zweiten Schritt. Beim dritten Beispiel gibt es keine Instanz. Weder die Pandemie noch das Marktumfeld können vor irgendeiner Instanz Rechenschaft ablegen. Beim Beispiel mit dem Familienvater oder den vier Seglern ist die Frage der Instanz nicht so eindeutig. Bei groben Verstößen (Verletzung der Aufsichtspflicht, unterlassener Hilfeleistung) ist die Instanz ein Richter in einem Strafverfahren, aber oftmals besteht eine Instanz im nichtjuristischen Bereich. Wenn ein Elternteil sein Kind ausschließlich mit Fastfood ernährt, so dass es übergewichtig wird, ist nicht zwingend damit zu rechnen, dass ein Richter darüber urteilt. Trotzdem wird dieses Handeln in den meisten Sittengemeinschaften als unmoralisch eingestuft. Die Instanz ist also sittlicher, moralischer oder ethischer Natur. Sehr häufig ist die Instanz das eigene Gewissen des Subjekts. Es läßt sich sogar argumentieren, dass ein Gewissen, also ein Bewusstsein, die Grundvoraussetzung für ein verantwortliches Wesen ist. Das intentionale Nachdenken über das eigene Tun und die Konsequenzen daraus schafft die Instanz der Verantwortung.

Für eine KI kann es keine Instanz geben, vor der sie Rechenschaft für ihr eigenes Tun ablegt, denn sie ist weder eine Person, die sich vor einem Gericht verantwortet, noch Teil einer sittlichen, moralischen oder ethischen Gemeinschaft, noch in ein System von Bewertungsmaßstäben eingebettet.

8.2.4 Normativer Bezugsrahmen der Verantwortung

In der Literatur gilt es als höchst umstritten, ob ein normativer Bezugsrahmen ein wirklich „*notwendiges Relatum der Verantwortungsrelation*“[587] ist. Der normative Bezugsrahmen beantwortet die Frage nach dem „Warum“ einer Verantwortlichkeit. Für dieses „Warum“ gibt es zwei Varianten:

- Warum ist das Subjekt verantwortlich für das Objekt und muss vor einer Instanz Rechenschaft darüber ablegen?
- Warum ist diese oder jene Handlungsfolge als „gut“ oder als „nicht gut“ anzusehen?

Einige Autoren sind der Meinung, dass dafür drei Relata ausreichend sind[588]. Buddeberg argumentiert, dass der normative Bezugsrahmen als Hintergrund erforderlich ist. Vor allem ist die Frage zu klären, „*warum Menschen überhaupt in einem solchen*

[587] Buddeberg (2011), S. 38

[588] Vgl. Bayertz (1995), insbesondere S. 15ff; Bayertz spricht von einem „*mehrstelligen Relationsbegriff*“ der Verantwortung, „*der mindestens drei Elemente in Beziehung zueinander bringt: a) ein **Subjekt** der Verantwortung, ein **Objekt** der Verantwortung und c) ein **System von Bewertungsmaßstäben“***

Rechtfertigungsverhältnis stehen und damit die Pflicht haben, anderen von ihrem Handeln betroffenen Menschen über sich und ihr Handeln Rechenschaft abzulegen"[589]. Dies kann nur das Ergebnis einer umfassenderen *moralphilosophischen Klärung* sein. Auch Bayertz, der an anderer Stelle die Instanz der Verantwortung mit dem normativen Bezugsrahmen vereint, vertritt klar die Auffassung, „*dass die Zurechnung von Verantwortung stets mit einem Werturteil verknüpft ist"*[590]. Die „*bloße Zuschreibung von Folgen ist ein deskriptiver Akt*" ohne „*moralische Bedeutung*: *Erst indem die Zuschreibung mit einem Werturteil verknüpft wird, kommt die genuin moralische Dimension des Verantwortungsproblems ins Spiel".*

8.3 Prinzip Verantwortung und KI

Der deutsch-amerikanische Philosoph Hans Jonas hat 1979 mit „***Prinzip Verantwortung***"[591] eine Zukunftsethik vorgelegt, die als „*Klassiker einer moralphilosophischen Verantwortungstheorie"*[592] gilt. Das Buch wurde jenseits der Philosophie, insbesondere auch in den Naturwissenschaften und der Technik, vielbeachtet und vor allem von der damals neugegründeten grünen Bewegung aufgegriffen. Jonas weitete den Objektbereich der menschlichen Verantwortung vom individuellen Menschen auf die Gegenwart und Zukunft der gesamten Menschheit und der Natur des Planeten aus.

Primär und enger gefasst ging es ihm damals um die Verantwortung für die langfristigen Folgen von Technologien und Umweltverschmutzung für das Überleben der Menschheit und der Natur insgesamt auf dem Planeten Erde. In weiteren Publikationen und Diskursen befasste er sich auch mit Technologien, die – mit einem utopischen Ansinnen – das „Menschsein" nachhaltig verändern. Die Künstliche Intelligenz mit all ihren Auswirkungen und Implikationen, so das Argument in diesem Abschnitt, gehört ebenfalls dazu. Jonas äußerte sich zu den beiden in der Einleitung skizzierten Verantwortungslücken, derjenigen im Nahbereich und derjenigen im Fernbereich der Entwicklung und (in einem Interview) des Einsatzes der KI.

In den folgenden Teilabschnitten sollen zunächst die wesentlichen Grundannahmen und logischen Schlussfolgerungen der Zukunftsethik nach Hans Jonas dargestellt werden. Danach erfolgt die Anwendung im Nah- und Fernbereich des KI-Einsatzes. Bezug genommen wird dabei auch auf Positionen von Philosophen und Zeitgenossen von Hans Jonas, die seine Methodik und sein Denken maßgeblich mit beeinflusst haben, wie z.B. Martin Heidegger und Hannah Arendt.

---

[589] Buddeberg (2011), S. 40

[590] Dieses und folgende Zitate: Bayertz (1995), S. 13f

[591] Jonas (1979)

[592] Buddeberg (2011), S. 45

8.3.1 Hans Jonas‘ Prinzip Verantwortung

Zu Beginn seines philosophischen Arguments, quasi als Basis für die Ableitung einer erweiterten Verantwortungsethik, stellt Jonas die Voraussetzungen aller bisherigen Ethik[593] zusammen:

1. *„Der menschliche Zustand, gegeben durch die Natur des Menschen und die Natur der Dinge, steht in den Grundzügen ein für alle Mal fest.*
2. *Das menschlich Gute lässt sich auf dieser Grundlage unschwer und einsichtig bestimmen.*
3. *Die Reichweite menschlichen Handelns und daher menschlicher Verantwortung ist eng umschrieben.“*[594]

In der heutigen Welt sind diese drei Voraussetzungen nicht mehr erfüllt. Historisch war „*aller Umgang mit der außermenschlichen Welt* […] *ethisch neutral“*[595]. Die bisherigen Ethiken waren anthropozentrisch und sowohl räumlich als auch zeitlich auf den Nahbereich ausgerichtet:

> *„Ethische Bedeutung gehörte zum direkten Umgang von Mensch mit Mensch, einschließlich des Umgangs mit sich selbst; alle traditionelle Ethik ist* ***anthropozentrisch*** [Hervorhebung von Jonas].
>
> *Das Wohl oder Übel, worum das Handeln sich zu kümmern hatte, lag nah bei der Handlung, entweder in der Praxis selbst oder in ihrer unmittelbaren Reichweite und war keine Sache entfernter Planung. Diese Nähe der Ziele galt für Zeit sowohl als Raum.“*[596]

Die oben genannten Relata Subjekt, Objekt und Instanz der Verantwortung befinden sich damit in „*einer gemeinsamen Gegenwart“*[597]. Innerhalb dieser gemeinsamen Gegenwart erfolgen moralische Urteile über gute und schlechte Handlungen und können sich Subjekte vor den jeweiligen Instanzen verantworten. Alles Wissen, das dafür erforderlich ist, liegt vor. Unbeabsichtigte spätere Auswirkungen des Handelns, ggf. außerhalb der Lebenszeit des Subjekts, liegen außerhalb des moralischen Urteils und der Verantwortung, insbesondere wenn sie (die Auswirkungen) den außermenschlichen Bereich der Natur betreffen.

Aber warum brauchen wir jetzt eine neue erweiterte Ethik? Die heutigen Techniken und Technologien haben in nie zuvor gekanntem Ausmaß Auswirkungen in der Nähe und weiter entfernt, in der Gegenwart, in der nahen und fernen Zukunft. Auch ist es immer seltener möglich, für Handlungen Individuen verantwortlich zu machen. Explizit und

---

[593] „*ob als direkte Anweisung, gewisse Dinge zu tun und andere nicht zu tun, oder als Bestimmung von Prinzipien für solche Anweisungen, oder als Aufweisung eines Grundes der Verpflichtung, solchen Prinzipien zu gehorchen“*; Jonas (1979), S. 15

[594] Jonas (1979), S. 15

[595] Jonas (1979), S.22; als einzige Ausnahme akzeptierte Jonas den medizinischen Bereich

[596] Jonas (1979), S.22

[597] Jonas (1979), S. 23

implizit gehen die wirkungsmächtigen Handlungen immer häufiger auf Kollektive zurück. Deswegen sieht Jonas die Notwendigkeit für drei „*neue Dimensionen der Verantwortung"*[598]. Die „*Verletzlichkeit der Natur"*, damit einhergehend „*das sittliche Eigenrecht der Natur"* und abschließend „*die neue Rolle des Wissens in der Moral"*[599].

Nie zuvor haben menschliche Handlungen in ihrem kumulativen Charakter die Natur stärker und an vielen Stellen unwiederbringlich verändert. Oftmals hat man Schäden und Veränderungen erst viel später nach den zugrundeliegenden menschlichen Handlungen an weit entfernten Orten festgestellt. Die Entdeckung dieser Auswirkungen menschengemachter Technologien hat nicht nur Philosophen in einen Schock versetzt:

> *„Sie bringt durch die Wirkungen an den Tag, dass die Natur menschlichen Handelns sich de facto geändert hat, und dass ein Gegenstand von gänzlich neuer Ordnung, nicht weniger als die gesamte Biosphäre des Planeten, dem hinzugefügt worden ist, wofür wir verantwortlich sein müssen, weil wir Macht darüber haben."*[600]

Gerade der oben angesprochene kumulative Charakter der Technikfolgen ist ein neues Phänomen. Jonas spricht von der „*kumulative*[n] *Selbstfortpflanzung technologischer Veränderung der Welt*, die *die Bedingungen ihrer beitragenden Akte"* fortwährend überholt und damit „*durch lauter präzedenzlose Situationen"* verläuft, „*für die die Lehren der Erfahrung ohnmächtig sind"*[601]. Für das, was wir derzeitig in der Natur beobachten, gibt es keine Präzedenz. Auch war die Kette der Entwicklungen zu Beginn bei den ersten Handlungen nicht zu erwarten und damit auch kein Gegenstand sittlicher Erwägungen.

Die Feststellung der Verletzlichkeit der Natur wirft die Frage nach der moralischen Relevanz dieser Erkenntnis auf. Hat die Natur ein sittliches Eigenrecht? Jonas stellt fest, dass in den bisherigen Ethiken die Natur an sich keinen moralischen Status innehatte. Dabei nimmt er explizit die Religion mit ihrem Schöpfungsgedanken heraus. In einer sehr umfassenden Argumentation leitet Jonas seine Zukunftsethik ontologisch her, d.h. er stellt eine direkte Verbindung zwischen dem Sein und Sein-Sollen der Natur her:

> *„Die* [anthropozentrische] *Verengung* [der Ethik; Ergänzungen DS] *auf den Menschen allein und als von aller übrigen Natur verschieden kann nur Verengung, ja Entmenschung des Menschen selbst bedeuten, die Verkümmerung seines Wesens auch im Glücksfall biologischer Erhaltung – widerspricht also ihrem vorgeblichen, eben von der Würde seines Wesens beglaubigten Ziel. Im wahrhaft menschlichen Blickpunkt bleibt der Natur ihre Eigenwürde, die der Würde unserer Macht entgegensteht. Als von ihr hervorgebracht schulden wir dem verwandten Ganzen ihrer Hervorbringung eine Treue, wovon die zu unserm eigenen Sein nur die höchste Spitze ist. Diese aber, recht verstanden, befasst alles andere unter sich."*[602]

---

598 Jonas (1979), S. 26

599 Jonas (1979), S. 26-30; Jonas reiht die Dimensionen etwas anders

600 Jonas (1979), S. 27

601 Jonas (1979), S. 28

602 Jonas (1979), S.245f

Damit argumentiert er klar gegen eine anthropozentrische Ethik, die nur den Menschen als Subjekt, Objekt und Instanz anerkennt. Die Würde des Menschen als Teil der Natur, aus der er hervorgebracht wurde, impliziert eine Eigenwürde der Natur, die es zu achten gilt.

Die dritte neue Dimension der Verantwortung befasst sich mit der neuen Rolle des Wissens in der Moral: Die Grundforderung, man solle nur dann Verantwortung übernehmen, wenn man alle Folgen des eigenen Handelns verstehen und absehen kann. Eine Mutter wird der Verantwortung für ihr Kind nur dann gerecht, wenn sie zum Beispiel weiß, welche Speisen für das Kind gesund sind und welche nicht. Jonas formuliert diesen Zusammenhang wie folgt: „*Das Wissen muss dem kausalen Ausmaß unseres Handelns größengleich sein*“[603]. Wenn wir mit Augenbinden ein Kraftfahrzeug bewegen, ist diese Größengleichheit nicht gegeben. Das Gleiche gilt für nächtliches Fahrradfahren ohne Beleuchtung. Bei den neuen Technologien ist diese Größengleichheit nicht mehr gegeben. Das „*vorhersagende Wissen“* über die Technologiefolgen bleibt „*hinter dem technischen Wissen, das unserm Handeln die Macht gibt“*, zurück:

> *„Die Kluft zwischen Kraft des Vorherwissens und Macht des Tuns erzeugt ein neues ethisches Problem. Anerkennung der Unwissenheit wird dann die Kehrseite der Pflicht des Wissens und damit ein Teil der Ethik, welche die immer nötiger werdende Selbstbeaufsichtigung unserer übermäßigen Macht unterrichten muss.“*

An einer anderen Stelle in seinem Buch fasst er dies noch etwas drastischer zusammen und wird dabei auch aufklärungskritisch. Er spricht vom *ethischen Vakuum*[604]:

> *„Erst wurde durch* [unser] *Wissen die Natur in Hinsicht auf Wert ‚neutralisiert‘, dann auch der Mensch. Nun zittern wir in einer Nacktheit eines Nihilismus, in der größte Macht sich mit größter Leere paart, größtes Können mit geringstem Wissen davon, wozu. Es ist die Frage, ob wir ohne die Wiederherstellung der Kategorie des Heiligen, die am gründlichsten durch die wissenschaftliche Aufklärung zerstört wurde, eine Ethik haben können, die die extremen Kräfte zügeln kann, die wir heute besitzen und dauernd hinzuerwerben und auszuüben beinahe gezwungen sind.“*

Jonas entwickelt als Axiom einen neuen Imperativ, der auf den „alten (kategorischen) Imperativ“ Kants aufbaut[605]:

> *„Ein Imperativ, der auf den neuen Typ menschlichen Handelns passt und an den neuen Typ von Handlungssubjekt gerichtet ist, würde etwa so lauten: ‚Handle so, dass die Wirkungen deiner Handlung verträglich sind mit der Permanenz echten menschlichen Lebens auf Erden.‘“*

Es ist überraschend, dass dieser neue kategorische Imperativ wiederum anthropozentrisch ist, zwar nicht im individuellen Sinne, sondern im kollektiven Sinne (die Menschheit

603 Dieses und folgende Zitate (einschließlich Blockzitat): Jonas (1979), S.28

604 Dieses und folgendes Blockzitat: Jonas (1979), S. 57

605 Jonas (1979), S. 35-38

umfassend). Jedoch hat er an anderer Stelle (wie oben zitiert) die Eigenwürde der Natur in einen engen Zusammenhang mit der Würde des Menschen gestellt.

Im Einklang mit dieser engeren Formulierung eines möglichen neuen kategorischen Imperativs in einer Zukunftsethik hat sich Jonas explizit mit dem „*Mensch als Objekt der Technik*“[606] auseinandergesetzt. Seine Sorge galt nicht nur den Auswirkungen der technologischen Zivilisation auf die außermenschliche Natur, sondern auch und insbesondere auf den Menschen an sich:

> *„Doch der Mensch selber ist unter die* ***Objekte der Technik*** [Hervorhebung DS] *geraten. Homo faber kehrt seine Künste auf sich selbst und macht sich dazu fertig, den Erfinder und Verfertiger alles Übrigen neu zu fertigen. Diese Vollendung seiner Gewalt, die sehr wohl die Überwältigung des Menschen bedeuten kann, diese letzte Einsetzung der Kunst über die Natur, fordert die letzte Anstrengung ethischen Denkens heraus, das nie zuvor wählbaren Alternativen zu dem, was für die definitiven Gegebenheiten der Menschenverfassung galt, ins Auge zu fassen hat.“*

In seinem Buch erläutert er drei Beispiele, die seine Sorge belegen, jedoch keinen Vollständigkeitsanspruch erheben: „*Lebensverlängerung“, „Verhaltenskontrolle“* (durch Drogen) „und *„genetische Manipulation“*. Bei den Beispielen vermeidet er das finale moralische Urteil, er weist aber sehr deutlich auf die Risiken des möglichen Verlustes des Einzigartigen des Menschseins, der Autonomie und Freiheit des Menschen und der Spontaneität des Lebens. Er stellt die Frage, ob es uns überhaupt zusteht, mit der Abschaffung der Sterblichkeit auch die Fortpflanzung abzuschaffen, mit sozialer Manipulation die individuelle Autonomie auszuschalten oder in die Evolution einzugreifen.

Hans Jonas fasste dieses sehr deutlich in den Schlussworten seiner Dankesrede anlässlich der Verleihung des Friedenspreises des Deutschen Buchhandels in Frankfurt am Main am 11. Oktober 1987 zusammen:

> *„Das bedeutet, dass wir wohl in alle Zukunft im Schatten drohender Kalamität leben müssen. Sich des Schattens bewusst sein aber, wie wir es jetzt eben werden, wird zum paradoxen Lichtblick der Hoffnung: Er lässt die Stimme der Verantwortung nicht verstummen. Dies Licht leuchtet nicht wie das der Utopie, aber seine Warnung erhellt unsern Weg – mit dem Glauben an Freiheit und Vernunft. So kommt am Ende doch das Prinzip Verantwortung mit dem Prinzip Hoffnung zusammen – nicht mehr die überschwängliche Hoffnung auf ein irdisches Paradies, aber die bescheidenere auf eine Weiterwohnlichkeit der Welt und* ***ein menschenwürdiges Fortleben unserer Gattung auf dem ihr anvertrauten, gewiss nicht armseligen, aber doch beschränkten Erbe****.* [Hervorhebung DS] „[607]

[606] Dieses und folgendes Blockzitat: Jonas (1979), S. 47f

[607] Rede abgedruckt in Böhler Herrmann (2017), Seiten 247 – 256; Zitat auf S. 256

8.3.2 Perspektive zur KI?

In den 1970er Jahren war die Künstliche Intelligenz noch nicht Gegenstand der breiten philosophischen Erörterungen. Auch war die Technologie noch nicht ausgereift genug, so dass die denkbaren Implikationen und Auswirkungen noch nicht bekannt waren. Insofern fand die Künstliche Intelligenz in Jonas' „Prinzip Verantwortung" noch keinen Platz.

Im Jahr 1991 diskutierte jedoch der Wissenschaftsjournalist Norbert Lossau mit Hans Jonas über das Thema Maschinen und Bewusstsein, wobei es auch um Künstliche Intelligenz ging[608]. In voller Übereinstimmung mit den in dieser Arbeit zitierten Philosophen (Bunge, Searle, Vossenkuhl) antwortete Hans Jonas auf die Frage nach einem möglichen Bewusstsein von Maschinen:

> *„Wer so etwas im Ernst erwägt, der setzt zuerst einmal Rechnen mit Denken und dann Denken mit Bewusstsein gleich. Rechnen und logisches Denken können zwar durch formale Regeln für die Abfolge objektiver Symbolzeichen ersetzt werden, und deren jeweiliges Ende kann – von einem es sinnlich wahrnehmenden Bewusstsein – in sein subjektives Äquivalent übertragen und als Lösung einer Denkaufgabe verstanden werden.*
>
> *Den fühlenden Vollzug der Schritte hat sich das Subjekt durch die Delegierung an unfühlende Objektprozesse erspart, das ‚Ergebnis' aber wird von einer Subjektivität aufgenommen und dem Ende eines regelmäßigen Gedankenganges gleichgesetzt. Wenn jemand da ist, der die Umsetzung des physikalischen Endpunktes in ein psychisches Datum und dessen Verständnis als Gedankensymbol vornimmt, dann hat die Maschine ‚rückblickend' gesehen ‚gedacht'.*
>
> *Ohne diese Einlösung ihres Symbolpotentials schließlich könnte sie ewig weiterlaufen, ohne dass sie je gedacht hätte, denn es besteht kein Grund anzunehmen, dass die einzelnen Schritte von Gefühl begleitet werden – ja es wäre eine sinnlose Annahme, denn der Zweck der ganzen Veranstaltung war doch gerade, die ganze Subjektivität zu umgehen und ohne sie auszukommen. Nur bliebe dies seinerseits sinnlos, wenn nicht zum Schluss wieder eine Subjektivität vom Erreichten Kenntnis nähme."*[609]

Eine KI wäre also ein *mechanisch arbeitender Erfüllungsgehilfe, ohne eigenen Spielraum oder gar ein Bewusstsein*[610]. Er deutet weiterhin die Gefahr an, dass mit der Technologie die Subjektivität umgangen wird, ein Gedankengang, auf den im Kontext der Individualität und Menschenwürde noch eingegangen wird.

Jonas sieht den Einsatz von Computern und Künstlicher Intelligenz generell positiv, insbesondere für sein Hauptanliegen der Vermeidung von ökologischen Katastrophen. Dagegen ist seine Meinung zur Übernahme von Verantwortung durch Maschinen (KI) klar:

> *„Entscheidungen werden von Subjekten getroffen, die für diese Entscheidung auch die Verantwortung tragen. Das wird ihnen niemals eine Maschine abnehmen können. Und dort,*

---

608 *Maschinen werden niemals ein Bewusstsein haben können – Gespräch mit Norbert Lossau*, veröffentlicht in Böhler Herrmann (2015), S. 609ff

609 Ebd.

610 Lossau im gleichen Interview

> *wo Entscheidungen von Kollektiven – zum Beispiel einem Staat – getroffen werden, trägt jedes einzelne Individuum Mitverantwortung. Das kann durchaus zu der paradoxen Situation führen, dass ein Individuum Verantwortung für etwas trägt, was es vielleicht gar nicht beeinflussen kann. Wenn der Einsatz von entsprechenden Computersystemen allerdings dazu führen sollte, dass die große Bedeutung des einzelnen Individuums zugunsten einer möglichst reibungslos arbeitenden gesellschaftlichen Maschinerie untergraben würde, wäre das schlimm: der Verlust des Respekts vor der Subjektivität wäre tatsächlich eine große Gefahr für die Menschheit."*[611]

Nach Jonas kann nur ein Mensch in seiner Subjektivität Verantwortung übernehmen. Ein rein objektiver Prozess kann keine Verantwortung übernehmen, er wäre nur instrumentell. Dies ist eine etwas ausführlichere Auslegung der durch die KI entstehenden Verantwortungslücke.

[611] Böhler Herrmann (2015), S. 611

8.4 Verantwortung in Relation zu Freiheit & Rationalität

Julian Nida-Rümelin versteht den *Begriff Verantwortung als Teil einer Trias – Rationalität, Freiheit, Verantwortung*[612]*:*

> „*Freiheit, Rationalität und Verantwortung sind drei Aspekte einer besonderen menschlichen Eigenschaft, nämlich der, sich von Gründen affizieren zu lassen. Dies scheint mir das Spezifikum der humanistischen Denktradition zu sein, Menschen zuzutrauen und ihnen zuzumuten, dass sie sich von Gründen leiten lassen*."

Für ihn beschreiben die Gründe das, was vom Lauf der Natur, getrieben vom freien Willen des Menschen, abweicht:

> *„Dort, wo Gründe keine Rolle mehr spielen, dort wo natürliche Tatsachen und Gesetze unser Verhalten bestimmen, gelten wir nicht mehr als verantwortlich.* […] *Die Person entwickelt und behauptet sich gegen andere Determinanten ihrer Existenz, indem sie Gründen Geltung und Wirkung verschafft."*

Grundsätzlich ist dies konsistent mit den oben dargestellten Relata, dass ein verantwortliches Subjekt sein Handeln begründen können muss. Allerdings ist es auch verantwortlich, wenn es sein Handeln nicht begründen kann, sich von anderen Eingebungen leiten lässt oder sich irrational verhält, wo Rationalität geboten wäre. Deswegen empfiehlt sich die folgende Positionierung:

> **Eine Person ist dann verantwortlich für ihr Tun, wenn sie einen Handlungsspielraum außerhalb dessen besitzt, was durch die Natur determiniert ist, und dieses auch begründen können sollte. Ob sie dies tatsächlich kann, ist zweitrangig.**

Damit ergeben sich ein direkter Zusammenhang und eine gegenseitige Bedingung von Freiheit und Verantwortung, nicht zwingend von Verantwortung und Rationalität. Kurt Bayertz schreibt dazu:

> *„Pointiert gesagt ist es nicht nur so, dass Verantwortung Freiheit voraussetzt; vielmehr wird zugleich auch Freiheit unterstellt, um Verantwortung zuschreiben zu können."*[613]

Dies weicht vom Verständnis des amerikanischen Philosophen Harry Frankfurt ab, der das „Principle of Alternative Possibilities" (Prinzip der alternativen Möglichkeiten) explizit als Indikator für den freien Willen ablehnt.

[612] Dieses und folgende Zitate: Nida-Rümelin (2011), S. 14

[613] Bayertz (1995), S. 12

8.5 Systemverantwortung

> *„Der Verantwortungsbegriff ist ein Bestandteil der Emanzipationsgeschichte des Menschen: Mit der Neuzeit wird die bestehende gesellschaftliche Ordnung als Gegenstand menschlicher Gestaltungsverantwortung bewusst. Die gegebenen Strukturen werden unter der Maßgabe beobachtet, ob und inwiefern sie der Freiheit und Entfaltung des Menschen dienlich sind.* ***Aber das Projekt der Aufklärung, die Emanzipation des Menschen, ist auf eine besondere Art und Weise hoch aktuell – die gewonnene Gestaltungsverantwortung droht wieder verloren zu gehen****.“*[614] [Hervorhebung DS]
>
> Günter Wilhelms

Das Problem des zunehmenden Verlusts der „*personengebundenen Verantwortung an Bedeutung und Legitimationskraft“*[615] besteht nicht ausschließlich bei neuen Technologien wie KI, sondern generell in den Prozessen, Strukturen und Institutionen der modernen Gesellschaft.

> *„Denn mit der Technisierung und Industrialisierung hat der Mensch Mechanismen ins Leben gerufen, die er in ihrer überbordenden Komplexität und Anonymität nicht mehr kontrollieren kann. Und weil die negativen Folgen von Technisierung und Industrialisierung nicht mehr Gott, der ‚unsichtbaren Hand‘ (A. Smith), der ‚List der Natur‘ (I. Kant) oder der ‚List der Vernunft‘ (G.W.F. Hegel) zugewiesen und überlassen werden können, wird die Frage der Zurechnung der Folgen menschlichen Handelns zum Problem.“*[616]

Dieses Problem wird in der Literatur als „kollektive Verantwortung“, „*korporative und kooperative Verantwortung“*[617], „politische Verantwortung“ und in einem übergreifenden Sinne als „*Systemverantwortung*“[618] behandelt. Letztlich beschäftigt sich die ganze Wirtschaftsethik mit dieser Problematik.

Die zentrale Frage lautet: Wie kann die Verantwortung des Systems gedacht werden, wenn die individuelle personale Verantwortung nicht mehr möglich ist, wenn die kognitiven Ressourcen des Einzelnen nicht mehr ausreichen, um die Folgen des Handelns des Systems abzuschätzen, und wenn Handlungs- und Verantwortungssubjekt nicht mehr übereinstimmen[619]. Genau dies sind auch Attribute des Verantwortungsproblems in der KI.

Wilhelms fordert die Ergänzung der „klassischen“ „*personengebundenen Verantwortung durch eine Regelverantwortung*, *die sich auf die strukturellen Möglichkeiten der Bedingungen von Verantwortung bezieht“:*

---

614 Wilhelms (2017), S. 502

615 Wilhelms (2017), S. 501

616 Wilhelms (2017), S. 502f

617 Vgl. Nida-Rümelin (2011), S. 123 ff

618 Vgl. Wilhelms (2017)

619 Wilhelms (2017), S. 514 - 517

> *„Der Begriff* [der Systemverantwortung] *hebt nicht mehr ab auf die Zurechnung von Handlungsfolgen, wie sollte er das unter komplexen Bedingungen, sondern auf Regelstrukturen, die Verantwortung erst ermöglichen, indem sie antizipatorisch potenzielle Handlungsfolgen ins eigene Kalkül aufzunehmen gestatten. Es geht darum, Subjekt und System in ein Verhältnis zu setzen"*[620].

In Bezug auf die in diesem Kapitel diskutierte Verantwortung für die KI bzw. die Verantwortung der KI kann die „Regelverantwortung", d.h. die Verantwortung für die Algorithmen, nur ein notwendiges Kriterium sein, aber niemals ein hinreichendes. Die Verantwortung für die komplexen Algorithmen der KI ist entweder zu allgemein und abgehoben von den tatsächlichen Anwendungssituationen der Technologie oder sie schränkt das System mit zu spezifischen Leitplanken so sehr ein, dass die Vorteile der Technologie nicht mehr genutzt werden können.

[620] Wilhelms (2017), S. 522

8.6 Zusammenfassung: Verantwortungslücken durch KI

> „*Denn in einer vollentwickelten Bürokratie gibt es, wenn man Verantwortung verlangt oder auch Reformen, nur den Niemand. Und mit dem Niemand kann man nicht rechnen, ihn kann man nicht beeinflussen oder überzeugen, auf ihn keinen Druck der Macht ausüben. Bürokratie ist diejenige Staatsform, in welcher es niemanden mehr gibt, der Macht ausübt; und wo alle gleichermaßen ohnmächtig sind, haben wir eine Tyrannis ohne Tyrannen.“*[621]
>
> Hannah Arendt, Macht und Gewalt

In der Philosophie werden oft und gerne Gedankenmodelle verwendet, um bestimmte Grundzusammenhänge zu illustrieren, Grenzwertbetrachtungen anzustellen oder auch das Allgemeine aus dem Speziellen abzuleiten. Zum Thema Verantwortung und Verantwortungslücke bietet sich das Nachdenken über eine Utopie oder eine Dystopie an, in der der Mensch sein komplettes Portfolio an Zuständigkeiten aus der Vita Activa gemäß Hannah Arendt an Maschinen abgegeben hat. Gemäß dem Szenario hat er dann eine Welt geschaffen, in der viel Gutes und viel Böses geschehen kann, dafür jedoch keine Person und auch keine Maschine Verantwortung übernehmen kann. Auch wird jedes zukünftige menschliche Wesen von Geburt an mit genau dieser Welt konfrontiert werden. Auch dafür übernimmt dann keiner die Verantwortung. Es ist schwer vorstellbar, dies als zu begrüßende Utopie zu verstehen.

Verantwortung setzt Freiheit voraus. Ohne Freiheit in ihren Ausprägungen als Gedanken-, Willens-, und Handlungsfreiheit kann es auch keine Verantwortung geben. Umgekehrt ist auch die Verantwortung Vorbedingung für die Freiheit. In einer verantwortungslosen Welt kann es auch keine Freiheit geben. In Hannah Arendts Tyrannis ohne Tyrannen herrscht keine Freiheit, nur grenzenlose Verantwortungslosigkeit.

Aus diesen Gründen ist es keine Kleinigkeit, wenn wir mit zunehmender Verbreitung der neuen Technologien, insbesondere der Künstlichen Intelligenz, mehr und mehr Räume ohne Verantwortung und ohne Möglichkeit der Verantwortungszuweisung schaffen. Notlösungen wie z.B. die Systemverantwortung oder kollaborative Verantwortung bieten keinen Ausweg. Von Anwendungssituation zu Anwendungssituation wird zu diskutieren sein, wie Verantwortlichkeit zu schaffen bzw. zu erhalten ist und in welchen Bereichen der Einsatz sogenannter autonomer Technologien nicht zu verantworten ist.

Wie oben bei Jonas dargestellt, tut sich bei der Künstlichen Intelligenz eine „*Kluft zwischen* [der] *Kraft des Vorherwissens und* [der] *Macht des Tuns“*[622] auf. Wir geraten wieder in eine Situation, in der wir nur noch unsere „*Unwissenheit“* in Bezug auf die Folgen unseres Handeln konstatieren können und eine „*Selbstbeaufsichtigung unserer übermäßigen Macht*“ unmöglich wird.

---

[621] Arendt (1970), S. 80

[622] Jonas (1979), S. 28

9 Unterminiert die KI die Person oder das Individuum?

„Die Krise der Vernunft manifestiert sich in der Krise des Individuums, als dessen Agens Vernunft sich entwickelt hat. Die Illusion, die die traditionelle Philosophie über das Individuum und die Vernunft gehegt hat - die Illusion ihrer Ewigkeit - , ist im Begriff zu zergehen. Das Individuum bestimmte einmal die Vernunft ausschließlich als ein Instrument des Selbst. Jetzt erfährt es die Kehrseite seiner Selbstvergottung. Die Maschine hat den Piloten abgeworfen; sie rast blind in den Raum. Im Augenblick ihrer Vollendung ist die Vernunft irrational und dumm geworden. Das Thema der Zeit ist Selbsterhaltung, während es kein Selbst zu erhalten gibt.“[623]

Max Horkheimer

Sowohl die Aufklärungskritik als auch zeitgenössische Kritiker der Künstlichen Intelligenz sehen das Individuum in Gefahr. Vor der Beurteilung dieser Hypothese steht die Frage nach dem Individuum. Was kennzeichnet das Individuum in der Bedeutung, wie z.B. Horkheimer oder Koenig es in Gefahr sehen? Bereits bei der Begrifflichkeit bestehen Unterschiede und Überlappungen. Immanuel Kant spricht von der „Person“, Johann Gottlieb Fichte lässt sich zur „Individualität“ aus, der Historiker Larry Siedentop widmet sein Hauptwerk der „*Erfindung des Individuums*“[624] und Ricarda Huch schreibt von der „*Persönlichkeit*“[625] und beklagt die „*Entpersönlichung*“.

Die unterschiedlichen Definitionen und Terminologien sollen in den folgenden Abschnitten dargestellt werden, mit dem Schwerpunkt darauf, was die Person, bzw. das Individuum oder die Persönlichkeit, wie wir sie heute verstehen, auszeichnet. Danach soll Gaspard Koenigs Argumentation zum „*Ende des Individuums*“[626] nachgezeichnet werden. Wie wir sehen werden, bezieht er sich dabei hauptsächlich auf das „Nudging“ des Menschen mithilfe der Künstlichen Intelligenz. Auch der durch Richard Thaler und Cass Sunstein geprägte Begriff „Nudge“[627] wurde unabhängig von der KI zum Gegenstand einer moralphilosophischen Debatte, auf die in einem weiteren Abschnitt eingegangen werden soll.

623 Horkheimer (1947), S. 146

624 Siedentop (2014), Siedentop (2015)

625 Safranski (2021), S. 167f

626 Koenig (2021)

627 Thaler Sunstein (2008)

9.1 Der Personenbegriff bei Immanuel Kant

> *„Dass der Mensch in seiner Vorstellung das **Ich*** [Hervorhebung DS] *haben kann, erhebt ihn unendlich über alle andere auf Erden lebende Wesen. Dadurch ist er eine **Person*** [Hervorhebung DS] *und, vermöge der Einheit des Bewusstseins, bei allen Veränderungen, die ihm zustoßen mögen, eine und dieselbe Person, d.i. ein von Sachen, dergleichen die vernunftlosen Tiere sind, mit denen man nach Belieben schalten und walten kann, durch Rang und Würde ganz unterschiedenes Wesen.“*[628]
>
> Immanuel Kant, Anthropologie in pragmatischer Hinsicht, §1

Kant macht hier den Personenbegriff an zwei *Fähigkeiten* fest. Zum einen geht es ihm um die „***Selbstbezüglichkeit“*** *(„das Ich haben können“)* und zum anderen um die „*Einheit des Bewusstseins bei allen Veränderungen“*, was Georg Mohr eine „***diachrone Identität“*** nennt. Gegen Ende der Anthropologie kommt noch das „***moralische Pflichtbewusstsein“***[629] als dritte definierende Eigenschaft der Person hinzu:

> *„Die moralische Anlage. Die Frage ist hier: ob der Mensch von Natur gut, oder von Natur böse, oder von Natur gleich für eines oder das andere empfänglich sei, nachdem er in diese oder jene ihn bildende Hände fällt (cereus in vitium flecti etc.). Im letztern Falle würde die Gattung selbst keinen Charakter haben. - Aber dieser Fall widerspricht sich selbst; denn ein mit praktischem Vernunftvermögen und Bewußtsein der Freiheit seiner Willkür ausgestattetes Wesen (eine Person) sieht sich in diesem Bewußtsein selbst mitten in den dunkelsten Vorstellungen unter einem Pflichtgesetze und im Gefühl (welches dann das moralische heißt), daß ihm, oder durch ihn anderen recht oder unrecht geschehe. Dieses ist nun schon selbst der intelligibele Charakter der Menschheit überhaupt und in so fern ist der Mensch seiner angebornen Anlage nach (von Natur) gut.“*[630]

In der Metaphysik der Sitten ergänzt Kant die **Zurechnungsfähigkeit** und die **Freiheit der Unterwerfung unter den kategorischen Imperativ**:

> *„Person ist dasjenige Subject, dessen Handlungen **einer Zurechnung fähig*** [Hervorhebung DS] *sind. Die moralische Persönlichkeit ist also nichts anders, als die Freiheit eines vernünftigen Wesens unter moralischen Gesetzen (die psychologische aber bloß das Vermögen, sich der Identität seiner selbst in den verschiedenen Zuständen seines Daseins bewußt zu werden), woraus dann folgt, **daß eine Person keinen anderen Gesetzen als denen, die sie (entweder allein, oder wenigstens zugleich mit anderen) sich selbst giebt, unterworfen*** [Hervorhebung DS] *ist.“*[631]

Zusammenfassend verfügen Personen über eine Selbstbezüglichkeit, eine diachrone Identität (ein integriertes und unveränderliches Bewusstsein), ein moralisches Pflichtbewusstsein, eine Zurechnungsfähigkeit und über die Freiheit, sich dem kategorischen Imperativ

---

[628] Mohr (2001), S. 103; Originalzitat Kant AA VII, Anthropologie in pragmatischer Hinsicht, S. 127; Kant (1798), S. 407 (BA 3,4)

[629] Mohr (2001), S. 103

[630] Originalzitat Kant AA VII, Anthropologie in pragmatischer Hinsicht, S. 324; Kant (1798), S. 677 (A320, B 318)

[631] Originalzitat Kant AA VI, Die Metaphysik der Sitten, S. 223; Kant (1797), S. 329 (AB 22)

zu unterwerfen. Darin begründen sich dann auch die ***„Selbstzweckhaftigkeit und die Würde des Menschen“***[632]:

> *„Ein jeder Mensch hat rechtmäßigen Anspruch auf Achtung von seinen Nebenmenschen, und wechselseitig ist er dazu auch gegen jeden Anderen verbunden.*
>
> *Die Menschheit selbst ist eine Würde; denn der Mensch kann von keinem Menschen (weder von Anderen noch sogar von sich selbst) blos als Mittel, sondern muß jederzeit zugleich als Zweck gebraucht werden, und darin besteht eben seine* ***Würde (die Persönlichkeit)*** [Hervorhebung DS]*, dadurch er sich über alle andere Weltwesen, die nicht Menschen sind und doch gebraucht werden können, mithin über alle Sachen erhebt. Gleichwie er also sich selbst für keinen Preis weggeben kann (welches der Pflicht der Selbstschätzung widerstreiten würde), so kann er auch nicht der eben so nothwendigen Selbstschätzung Anderer als Menschen entgegen handeln, d. i. er ist verbunden, die Würde der Menschheit an jedem anderen Menschen praktisch anzuerkennen, mithin ruht auf ihm eine Pflicht, die sich auf die jedem anderen Menschen nothwendig zu erzeigende Achtung bezieht.“*[633]

Nach Kants Definition der Person und der Persönlichkeit ist jeder Mensch (und nur der Mensch, sonst niemand und nichts) eine Person und verfügt über eine Würde, die ihn zu einer Persönlichkeit macht.

Eine KI erfüllt keines der hier aufgeführten Attribute einer Person oder Persönlichkeit. Sie verfügt weder über die Selbstbezüglichkeit, die diachrone Identität, ein moralisches Pflichtbewusstsein, eine Zurechnungsfähigkeit und die Freiheit zum Befolgen eines kategorischen Imperativs. Vor allem genießt sie keine Selbstzweckhaftigkeit und keine der Menschenwürde vergleichbare Würde.

Im weiteren Verlauf dieses Kapitels wird zu diskutieren sein, inwieweit die zunehmende Anwendung der KI die Würde des Menschen gemäß des dargestellten Personenbegriffs einschränkt.

632 Mohr (2001), S. 115

633 Kant AA VI, Die Metaphysik der Sitten, Tugendlehre § 38, S. 462; Kant (1797), S. 600f (A 140)

9.2 Personalität und Individualität bei Johann Gottfried Fichte

Johann Gottfried Fichte hat das Konzept der Person bzw. des Individuums in einer für die Zwecke des Arguments hilfreichen Weise in dieser Arbeit weiterentwickelt, in Bezug auf zwei Aspekte: das Urrecht des Individuums nur Ursache zu sein und die Einbindung der Freiheit des Individuums in die Freiheitsansprüche anderer Personen.

Fichtes Personenbegriff baut unmittelbar auf denjenigen von Kant auf. Nach ihm „*ist eine Person* […] *ein durch seinen Willen in der Sinnenwelt wirksames, leibhaft-vernünftiges Individuum, das sich eine begrenzte Sphäre der Freiheit im Handeln zuschreibt, in reziproken Anerkennungsbeziehungen mit anderen Personen steht und diesen nach einem allgemeinen Rechtsgesetz ebenso jeweils begrenzte Freiheitssphären einräumt"*[634]. Begrifflich sind darin enthalten „*der Freiheitsbegriff, der Begriff des Individuums bzw. der Individualität, der Leibbegriff, die Interpersonalitätsthese und die These vom Recht (`Unrecht') als wechselseitige Anerkennung"*[635].

In seiner Freiheitstheorie verbindet Fichte die Willensfreiheit mit der Handlungsfreiheit:

> *„Als frei erfahre ich mich in meinen Handlungen. Als meine freien Handlungen und somit als Realisierungen meiner Freiheit schaue ich nur solche Handlungen an, deren Ursache ich selbst (mein Wille) bin. Die Identität der Person ist die Identität der Sphäre ihrer möglichen freien Handlungen."*[636]

Die „*Individualität von Personen* ist *durch deren Wollen bestimmt"*[637]:

> *„Mein Wille ist ursprünglich bestimmt, diese Bestimmtheit meines Willens, macht meinen wahren* ***Charakter*** *als Vernunftswesen aus.* […] *In diesem Wollen nun in der letzten Rücksicht ist nun mein ganzes Sein und Wesen bestimmt für einmal in alle Ewigkeit; ich bin nichts als ein so wollendes, und mein Sein ist nichts als ein so wollen. Dieß ist die ursprüngliche Realität des Ich."*[638]

Der Wille beschreibt den Charakter des Individuums im Kontext der allgemeinen Vernunft. Fichte spricht davon, dass das Individuum seine individuelle Vernunft aus dem Sittengesetz der allgemeinen Vernunft „*herausgreifen*" muss.

Aus seiner Sicht ist es zwingend, dass sich das Individuum als „*Eins, unter mehreren vernünftigen Wesen"* begreift und „*auf andere Individuen"*[639] bezieht. Dies „*impliziert*

---

634 Mohr (2001), S. 119

635 Mohr (2001), S. 120

636 Zitat von Mohr über Fichte: Mohr (2001), S. 121

637 Mohr (2001), S. 125

638 Mohr (2001), S. 125; Originalzitat: Fichte (1798/99), S. 15; Hervorhebung DS

639 Zitat von Mohr über Fichte: Ebd., S. 125; Vgl. Fichte (1796), S.8: *Es findet sich in Absicht dieses Begriffs, daß er notwendig werde dadurch, daß das vernünftige Wesen sich nicht als ein solches mit Selbstbewußtsein setzen kann, ohne sich als ein Individuum, als Eins, unter mehreren vernünftigen Wesen zu setzen, welches es außer sich annimmt, so wie es sich selbst annimmt.*

[…] *die wechselseitige Anerkennung begrenzter personaler Freiheitssphären"*[640]. Personen verstehen ihre eigenen Freiheiten, Rechte und Begrenzungen und respektieren diese gleichzeitig bei anderen Personen. Fichte schreibt dazu in seiner Grundlage des Naturrechts[641]:

> *„Ich setze mich als vernünftig, d.h. als frei. Es ist in mir bei diesem Geschäfte die Vorstellung der Freiheit. Ich setze in der gleichen ungeteilten Handlung zugleich andere freie Wesen. Ich beschreibe sonach durch meine Einbildungskraft eine Sphäre für die Freiheit, in welche mehrere Wesen sich teilen. Ich schreibe mir selbst nicht alle Freiheit zu, die ich gesetzt habe, weil ich auch noch andere freie Wesen setzen, und denselben einen Teil der Freiheit zuschreiben muß."*

Daraus entwickelt Fichte seine gesamte Rechtsphilosophie. Diese beinhaltet zwei fundamentale Grundideen: diejenige der „*fundamentalen Rechte der Person"* (auch „***Urrecht***"[642] genannt) und diejenige „*der Kompatibilität von gegenseitig eingeräumten Sphären freier Handlungen"*[643], wie im obigen Zitat beschrieben.

Zusammenfassend lässt sich festhalten, dass nach Fichte Individuen als Vernunftswesen im Kontext der allgemeinen Vernunft über Willensfreiheit in Kombination mit Handlungsfreiheit bei gleichzeitiger Achtung der Freiheiten anderer Personen verfügen. Nur ein Individuum hat das Recht, ausschließlich Ursache zu sein. Jede Einschränkung dieser Kriterien ist eine Einschränkung der Individualität.

---

640 Mohr (2001), S. 129

641 Fichte (1796), S. 8

642 Zum Urrecht in Mohr (2001), S. 129f: „*Ein Recht das unmittelbar Bedingung der Möglichkeit des Personseins ist, nennt Fichte „Urrecht". Ein Urrecht ist ein Recht, „daß jeder Person, als einer solchen absolut zukommen soll".* […] *In bezug auf jede Person für sich betrachtet ist das Urrecht das absolute Recht der Person, in der Sinnenwelt nur Ursache zu sein. (Schlechthin nie Bewirktes)." Das Urrecht ist das Recht auf eine „fortdauernde, lediglich vom Willen der Person abhängige, Wechselwirkung derselben mit der Sinnenwelt außer ihr"."*

643 Mohr (2001), S. 129f

9.3 Künstliche Intelligenz und das Individuum

> *„Die Bedrohung besteht nicht so sehr im Auftauchen der KI in der realen Welt, sondern in der Veränderung der realen Welt, die der KI entgegenkommen will.“*[644]
>
> Gaspard Koenig, „Das Ende des Individuums“

Der französische Philosoph Gaspard Koenig hat sich in seinen Studien ausführlich mit den Auswirkungen der KI auf das Individuum[645] befasst. In seinen Schlussfolgerungen aus mehr als hundert Interviews mit KI-Experten aus aller Welt setzt er beim Individuum, wie von Kant und Fichte definiert, an. Er stimmt mit Fichte überein, dass ein Individuum in der Lage sein muss, „*eine tiefgreifende und überlegte Wahl zu treffen, die vielleicht suboptimal hinsichtlich des Nutzens* [für das Individuum selbst oder die Gesamtheit aller Individuen] *ist, aber unabdingbar zur Herausbildung einer Individualität“*[646]. Genau dort sieht er das Problem: Die KI hindert uns daran, „*wir selbst zu werden und unsere Willensfreiheit in der Tätigkeit des Überlegens zu begründen“.*

Das Problem liegt nicht oder zumindest nicht primär darin, dass die KI die Individuen in einer für die Individuen selbst oder für die Gesamtgesellschaft nachteiligen Art und Weise beeinflusst, sondern dass sie nach und nach die *„Fähigkeit* [der] *Ausübung der Willensfreiheit als Überlegung* […] *an die Maschine“* delegiert. An dieser Stelle grenzt er sich klar von Yuval Harari ab:

> *„Wo uns also Harari erklärt, dass die KI uns besser kennt als wir uns selbst und dass es besser wäre, unsere Entscheidungen mangels freiem Willen an sie zu delegieren, dort meine ich, dass die KI uns daran hindert, wir selbst zu werden und unsere Willensfreiheit in der Tätigkeit des Überlegens zu begründen. Wo sich Harari damit brüstet, dem Individuum zu entsagen und das Ich in der Meditation aufzulösen, dort schlage ich vor, das einzigartige Wesen, das wir sind, zu bestärken oder ihm zumindest die Möglichkeit einer Zukunft zu lassen. Wenn schon nicht für uns, dann wenigstens für den Fortbestand der Art.“*[647]

Koenig sieht die systematische Untergrabung des freien Willens im „Nudging“, wie es von den Verhaltensökonomen Richard Thaler, Cass Sunstein und auch Daniel Kahneman diskutiert wird.

> *„Ich bin gewiss nicht ‚gegen‘ die KI – ich muss trotzdem feststellen, dass ihre industriellen Anwendungen, so wie sie heute konzipiert sind, eine Gefahr für die Entscheidungsfähigkeit des Individuums darstellen, indem sich auf utilitaristischer Basis die Techniken des* ***nudge*** *vervielfältigen. Die Liberalen haben immer Recht, den Fortschritt zu rühmen und die tausendfache List des Konservativismus zu durchkreuzen. Aber sie müssen sich heute ernsthaft selbst befragen.* ***Ist nicht das Prinzip Innovation, für das sie so sehr gekämpft***

644 Koenig (2021), S. 122

645 Koenig (2021)

646 Dieses und folgende Zitate: Koenig (2021), S. 318

647 Koenig (2021), S. 318

> ***haben, darin begriffen, das Fundament ihres eigenen Systems zu untergraben?“*** [648] [Hervorhebungen durch DS]

Verfechter der KI bestreiten dies vehement. Im Gegenteil, so argumentieren sie, erlaube die KI eine Optimierung und maximale personalisierte Anpassung des Produkt- und Dienstleistungsangebots an die Bedürfnisse des Individuums. Genau hier sieht Koenig ein Paradox:

> *„Die Wissensdispositive sind aus den Fugen geraten, wie Foucault es vorweggenommen hat: Der Mensch ist nicht mehr ein bewusster Produzent von Erkenntnis, sondern ein passiver Empfänger von Informationen.* ***Paradoxerweise führt die Hyper-Personalisierung der Produkte zu einer Deinidividualisierung des Subjekts****. Man konsumiert Maßarbeiten, verliert sein* ***eigenes Maß****. Dem beständigen Strom der Verlockung ausgesetzt und vor jeder Introspektion bewahrt, verliert der digitale Mensch seine intellektuelle und moralische Autonomie.“*[649] [Hervorhebungen durch DS]

Nur, wenn der Mensch sich in der Erkenntnis seiner eigenen Bedürfnisse überlegen kann, was er will, und darauf basierend seine Auswahl treffen kann, wird er zum Individuum. Wenn dieses systematisch verweigert wird, verliert er seine Autonomie und seine Individualität.

In ihrem Buch „*Das Zeitalter des Überwachungskapitalismus*“[650] zitiert Shoshana Zuboff den MIT-Informatikprofessor und Begründer der „*Theorie der instrumentären Gesellschaft*“[651] Alex Pentland, der in seinem Artikel „*Der Tod der Individualität*“ recht unverblümt genau dies begrüßt:

> *„Statt von rationalen Individuen scheint unsere Gesellschaft von einer kollektiven Intelligenz regiert zu werden, die aus dem Fluss von Ideen und Beispielen aus der Umgebung kommt ... Es wird Zeit, dass wir die Fiktion vom Individuum als Basiseinheit der Rationalität fallen lassen und erkennen, dass unsere Rationalität größtenteils vom sozialen Gefüge um uns herum bestimmt wird.“*[652]

Pentland spricht in einem anderen Aufsatz von der „*Gottesperspektive unserer selbst*“[653], die dadurch entsteht, dass das Verhalten aller Menschen sichtbar wird, so dass sich alle aufeinander abstimmen können. Die Voraussetzung dafür ist, dass sich „*jeder von uns* [...] *willig in ein bis ins Letzte vermessenes Leben instrumentärer Ordnung“* fügt:

> *„Für die Gesellschaft ergibt sich die Hoffnung, dass wir dieses tiefe Verständnis individuellen Verhaltens dazu nutzen können, Effizienz und Reaktionsfähigkeit von Industrie*

---

[648] Koenig (2021), S. 171f

[649] Koenig (2021), S. 176

[650] Zuboff (2018)

[651] Zuboff (2018), S. 481

[652] Zuboff (2018), S. 505; Originalzitat: Pentland (2014), S. 31

[653] Zuboff (2018), S. 494; Originalzitat: Alex Pentland, “*Society’s Nervous System: Building Effective Government, Energy, and Public Health Systems*”, MIT Open Access Articles, October 2011, http://dspace.mit.edu/handle/1721.1/66256

> *und Staat zu erhöhen. Für Individuen liegt die Attraktivität in der Möglichkeit einer Welt, in der alles bequem zu haben ist – Ihr Check-up beim Arzt fällt mit dem Einsetzen der Krankheit zusammen, der Bus kommt genau in dem Augenblick, in dem Sie an die Haltestelle kommen, und auf Ämtern brauchen Sie nie wieder Schlange zu stehen. Da diese Fähigkeiten durch den Einsatz immer raffinierterer statistischer Modelle und Sensoren ständig verfeinert werden, könnten wir sehr gut das Entstehen einer quantitativen prädiktiven Wissenschaft menschlicher Organisationen und der menschlichen Gesellschaft erleben."*[654]

Er reflektiert nicht über den Preis, den wir alle für diesen Komfort und die Bequemlichkeit zahlen, nämlich die Preisgabe unserer Freiheit, Autonomie und Individualität. In der von ihm anvisierten Welt treiben nicht unser individuelles Denken, Wollen und Urteilen unser Verhalten, sondern die statistischen Modelle, die das Verhalten der gesamten Gesellschaft auswerten.

Man muss Pentland dennoch zugutehalten, dass er am Ende seines Aufsatzes „*The death of individuality*" auf die gewaltige Gefahr einer massenhaften Verhaltensmanipulation hinweist:

> *"This power of the social fabric on individual decision-making is, in fact, the real reason that privacy is so important. As Stanley Milgram's work on social conformity demonstrated many years ago, the power of social influence can lead people to both good and terrible behaviours, and can transform transform our behaviour to an extent that is scarcely believable. Without privacy, the power of corporations or government to manipulate our behaviour becomes virtually unlimited*."[655]

Als Lösung dieses Problems schlägt Pentland „trust networks" (Vertrauensnetzwerke)[656] vor. Dabei soll es sich um Netzwerke handeln, die es dem Individuum gestatten, die Verwendung seiner Daten zu kontrollieren. In der Theorie erscheint dies als interessante Idee. In der Praxis wird es viele Umgehungen dieser Netzwerke geben. Vor allem erkennen viele Menschen die inhärente Gefahr nicht und daher auch nicht die Notwendigkeit, ihre Daten über ein derartiges Netzwerk zu schützen.

Zuboff sieht in der instrumentären Gesellschaft eine „*Aufgabe der Freiheit zugunsten des Wissens"*[657] und verweist auf das Werk des Behavioristen B. F. Skinner, der in seinem 1971 veröffentlichten Buch „*Beyond Freedom and Dignity*" (dt.: *Jenseits von Freiheit und Würde*) schrieb:

> *„Was im Begriff ist, abgeschafft zu werden, ist der „autonome Mensch" – der innere Mensch, der Homunkulus, der besitzergreifende Dämon, der Mensch, der von der Literatur der Freiheit und Würde verteidigt wird. Seine Abschaffung ist seit langem überfällig ... Er ist ein Produkt unserer Unwissenheit, und während unser Wissen wächst, löst sich die*

---

654 Zuboff (2018), S. 495

655 Pentland (2014), S. 31

656 Vgl.: https://www.media.mit.edu/articles/alex-sandy-pentland-discusses-trust-networks-at-the-world-economic-forum-2/

657 Zuboff (2018), S. 505

> *Substanz, aus der er gemacht wird, immer mehr in Nichts auf. Die Wissenschaft ... „dehomunkulisiert" ihn, und es bleibt ihr nichts anderes übrig, wenn sie der Abschaffung der menschlichen Spezies vorbeugen will: Wir können froh sein, wenn wir uns von diesem Menschen im Menschen befreit haben. Nur wenn wir ihn seiner Rechte entsetzen, können wir ... vom Abgeleiteten zum Beobachteten, vom Wunderbaren zum Natürlichen, vom Unzulänglichen zum Beeinflussbaren."*[658]

Hannah Arendt hat genau diese Entwicklung bereits 1958 im englischen Original ihres 1967 in deutscher Übersetzung erschienenen Buches „*Vita Activa oder vom tätigen Leben*" vorausgesehen und beklagt:

> *„Vergleicht man die moderne Welt mit den Welten, die wir aus der Vergangenheit kennen, der drängt sich vor allem der enorme* ***Erfahrungsschwund*** [Hervorhebung DS] *auf, der dieser Entwicklung inhärent ist. Nicht nur, daß die anschauende Kontemplation keine Stelle mehr hat in der Weite spezifisch menschlicher und sinnvoller Erfahrungen, auch das Denken, sofern es im Schlußfolgern besteht, ist zu einer Gehirnfunktion degradiert, welche die elektronischen Rechenmaschinen erheblich besser, schneller und reibungsloser vollziehen als das menschliche Gehirn."*[659]

Es ist der individuelle Erfahrungsschwund zumindest für die meisten Individuen, der einher geht mit dem kollektiven Wissenszuwachs, der in den elektronischen Rechenmaschinen in der Sprache Arendts bzw. in den Algorithmen der KI abgelegt ist. Arendt erkannte bereits damals, was dies für die Arbeitsgesellschaft bedeuten würde:

> *„In ihrem letzten Stadium verwandelt sich die Arbeitsgesellschaft in eine Gesellschaft von Jobholdern, und diese verlangt von denen, die ihr zugehören, kaum mehr als ein automatisches Funktionieren, als sei das Leben des Einzelnen bereits völlig untergetaucht in den Strom des Lebensprozesses, der die Gattung beherrscht, und als bestehe die einzige aktive, individuelle Entscheidung nur noch darin, sich selbst gleichsam loszulassen,* ***seine Individualität aufzugeben*** *bzw. die Empfindungen zu betäuben, welche noch die Mühe und Not des Lebens registrieren, um dann völlig „beruhigt" desto besser und reibungsloser funktionieren zu können."*[660] [Hervorhebung DS]

Arendt bezog sich auf die zu ihrer Zeit herrschende Arbeitsgesellschaft. Sie kannte die Entwicklungen der KI in den ersten Jahren des neuen Millenniums noch nicht. Heute würde sie diese Schlussfolgerungen für die Individualität auch auf Bereiche jenseits der Arbeitsgesellschaft ausweiten. Genau wie Zuboff verknüpft sie diese Entwicklung mit dem Behaviorismus, ohne Skinner namentlich zu nennen:

> *„Das Beunruhigende an den modernen Theorien des Behaviorismus ist nicht, daß sie nicht stimmen, sondern daß sie im Gegenteil sich als nur zu richtig erweisen könnten, daß sie vielleicht nur in theoretisch verabsolutierender Form beschreiben, was in der modernen Gesellschaft wirklich vorgeht."*[661]

---

658 Zuboff (2018), S. 505; Skinner (1971), S. 205f

659 Arendt (1967), S. 410

660 Arendt (1967), S. 410-411; auch zitiert in Zuboff (2018), S. 445

661 Arendt (1967), S. 410-411

Vor mehr als siebzig Jahren hat Skinner in seiner Utopie bereits beschrieben, wie das vollständige Wissen über die Verhaltensweisen des Menschen, dessen technische Verwirklichung er damals noch nicht kannte. Auf Basis von „*Reiz-Reaktions-Schemata*" wird eine präzise Voraussage künftiger Verhaltensweisen ermöglicht. Die „*Psychologie*" wird „*damit in eine exakte Wissenschaft*"[662] verwandelt. Nach diesem Verständnis verliert der Mensch seine Individualität und Autonomie und wird vom Subjekt zum Objekt degradiert.

Auch Hannah Arendt konnte Skinners Sichtweise einige Jahre später nur zustimmen. Und auch einige Informatiker der heutigen Zeit, wie z.B. Alex Pentland, würden beipflichten. Trotzdem sind sich die drei nicht einig über die Richtung und Ausprägung der normativen Konsequenzen[663].

Die Technik, hier die KI, kann die Schranke zwischen Objekt und Subjekt nicht überschreiten; es bleibt nur die Abschaffung des Subjekts, der Verlust der originären Individualität oder des Individuums – oder in Zuboffs Worten „*Die Automatisierung des Selbst als Vorbedingung der Automatisierung der Gesellschaft"*[664].

662 Deutscher Ethikrat (2023), S. 19

663 Der Vollständigkeit halber sollte auch angemerkt werden, dass Pentland durchaus vor negativen Auswirkungen gewarnt hat: „*As these new abilities become refined by the use of more sophisticated statistical models and sensor capabilities, we could well see the creation of a quantitative, predictive science of human organizations and human society. At the same time, these new tools have the potential to make George Orwell's vision of an all-controlling state into a reality. **As a consequence, we need to think carefully about the growth and increasingly broad usage of personal data to drive societies systems, and particularly about the safety, stability, and fairness of their design**."*; Quelle: Alex Pentland, "*Society's Nervous System: Building Effective Government, Energy, and Public Health Systems*", MIT Open Access Articles, October 2011, http://dspace.mit.edu/handle/1721.1/66256, [Hervorhebung DS]

664 Zuboff (2018), S. 445

9.4 Die Ethik des Nudging

Die Art und Weise wie bei Nutzung der Künstlichen Intelligenz in den sozialen Medien das Verhalten der Menschen beeinflusst wird, rangiert häufig unter dem Begriff „Nudging“ (englisch für „Schubsen“, „Anstoßen“). Die KI bietet dem Menschen bestimmte Optionen, auf die er selbst in den allermeisten Fällen nicht gekommen wäre, und nimmt so Einfluss auf sein Verhalten. In diesem Abschnitt soll der verhaltensökonomische Hintergrund erläutert und der ethische Diskurs zum Thema dargestellt werden.

Seit Ende des letzten Jahrhunderts hat die Volks- und Betriebswirtschaftslehre über die Verhaltensökonomie eine signifikante Umwälzung erfahren, in der man sich vom Menschenbild der klassischen Ökonomie, dem Homo Oeconomicus, entfernt hat, das sich auf die „*drei zentralen Annahmen der unbegrenzten Rationalität*, der *unbegrenzten Willenskraft* und *dem unbegrenzten Eigennutzstreben“* begründete[665]. Mit der Entwicklung der Verhaltensökonomie (Behavioral Economics) hat sich zunehmend abgezeichnet, dass die „*menschliche Realität“* von diesem traditionellen Bild erheblich abweicht[666]:

1. ***„Begrenzte Rationalität****: Menschen machen Fehler bei der Informationsaufnahme und -Verarbeitung. Aufgrund ihrer begrenzten Fähigkeiten, Informationen aufzunehmen und zu verarbeiten, greifen sie zu sogenannten Heuristiken, also einfachen Problemlösungsmechanismen. Bei der Anwendung dieser Heuristiken, so eine Idee der Behavioral Economics, kann es zu Verhaltensweisen kommen, die von der ökonomischen Rationalität abweichen und unter Umständen zu systematischen Fehlern führen können.* […]
2. ***Begrenzte Willenskraft****: Menschen vertagen unbequeme Entscheidungen, sie verschieben Diäten, vernachlässigen ihre Altersvorsorge und drücken sich vor unangenehmen Entscheidungen* [oder setzen rational gebotene Maßnahmen aus Mangel an Disziplin oder Willenskraft nicht um. …]
3. ***Begrenzter Eigennutz****: Menschen sind nicht ausschließlich egoistisch, sie sorgen sich um andere Menschen, sind auf Fairness bedacht und bereit, für die Bestrafung unfairer Mitmenschen zu bezahlen* [und auf eigene Vorteile zu verzichten.]*“*

Eine Vielzahl von Ökonomen hat sich um die Entwicklung der Verhaltensökonomie verdient gemacht. Zu nennen sind insbesondere die Nobelpreisträger von 2002, Daniel Kahneman und Vernon Smith, sowie der Nobelpreisträger von 2017, Richard Thaler. Letzterer hat sich gemeinsam mit Cass Sunstein mit dem Thema befasst, um das es in diesem Abschnitt gehen soll: das Nudge Concept.

[665] Vgl. Beck (2014), S. 2

[666] Die folgenden Punkte zitiert aus Beck (2014), S. 2f

In ihrem Buch „*Nudge – Wie man kluge Entscheidungen anstößt*“[667] haben sie „*ein Konzept entwickelt, wonach der Staat das Verhalten der Bürger beeinflussen können soll, ohne Ge- und Verbote, Steuern und Subventionen einzusetzen, und wonach er das Leben der Bürger `länger, gesünder und besser´ machen können soll, ganz* [so behaupten sie; Anmerkung DS] „*ohne deren Freiheit zu beschneiden*“[668].

Thaler und Sunstein definieren „Nudge“ wie folgt:

> *„Unter Nudge verstehen wir* […] *alle Maßnahmen, mit denen Entscheidungsarchitekten das Verhalten von Menschen in vorhersehbarer Weise verändern können, ohne irgendwelche Optionen auszuschließen oder wirtschaftliche Anreize stark zu verändern.“*[669]

In Studien der Verhaltensökonomie wurden Heuristiken identifiziert und empirisch bestätigt, die streng genommen nicht rational sind. Bei diesen setzt dann das Nudging an[670]. Ein Beispiel ist die Ausgestaltung einer Ausgangssituation als Standardvorgabe, von der man dann – falls man sie explizit nicht wünscht – aktiv abweichen muss. Dies geschieht in vielen Ländern bei der Organspende. Die Standardvorgabe ist, dass man Organspender ist und als solcher in den Datenbanken geführt wird. Nur wenn man das nicht möchte und aktiv ausschließt („*opt out*“), wird man in eine Negativliste aufgenommen. Damit erreicht man eine deutlich höhere Spenderquote als mit einem Modell, in dem man sich aktiv für das Organspenden entscheiden muss. Das Gleiche gilt für Altersvorsorgeprogramme oder die Teilnahme an Krebsvorsorgeuntersuchungen. Ein anderes Beispiel sind sogenannte „Soziale Nudges“. Hier wird sozialer Druck erzeugt, in dem darüber informiert wird, wie sich die anderen Bürger (mehrheitlich) entschieden haben, z.B. beim Recycling von Wertstoffen oder beim Sparen von Energie und Wasser. In die gleiche Kategorie passt die Bereitstellung von (selektiven) Informationen für die zu treffende Entscheidung (z.B. Reduzierung des CO2-Ausstoßes beim Kauf eines bestimmten Produktes).

Thaler und Sunstein geht es in ihrem ursprünglichen Buch darum, wie staatliche Stellen den Bürger zu einem gesünderen, nachhaltigeren und sozialeren Verhalten motivieren können. Die beschriebenen Prinzipien und Wirkungsweisen gelten aber auch für kommerzielle Unternehmen, die die Verhaltensweisen ihrer Kunden beim Einkauf von Produkten und Dienstleistungen zum eigenen Nutzen beeinflussen möchten. Weiterhin lassen sich Nudges für den massenhaften Einsatz mit Hilfe der Künstlichen Intelligenz personalisieren. Dies geschieht, wenn etwa Amazon dem Kunden nach einer Bestellung eine Liste von Angeboten präsentiert, die sich andere Kunden im Zusammenhang mit dem ersten Produkt auch gekauft haben. Dies geschieht bei Netflix, wenn eine Liste ähnlicher Filme vorgelegt wird („Weil Sie Harry Potter gesehen haben ...“). Mit Hilfe der KI erfährt

667 Thaler Sunstein (2008); Thaler Sunstein (2009)

668 Wolff (2015), S. 195

669 Thaler Sunstein (2009), S. 15

670 Vielzahl der folgenden Beispiele aus Wolff (2015), S. 200f

der Bürger oder der Kunde immer wieder einen Schubs in eine bestimmte Richtung, die aus Sicht einer übergeordneten Rationalität sinnvoll erscheint, sei es für die breitere Gesellschaft oder auch für die einzelne Person (z.B. bei Fragen der Gesundheit) oder auch für das anbietende Unternehmen.

Thaler und Sunstein bezeichnen dies entsprechend als eine Form des „***Libertären Paternalismus***“:

> *„Libertäre Paternalisten wollen es den Menschen leichtmachen, ihren eigenen Weg zu gehen; sie möchten niemanden daran hindern, von seinen Freiheitsrechten Gebrauch zu machen. Paternalismus ist deshalb wichtig, weil es unserer Überzeugung nach für Entscheidungsarchitekten legitim ist, das Verhalten der Menschen zu beeinflussen, um ihr Leben länger, gesünder und besser zu machen. Anders gesagt, wir sind dafür, dass private Institutionen, Behörden und Regierungen bewusst versuchen, die Entscheidungen der Menschen so zu lenken, dass sie hinterher besser dastehen – und zwar gemessen an ihren eigenen Maßstäben.“*[671]

Trotzdem können diese „Schubse“ dem Interesse der Einzelperson auch zuwiderlaufen, ohne dass der Betreffende dies wahrnimmt. Aus einer moralphilosophischen und grundrechtlichen Erwägung erhebt sich die Frage, inwieweit dieser durch das Nudging geförderte Paternalismus noch dem Menschenbild des aufgeklärten und mündigen Bürgers entspricht. Wo ist der Unterschied zum Pfarrer, der von der Kanzel den Gemeindemitgliedern sagt, was sie zu ihrem eigenen Seelenheil oder im Hinblick auf das Jüngste Gericht zu tun haben?

Die Kritik richtet sich hier zum einen gegen die Idee des Nudging als Programm der Massenbeeinflussung ohne demokratische Legitimation und zum anderen gegen die massenhafte Ausbreitung desselben mit den technologischen Möglichkeiten der KI. Dies gilt primär für den Einsatz in autoritären Regimes wie China oder Russland, jedoch auch in den westlichen Demokratien.

Karen Yeung prägte für das systematische dynamische Nudging mit Methoden der KI und Big Data den Begriff „*Hypernudge*“ :

> „Big Data analytic nudges are extremely powerful and potent due to their networked, continuously updated, dynamic and pervasive nature (hence ‘**hypernudge**’).” [...] [D]igital decision-‚guidance‘ processes are designed so that it is not the machine, but the targeted individual, who makes the relevant decision. These technologies seek to direct or guide the individual’s decision making processes in ways identified by the underlying software algorithm as ‘optimal’, by offering ‘suggestions’ intended to prompt the user to make decisions preferred by the choice architect”[672], [Hervorhebung DS]

[671] Thaler Sunstein (2009), S. 14f

[672] Yeung (2017), S. 118f

Mit "*datenbasierten algorithmischen Systemen"* wird eine "*hochgradig personalisierte Informationsumgebung"* geschaffen, der man sich „*nur schwer entziehen kann*"[673]. Damit wird die menschliche Autonomie und Individualität nicht gestärkt, sondern eingeschränkt.

Aus moralphilosophischer Sicht werden verschiedene wiederkehrende Argumente gegen das Nudging und damit auch gegen den Einsatz von KI zwecks Nudging vorgebracht[674]:

1. ***Einschränkung der Wahlfreiheit des autonomen Individuums*** *(„Freedom of choice")*: Es wird argumentiert, dass Nudges die Wahlfreiheit einschränken und nicht „leicht abwendbar" sind, wie die Unterstützer des Nudging behaupten.
2. ***Einschränkung der Willensautonomie*** *(„Volitianal autonomy")*: Das Argument lautet, dass Menschen, die einem Nudging unterworfen werden, nicht mehr die wahren Urheber ihrer Entscheidungen sind:
   „We are no longer the `authors´ of our choices: They are not really our own anymore in that they do not reflect our own autonomous desires. Nudgers pull our strings and employ tricks to get us to do what <u>they</u> want. When subjected to nudges, we may be influenced so that our resulting actions are no longer genuinely our own."[675]
3. ***Unterminieren des rationalen Handelns*** *("Rational agency")*: Dadurch, dass die Nudges offen die Entscheidungsprozesse und -mechanismen nutzen, die außerhalb der Rationalität des Homo Oeconomicus liegen, und teilweise irrational oder arational sind, geht der Respekt vor die Rationalität des Menschen verloren:
   "Even if nudges respect people's freedom and promote their goals and well-being, that nudgers tap into our irrational or arational heuristics and biases means they fail to treat us like rational human beings and thereby condescend and infantilize us."[676]
   Damit zusammenhängend nimmt man den Menschen die Möglichkeit, auch einmal falsche Entscheidungen zu treffen und aus Fehlern zu lernen:
   „… nudging can `deprive people of the capacity for making wrong choices´ and erodes their responsibility for their own choices. […] *If responsibility for making choices is moved away from individuals, their practical judgement and decision-making capacities cannot develop, which in turn undermines their moral independence".*

[673] Deutscher Ethikrat (2023), S. 135

[674] Vgl. Schmidt Engelen (2020), S. 4f

[675] Schmidt Engelen (2020), S. 4f

[676] Schmidt Engelen (2020), S. 5; sowie folgende Zitate

4. ***Manipulation und Unterminierung der Würde*** *(„Domination, manipulation and dignity“)*: Dieses Argument geht einher mit den vorherigen Überlegungen. Die Menschen werden mit einem Prozess in eine Richtung manipuliert, ohne dass darüber Transparenz herrscht und ohne explizite demokratische Legitimation.
5. ***Unzulässige Zwecke*** *(„Illicit ends“)*: Generell lässt das Nudging keinen Raum für einen Diskurs über Zwecke oder Ziele. Vielleicht will das Individuum keine großzügige Altersversorgung (weil es vielleicht mit einer Erbschaft rechnet). Dies kann sehr stark in die Richtung eines „***exzessiven Paternalismus“*** gehen.

Für jedes dieser Argumente gibt es eine Reihe von Gegenargumenten. Der Diskurs ist weit davon entfernt, abgeschlossen zu sein. Alle Argumente wiegen aber umso schwerer, wenn es nicht um Nudging durch einen fürsorglichen Staat geht, sondern um Nudging durch kommerzielle Wirtschaftsunternehmen oder totalitäre Regime.

9.5 Zusammenfassung: KI und Individualität

Zu keinem der vier in dieser Arbeit gewählten Beurteilungsdimensionen der KI gibt es kontroversere Diskussionen darüber, dass der gewählte Aspekt der normativen Betrachtung schützenswert ist oder nicht, als bei der Dimension der Individualität. Kaum ein Gelehrter stellt Autonomie und Freiheit in Frage, keiner plädiert für eine verantwortungslose Welt und auch das Ziel des Schutzes der Menschenwürde wird nicht infrage gestellt. Kontrovers wird die Diskussion darüber, ob diese Werte durch die KI bedroht sind oder nicht, ob die KI autonom sein kann oder nicht, ob die KI die Menschenwürde stärkt oder nicht und ob sie Verantwortung übernehmen kann oder nicht. In Bezug auf die Individualität sieht die Konfliktlinie anders aus. Bei ihr herrschen kontroverse Meinungen darüber, ob sie tatsächlich durch die KI eingeschränkt oder unterdrückt wird, ähnlich wie bei den anderen Aspekten. Zusätzlich gibt es eine Reihe von impliziten und expliziten Positionen, die sie nicht mehr für zeitgemäß und vor allem, aus einer konsequentialistischen Logik, nicht mehr für hilfreich für die menschliche Entwicklung halten. Behavioristen wie Skinner und Pentland fordern unverhohlen „Verhaltenstechnologien“ (Skinner) oder setzen sich für „Social Physics“ (Pentland) ein. B. F. Skinner schreibt in der Zusammenfassung von „*Jenseits von Freiheit und Würde*“:

> ***„Hilfe durch Verhaltenstechnologie.*** *Physikalische und biologische Technologien haben uns befreit von Seuchen, Hungersnöten und vielen schmerzhaften, gefährlichen und kräftezehrenden Merkmalen des Alltags. Nun kann eine Verhaltenstechnologie beginnen, uns von anderen Arten von Übeln zu befreien. Bei der Analyse menschlichen Verhaltens befinden wir uns möglicherweise in einem Stadium, das demjenigen entspricht, in dem sich Newton*

> *bei der Analyse des Lichts befand; denn wir haben gerade erst angefangen, Technologien praktisch anzuwenden.“*[677]

Inzwischen sind fünfzig Jahre vergangen. Mittlerweile wurden Technologien entwickelt, die sich Skinner in den früher 1970er Jahren noch gar nicht vorstellen konnte. Zudem haben sich andere wissenschaftliche Disziplinen, wie die Wirtschaftswissenschaften, hin zu einer Verhaltensbeeinflussung des Individuums weiterentwickelt (wie z.B. Thaler Sunstein in „The Nudge“). Sowohl die Technologien (insbesondere der KI) als auch die Verhaltensökonomie befinden sich noch in frühen Phasen einer Verwirklichung ihres Entwicklungs- und Gestaltungspotentials.

Mit der Künstlichen Intelligenz wird die Realisierung der beunruhigenden und provokativen Ideen der Behavioristen innerhalb und außerhalb der KI-Industrie und auch einiger Verhaltensökonomen ermöglicht. Es wird vielfach übersehen, dass damit auch Tür und Tor geöffnet wird für einen neuen Totalitarismus durch die Hintertür oder eine Stärkung des Totalitarismus von oben (wie z.B. in China).

Es soll hier nicht in Frage gestellt werden, dass die größten und wirksamsten Neuerungen und Fortschritte in Gesellschaft, Wirtschaft, Technik, Wissenschaft und Kultur durch das Zusammenwirken von Mensch und Technologie zustandegekommen sind und dass jeder Beitrag, den eine Technologie wie die KI dabei leisten kann, die Zusammenarbeit zu verbessern, dem weiteren Fortschritt dient. Allerdings darf dies niemals zulasten der Individualität und der Gedanken-, Willens- und Handlungsfreiheit des Einzelnen gehen. Das Individuum selbst muss entscheiden können, wann und wie es sich in kollektive Vorhaben einbringen möchte und wann nicht.

Es gilt nicht nur, Freiheit und Autonomie sowie Menschenwürde und die menschliche Verantwortung zu wahren, sondern auch explizit das Individuum mit seinen Möglichkeiten zu schützen. Dies bedeutet – wie oben bei der Analyse des Personenbegriffs von Fichte festgestellt – letztlich den Schutz des Menschen vor der Charakterlosigkeit und die Wahrung des Urrechts der Person.

[677] Skinner (1971), S. 219

10 Verletzt die KI die Menschenwürde?

> *„Die Würde des Menschen ist unantastbar. Sie zu achten und zu schützen ist Verpflichtung aller staatlichen Gewalt."*
>
> Art. 1 Abs. 1 Grundgesetz der Bundesrepublik Deutschland

Die Künstliche Intelligenz reiht sich nicht zwangsläufig in die Liste der offensichtlichen und in der moralphilosophischen und juristischen Literatur ausführlich erörterten Verstöße gegen die Menschenwürde ein. Gleichwohl wird die Künstliche Intelligenz zusammen mit anderen Technologien, wie z.B. *Eingriffen in das Gehirn* und *Eingriffen in das Genom,* in letzter Zeit als eine potenzielle Bedrohung der Menschenwürde thematisiert, wie etwa 2018 im Deutschen Ethikrat[678]. Auch wenn man sich in diesen Diskussionen darüber einig ist, dass die KI die Menschenwürde „*tangieren könnte"* und möglicherweise „*unsere Fähigkeit, als moralisch verantwortlich handelnde Akteure zu handeln, schwächt"*[679], bleibt die Argumentation vergleichsweise unpräzise und oberflächlich. Dies zeigt auch das Schlusswort von Peter Dabrock, dem damaligen Vorsitzenden des Deutschen Ethikrates:

> *„Diesseits der Richtigkeit des Instrumentalisierungsverbotes, des Nicht-demütigen-Sollens und mit Blick auf Technologien, die einerseits auf leisen Sohlen daherkommen und andererseits einen zunehmend starken Einfluss auf unser Leben gewinnen werden, ist deutlich geworden, dass man, gerade wenn man Menschenwürde als ein Achtungskonzept und Schutzkonzept begreifen möchte, es einbetten muss in eine Kultur der Achtung voreinander und in eine politische und demokratische Kultur."*[680]

Auch jenseits der Künstlichen Intelligenz hat sich der Umgang mit dem Begriff der Menschenwürde als schwierig und unbefriedigend erwiesen, wie Ralph Stoecker erklärt:

> *„Der Rückgriff auf die Menschenwürde ist in der modernen angewandten Ethik verbreitet, sein Stellenwert ist aber,* [...] *moraltheoretisch höchst umstritten. Es ist nicht nur unklar, ob die Menschenwürde tatsächlich ein eigenständig zu respektierendes Gut ist, sondern vor allem, was Menschenwürde überhaupt ist und welches Verhalten sie von uns fordert. Diese Unklarheiten wecken wiederum Zweifel daran, ob es sinnvoll ist, der Menschenwürde einen prominenten Platz in der Ethik zuzuweisen ..."*[681]

---

678 Vgl. Hering et al. (2018), S. 381

679 Hering et al. (2018), S. 318; Zitat über den Beitrag von Paula Boddington: „*In ihrer anschließenden Analyse der Berührungspunkte zwischen Menschenwürde und künstlicher Intelligenz formulierte die britische Philosophin Paula Boddington ein klares Plädoyer dafür, schon gegenüber gegenwärtigen Anwendungen wachsam zu sein. Bemühungen von Konzernen und Regierungen, mithilfe lernender Algorithmen menschliche Kommunikation zu beeinflussen, statt Menschen zu ermuntern, ihre eigene intellektuelle und moralische Urteilskraft zu entwickeln, seien ein besonders eindrückliches Beispiel dafür, wie künstliche Intelligenz nicht eingesetzt werden sollte.* ***‚Eine solche Verwendung künstlicher Intelligenz könnte die Menschenwürde bedrohen, indem sie unsere Fähigkeit, als moralisch verantwortliche Akteure zu handeln, schwächt'****", so Boddington.* [Hervorhebung DS]

680 Zitiert in Hering et al. (2018), S. 383

681 Stoecker (2019), S. 33

Er kommt allerdings zu dem Ergebnis, „*dass dies nicht nur sinnvoll, sondern sogar notwendig ist, weil man ohne Rückgriff auf die Menschenwürde dem besonderen ethischen Charakter bestimmter Handlungen nicht gerecht werden kann, deren Verwerflichkeit darin liegt, ihre Opfer zu demütigen, zu erniedrigen, zu entwürdigen"*[682].

Nach einigen gescheiterten Versuchen, eine in sich geschlossene positive Beschreibung zu formulieren, stehe fest, „*dass es sich lohnt, weniger auf die Menschenwürde selbst, als vielmehr auf Menschenwürdeverletzungen zu schauen, denn diese sind es meistens, die die Leute dazu veranlassen, sich auf die Würde des Menschen zu berufen"*[683]. Weiterhin könne man nicht bestreiten, dass der *„Begriff eine bemerkenswerte historische Karriere gemacht"* habe. Er schlägt deshalb einen induktiven, negativen und historischen Angang vor:

> *„Induktiv soll man den Begriff in verschiedenen Anwendungskontexten untersuchen und jeweils fragen, worin dort das Interesse und der Stellenwert der Menschenwürde liegen. Negativ ist die Strategie insofern, als sie davon ausgeht, dass es philosophisch am interessantesten ist, Menschenwürde von ihren Verletzungen her zu verstehen. Und das Bewusstsein der Geschichte des Begriffs ist wichtig, weil das moderne Verständnis der Menschenwürde zu großen Teilen auf ihre Verwendung außerhalb der Philosophie, im öffentlichen Raum in den letzten zweihundertfünfzig Jahren zurückgeht."*

Dieser Ansatz soll auch der Leitfaden des Herausarbeitens konkreter Risiken der Menschenwürdeverletzung durch die Künstliche Intelligenz sein, beginnend mit dem historischen Angang (Kapitel 10.1). Dann soll aus unterschiedlichen Perspektiven versucht werden, die Menschenwürde von ihren KI-induzierten Verletzungen her zu verstehen, veranschaulicht durch Beispiele aus spezifischen Anwendungskontexten. Konkret geht es um die folgenden Ansätze:

- Menschenwürde und der Subjektcharakter des Menschen (Kap. 10.2)
- Menschenwürde und Demütigung (Kap. 10.3)
- Menschenwürde und Machtausübung (Kap. 10.4)
- Menschenwürde und das „Reich ohne Notwendigkeit" (Kap. 10.5)
- Menschenwürde und der „Fähigkeitenansatz" (Kap. 10.6)

Abschließend werden die verschiedenen Zugänge zur Materie zusammengefasst.

[682] Stoecker (2019), S. 33

[683] Dieses und folgende Zitate (einschließlich Blockzitat): Stoecker (2019), S. 9f

10.1 Geschichte des Menschenwürdebegriffs

Anders als viele andere Termini der Moralphilosophie und Ethik finden sich die Ursprünge des Begriffes der Menschenwürde nicht in der griechischen Antike, „*sondern erst in der römischen Antike, und zwar in dem lateinischen Ausdruck ‚dignitas'*"[684]. Schon damals wurde er in zweifacher Bedeutung verwendet:

> *„Einerseits bezeichnet die Würde des Menschen seine besondere Stellung im Universum, andererseits eine von ihm eingenommene gesellschaftliche Stellung. Die eine Auffassung ist ontologisch, die andere ist wertend. Der Ausdruck `Würde´ bezieht sich sowohl auf die Tatsache, dass der Mensch sich vom Rest der Natur unterscheidet, weil er das einzige animal rationale ist, als auch auf die aktive Rolle eines Menschen im öffentlichen Leben, die ihn von anderen Individuen und ihm besonderen Wert verleiht. Im Sinne der ersten Bedeutungsvariante hat der Mensch als solcher Würde, weil er an der Spitze der Hierarchie der Natur steht; im Sinne der zweiten bemisst sich Würde nach seiner Stellung in der gesellschaftlichen Hierarchie."*[685]

Beide Bedeutungen blieben bis heute erhalten. Im Sinne des Diskurses in dieser Abhandlung soll es um die Menschenwürde in der universalistischen, „*im Kern unveränderlichen, allgemeinen Eigenschaft des Menschen"*[686], gehen. Nach der römischen Antike wurde der Begriff der Menschenwürde zunächst im frühen Christentum weiterentwickelt, u.a. von Thomas von Aquin. Die christlichen Denker gaben der Menschenwürde als Merkmal der unsterblichen Seele eine innere Begründung, die hingegen nachrangig zu der letzten göttlichen Begründung für die menschliche Seele als vernünftiger, unsterblicher Substanz nach dem Bilde Gottes war[687]. In der italienischen Renaissance begann eine Säkularisierung des Verständnisses der Menschenwürde, insbesondere mit Giovanni Pico della Mirandolas Rede von 1486 „*De hominis dignitate*" („*Über die Würde des Menschen*"), in der er „*die Idee des Menschen als Herr seines Schicksals"*[688] entwickelte:

> *„Also nahm* [Gott] *den Menschen hin als Schöpfung eines Gebildes ohne besondere Eigenart, stellte ihn in den Mittelpunkt der Welt und redete ihn so an: „Keinen bestimmten Platz habe ich dir zugewiesen, auch keine bestimmte äußere Erscheinung und auch nicht irgendeine besondere Gabe habe ich dir verliehen, Adam, damit du den Platz, das Aussehen und alle die Gaben, die du dir selber wünscht,* ***nach deinem eigenen Willen und Entschluss*** *erhalten und besitzen kannst.* […] *Weder als einen Himmlischen noch als einen Irdischen habe ich dich geschaffen und weder sterblich noch unsterblich dich gemacht, damit du wie ein Former und Bildner deiner selbst* ***nach eigenem Belieben und eigener Macht zu der Gestalt dich ausbilden kannst, die du bevorzugst****."*[689] [Hervorhebungen DS]

---

684 Pfordten (2016), S. 8

685 Becchi (2016), S. 11

686 Pfordten (2016), S. 8

687 Vgl. Pfordten (2016), S. 23f

688 Becchi (2016), S. 14

689 Mirandola (1486), S. 84f

Damit nahm der Siegeszug des „*Homo Faber*“[690] seinen Anfang, „*der bewusst seiner Fähigkeiten keinen Gott mehr braucht, um sich zu verstehen“*[691]. Weiterhin muss herausgestellt werden, „dass Pico *mit der Radikalität seines Freiheitsbegriffes die menschliche Subjektivität in bis dahin unbekanntem Maße hervorhebt“*[692]. Zum Subjektcharakter des Menschen als einer Begründung für Würde kommen wir noch an späterer Stelle zurück.

Grundsätzlich neubestimmt wurde der Menschenwürdebegriff durch Immanuel Kant, dessen Verständnis der Menschenwürde im Jahr 1948 in der Allgemeinen Erklärung der Menschenrechte der vereinten Nationen, im Jahr 1949 in Artikel 1 des bundesdeutschen Grundgesetzes berücksichtigt wurde und seither in diversen anderen Verfassungen aufgegriffen wurde. Horst Dreier schreibt in seinem Kommentar zu Artikel 1 des Grundgesetzes dazu[693]:

> *„Von all dem löst sich die* ***Philosophie Immanuel Kants*** [Hervorhebung HD]*, dessen Würdekonzeption von bleibender und denkbar großer Bedeutung auch und gerade für die modernen Debatten geworden ist. Kants Anschlussfähigkeit beruht neben der Freiheitsorientierung seiner praktischen Philosophie maßgeblich auf ihrem universellen Begründungsmodus,* ***der auf Glaubensinhalte einer partikularen Religion ganz verzichten kann*** [Hervorhebung DS]*. Sie findet ihren letzten Grund vielmehr in der Idee reiner, praktischer Vernunft sowie im streng formalen, von empirischen Bedingungen mit ihren Kontingenzen unabhängigen und in diesem Sinne transzendentalen Begründungsanspruch.“*

Nach Kant besitzt der Mensch „*als aus der Natur herausgehobenes Vernunftwesen einen unverrechenbaren, durch nichts zu ersetzenden Eigenwert“*[694]. Er versteht Würde als einen unbedingten, absoluten Wert:

> *„Im Reiche der Zwecke hat alles entweder einen Preis, oder eine Würde. Was einen Preis hat, an dessen Stelle kann auch etwas anderes als Äquivalent gesetzt werden; was dagegen über allen Preis erhaben ist, mithin kein Äquivalent verstattet, das hat eine Würde. Was sich auf die allgemeinen menschlichen Neigungen und Bedürfnisse bezieht, hat einen Marktpreis; das, was, auch ohne ein Bedürfniß vorauszusetzen, einem gewissen Geschmacke, d. i. einem Wohlgefallen am bloßen zwecklosen Spiel unserer Gemüthskräfte, gemäß ist, einen Affectionspreis; das aber, was die Bedingung ausmacht, unter der allein etwas Zweck an sich selbst sein kann, hat nicht bloß einen relativen Werth, d. i. einen Preis, sondern einen innern Werth, d. i. Würde.“*[695]

---

690 Erläuterung bei Hannah Arendt in Vita Activa, Arendt (1967), S. 451: „*Das lateinische Wort* ***faber****, das vermutlich mit facere im Sinne des hervorbringenden Machens zusammenhängt, bezeichnet den Künstler oder Handwerker, der hartes Material bearbeitet – Holz, Stein oder Metall.* […] *Der Plural* ***fabri*** *ist häufig in* ***fabri tignarii*** *für Bauhandwerker und Zimmerleute. Es war mir unmöglich festzustellen, wann der Begriff Homo Faber zuerst auftaucht oder wer ihn geprägt hat. Sicher ist nur, daß er ganz modernen Ursprungs ist: Jean Leclercq meint, daß Bergson sein Urheber ist.“*

691 Becchi (2016), S. 14

692 Gisbertz (2017), S. 36; Hervorhebung DS

693 Dreier (2013), S. 167f

694 Wetz (2019), S. 108

695 Kant (1785a), Grundlegung der Metaphysik der Sitten, S. 434f; Kant (1785b), S. 68 (BA 78)

Oder etwas knapper:

> *„Autonomie ist also der Grund der Würde der menschlichen und jeder vernünftigen Natur.“*[696]

Die Autonomie, oder etwas breiter gefasst, die Freiheit ist die Grundbedingung der Menschenwürde[697].

Mit der Formel „*Die Würde des Menschen ist unantastbar. Sie zu achten und zu schützen ist Verpflichtung aller staatlichen Gewalt*“[698] hat die Menschenwürde gemäß der von Kant vorgezeichneten Positionierung ihren Weg ins Grundgesetz der Bundesrepublik Deutschland gefunden. Aber was ist die Menschenwürde konkret? Wie kann man konkret prüfen, ob etwas gegen die Menschenwürde verstößt oder nicht, zum Beispiel die hier diskutierte Künstliche Intelligenz? Allein die Tatsache, dass der Kommentar zu Artikel 1 des Grundgesetzes im Werk von Dreier mehr als 170 Seiten umfasst, deutet darauf hin, dass die Beantwortung dieser Frage alles andere als offensichtlich ist. Die nächsten Abschnitte sollen genau diese Frage aus unterschiedlichen Perspektiven aufgreifen und den Zusammenhang von KI und Menschenwürde beleuchten.

---

[696] Kant (1785a), S. 436; Kant (1785b), S. 69 (BA 79, 80); Eingebettet in das folgende Argument: „*Denn es hat nichts einen Werth als den, welchen ihm das Gesetz bestimmt. Die Gesetzgebung selbst aber, die allen Werth bestimmt, muß eben darum eine Würde, d. i. unbedingten, unvergleichbaren Werth, haben, für welchen das Wort Achtung allein den geziemenden Ausdruck der Schätzung abgiebt, die ein vernünftiges Wesen über sie anzustellen hat. Autonomie ist also der Grund der Würde der menschlichen und jeder vernünftigen Natur.*“

[697] Vgl. Nida-Rümelin (2005), S. 156f: „*Die Freiheit, die für ein Leben in Selbstachtung und in Achtung gegenüber anderen unverzichtbar ist, nennen wir sie die „kantische Freiheit“, verlangt, dass das Abwägen von Gründen für unser Handeln und Urteilen autonom und nicht heteronom ist, dass mein Handeln und Urteilen das Ergebnis eigener Abwägung ist und nicht lediglich kausale Folge von Vorprägungen genetischer und biographischer Art sowie von Umweltbedingungen. Die kantische Freiheit, die wir für ein Leben in Selbstachtung und ich Achtung gegenüber Anderen voraussetzen müssen, ist mit einer vollständigen naturalistischen Determination unseres Handelns und unseres Urteilens unvereinbar.“*

[698] Art 1 GG

10.2 Menschenwürde und der Subjektcharakter des Menschen

Der Philosoph Hans Wagner (1917–2000) beschäftigte sich in seinem letzten Werk mit genau dieser Frage:

> *„Was ist die unverletzliche Würde des Menschen, worin besteht sie, worauf ist sie gegründet? Gibt es eine verlässliche und zwingende philosophische Begründung für sie?“*[699]

In einer umfassenden Erörterung kommt der Neukantianer zu dem Ergebnis, dass der Mensch „*das einzige Wesen auf der Welt ist, das Subjekt ist*“[700]. Für ihn ist der Subjektcharakter des Menschen die Begründung für dessen Würde:

> *„Wäre der Mensch nicht Subjekt, so gäbe es in allem unserem Wissen nicht einen einzigen Grund, warum wir irgendein menschliches Exemplar, wenn wir dazu Anlass haben sollten, nicht genauso behandeln dürften (ja sollten) wie jedes beliebige andere materielle Gebilde oder wie jedes beliebige andere Lebewesen auch. Das Wissen um das Subjektsein des Menschen (und des Menschen allein) ist der einzige in unserem Wissen liegende Grund für die Anerkennung der einzigartigen Würde des Menschen, jedes Menschen.“*[701]

Wichtig ist die Ergänzung, dass der Mensch immer Objekt und Subjekt ist, während alle anderen Lebewesen und „materiellen Gebilde“ (also auch Computer und Roboter) nur Objekte sind und sein können, niemals aber Subjekte. Für Wagner ist der Mensch gleichzeitig „*Objekt der empirischen Wissenschaften* und *Subjekt alles Wissens und Denkens*“[702]:

> *„Weil der Mensch Subjekt ist, ist er zu objektiver Erkenntnis, zu Wissen und Wissenschaft befähigt. Und nur er ist dazu befähigt, und zwar deshalb, weil eben nur er unter allem, was in der Welt ist, Subjekt ist.“*[703]

Dies korrespondiert auch sehr gut mit der sogenannten „Objektformel“ des Tübinger Staatsrechtslehrers Christoph Dürig von 1956:

> *„Jeder Mensch ist Mensch kraft seines Geistes, der ihn abhebt von der unpersönlichen Natur und ihn aus eigener Entscheidung dazu befähigt, seiner selbst bewusst zu werden, sich selbst zu bestimmen und sich und die Umwelt zu gestalten.“*[704]
>
> *„Die Menschenwürde als solche ist getroffen, wenn der konkrete Mensch zum Objekt, zum bloßen Mittel zur vertretbaren Größe herabgewürdigt wird.“*[705]

Die Argumentation ist also rückwärtskompatibel mit den Argumenten Kants, aber zugleich auch vorwärtskompatibel mit vielen die Menschenwürde betreffenden Auseinandersetzungen der Gegenwart und Zukunft.

---

699 Wagner (1992), S. 137
700 Wagner (1992), S. 187
701 Wagner (1992), S. 187
702 Wagner (1992), S. 166
703 Wagner (1992), S. 178
704 Zitiert in Pfordten (2016), S. 49 und in Wetz (2019), S. 215
705 Wetz (2019), S. 218

Hans Wagner beklagt sehr leidenschaftlich, dass durch die (verdienten) Erfolge der empirischen Wissenschaften der Respekt vor dem Subjektsein des Menschen (die „Subjekttheorie“) und damit in Konsequenz auch der Respekt vor der Menschenwürde verloren geht:

> *„Wir müssen uns im folgenden davon überzeugen, daß das, was der Mensch ist und was er wert ist, in zulänglicher Form allein in einer Theorie eruiert werden kann, die den Menschen im wohldefinierten Sinn des Wortes als* ***Subjekt*** *zu fassen vermag. Damit treten wir eine Fahrt an, auf der wir den Strom der meisten heutigen Denkweisen in der Philosophie und den Wissenschaften mit aller Macht gegen uns haben. Obwohl nämlich in den ersten Jahrzehnten unseres Jahrhunderts noch in einigen starken philosophischen Bewegungen lebhaft und eindringend um den Begriff des menschlichen Subjektseins gerungen worden ist* […]*, dominieren heute philosophische Richtungen, die entweder zur Subjekttheorie nichts beitragen* […] *oder jede Subjekttheorie eliminieren* […]*. Und indem sie auf diese Weise ein Problemfeld philosophisch unbearbeitet lassen, überlassen sie die Frage nach dem Menschen und seinem Wert gänzlich den empirischen Wissenschaften – von denen doch von Anfang an feststeht, daß sie, woimmer sie sich an die Methodik binden, der sie ihre Erfolge und Einsichten verdanken, den Menschen, selbst wenn er sich darin unmöglich erschöpfen würde, ausschließlich als bloßes Stück von Natur und Welt erforschen und auffassen können.“*[706]

Die von Wagner beschriebene Abwendung vom Subjekt Mensch hin zu einer Objektivierung des Menschseins bedroht die Würde des Menschen. Insofern – das ist das Argument in dieser Arbeit – ist die Künstliche Intelligenz als eine menschengemachte Technologie, die selbst immer nur Objekt sein kann und die zur drastischen Objektivierung des Menschen beiträgt, als Bedrohung der Menschenwürde zu verstehen.

Die Nutzung der KI wird aus Sicht des Schutzes der Menschenwürde so lange neutral sein, wie sie lediglich als Werkzeug des Menschen für den Menschen eingesetzt wird. Immer dann, wenn sie den Menschen z.B. in statistischen Modellen objektiviert und ihn mit Hilfe des KI-gestützten Nudging gängelt und sein Subjektsein einschränkt oder gar negiert, ist von einer Menschenwürdeverletzung im Sinne von Wagner und Dürig auszugehen.

10.3 Menschenwürde und Demütigung

> *„Was ist eine anständige Gesellschaft? Die Antwort, die ich vorschlage, lautet in groben Zügen so: Eine Gesellschaft ist dann anständig, wenn ihre Institutionen die Menschen nicht demütigen.“*[707]
> Avishai Margalit

Der für viele Menschen offensichtliche Gegenpol der Menschenwürde ist die Demütigung. Es ist allgemein unstrittig, dass alles, was Menschen demütigt, auch deren Würde

[706] Wagner (1992), S. 167

[707] Margalit (2012), S. 13

beschädigt. In diesem Abschnitt soll dieser Gedankengang vertieft und vor allem geklärt werden, ob die KI zu einer Demütigung führt oder führen kann.

Für die Demütigung gilt die gleiche Frage, wie für die Menschenwürde. Was ist eine Demütigung konkret? Wie wird sie definiert und abgegrenzt? Jeder Mensch weiß, wann und wie er gedemütigt wurde, wenn er dies schon einmal erlebt hat. Wir wissen von historischen Situationen, z.B. aus dem Dritten Reich, in denen Menschen gedemütigt wurden. Trotzdem fällt uns die Definition derselben schwer.

Der oben zitierte israelische Philosoph Avishai Margalit hat sich in seinem Buch zur „*Politik der Würde – Über Achtung und Verachtung*“ mit dem Begriff der Demütigung auseinandergesetzt. Er versteht darunter „*alle Verhaltensformen und Verhältnisse, die einer Person einen rationalen Grund geben, sich in ihrer Selbstachtung verletzt zu sehen“*[708]. Dazu macht er eine Reihe von Präzisierungen: Nicht jede Person, „*die sich gedemütigt fühlt,* [hat] *tatsächlich einen berechtigten Grund für dieses Gefühl“*[709]. Andererseits muss nicht jede Person, die eigentlich einen berechtigten Grund hätte, sich auch tatsächlich so fühlen. Demütigungen ergeben sich nur als „*Folge des Verhaltens anderer Menschen uns gegenüber“*, direkt oder indirekt und unabhängig davon, ob die Demütigung beabsichtigt war.

Julian Nida-Rümelin leitet aus der deontologischen Nicht-Verrechenbarkeit von menschlichem Leben diejenige der Demütigung von Personen ab:

> *„Man kann diesen deontologischen normativen Sachverhalt auch als den des absoluten Wertes menschlichen Lebens charakterisieren. Auch die Menschenwürde hat einen absoluten Wert in dem gleichen Sinne der Nicht-Verrechenbarkeit. Die Demütigung einer Person wird nicht dadurch gerechtfertigt, dass damit die Demütigung anderer Personen verhindert werden kann.“*[710]

Weder kann ein Menschenleben gegen ein anderes oder viele andere Menschenleben aufgerechnet werden, noch gilt dies für die Verletzung der Würde eines Menschen zur Vermeidung der Würdeverletzung anderer Menschen, noch ist die Demütigung eines Menschen vertretbar, wenn die Demütigung anderer Menschen damit vermieden werden kann.

In Bezug auf die Künstliche Intelligenz ist die Frage zu stellen, inwieweit die Einführung von Systemen der KI zu einer Demütigung von Menschen führen kann bzw. inwieweit von einer begründeten Demütigung von Menschen auszugehen ist, selbst wenn einzelne Individuen sich nicht gedemütigt fühlen.

Anhand von drei Beispielen soll die Relevanz aufgezeigt werden:

---

[708] Margalit (2012), S. 21

[709] Margalit (2012), S. 21; dieses und die folgenden Zitate

[710] Nida-Rümelin (2005), S. 151

1. **Roboter in der Pflege**
   Für die Altenpflege wurden Roboter entwickelt, die optisch eher Stofftieren ähneln oder humanoid ausgeführt sind, die mit Patienten kommunizieren, einfache Gespräche führen und über eine Sensorik auf Reizungen wie z.B. Streicheln reagieren. Sie können bei der Pflege von Demenzkranken und Kindern im Krankenhaus eingesetzt werden. Es ist zu diskutieren, ob das Simulieren einer menschenähnlichen Kommunikation mit dem Vortäuschen von Empathie, Interesse und Emotion, die insgesamt einen positiven Effekt beim Patienten herbeiführt und gleichzeitig das reguläre Pflegepersonal entlastet und von den meisten Patienten nicht als Demütigung empfunden wird, dennoch so zu bewerten ist.
2. **Ersetzen von menschlichen Arbeitsplätzen durch KI-Systeme**
   Automatisierung begleitet die Menschheit seit Beginn der industriellen Revolution. Immer wieder hat es Innovationsschübe gegeben, wie bei der Einführung maschineller Webstühle (so dass keine Weber mehr gebraucht wurden), automatischer Fahrstühle (so dass Fahrstuhlführer entbehrlich wurden) oder modernster Industrieroboter, die zum Abbau von Arbeitsplätzen und Strukturwandel geführt haben. Über Um- und Weiterqualifikation konnten i.d.R. die Menschen weiter beschäftigt werden. Größtenteils waren in der Vergangenheit Arbeitsplätze von geringqualifizierten Mitarbeitern und meistens Tätigkeiten betroffen, die entweder körperlich anstrengend oder eintönig waren. Mit der KI sind allerdings zunehmend Arbeitsplätze betroffen, die ein hohes Qualifikationsniveau erfordern, oftmals sogar Hochschulabschlüsse, wie z.B. Fachkräfte in der Versicherungsbranche, Rechtsanwälte, Sekretärinnen und technische Berater. Die Diskussion über die Demütigung der von Maschinen ersetzten Arbeitskräfte besteht schon lang, sie gewinnt jedoch mit dem höheren Qualifikationsniveau der aktuell von Robotern ersetzten Mitarbeiter an Relevanz.
3. **Einsatz von KI in Waffensystemen**
   Der moralphilosophische Dialog über Krieg und Kriegsführung ist grundsätzlich in einem ethisch-moralischen Grenzbereich angesiedelt. Auch ohne jeglichen Einsatz von modernen Technologien oder Massenvernichtungswaffen kommt es in Kriegen zu Kriegsverbrechen, Verletzungen der Menschenrechte und der Menschenwürde. Trotzdem ist zu diskutieren, ob die KI nicht doch zu einer weiteren Dehumanisierung und damit auch zu einer zusätzlichen Demütigung der beteiligten zivilen und militärischen Personen beiträgt, oder aber – was einige Experten behaupten – zu einer Verbesserung beitragen: Roboter begehen keine Kriegsverbrechen, stehlen und vergewaltigen nicht.

Einige dieser Überlegungen werden in den Fallstudien in Kapitel 12 weiter vertieft.

10.4 Menschenwürde und Machtausübung

> „*Macht bedeutet jede Chance, innerhalb einer sozialen Beziehung den eigenen Willen auch gegen Widerstreben durchzusetzen, gleichviel, worauf diese Chance beruht.“*[711]
>
> Max Weber

Wenn man Macht als das „*Einflussnehmen auf Denk- und Verhaltenswahrscheinlichkeiten“*[712] versteht, kann man die Künstliche Intelligenz in diversen Anwendungen als ein Machtinstrument verstehen. Die obige Definition von Max Weber ist etwas enger gefasst: Durchsetzung des eigenen Willens auch gegen Widerstand. Hier wären in Bezug auf die KI in sozialen Medien zwei Fragen zu stellen: Wessen Wille wird konkret durchgesetzt? Und wird er gegen Widerstreben durchgesetzt? Beides ist beim KI-gestützten Nudging nicht offensichtlich. In diesem Abschnitt soll das Phänomen der Macht näher beleuchtet und die KI als Machtinstrument diskutiert werden.

In der Philosophie und Soziologie haben sich z.B. Niccolo Machiavelli, Thomas Hobbes, Friedrich Nietzsche, der oben zitiert Max Weber, Hannah Arendt und Michel Foucault zum Teil ausführlich mit dem Phänomen der Macht aus verschiedenen Perspektiven auseinandergesetzt. Eine in sich geschlossene und allgemein anerkannte Darstellung des Phänomens Macht[713] stammt von dem Soziologen Heinrich Popitz.

Popitz geht zunächst von drei Prämissen der Wahrnehmung von Machtphänomenen[714] aus. An erster Stelle steht der „*Glaube an die Machbarkeit von Machtordnungen“*[715], die weder „gottgegeben“ noch „*naturnotwendig“* oder „*durch unantastbare Traditionen geheiligt“* sind. Die Machtordnungen sind „*Menschenwerk“* und damit „*Produkt menschlichen Könnens“* und Wollens. Dies haben bereits Platon und Aristoteles erkannt, als sie ihre „*vergleichenden Theorien der Verfassungsformen“* verfassten.

> *„Zu den selbstverständlichen Prämissen unseres Verständnisses von Macht gehört die Überzeugung, daß Macht `gemacht´ ist und anders als sie ist, gemacht werden kann.“*[716]

Die zweite Prämisse betrifft die *Omnipräsenz von Macht*. Kaum ein Lebensbereich der modernen Welt ist vollkommen machtfrei. Popitz zitiert hierzu Max Weber:

> *„Alle denkbaren Qualitäten eines Menschen und alle denkbaren Konstellationen können jemand in die Lage versetzen, seinen Willen in einer gegebenen Situation durchzusetzen.“*[717]

---

711 Weber (1922), S. 38

712 Paulick (2018)

713 Popitz (1986)

714 Vgl. Popitz (1986), S. 12 ff

715 Dieses und die folgenden Zitate: Popitz (1986), S. 12 ff

716 Popitz (1986), S. 15

717 Zitiert in Popitz (1986), S. 17; Originalzitat: Weber (1922), S. 38

Popitz' „*dritte Prämisse des Macht-Verständnisses beruht auf der Konfrontation von Macht und Freiheit:* ***Alle Machtanwendung ist Freiheitsbegrenzung****. Jede Macht ist daher rechtfertigungsbedürftig"*[718]. Machtausübung ist also immer ein Eingriff in die Autonomie und Freiheit des Menschen und muss begründet werden.

Er fasst zusammen:

> *„Macht ist machbar, Machtordnungen sind veränderbar, eine gute Ordnung entwerfbar: es kann getan werden. Macht ist omnipräsent, eindringend in soziale Beziehungen jeden Gehalts: sie steckt überall drin. Macht ist freiheitsbegrenzend, als Eingriff in die Selbstbestimmung anderer begründungsbedürftig:* ***alle Macht ist fragwürdig****.* [Hervorhebung DS]"[719]

Popitz unterscheidet vier „*anthropologische Grundformen der Macht"*[720]*: „die verletzende Aktionsmacht, die instrumentelle Macht, die autoritative Macht und die datensetzende Macht".*

Die „***verletzende Aktionsmacht"*** nutzt die Vulnerabilität von Menschen (und auch von Tieren) aus. Neben der offensichtlichen körperlichen Verletzbarkeit spielen auch die ökonomische und soziale Angreifbarkeit eine Rolle. Die Aktionsmacht ist einseitig verteilt: Sie ist die Macht des Stärkeren über den Schwächeren: „*Menschen können über andere Macht ausüben, weil sie andere verletzen können."*

„Instrumentelle Macht" entsteht durch die Instrumentalisierung einer „*glaubhaften Gefahr"* oder auch einer „*glaubhaften Chance"* (z.B. einer Belohnung). Das ist die Begründung einer „*permanenten Unterwerfung"*. Der „*Aufbau"* und das „*Bewahren dieser Glaubwürdigkeit"* kennzeichnen die „*Strategie instrumenteller Machtausübung"*. Es besteht ein dauerhaftes „quid pro quo". Beispiele für die Anwendung instrumenteller Macht im alltäglichen Geschäft sind Bonus/Malus-Regelungen. Menschen werden damit unablässig zum „*Werkzeug fremden Willens"*.

Popitz unterscheidet zwischen äußerer Macht und innerer Macht. Mit der „*äußeren Macht"* meint er im Kontext der instrumentellen Macht konkrete Drohungen (glaubhafte Gefahr) und konkrete Belohnungen (glaubhafte Chancen). Bei der „*inneren Macht*" geht es zum Beispiel um Konformität mit normativen (sittlichen, moralischen oder ethischen) Vorgaben, die nicht zwingend in jedem Fall überprüft werden kann. Im Zentrum steht die Anerkennung durch andere Personen oder soziale Gruppen oder der Entzug von Anerkennung. Beim bewussten Einsatz der Alternative zwischen „*erhoffter Anerkennung* und *befürchteten Anerkennungsentzügen* wird ***autoritative Macht*** ausgeübt".

---

[718] Popitz (1986), S. 17

[719] Popitz (1986), S. 20

[720] Popitz (1986), S. 22 ff (einschließlich aller folgenden Zitate)

Der Mensch übt aber auch Macht durch die „*Ausnutzung und das Zurechtstutzen der Natur"* zu seinem Nutzen aus. Indem „*Freiräume und Zwänge für viele Menschen"* entstehen können, ergibt sich hier die vierte Grundform der Macht: die „***datensetzende Macht"***:

> *„Die Macht des Datensetzens ist eine objektvermittelte Macht. Sie wird gleichsam in materialisierter Form auf die Betroffenen übertragen. Das heißt: sie ist keineswegs eine Macht der Dinge über den Menschen – obwohl sie die Ideologie „verdinglichter" Macht nahelegt -, sondern eine Macht des Herstellens und der Hersteller; eine vom Hersteller in das Ding eingebaute, häufig längere Zeit latente Macht, die jederzeit manifestiert werden kann. Solche Macht-Minen können wir heute für kommende Generationen für zehntausende von Jahren vergraben. Es besteht also wohl Grund, den doppelten Machtcharakter technischen Handelns zu reflektieren:* ***die Macht über die Kräfte der Natur und die objektvermittelte Entscheidungsmacht über die Lebensbedingungen anderer Menschen****.* [Hervorhebung DS][721]

Die Entwicklung von Technologien, insbesondere die der Künstlichen Intelligenz, können „*im Prinzip zu einer unbegrenzten Steigerung von Macht führen"*[722]. Heinrich Popitz schreibt dazu:

> *„Die objektiven, objektivierten Bedingungen der menschlichen Existenz verändern sich in hochtechnisierten Gesellschaften radikal mit dem Umblättern des Kalenders. Wer heute über die technische Gestaltung unserer Lebensumwelt entscheidet, wer datensetzende Macht hat, kann in kürzester Frist ein unermessliches Ausmaß von Macht über unermesslich viele Menschen und eventuell* [...] *unermesslich lange Zeiträume ausüben. Wir können den technischen Progress rückblickend konstatieren. Aber wir können nicht entscheiden, wie lange und wie weitgehend sich die Effizienz technischen Handelns noch weiter steigern wird. Unsere bisherige Erfahrung verweist uns auf kein Prinzip, auf dem sich eine solche Prognose ableiten ließe. Technisches Handeln scheint eine prinzipiell offene Fähigkeit des Menschen zu sein. Entsprechend können wir auch noch nicht wissen, bis in welche namenlose Regionen sich das Potential sozialer Macht weiter auftürmen lässt. Wenn technisches Handeln prinzipiell offen ist, dann ist auch die potentielle Gefährlichkeit des Menschen für den Menschen prinzipiell offen*. [...] *Immerhin können wir mit großer Wahrscheinlichkeit für eine absehbare Zeit mit einer weiteren Zunahme des Machtpotentials rechnen, und zwar in dem beschriebenen dreifachen Sinne. Damit werden aber die Probleme der Machtkontrolle immer schwerer zu lösen. Zugleich wird immer gewisser: Der Angelpunkt jeder Machtkontrolle in modernen Gesellschaften ist die Kontrolle technischen Handelns."*[723]

Die Künstliche Intelligenz bildet in dreifacher Hinsicht ein Beispiel für ein datensetzendes Machtinstrument, das die angesprochene Machtkontrolle erforderlich macht: unermessliches Ausmaß der Macht über unermesslich viele Menschen und eventuell unermesslich lange Zeiträume.

---

721 Popitz (1986), S. 31

722 Nemitz Pfeffer (2020), S. 22

723 Popitz (1986), S. 180f; auch zitiert in Nemitz Pfeffer (2020), S. 22f

Ein Beispiel für den potenziell unermesslichen Machtumfang der KI ist deren Einsatz in der Waffentechnik, z.B. in selbststeuernden Drohnen, in sogenannten „Kampfrobotern" oder in digitalen Programmen des Internetkrieges bzw. „Cyber-Wars"[724]. Unermesslich viele Menschen werden über Applikationen und Plattformen des Internets erreicht, wie etwa durch soziale Medien, die allesamt die KI nutzen. Die Technik ist in der Regel datensetzend, oftmals sogar machtausübend, wie zum Beispiel im Sozialkredit-System (Social Scoring) der chinesischen Regierung[725].

Vertreter großer Hightech-Unternehmen bestreiten das drastisch wachsende Machtpotential ihrer Technologie nicht einmal; dies zeigt ein Statement von Eric Schmidt, dem ehemaligen Vorstandvorsitzenden von Google:

> *„Wir sind überzeugt, dass Portale wie Google, Facebook, Amazon und Apple weitaus mächtiger sind, als die meisten Menschen ahnen. Ihre Macht beruht auf die Fähigkeit, exponentiell zu wachsen. Mit Ausnahme von biologischen Viren gibt es nichts, was sich mit derartiger Geschwindigkeit, Effizienz und Aggressivität ausbreitet wie diese Technologieplattformen, und dies verleiht auch ihren Machern, Eigentümern und Nutzern neue Macht*."[726]

Die durch die KI ermöglichte Machtkonzentration in Kombination mit privater Herrschaft den Internetgiganten bedroht die „*Funktionsfähigkeit der wesentlichen Steuerungssysteme unserer Gesellschaft: Demokratie und Markt"*[727]. Die auf den Plattformen der Technologieunternehmen genutzten Algorithmen haben zunehmend einen Charakter von Regeln, die Gesetzen ähneln, die allerdings weder demokratisch reglementiert sind noch juristisch überprüft werden können.

In den Händen autoritärer Regimes wird die KI zum Unterdrückungsinstrument.

Da jede Form von Macht, insbesondere jenige, die quasi versteckt und „auf Samtschuhen" daherkommt, die menschliche Freiheit und Autonomie einschränkt, ist sie immer potenziell Menschenwürde-einschränkend und somit begründungsbedürftig.

---

[724] Nemitz Pfeffer (2020), S. 23

[725] Beim Sozialkredit-System Chinas handelt es sich um ein Punkte-System, nachdem Bürger für wünschenswertes Verhalten Punkte bekommen. Für negatives Verhalten werden Punkte entzogen. Diejenigen Bürger, die ein zu geringes oder gar negatives Punkte-Niveau erreichen, erfahren Einschränkungen in Bezug auf Reisefreiheit, Zugang zu Arbeitsplätzen, Ausbildung oder Wohnraum. Für die Erfassung und Analyse der Daten wird klassische Überwachungstechnik im öffentlichen Raum (Kameras) und im Internet in Kombination mit KI eingesetzt. Vgl. https://www.heise.de/newsticker/meldung/34C3-China-Die-maschinenlesbare-Bevoelkerung-3928422.html

[726] Zitiert von Nemitz Pfeffer (2020), S. 24; Originalzitat aus dem FAZ-Artikel vom 3.4.2014: Angst vor Google. https://www.faz.net/aktuell/feuilleton/debatten/weltmacht-google-ist-gefahr-fuer-die-gesellschaft-12877120.html

[727] Nemetz Pfeffer, S. 24

10.5 Menschenwürde und das „*Reich ohne Notwendigkeit*“

> *„Es ist durchaus denkbar, daß die Neuzeit, die mit einer so unerhörten und unerhört vielversprechenden Aktivierung aller menschlichen Vermögen und Tätigkeiten begonnen hat, schließlich in der tödlichsten, sterilsten Passivität enden wird, die die Menschheit je gekannt hat.“*[728]
>
> Hannah Arendt, Vita Activa oder Vom tätigen Leben, 1967

Die Verfechter der Künstlichen Intelligenz beschreiben die von ihr dominierte Welt gern als eine, in der die Menschen von allen banalen (Routine-)Aufgaben am Arbeitsplatz oder im Privaten befreit sind[729]. Es stellt sich jedoch die Frage: Wo enden banale Routineaufgaben? Und was machen die Menschen in ihrer „gewonnenen“ Zeit? Es ist kein Zufall, dass die gleichen KI-Enthusiasten auch davon überzeugt sind, dass ein stark wachsender Anteil der Bevölkerung von einem bedingungslosen Grundeinkommen leben wird. Wenn man dieser Vorstellung folgt, entfernt sich die Menschheit mit der Verbreitung der neuen Technologien und insbesondere der Künstlichen Intelligenz sukzessive von der Notwendigkeit zu arbeiten. Einige wenige Menschen werden weiterhin politisch, unternehmerisch oder künstlerisch tätig sein, die meisten anderen Menschen geraten in ein „Reich ohne Notwendigkeit“, eine Welt des Konsums, des Bedientwerdens und des Entertainments. In diesem Kapitel soll es um die Frage gehen, ob eine derartige Welt eher eine anzustrebende Utopie ist oder eine Dystopie. Auf Basis erster Überlegungen von Karl Marx haben sich zwei Philosophen mit dem besagten „Reich ohne Notwendigkeit“ auseinandergesetzt, Ernst Bloch und Hans Jonas.

Ausgangspunkt von Jonas‘ Kritik an der marxistischen Utopie des „Reiches ohne Notwendigkeit“ ist der folgende Abschnitt aus dem dritten Band von „*Das Kapital. Kritik der politischen Ökonomie*“ von Karl Marx:

> *„Das Reich der Freiheit beginnt in der That erst da, wo das Arbeiten, das durch Noth und äußere Zweckmäßigkeit bestimmt ist, aufhört; es liegt also der Natur der Sache nach jenseits der Sphäre der eigentlichen materiellen Produktion. Wie der Wilde mit der Natur ringen muß, um seine Bedürfnisse zu befriedigen, um sein Leben zu erhalten und zu reproduciren, so muß es der Civilisirte, und er muß es in allen Gesellschaftsformen und unter allen möglichen Produktionsweisen. Mit seiner Entwicklung erweitert sich dies Reich der Naturnothwendigkeit, weil die Bedürfnisse; aber zugleich erweitern sich die Produktivkräfte, die diese befriedigen. Die Freiheit in diesem Gebiet kann nur darin bestehn, daß der vergesellschaftete Mensch, die associirten Producenten, diesen ihren Stoffwechsel mit der Natur rationell regeln, unter ihre gemeinschaftliche Kontrolle bringen, statt von ihm als von einer blinden Macht beherrscht zu werden; ihn mit dem geringsten Kraftaufwand und unter den, ihrer menschlichen Natur würdigsten und adäquatesten Bedingungen vollziehn. Aber es bleibt dies immer ein Reich der Nothwendigkeit. Jenseits desselben beginnt die*

[728] Arendt (1967), S. 411

[729] Vgl. Daniela Rus, Leiterin des Computer Science and Artificial Intelligence Laboratory (CSAIL) am MIT in Ford (2019), S. 257f

> *menschliche Kraftentwicklung, die sich als Selbstzweck gilt, das wahre Reich der Freiheit, das aber nur auf jenem Reich der Nothwendigkeit als seiner Basis aufblühn kann. Die Verkürzung des Arbeitstags ist die Grundbedingung.“*[730]

Jenseits des Reiches der Notwendigkeit beginnt das Reich der Freiheit als Folge der schrittweisen Verkürzung des Arbeitstages. In Blochs Buch „*Prinzip Hoffnung*“ findet Jonas das Bild dessen, wie sich die Marxisten das „*irdische Paradies der tätigen Muße vorstellen*“[731]. Fast zynisch oder sarkastisch kommentiert er, wie Bloch „tätige Muße“ als das Paradies der Zukunft feiert. Sein Widerspruch könnte nicht vehementer sein:

> *„Die Abscheidung vom Reiche der Notwendigkeit entzieht der Freiheit ihren Gegenstand, sie wird ohne ihn ebenso nichtig wie Kraft ohne Widerstand. Leere Freiheit, wie leere Macht, hebt sich selber auf – und das echte Interesse am dennoch unternommenen Tun.*[732] *Oder: Es gibt gar kein `Reich der Freiheit´ außerhalb des Reiches der Notwendigkeit!“*[733]

Jonas will hiermit klarstellen, dass der Mensch ohne Aufgaben, Pflichten und Verantwortlichkeiten keine Freiheit mehr besitzt und vollständig determiniert ist durch seine Triebe, Bedürfnisse und Leidenschaften. Noch schlimmer sind aus seiner Sicht die Auswirkungen auf die Menschenwürde, die – wie wir schon in vorherigen Beiträgen gesehen haben – im engen Zusammenhang mit der Freiheit steht:

> *„Aber bleiben wir beim ganz ungewaltsamen Schicksal der Menschenwürde in der scheintätigen Muße des utopischen Paradieses, auch ohne dass sein Friede von solchen Capricen des Menschenherzens gestört wird. Ihr friedlicher Tod ist nicht weniger eine Katastrophe. Mit dem Ernst der Wirklichkeit, die immer auch Notwendigkeit ist, schwindet auch die Würde dahin, die den Menschen eben im Verhältnis zum Wirklich-Notwendigen auszeichnet. Das Spiel als Lebensberuf, weit entfernt, das Menschenwürdige darzustellen, schließt von ihm aus.“*[734]

Jonas hat den Marxismus insgesamt „*einer ausgiebigen Kritik unterzogen, insbesondere die Utopien von Marx bis Bloch“*[735] und in dem Zusammenhang auch den ***technologischen Fortschritt***, von dem Marx eine Befreiung des Proletariats erwartet hatte. Dazu schreibt Jürgen Nielsen-Sikora:

> *„Damit einher geht eine Kritik an einem im Marxismus angelegten deterministischen Weltbild der Geschichte, die letztlich zum utopischen Ideal hinführen müsse. Zugleich bedeutet die Absage an den Determinismus die Möglichkeit von Freiheit und somit von Verantwortung.“*

Die Konsequenz in Bezug auf die Künstliche Intelligenz ist erheblich. Wenn wir mit der KI ein „Reich ohne Notwendigkeit“ schaffen, vergrößern wir damit nicht die Freiheit des

730 Marx (1894), S. 794f

731 Jonas (1979), S. 348

732 Jonas (1979), S. 364

733 Jonas (1979), S. 365

734 Jonas (1979), S. 365

735 Dieses und die folgenden Zitate (einschließlich Blockzitat): Nielsen-Sikora (2017), S. 261

Menschen, wie etwa von Marx und Bloch geglaubt, sondern schränken diese ein, genauso wie die Menschenwürde. Wir verkleinern den Bereich dessen, wofür der Mensch Verantwortung übernehmen kann, bis zur völligen Verantwortungslosigkeit.

Jonas erkennt „*die von Ernst Bloch in die Diskussion eingebrachte Hoffnung*“[736] als „*eine Bedingung jeden Handelns*“ an. In der Tat genügt die Hoffnung nicht für die Gestaltung einer menschlichen Zukunft:

> *„Hierzu sei vielmehr `Mut zu Verantwortung´ vonnöten: „Verantwortung ist die als Pflicht anerkannte Sorge um ein anderes Sein, die bei Bedrohung seiner Verletzlichkeit zur `Besorgnis´ wird“. Diese Art der Sorge fragt danach, was passieren kann, wenn ich mich bestimmter Aufgaben nicht annehme, bestimmten Dingen nicht zuwende. Je unklarer ist, was passieren kann, desto größere Behutsamkeit sei erfordert. Von einer begründeten `Hellsicht der Einbildungskraft´ spricht Jonas hier.“*[737]

Die führt zu einem kleinen Vorgriff auf das Schlusswort dieser Arbeit: Beim Umgang mit Künstlicher Intelligenz ist eine „Hellsicht der Einbildungskraft“ gefordert.

In Zeiten der Vollbeschäftigung erscheint der Gedanke an ein „*Reich ohne Notwendigkeit*“ weit hergeholt. Die fortschreitende Automatisierung und Digitalisierung hat bisher zwar dazu geführt, dass viele Arbeits- und Tätigkeitsvolumina von Menschen an die Technik übertragen wurden. Bisher haben sich immer neue Aufgaben und Anforderungen an den Menschen ergeben, so dass bisher von einem „*irdischen Paradies der tätigen Muße*“[738] keine Rede sein kann. Mit dem sich beschleunigenden Innovationstempo und der der Nutzung von KI und Maschinenlernen in nahezu allen Bereichen menschlicher Tätigkeit kann sich dies in der nahen bis mittelfristigen Zukunft ändern. Man wird frühzeitig darüber nachdenken müssen (im Sinne einer „*Hellsicht der Einbildungskraft*“), welche humanitären, sozialen und Menschenwürde-bezogenen Auswirkungen sich daraus ergeben.

[736] Nielsen-Sikora (2017), S. 261; dieses und die folgenden Zitate

[737] Nielsen Sikora zitiert hier Jonas (1979), S. 391f

[738] Jonas (1979), S. 348

10.6 Der Capability Approach (CA) zur Menschenwürde

Wie schon mehrfach festgestellt, ist man sich bezüglich der Bedeutung der Menschenwürde überwiegend einig, trotzdem fällt es immer wieder schwer, diese klar zu definieren und nachvollziehbar zu benennen, was spezifisch zur Wahrung der Menschenwürde beiträgt und wo eine Verletzung derselben droht. In den vorherigen Abschnitten wurden – wie von Ralph Stoecker vorgeschlagen – verschiedene negative Angänge des Verständnisses der Menschenwürde über ihre Verletzungen verfolgt. Damit ist es gelungen, ein solides Verständnis für die grundsätzlichen Menschenwürdeverletzungen beim Einsatz der Künstlichen Intelligenz zu gewinnen, nämlich über die Unterminierung des menschlichen Subjektseins, über Demütigungen, Machtausübung und die Schaffung eines „Reiches ohne Notwendigkeiten". Hingegen fehlt immer noch ein Rahmen, mit dem eine detaillierte Prüfung und Analyse der „Menschenwürdigkeitstauglichkeit" im einzelnen Anwendungsfall möglich ist.

In diesem Abschnitt soll geprüft werden, inwieweit der „Fähigkeitenansatz" (auch „Befähigungsansatz, englisch: „Capability Approach") nach Amartya Sen und Martha Nussbaum auf die Diskussion der Künstlichen Intelligenz angewendet werden kann.

Beim Fähigkeitenansatz handelt es sich um „*eine grundlegende Theorie der Gerechtigkeit, die das Ziel hat, jedem Menschen ein Leben in Würde zu ermöglichen*"[739]. Martha Nussbaum versucht mit ihrem Ansatz, die folgende Frage zu beantworten:

> *„Was ist für ein der menschlichen Würde angemessenes Leben erforderlich?"*[740]

Sie hat sich bereits seit den 1980er Jahren gemeinsam mit Amartya Sen dieser Frage gewidmet. Sen erhielt für seine Arbeiten an einer Gerechtigkeitstheorie und deren Umsetzung z.B. im Human Development Index der Vereinten Nationen 1998 den Nobelpreis der Ökonomie. Die Theorie lehnt sich an die Tugendlehren von Aristoteles und die Gerechtigkeitstheorie von John Rawls an.

Ausgangspunkt ist die Würde des Menschen:

> *„Der Begriff der Würde steht in enger Beziehung zur Idee aktiven Strebens. Er ist somit ein enger Verwandter des Begriffs der grundlegenden Fähigkeit, und bezeichnet etwas, das einer Person innewohnt und das beansprucht, entwickelt zu werden."*[741]

Sie definiert ein Minimum von zehn zentralen Fähigkeiten, die zu sichern sind[742]:

1. ***„Leben**: Fähig zu sein, ein Menschenleben normaler Dauer zu leben; nicht verfrüht zu sterben oder bevor das Leben so eingeschränkt ist, dass es nicht mehr lebenswert ist.*

739 https://www.socialnet.de/lexikon/Capability-Approach

740 Nussbaum (2015), S. 40

741 Nussbaum (2015), S. 39

742 Nussbaum (2015), S. 41f

2. ***Körperliche Gesundheit****: Sich einer guten Gesundheit, einschließlich der der reproduktiven Gesundheit erfreuen zu können; ausreichend ernährt zu sein und eine angemessene Unterkunft zu besitzen.*
3. ***Körperliche Unversehrtheit****: Fähig zu sein, sich frei zu bewegen; vor gewalttätigen, einschließlich sexuellen Übergriffen und häuslicher Gewalt geschützt zu sein* [...].
4. ***Sinne, Vorstellungskraft, Denken****: In der Lage zu sein, die Sinne zu benutzen, Vorstellungen zu entwickeln, zu denken und zu argumentieren – und all dies auf ‚wirklich menschliche' Weise zu tun, d.h. geprägt und kultiviert durch eine hinreichende Bildung, die Lese-, Schreibfähigkeit und Grundkenntnisse der Mathematik und Wissenschaft einschließt, sich darauf aber nicht beschränkt; Vorstellungskraft und Denken im Zusammenhang mit dem Erleben und Erzeugen von Werken der eigenen Wahl, u.a. religiöser, literarischer, musikalischer Art, nutzen zu können; befähigt zu sein, den eigenen Verstand auf eine Weise zu nutzen, die durch Garantien politischer und künstlerischer Meinungsfreiheit sowie der freien Religionsfreiheit geschützt ist; fähig zu sein, angenehme Erfahrungen zu machen und unnötigen Schmerz zu vermeiden.*
5. ***Gefühle****: Fähig zu sein, Bindungen zu Dingen und Personen außerhalb unserer selbst zu entwickeln; die zu lieben, von denen man geliebt wird und die sich um einen sorgen; ...* [...].
6. ***Praktische Vernunft****: Fähig zu sein, eine Vorstellung vom Guten zu bilden und über die eigene Lebensplanung in kritischer Weise nachzudenken* [...].
7. ***Zugehörigkeit****: (A) Fähig zu sein, mit anderen und für andere zu leben, andere Menschen anzuerkennen und sich um sie zu kümmern, sich an vielfältigen Formen gesellschaftlicher Interaktion zu beteiligen; sich in die Lage eines anderen hineinversetzen zu können* [...]. *(B) Über die gesellschaftlichen Grundlagen der Selbstachtung und der Nichtdemütigung zu verfügen; fähig zu sein, mit einer Würde behandelt zu werden, die der anderer gleich ist. Hierzu gehören Regelungen, die die Diskriminierung* [...] *ausschließen.*
8. ***Andere Gattungen****: Fähig sein, in Rücksicht auf Tiere, Pflanzen und Natur und in Beziehungen zu diesen zu leben.*
9. ***Spiel****: Lachen, spielen und sich an Freizeitaktivitäten erfreuen zu können.*
10. ***Kontrolle über die eigene Umwelt****: (A) Politisch: Fähig zu sein, sich effektiv an den politischen Entscheidungsprozessen zu beteiligen, die das eigene Leben bestimmen; das Recht zu politischer Teilnahme zu besitzen, den Schutz der freien Rede und der Versammlungsfreiheit zu genießen. (B) Materiell: Über Eigentum* [...] *verfügen zu können und Eigentumsrechten gleich anderen Menschen zu besitzen; das Recht, gleich anderen eine Beschäftigung zu suchen;* [...]. *Fähig zu sein, als Mensch zu arbeiten, die praktische Vernunft einzusetzen und in sinnvollen Beziehungen zu anderen Beschäftigten auf der Basis gegenseitiger Anerkennung zu treten."*

Nussbaum fordert, dass „*die Achtung der menschlichen Würde verlangt, die Bürger über eine hinreichende (und je spezifische) Schwelle an Fähigkeiten in allen zehn Bereichen zu heben"*[743]. Zwei Fähigkeiten besitzen eine Querschnittsbedeutung oder

[743] Nussbaum (2015), S. 43

„architektonische Rolle“[744]: die der Zugehörigkeit (#7) und die der praktischen Vernunft (#6). Sie stellen je für sich selbst sehr wichtige Fähigkeiten dar, ermöglichen aber viele der anderen Fähigkeiten.

Der dargestellte Rahmen unverzichtbarer Fähigkeiten zur Sicherstellung und Wahrung der Menschenwürde passt zu den vorher entwickelten Perspektiven aus unterschiedlichen negativen Angängen und ist vor allem auch konsistent mit Kants Grundidee, obwohl Kant dies vermutlich nie so bestätigen würde und Nussbaum auch die Verbindung so nicht aufbaut. An einer Stelle grenzt sie sich sogar eindeutig von Kant ab: Sie interpretiert Kants Philosophie so, dass er kognitiv stark eingeschränkten Menschen nicht die gleiche Würde zuschreiben würde, wie Menschen, die im vollen Besitz ihrer Vernunftsbegabung sind.

Für die Beurteilung der Frage, ob bestimmte Technologien wie z.B. die Künstliche Intelligenz die Menschenwürde beeinflussen, ist der Rahmen der zehn Fähigkeiten sehr gut geeignet. Es sind Situationen denkbar, in denen eine Technologie wie die KI bei den allermeisten Fähigkeiten förderlich ist, allerdings zum Preis eines Kompromisses bei anderen Fähigkeiten, was nach Nussbaum nicht zulässig ist.

10.7 Zusammenfassung: Menschenwürdeverletzung durch KI

In diesem Kapitel wurde herausgearbeitet, warum eine begründete Sorge um Menschenwürdeverletzungen durch die Künstliche Intelligenz geboten ist. Ralph Stoeckers Vorschlag folgend wurde versucht, die Gefahr der KI für die Menschenwürde „negativ“ von den möglichen Verletzungen her zu verstehen. Aus fünf Perspektiven wurden die allesamt signifikanten Gefahren der Menschenwürdeverletzung zutage gefördert: erstens die Argumentation über den Verlust des Subjektseins des Menschen nach Hans Wagner, zweitens die Argumentation über Demütigungen nach Avishai Margalit, drittens die Argumentation über Menschenwürdeverletzungen durch Machtausübungen, viertens die Argumentation über das mittels KI geschaffene Reich ohne Notwendigkeit nach Hans Jonas und fünftens und abschließend die Anwendung des Fähigkeitenansatzes nach Martha Nussbaum (und Amartya Sen).

Die Meinungen darüber, ob das Gebot der Unantastbarkeit der Menschenwürde überhaupt begründbar ist, gehen weit auseinander[745]. Hans Wagner hat allerdings eine sehr plausible Fundierung der Würde des Menschen vorgelegt, die für die Analysen in dieser Arbeit von großer Bedeutung sind: die Begründung der einzigartigen Würde unserer Gattung durch unseren Subjektcharakter. Dieses Argument erlaubt eine klare Abgrenzung

[744] Nussbaum (2015), S. 46

[745] Vgl. Vossenkuhl (2021), S. 269 ff; Zitat (S. 271): „*Kant versteht* […] *die Würde nicht als Prinzip“*; generell argumentiert Vossenkuhl, dass die Geltung der Menschenwürde und des Satzes des Widerspruches nicht begründbar seien

gegenüber allen anderen Wesen in unserer Welt und auch gegenüber von uns geschaffenen Artefakten. Vor allem hilft sie bei der Suche nach Gründen für die Feststellung einer Verletzung der Menschenwürde. Unsere Würde ist immer dann bedroht, wenn wir nicht mehr Subjekt sind und zum Objekt degradiert werden. Gerade beim Einsatz der KI ergeben sich dafür viele Hinweise.

Die Tatsache, dass Menschen gedemütigt werden, bedeutet nicht zwingend, dass eine Demütigung beabsichtigt ist, und ebenso wenig, dass die betroffenen Personen sich gedemütigt fühlen. In vielen Anwendungsbereichen der KI finden sich unzählige Beispiele für Demütigungen, die selten beabsichtigt sind und oftmals von den betroffenen Personen nicht als solche empfunden werden, von Dritten häufig sehr wohl.

Die Künstliche Intelligenz ist bestens dafür geeignet, Macht auf Individuen auszuüben, oftmals, ohne dass die Objekte es so wahrnehmen. Die Möglichkeiten der Skalierung dieser Macht durch Herrschaftsorgane und privatwirtschaftliche Unternehmen sind „unermesslich".

Auch die Utopie einer durch die Technik veränderten Welt, die dem Menschen immer weniger Aufgaben übrig lässt, das sogenannte „Reich ohne Notwendigkeit", muss hinsichtlich der Würdeauswirkungen kritisch hinterfragt werden. Auf den zweiten Blick offenbart sich hier eine Dystopie.

Martha Nussbaum[746] hat eine Gerechtigkeitstheorie auf Basis von Fähigkeiten entwickelt und diese direkt an die Menschenwürde angebunden. Eine sehr konkrete Liste von zehn Einzelfähigkeiten, die die Würde des Menschen ausmachen, gestattet eine detaillierte Identifikation und Analyse von Menschenwürdeverletzungen im Allgemeinen und im Bereich der KI im Besonderen.

Insgesamt ergibt sich damit ein breites Spektrum von Kriterien und Ansätzen, um mögliche Menschenwürdeverletzungen durch die KI zu identifizieren und zu analysieren. Die fünf Perspektiven sind überlappend und an vielen Stellen redundant, doch sie schärfen den Blick im Diskurs darüber, wie sich die Künstliche Intelligenz auf die Würde unserer Gattung auswirkt.

Qualitativ ist der Begriff der „Sorge" mehr als angemessen. In vielen Bereichen zeichnen sich bereits jetzt erhebliche Verletzungen der Menschenwürde durch die KI ab.

[746] Zusammen mit Amartya Sen

11 Vollendet oder beendet die KI die Aufklärung?

Leitgedanke für die Grundhypothese dieser Arbeit ist die Sorge, dass mit der Einführung und zunehmenden Verbreitung der Künstlichen Intelligenz viele der Errungenschaften des letzten Vierteljahrtausends für die Menschheit wieder zunichte gemacht werden und sich die Aufklärung quasi gegen sich selbst wendet. In den vorherigen Kapiteln wurden die Auswirkungen der KI auf die Themenkomplexe Autonomie, Verantwortung, Individualität und Menschenwürde untersucht. Gegenstand des Arguments dieses Kapitels ist die Kernhypothese der vorliegenden Arbeit:

> **Der Mensch begibt sich mit der KI (erneut) in eine selbstverschuldete Unmündigkeit – die durch die Aufklärung freigesetzten Mechanismen wenden sich gegen sie selbst.**

Es geht um die Aufklärung insgesamt, die nach Steven Pinker Vernunft, Wissenschaft, Humanismus und Fortschritt in nie zuvor gekannten Maßen ermöglichte[747], allerdings mit einem Fokus auf das Kant'sche Verständnis derselben.

Dafür soll im ersten Abschnitt die Geschichte der Aufklärung mit einem Überblick ihrer Erscheinungsformen und regionalen Ausprägungen dargestellt werden.

Danach werden Immanuel Kants berühmter Aufsatz „*Was ist Aufklärung?*" resümiert und die von ihm verwendeten Begriffe und Zusammenhänge vertieft. Besonders bedeutsam sind dabei die Begriffe der Mündigkeit sowie der Unmündigkeit und deren Selbstverschuldung und auch das Wie der Herausführung des Menschen aus der Unmündigkeit.

Die Aufklärung war immer wieder Gegenstand heftiger Kontroversen, die im dritten Abschnitt diskutiert werden. Eine Reihe von Argumenten der Aufklärungskritik und Gegenaufklärung bezieht sich zwar auf Entwicklungen und Wahrnehmungen aus dem 19. und 20. Jahrhundert, weisen aber auch auf Phänomene, die das zentrale Argument dieser Arbeit untermauern.

Abschließend werden die wesentlichen Erzählstränge wieder aufgegriffen und in einem Argument zusammengeführt. Dabei ist zunächst zu klären, ob und wie die vom Menschen geschaffene Künstliche Intelligenz die Mündigkeit des Menschen einzuschränken droht. Weiterhin ist zu prüfen, ob und warum dies nicht nur ein Thema für jeden einzelnen Menschen ist, sondern auch Gegenstand von gesellschaftlichem und politischem Handeln sein sollte.

[747] Vgl. Pinker (2018), S. 8ff

11.1 Eckpunkte der Aufklärungsgeschichte

Eine umfängliche Darstellung der Geschichte der Aufklärung würde den Rahmen dieser Dissertation sprengen. Hier soll es um eine zeitliche, regionale und inhaltliche Abgrenzung gehen, bevor dann das zentrale Argument entwickelt wird.

Die meisten der mit der Aufklärung verbundenen Neuerungen und Umwälzungen können im 18. Jahrhundert angesiedelt werden, das deswegen immer wieder als „*Zeitalter der Aufklärung*“[748] bezeichnet wird. In der Tat sind wichtige Entwicklungen, die die Aufklärung erst ermöglichten, in den zwei Jahrhunderten davor zu verorten. Die historische Ausgangslage der Aufklärung leitet sich maßgeblich aus der Reformation im 16. Jahrhundert, dem Dreißigjährigen Krieg (1618–1648) und der Glorious Revolution in England (1688) im 17. Jahrhundert ab. Ihren historischen Abschluss fand die Aufklärung mit der Französischen Revolution (1789) und der Amerikanischen Revolution dreizehn Jahre zuvor (1776).

Die Aufklärung ist untrennbar verbunden mit Fortschritten in den mathematisch-mechanischen Naturwissenschaften (Issaac Newton, Principia Mathematica, 1687) und bahnbrechenden Erfindungen wie z.B. Spinnmaschine (1764), Dampfmaschine (1765) und mechanischer Webstuhl (1785)[749]. Die Erkenntnisse in den Wissenschaften wurden zum Treibriemen der politischen und gesellschaftlichen Transformation und auch die Entwicklung zu einer Philosophie mit eigener Macht und Autorität auf der Grundlage ihrer eigenen Prinzipien:

> *“The rise of the new science progressively undermines not only the ancient geocentric conception of the cosmos, but also the set of presuppositions that had served to constrain and guide philosophical inquiry in the earlier times. The dramatic success of the new science in explaining the natural world promotes philosophy from a handmaiden of theology, constrained by its purposes and methods, to an independent force with the power and authority to challenge the old and construct the new, in the realms both of theory and practice, on the basis of its own principles*.”[750]

Die Philosophen Francis Bacon, Thomas Hobbes, René Descartes, John Locke, Baruch Spinoza und Gottfried Wilhelm Leibniz schufen die Grundlagen der Aufklärungsphilosophie, insbesondere in Bezug auf die Vertragstheorien der politischen Philosophie (Hobbes, Locke), die naturwissenschaftlichen Erkenntnistheorien (Bacon), die Geistesphilosophie (Descartes, Spinoza, Leibniz) und die Ethik (Locke, Spinoza)[751].

---

[748] Schneiders (1997), S. 7

[749] Vgl. Schneiders (1997), S. 23

[750] Quelle: „Enlightenment“ in Stanford Encyclopedia of Philosophy; Bristow, William, "Enlightenment", The Stanford Encyclopedia of Philosophy (Fall 2017 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/fall2017/entries/enlightenment/>.

[751] Vgl. Schneiders (1997)

Die Liste der Philosophen und Dichter, die dann im 18. Jahrhundert die Aufklärung maßgeblich in England/Schottland, Frankreich und Deutschland vorangetrieben haben ist lang: Frances Hutcheson, David Hume, Adam Smith, Thomas Reid, Voltaire (Francois-Marie Arouet), Jean Le Rond d'Alembert, Denis Diderot, Jean-Jacques Rousseau, Christian Wolff, Moses Mendelssohn, G.E. Lessing und Immanuel Kant[752]. Mit Ausnahme von Kant soll auf die einzelnen Beiträge dieser Philosophen hier nicht eingegangen werden.

Generell lassen sich zwei Begriffe unterscheiden, die in ihrer Kombination die Aufklärung ausmachen: der auf „*Wahrheit durch Klarheit*"[753] zielende „*rationalistische Aufklärungsbegriff"* und der auf „*Freiheit und Selbständigkeit"* zielende „*emanzipatorische Aufklärungsbegriff"*. Die Aufklärung wird bisweilen auch als „*Zeitalter der Kritik*" bezeichnet, was sich auf die „*Kritik des Aberglaubens und der Vorurteile, des Fanatismus und der Schwärmerei"* bezieht.

Religion, Politik und Wissenschaft sind gleichermaßen betroffen:

> *„Der Blick auf Religion, Politik und Wissenschaft macht die Aufklärung von ihrem Anfang her als Reaktion auf eine geschichtliche Situation verständlich. Angesichts der Erfahrung der Unvernunft in Religion und Politik und der Erfolge von Verstand oder Vernunft in den neuen Wissenschaften setzen die Aufklärer auf Verstand und Vernunft, und zwar auf die Aktivität des eigenen Denkens (Selbstdenkens), für das folglich mehr und mehr auch Freiheit (Mündigkeit) gefordert werden muss."*[754]

Die Aufklärung entwickelte sich also von einer Reaktion auf die geschichtliche Situation hin zu einem Prozess der Transformation und wurde dann zu einer Weltanschauung, „*zu einem variablen Ensemble von Ideen in einem relativ festen Denkrahmen zu einem von der Wirklichkeit sich abkoppelnden Gedankengebäude*"[755]. Letzteres zeigt sich schon in aller Deutlichkeit in der Französischen Revolution, von der man sich zunächst die „*Verwirklichung grundlegender Forderungen der Aufklärung"*[756] versprach und die nicht nur aufklärungsnahe Intellektuelle begeisterte. Nach der Hinrichtung des französischen Königs und seiner Gattin und dem folgenden Terror auf den Straßen von Paris und im ganzen Land änderten sich die Sichtweise und das „*geistige und politische Klima"* schnell. Eine endgültige Ernüchterung stellte sich dann bei und nach den Eroberungskriegen Napoleons ein.

752 Vgl. Schneiders (1997), S. 30f

753 Dieses und die folgenden Zitate: Schneiders (1997), S. 7

754 Schneiders (1997), S. 10

755 Schneiders (1997), S. 12

756 Schneiders (1997), S. 128

11.2 Kant: Was ist Aufklärung?

Immanuel Kants aktive Schaffensperiode überlappt nur mit der späten Phase der Aufklärung. Seine wesentlichen Veröffentlichungen haben die eigentliche Aufklärung nur wenig beeinflusst. Andere Philosophen hatten einen deutlich größeren Einfluss auf die Ausgestaltung und den Verlauf der Epoche. Trotzdem hat Kant mit seinem Aufsatz zur Beantwortung der Frage „*Was ist Aufklärung?*“ sicherlich ein außerordentlich prägnantes, wirkmächtiges und – wie wir sehen werden – zeitunabhängiges Dokument geschrieben, an dem sich die Gelehrten bis heute abarbeiten, weltweit. Samuel Fleischacker schreibt dazu:

> *“... his piece has become an emblem of the entire Enlightenment, an essay by which students are introduced to the intellectual world of the eighteenth century”*[757]

Bei dem im Dezember 1784 erschienenen Aufsatz[758] handelt es sich um Kants Antwort auf die in der Zeitschrift „Berlinische Monatsschrift“ vom Prediger und Theologen Johann Friedrich Zöllner im Dezember 1783 gestellte Frage „*Was ist Aufklärung?*“. Sein im Dezember 1784 veröffentlichter Text steht im Zusammenhang mit der in der September 1784 in der „Berlinischen Monatsschrift“ veröffentlichten Antwort des Philosophen Moses Mendelssohn und zwei weiteren Aufsätzen von Kant, einerseits sein Aufsatz „*Idee zu einer allgemeinen Geschichte in weltbürgerlicher Absicht*“[759] in der Ausgabe vom November 1784 der „Berlinischen Monatsschrift“ und andererseits sein Aufsatz „*Was heißt: sich im Denken orientieren?*“[760] vom Oktober 1786. Die beiden letztgenannten Schriften geben interessante zusätzliche Hinweise auf Kants Verständnis der Aufklärung.

Bereits der erste Satz thematisiert die beiden zentralen Begriffe seiner Position: die Unmündigkeit und die Selbstverschuldung derselben:

> ***„Aufklärung ist der Ausgang des Menschen aus seiner selbstverschuldeten Unmündigkeit***. [Hervorhebung DS] *Unmündigkeit ist das Unvermögen, sich seines Verstandes ohne Leitung eines anderen zu bedienen. Selbstverschuldet ist diese Unmündigkeit, wenn die Ursache derselben nicht am Mangel des Verstandes, sondern der Entschließung und des Mutes liegt, sich seiner ohne Leitung eines andern zu bedienen. Sapere aude! Habe Mut, dich deines eigenen Verstandes zu bedienen! ist also der Wahlspruch der Aufklärung.“*[761]

Aber was meint Kant nun konkret damit? Unter Unmündigkeit versteht er das „*Unvermögen, sich seines Verstandes ohne Leitung eines anderen zu bedienen*.“ Dies erläutert er an einer späteren Stelle mit Beispielen:

---

[757] Fleischacker (2013), S.12

[758] Kant (1784b)

[759] Kant (1784a)

[760] Kant (1786)

[761] Kant (1784b), S. 20

> *„Es ist so bequem unmündig zu sein. Habe ich ein Buch, das für mich Verstand hat, einen Seelsorger, der für mich Gewissen hat, einen Arzt, der für mich die Diät beurteilt usw. so brauche ich mich ja nicht selbst zu bemühen.*“[762]

Es drängt sich die Auslegung auf, dass man niemals auf Autoritäten in Form von Personen oder Büchern hören solle[763]. Das meinte Kant allerdings nicht. Er argumentierte vielmehr so: Falls man zu dem Schluss komme, dass der Arzt ein Scharlatan sei, solle man seinem Rat nicht mehr folgen. Man solle jederzeit in der Lage sein, Autoritäten kritisch zu hinterfragen. In seiner Liste der obigen Beispiele widmet er sich primär der Autorität und geistigen Führungsanspruch von Seelsorgern. Ferner spricht er den weltlichen Führungsanspruch weltlicher Institutionen an. War er also nur „politisch korrekt“ und hat bewusst den Unmündigkeitsbegriff allgemeingehalten? Vermutlich hat dies eine Rolle gespielt, insbesondere da der Artikel in einer prominenten Zeitschrift erschien. Durch den Verzicht auf eine Präzisierung ist die ganze Aussage *zeitunabhängig* geworden[764].

Der zweite Begriff ist derjenige der Selbstverschuldung. Seine Erläuterung dazu ist sehr deutlich:

> *„Faulheit und Feigheit sind die Ursachen, warum ein so großer Teil der Menschen, nachdem sie die Natur längst von fremder Leitung freigesprochen (naturaliter maiorennes), dennoch gerne zeitlebens unmündig bleiben; und warum es anderen so leicht wird, sich zu deren Vormündern aufzuwerfen.*[765]

Faulheit und Feigheit sind die Gründe, also fehlender Antrieb und fehlender Mut. An anderer Stelle stellt er die Verbindung zur Freiheit und insbesondere zur Willensfreiheit her:

> *„Dass aber ein Publikum sich selbst aufkläre, ist eher möglich; ja es ist, wenn man ihm nur Freiheit lässt, beinahe unausbleiblich.“*[766]

Der Mensch ist frei und kann sich gegen die Unmündigkeit entscheiden. Kant geht noch weiter und schreibt in seinem späteren Aufsatz „*Was heißt: sich im Denken orientieren?*“:

> *„Selbstdenken heißt den obersten Probierstein der Wahrheit in sich selbst (d.i. in seiner eigenen Vernunft) suchen; und die Maxime, jederzeit selbst zu denken, ist die Aufklärung.“*[767]

---

[762] Kant (1784b), S. 20

[763] Vgl. Fleischacker (2013), S. 14

[764] Vgl. Villhauer (2009), S. 16: *„Er vermeidet eine inhaltliche Aufzählung von Kriterien der Mündigkeit, zeigt vielmehr nur, wie und gegen was Mündigkeit sich entwickeln lässt. Genauso wird die Aufklärung nur **ex negativo** bestimmt, eben als Ausgang aus der Unmündigkeit. Das ist ein entscheidender Punkt, der beispielsweise verhindert, dass Kants Theorie zu sehr **zeitgebunden** bleibt.“;* Hervorhebungen durch DS

[765] Villhauer (2009), S. 16

[766] Kant (1784b), S. 21

[767] Kant (1786), S. 60

Kant fordert das Selbstdenken und das damit kohärente Handeln des Menschen, allerdings nicht für für alle Lebenslagen und Situationen fordert. Er unterscheidet eindeutig zwischen öffentlichem und privatem Vernunftsgebrauch[768]:

> *„Der öffentliche Gebrauch seiner Vernunft muss jederzeit frei sein, und der allein kann Aufklärung unter Menschen zustande bringen; der Privatgebrauch derselben darf öfters sehr enge eingeschränkt sein, ohne doch darum den Fortschritt der Aufklärung sonderlich zu hindern. Ich verstehe aber unter dem öffentlichen Gebrauche seiner eigenen Vernunft denjenigen, den jemand als Gelehrter von ihr vor dem ganzen Publikum der Leserwelt macht. Den Privatgebrauch nenne ich denjenigen, den er in einem gewissen ihm anvertrauten bürgerlichen Posten oder Amte von seiner Vernunft machen darf.“*[769]

Diese Position spiegelt Kants Sichtweise auf die Französische Revolution wider. Er glaubte an die Aufklärung über den öffentlichen Diskurs und nicht über Revolutionen und zivilen Ungehorsamkeit. Samuel Fleischacker bringt es auf den Punkt:

> *“The realm of argument, of free debate, must be separated from the realm of obedience.”*[770]

Nach Fleischacker ist der öffentliche Bereich derjenige, in dem wir uns von unseren privaten Identitäten als Anwälte, Ärzte, Juden oder Christen lösen und sie zum Gegenstand unserer Untersuchungen und Beurteilungen machen und ein umfassenderes menschliches Urteil fällen können:

> *“More broadly, Kant thinks that the public or general point of view can serve as a test for the correctness of our beliefs even on ordinary empirical matters. It can of course happen that everyone’s view on a certain subject are mistaken or corrupt, and we shouldn’t overlook the importance of individuals like Copernicus, who defy common sense correctly on some issue. But for the most part Kant is surely right that the understanding of those around us is a healthy corrective for our private judgements, and that one who refuses to check in with the judgements of others, when he thinks he sees or hears something, is on the way to madness.”*[771]

Keinesfalls – so Fleischacker – sollte man davon ausgehen, dass Kant mit „Selbstdenken“ nur das individuelle Denken jedes Einzelnen für sich allein gemeint hat. Die Beiträge und Aussagen anderer Menschen spielen eine Rolle und können und dürfen nicht von unseren eigenen Wahrnehmungen und rationalen Überlegungen getrennt werden:

> *“Consequently, we must regard the word of others as an independent source of knowledge, right up there with perception and our various modes of reasoning”*[772]

Trotzdem sollte das Urteil anderer niemals ungeprüft und unwidersprochen übernommen werden:

---

768 Vgl. Villhauer (2009), S. 17

769 Kant (1784b); S. 22; A485

770 Fleischacker (2013), S. 15

771 Fleischacker (2013), S. 19

772 Ebd.

> *"We should not be cowed by the aura of superiority with which certain people or institutions appear to us. We should realize instead that we are responsible for the power that that aura has over us, and have the courage to resist that power"*[773]

Der aufgeklärte Mensch nutzt seinen Verstand und seine Vernunft, um die Mächte zu erkennen, die Einfluss auf ihn nehmen, und widersteht dort, wo es erforderlich scheint.

Auch Michel Foucault betont die Einschränkung der Aufklärung auf den öffentlichen Gebrauch der Vernunft:

> *„Aufklärung ist ... nicht bloß der Prozess, in dem die Individuen ihre persönliche Meinungsfreiheit garantiert sehen.* ***Aufklärung gibt es dort, wo sich der universale, der freie und der öffentliche Gebrauch der Vernunft überlagern****."*[774]

Die Kombination von Gehorsamkeit im Privaten und freiem Gebrauch der Vernunft kann hingegen nur konstruktiv wirken, wenn der öffentliche Diskurs Konsequenzen in der Gesetzgebung hat:

> *„Der öffentliche und freie Gebrauch der autonomen Vernunft wird* ***die beste Garantie des Gehorsams*** *sein, jedoch unter der Bedingung, dass das politische Prinzip, dem gehorcht werden muss, selbst mit der universalen Vernunft übereinstimmt."*[775]

Oberste Autorität ist die Vernunft, die sich im freien, universalen und öffentlichen Diskurs offenbart. Onara O'Neill schreibt dazu:

> *"A public use of reason,* […], *is in the first place one that could reach the world at large if suitably publicized. It must therefore assume no authority that could not be accepted by an unrestricted audience. Since the "world at large" accepts no common external authority, the only authority the communication can assume must be internal to the communication.* […] *The only authority internal to communication is, in Kant's view, reason."*[776]

Kants Präzedenz des öffentlichen über den privaten Vernunftsgebrauch in der Aufklärung, die auf den ersten Blick liberalen Grundprinzipien widerspricht, erschließt sich also aus der Forderung nach einer Abwesenheit jeglicher anderen Autorität als derjenigen der Vernunft[777]. Die sich daraus ergebenen Anforderungen an den aufgeklärten öffentlichen Diskurs sind nach O'Neill klar:

> *"What is spoken or written cannot count as a public use of reason merely because it is noised or displayed or broadcast to the world at large. Communication has also to meet sufficient standards of rationality to be interpretable to audiences who share no other, rationally ungrounded, authorities."*[778]

---

[773] Fleischacker (2013), S. 21

[774] Foucault (1990), S. 40; Hervorhebung DS

[775] Foucault (1990); Hervorhebung DS

[776] O'Neill (1989), S.35; Hervorhebung DS

[777] Vgl. auch O'Neill (1989), S. 35; Zitat von Ronald Beiner: „*This precedence accorded to public over private prerogatives may appear as something of an inversion of traditional liberal priorities on the part of one of the fountainheads of liberal thought.*"

[778] O'Neill (1989), S. 35

Kant dazu in der Kritik der reinen Vernunft:

> *„Denn es ist sehr was Ungereimtes, von der Vernunft Aufklärung zu erwarten und ihr doch vorher vorzuschreiben, auf welche Seite sie nothwendig ausfallen müsse. Überdem wird Vernunft schon von selbst durch Vernunft so wohl gebändigt und in Schranken gehalten, daß ihr gar nicht nöthig habt, Scharwachen aufzubieten, um demjenigen Theile, dessen besorgliche Obermacht euch gefährlich scheint, bürgerlichen Widerstand entgegen zu setzen.“*[779]

Daraus leitet O’Neill auch Kants Toleranzgedanken ab. Voraussetzung der Autorität der Vernunft ist die Toleranz. Die Entwicklung der Vernunft und der Toleranz sind miteinander verwoben und bedingen sich gegenseitig:

> *“Such instrumental justification of toleration all presuppose that we have independent standards of rationality and methods of reaching truth. Kant’s thought is rather a degree of toleration must characterize ways of life in which presumed standards of reason and truth can be challenged, and so acquire the only sort of vindication of which they are susceptible. The development of reason and of toleration is interdependent: a measure of publicizability is needed for publicity; and publicity in turn is needed for further development of standards of publicizability. Practices of toleration help constitute reason‘s authority.“*[780]

Dies ist ein Gedankengang, auf den im weiteren Verlauf des Arguments in dieser Arbeit noch eingegangen werden soll. Schon Kant hat unter Aufklärung nicht nur das Selbstdenken von Mündigen verstanden, sondern auch die Notwendigkeit eines öffentlichen Forums herausgestellt, in dem nur eine Autorität akzeptiert wird, die der Vernunft, und ein hohes Maß an Toleranz herrscht.

Ernst Cassirer schrieb 1918 dazu:

> *„Denn seine Ethik verweist ihn zwar auf das Individuum und auf den Grundbegriff der sittlichen Persönlichkeit und ihrer Selbstgesetzgebung; aber seine geschichtliche und geschichtsphilosophische Einsicht führt* [ihn] *auf die Überzeugung, dass nur durch das Medium der Gesellschaft hindurch die ideelle Aufgabe des sittlichen Selbstbewusstseins ihre tatsächliche empirische Erfüllung finden könne.“*[781]

Der indisch-amerikanische Philosoph Amartya Sen hat in seinem Werk „***Die Idee der Gerechtigkeit***“ den öffentlichen Gebrauch der Vernunft als Grundvoraussetzung für eine – aus derzeitiger Sicht nicht erwartbare – globale Demokratie in der Zukunft benannt und bezieht sich dabei auf Kant und Mill sowie auf Rawls und Habermas:

---

[779] Kant (1787a), III AA 489; Kant (1787b), S. 636 – 637 (B775, A 747)

[780] O’Neill, S. 39

[781] Cassirer (1918), S. 238; Ebenfalls von Interesse ist Cassirers ausführlichere Erläuterung in Cassirer (1918), S. 240: „*So ist es das Böse selbst, das im Lauf und Fortgang der Geschichte zum Quell des Guten werden muss: so ist es die Zwietracht, aus der allein die wahrhafte, ihrer selbst sichere sittliche Eintracht sich herstellen kann. Die eigentliche Idee der sozialen Ordnung besteht darin, die Einzelwillen nicht in einer allgemeinen Nivellierung untergehen zu lassen, sondern sie in ihrer Eigenart und somit in ihrem Gegensatz zu erhalten; - zugleich aber die Freiheit jedes Individuums derart zu bestimmen, dass sie an der des anderen ihre Grenze findet.*“

> *„In diesem Buch wird Demokratie am öffentlichen Vernunftsgebrauch gemessen, das heißt, als `Regierung durch Diskussion´ verstanden (eine Vorstellung, die John Stuart Mill sehr gefördert hat). Aber Demokratie muss auch allgemeiner gesehen werden, im Rahmen ihrer Fähigkeit, durchdachtes Engagement zu fördern, indem sie für mehr Informationen sorgt und interaktive Diskussionen möglich macht.“*[782]

> *„Häufig und offenbar überzeugend wird darauf hingewiesen, dass es in absehbarer Zukunft keinen globalen Staat geben kann und dass deshalb ein globaler demokratischer Staat erst recht undenkbar ist. Das ist richtig; und doch muss die Hoffnung auf eine globale Demokratie nicht notwendig auf unbestimmte Zeit auf Eis gelegt werden,* ***wenn Demokratie als öffentlicher Gebrauch der Vernunft verstanden wird****.“*[783] [Hervorhebung DS]

Fleischacker unterscheidet eine minimalistische und eine maximalistische Lesart von Kants Aufklärungsbegriff. Die minimalistische Version ist eine flexible Fassung des individuellen Selbstdenkens, die klar vorgibt, *wie* zu denken sei und nicht *was* zu denken sei. In einer stärkeren maximalistischen Version lässt sich interpretieren, dass Kant die traditionellen Religionen durch eine neue universelle moralische Religion ersetzt:

> *“The minimalist enlightenment that consists in the free expression of responsible views will then simply pave the way for a maximalist enlightenment in which traditional religions, and other authority-bound worldviews, fade away and everyone is committed to a rational science and morality alone.”*[784]

Die minimalistische Interpretation überlässt es jedem Einzelnen, zu seinem eigenen Urteil zu kommen, das sich durchaus von den Urteilen anderer Personen unterscheiden kann. Erst die maximalistische Auslegung beinhaltet die Erwartung, dass unterschiedliche Personen beim jeweiligen Selbstdenken und bei der jeweiligen Anwendung des kategorischen Imperativs zu dem gleichen Urteil gelangen.

Entlang diverser Textstellen in Kants Schriften, die in dieser Arbeit nicht detailliert nachgezeichnet werden sollen, lassen sich beide Lesarten belegen. Die Ambivalenz kann in Kants Originalquellen nicht ausgeräumt werden. Sie spielt jedoch eine große Rolle bei den Philosophen, die sich auf Kant berufen und die Aufklärung in der weiteren Entwicklung ausgelegt haben, und in der Aufklärungskritik.

Ist die universale Vernunft im obigen Zitat von Foucault, die die politischen Prinzipien diktiert, nicht ebenso autoritär wie die religiöse und die weltliche Autorität der Voraufklärung und steht damit im Widerspruch zu den Interessen des Individuums?

---

[782] Sen (2010), S. 13

[783] Sen (2010), S. 436

[784] Fleischacker (2013), S. 33

11.3 Aufklärungskritik

Schon seit ihrem Beginn im 18. Jahrhundert war die Aufklärung Gegenstand der Kritik und Ablehnung. Auslöser, Argumente und Zielsetzung der Aufklärungskritik in diesen rund 250 Jahren waren höchst unterschiedlich. Staat und Kirche hatten in den frühen Jahren ein großes Interesse daran, das Rad zurückzudrehen. Einige Gelehrte wie zum Beispiel Edmund Burke und Johann Georg Hamann lieferten dafür die intellektuelle Argumentation. Im 19. Jahrhundert begann eine Bewegung gegen den aufklärerischen Geist spätestens nach dem Sturz Napoleons im Zeitalter der Restauration und der geistig-kulturellen Romantik. Friedrich Nietzsche rief sogar offen zur Gegenaufklärung auf.

Die Aufklärungskritik wurde im späten 19. und frühen 20. Jahrhundert deutlich präziser, etwa bei Max Weber.

11.3.1 Aufklärungskritik bei Max Weber: das stahlharte Gehäuse

Max Weber stand dem mit der Aufklärung einhergehenden „*Prozess der Entzauberung und Rationalisierung"* eher skeptisch gegenüber, wenngleich er ihn für „*unabwendbar* hielt"[785]. In diversen Schriften beklagte er, dass das „*stahlharte Gehäuse*" der „*gigantischen Verwaltungsmaschinerie die individuelle Freiheit ernsthaft bedroht"*[786].

Darin ist auch der für die Untersuchungen dieser Arbeit so wichtige Gedanke enthalten, dass die von der Aufklärung freigesetzten Mechanismen die Freiheit und Autonomie des Menschen wieder einschränken:

> *„Der Puritaner wollte Berufsmensch sein, - wir müssen es sein. Denn indem die Askese aus den Mönchszellen heraus in das Berufsleben übertragen wurde und die innerweltliche Sittlichkeit zu beherrschen begann, half sie an ihrem Teile mit daran, jenen mächtigen Kosmos der modernen, an die technischen und ökonomischen Voraussetzungen mechanisch-maschineller Produktion gebundenen, Wirtschaftsordnung erbauen, der heute den Lebensstil aller einzelnen, die in dies Triebwerk hineingeboren werden - nicht nur der direkt ökonomisch Erwerbstätigen -, mit überwältigendem Zwange bestimmt und vielleicht bestimmen wird, bis der letzte Zentner fossilen Brennstoffs verglüht ist. Nur wie 'ein dünner Mantel, den man jederzeit abwerfen könnte', sollte nach Baxters[787] Ansicht die Sorge um die äußeren Güter um die Schulter seiner Heiligen liegen. Aber aus dem Mantel ließ das Verhängnis ein* ***stahlhartes Gehäuse*** *werden. Indem die Askese die Welt umzubauen und in der Welt sich auszuwirken unternahm, gewannen die äußeren Güter dieser Welt zunehmende und schließlich unentrinnbare Macht über den Menschen, wie niemals zuvor in der Geschichte. Heute ist ihr Geist - ob endgültig, wer weiß es? - aus diesem Gehäuse entwichen. Der siegreiche Kapitalismus jedenfalls bedarf, seit er auf mechanischer Grundlage ruht, dieser Stütze nicht mehr.* ***Auch die rosige Stimmung ihrer lachenden Erbin: der***

[785] Müller Sigmund (2020), S. 63

[786] Müller Sigmund (2020), S. 144

[787] Baxter war ein puritanischer Prediger, dessen Schriften Weber in seiner Argumentation des Öfteren heranzog; Vgl. Heinemann (2011), S.26

> ***Aufklärung, scheint endgültig im Verbleichen** und als ein Gespenst ehemals religiöser Glaubensinhalte geht der Gedanke der 'Berufspflicht' in unserm Leben um. (...) Niemand weiß noch, wer künftig in jenem Gehäuse wohnen wird, und ob **am Ende dieser ungeheuren Entwicklung ganz neue Propheten oder eine mächtige Wiedergeburt alter Gedanken und Ideale** stehen werden, oder aber - wenn keins von beiden - mechanisierte Versteinerung, mit einer Art von krampfhaftem Sich-wichtig-nehmen verbrämt. Dann allerdings könnte für die 'letzten Menschen' dieser Kulturentwicklung das Wort zur Wahrheit werden: **'Fachmenschen' ohne Geist, Genußmenschen ohne Herz**: dies Nichts bildet sich ein, eine vorher nie erreichte Stufe des Menschentums erstiegen zu haben."*[788]

Weber sah einerseits das „*Verbleichen der Aufklärung*" durch Einengung der Freiheit im „*stahlharten Gehäuse*" im *Kosmos der Wirtschaftsordnung*, die Transformation des Menschen zum „*Fachmenschen ohne Geist*" oder zum „*Genussmenschen ohne Herz*". Damit einhergehend konstatierte er in seinem Vortrag „*Wissenschaft als Beruf*" die „*Entzauberung der Welt*":

> *„Die zunehmende Intellektualisierung und Rationalisierung bedeutet also nicht eine zunehmende allgemeine Kenntnis der Lebensbedingungen, unter denen man steht. Sondern sie bedeutet etwas anderes: das Wissen davon oder den Glauben daran: daß man, wenn man nur wollte, es jederzeit erfahren könnte, daß es also prinzipiell keine geheimnisvollen unberechenbaren Mächte gebe, die da hineinspielen, daß man vielmehr alle Dinge – im Prinzip – durch Berechnen beherrschen könne. Das aber bedeutet: die **Entzauberung der Welt**. Nicht mehr, wie der Wilde, für den es solche Mächte gab, muss man zu magischen Mitteln greifen, um die Geister zu beherrschen oder zu erbitten. Sondern technische Mittel und Berechnung leisten das. Dies vor allem bedeutet die Intellektualisierung als solche."*[789]

Für Weber, so schreibt Gunzelin Schmid Noerr, „*bestand die wichtigste Konsequenz dieses Entzauberungsprozesses im Rückzug der Wissenschaften von den Fragen des Lebenssinns*, die *allein den privaten Vorlieben und Intuitionen überlassen"*[790] werden.

Diese beiden Kerngedanken der „Entzauberung der Welt" in Kombination mit der zunehmenden schöpferischen und geistigen Einengung des Menschen durch das „stählerne Gehäuse" sollten auch viele der ihm folgenden Kritiker des aufgeklärten Zeitalters bzw. der Moderne beschäftigen.

---

[788] Zitiert in Heinemann (2011), S. 26; Originalzitat: Weber (1920), S. 203f; Hervorhebungen durch DS

[789] Weber (1919a), S. 19; Hervorhebung DS

[790] Schmid Noerr (2019), S. 26

11.3.2 Die Aufklärungskritik der Frankfurter Schule

> *„Seit je hat Aufklärung im umfassendsten Sinn fortschreitenden Denkens das Ziel verfolgt, von den Menschen die Furcht zu nehmen und sie als Herren einzusetzen. Aber die vollends aufgeklärte Erde strahlt im Zeichen triumphalen Unheils.“*[791]

Die „*während des Zweiten Weltkriegs im US-Exil von Max Horkheimer und Theodor Adorno* geschriebene „*Dialektik der Aufklärung*“ (Jürgen Habermas: „*Die Dialektik der Aufklärung ist ein merkwürdiges Buch.*“[792]) *gehört unter den sozialphilosophischen Werken, die auf die Katastrophen des 20. Jahrhunderts reagieren, zu den bedeutendsten“*[793]. Gegenstand des Buches ist der „***Selbstzerstörungsprozess der abendländischen Rationalität***“[794].

Horkheimer und Adorno schließen ihre Argumentation nahtlos an die „Entzauberung der Welt“ und die reine Zweckrationalität von Max Weber an. Sie beklagen primär die Reduzierung der aufklärerischen Vernunft auf eine „instrumentelle Rationalität“. Sie formulieren noch zwei sekundäre Thesen, die an dieser Stelle nicht weiterverfolgt werden sollen:

- Schon die Mythen beinhalten Aufklärung
- Die Aufklärung verstrickt sich in Mythologie

Im Original:

> *„Wie die Mythen schon Aufklärung vollziehen, so verstrickt Aufklärung mit jedem ihrer Schritte tiefer sich in Mythologie.“*[795]

Im Folgenden soll die Kritik von Kants Aufklärungsbegriff vertieft werden:

> *„Kants Begriffe sind doppelsinnig. Vernunft als das transzendentale überindividuelle Ich enthält die Idee eines freien Zusammenlebens der Menschen, in dem sie zum allgemeinen Subjekt sich organisieren und den Widerstreit zwischen der reinen und empirischen Vernunft in der bewussten Solidarität des Ganzen aufheben. Es stellt die Idee der wahren Allgemeinheit dar, die Utopie. Zugleich jedoch bildet Vernunft die* ***Instanz des kalkulierenden Denkens****, das die Welt für die Zwecke der Selbsterhaltung zurichtet und keine anderen Funktionen kennt als die Präparierung des Gegenstandes aus bloßem Sinnenmaterial zum Material der Unterjochung.“*[796]

Die Autoren interpretieren das „*Selbstdenken als Verfahren von Subsumtion und Formalismus und als Erzeugung von Einstimmigkeit und System“*[797], heute würde man sagen: Die Reduktion auf den kleinsten gemeinsamen Nenner. „*Aufklärerisches Denken wird*

---

[791] Horkheimer Adorno (1944), S. 9

[792] Habermas (1988), S. 130

[793] Schmid Noerr (2019), S. 1

[794] Schmid Noerr (2019), S. 1

[795] Horkheimer Adorno (1944), S. 18

[796] Horkheimer Adorno (1944), S. 90; Hervorhebung DS

[797] Schmid Noerr (2019), S. 26

auf `instrumentelle Vernunft´ [oder auf ‚kalkulierendes Denken'; Ergänzung DS] *regrediert, die zur universell einsetzbaren Verfügungsmacht über äußere Natur, Gesellschaft und innere Natur wird"*[798]. Oder wie Fleischacker schreibt:

> *"Horkheimer and Adorno allege that there is and can be no content to Kant's moral system. They say that reason for Kant is and must be wholly instrumental, helping us fulfill purposes without telling us anything about the purposes we ought to have.*"[799]

Nicht nur Fleischacker interpretiert dies schon fast als böswilliges Missverstehen Kants durch Adorno und Horkheimer. Unabhängig von der (Fehl-) Auslegung von Kants originalen Schriften bleibt der Vorwurf hinsichtlich des Rückzugs der Vernunft auf die rein instrumentelle Rationalität in der modernen Welt bestehen und ist auch für die Zwecke dieser Arbeit erwägenswert.

Mit der „*Zweckrationalität*" Webers, die Horkheimer „*Instrumentelle Vernunft"* nannte, hat er sich in seiner 1947 herausgebrachten Schrift „*Eclipse of Reason*"[800] zur Kritik derselben auseinandergesetzt. Gemäß Henning Ottmann meint er damit „*eine Vernunft, welche die Mittel, nicht jedoch die Ziele des Handelns reflektiert"*[801]. In anderen Worten: Die menschliche Vernunft verliert ihre Zwecksetzungskompetenz, sobald sie ausschließlich der Bereitstellung technischer und ökonomischer Mittel dient.

Horkheimer unterscheidet eine subjektive Vernunft der Selbsterhaltung des Individuums oder der Gemeinschaft von der objektiven Vernunft, die darauf abzielt*, „ein umfassendes System oder eine Hierarchie alles Seienden einschließlich des Menschen und seiner Zwecke zu entfalten"*[802]. In der Tat unterstellt Horkheimer der objektiven Vernunft, dass auch sie letzten Endes „*im Dienst subjektiver Interessen"* stehe.[803] Er argumentiert, dass „*Gerechtigkeit, Glück, Gleichheit und Toleranz"* Begriffe geworden seien, die „*ihre geistigen Wurzeln verloren"* hätten. Sie seien noch Ziele und Zwecke, allerdings gebe es „*keine rationale Instanz, die befugt wäre, ihnen einen Wert zuzusprechen und sie mit einer objektiven Realität zusammenzubringen"*[804]:

> *„Das Fortschreiten der Aufklärung löst die Idee der objektiven Vernunft auf, den Dogmatismus und den Aberglauben, aber oft ziehen Reaktion und Obskurantismus den größten Vorteil aus dieser Entwicklung."*[805]

Jürgen Habermas kommentierte diesen Gedanken in seinen Vorlesungen zum Diskurs der Moderne wie folgt:

---

798 Schmid Noerr (2019), S. 27

799 Fleischacker (2018), S. 128

800 Auf deutsch: „Zur Kritik der instrumentellen Vernunft", Horkheimer (1947)

801 Ottmann (2012), S. 69

802 Horkheimer (1947), S. 17

803 Vgl. Ottmann (2012), S. 69

804 Horkheimer (1947), S. 36f; einschließlich der vorherigen Zitate

805 Ebd.

> *„Die Argumentation folgt also in Ansehung der Wissenschaft, der Moral und der Kunst derselben Figur: bereits die Trennung der kulturellen Bereiche, der Zerfall der in Religion und Metaphysik noch verkörperten substantiellen Vernunft, entmächtigt die isolierten, ihres Zusammenhalts beraubten Vernunftsmomente so sehr, dass diese zur Rationalität im Dienste wildgewordener Selbsterhaltung regredieren. Vernunft wird in der kulturellen Moderne endgültig ihres Geltungsanspruchs entkleidet und an schiere Macht assimiliert. Die kritische Fähigkeit, mit „Ja" oder „Nein" Stellung zu nehmen, zwischen gültigen und ungültigen Aussagen zu unterscheiden, wird unterlaufen, indem Macht- und Geltungsansprüche eine trübe Fusion eingehen."*[806]

Aus der Vernunft wird eine ausschließlich der Selbsterhaltung dienende Rationalität, in der das gilt, was der Macht dient.

Das von den beiden Autoren in ihrem Buch von 1947 gezeichnete Bild der Aufklärung ist unabhängig von diversen Einsprüchen und Widersprüchen bis heute einflussreich[807], auch bei der Beurteilung aktueller Phänomene.

11.3.3 Kritik des Glaubens an Technik und Kybernetik

Martin Heidegger hat sich insbesondere in seinem Spätwerk kritisch mit der Technik und der Kybernetik und vielen Aspekten der Aufklärung auseinandergesetzt. Ein wichtiges Dokument seiner diesbezüglichen Überlegungen ist die Gedenkrede „*Gelassenheit*"[808], die er 1955 in Meßkirch hielt. Einige Kernaussagen sollen hier herausgestellt werden. Generell beklagt er „*die zunehmende Gedankenlosigkeit"* des heutigen (Stand 1955) Menschen:

> *„Der heutige Mensch ist auf der Flucht vor dem Denken."*[809]

Heidegger unterscheidet zwischen zwei Arten des Denkens: „*das rechnende Denken und das besinnliche Nachdenken"*[810]. Das rechnende Denken, das der subjektiven Vernunft von Horkheimer ähnelt, dominiert alles und entzieht dem Nachdenken den Boden:

> *„Das rechnende Denken kalkuliert. Es kalkuliert mit fortgesetzt neuen, mit immer aussichtsreicheren und zugleich billigeren Möglichkeiten. Das rechnende Denken hetzt von einer Chance zur nächsten. Das rechnende Denken hält nie still, kommt nicht zur Besinnung. Das rechnende Denken ist kein besinnliches Denken, kein Denken, das dem Sinn nachdenkt, der in allem waltet, was ist.* […]
>
> *Dieses Nachdenken aber meinen wir, wenn wir sagen, der heutige Mensch sei auf der Flucht vor – dem Denken. Allein, so entgegnet man, das bloße Nachdenken schwebt doch unversehens über der Wirklichkeit. Es verliert den Boden. Es taugt nichts für die*

806 Habermas (1988), S. 137

807 Vgl. Lavaert Schröder (2018), S. 3

808 Heidegger (1955)

809 Heidegger (1955), S. 11f

810 Heidegger (1955), S. 13

> *Bewältigung der laufenden Geschäfte. Es bringt nichts ein für die Durchführung der Praxis.“*[811]

In seinen weiteren Ausführungen macht Heidegger den Mangel an Nachdenken für das damals aktuelle Thema der Atomphysik, der Nuklearenergie und – etwas breiter gefasst – das Atomzeitalter verantwortlich. Auf dieses bezogen macht er eine pessimistische Aussage, die man ohne weiteres auch auf andere Technologien mit Potential für umfassende Veränderungen übertragen könnte:

> *„Kein einzelner Mensch, keine Menschengruppe, keine Kommission noch so bedeutender Staatsmänner, Forscher und Techniker, keine Konferenz von führenden Leuten der Wirtschaft und Industrie vermag den geschichtlichen Verlauf des Atomzeitalters zu bremsen oder zu lenken. Keine nur menschliche Organisation ist imstande, sich der Herrschaft über das Zeitalter zu bemächtigen.“*[812]

Heidegger war nicht kategorisch gegen den technischen Fortschritt, er unterschied aber zwischen Technologien, die wir verstehen und zu denen wir „ja“ oder „nein“ sagen können, und Technologien, deren Sinn sich uns verbirgt[813]. Er kontrastiert die „*Gelassenheit zu den Dingen*“ (daher der Titel des Vortrages) beim ersten Typus von Technik mit der bewussten „*Offenheit für das Geheimnis*“ bei den Technologien mit unbekanntem Sinn. Wie so oft bei Heidegger, ist nicht vollständig klar, was er konkret mit dieser Formulierung meint. Die Erläuterung von Holger Zaborowski konkretisiert Heideggers Gedanken:

> „*Heidegger ist sich bewusst, wie leicht das Wort ‚Geheimnis‘ missverstanden werden kann. Zur ‚Offenheit für das Geheimnis‘ gehört es allerdings, dass er auch die Bedeutung von ‚Geheimnis’ offen lässt. Das ‚Geheimnis‘ ist nicht etwas was nicht bekannt ist oder das die Wissenschaft noch nicht herausgefunden hat. Auch handelt es sich nicht um ein Geheimwissen. Heidegger spricht keiner Esoterik, die sich nur an wenige Eingeweihte richtet, das Wort. ‚Geheimnis‘ ist genauso wenig wie ‚Gelassenheit‘ ein Begriff, der in seiner Bedeutung klar umrissen werden könnte.* […] *Geheimnis‘ ist wie ‚Gelassenheit‘ so etwas wie ein Wegweiser.* […] *Was im Geheimnis erfahren werden kann, so zeigt sich, ist gerade nicht das Unbekannte, sondern das Vertraute, das Heimatliche. ‚Geheim‘ bedeutet ursprünglich nämlich auch: vertraut, auf das eigene Heim, die Heimat bezogen. Die ‚Offenheit für das Geheimnis‘ verweist daher auf jenes Offensein für das, was einem immer schon vertraut gewesen ist, für das, was hier und heute sich zeigt. Heideggers Denken ist daher immer auch eine Erinnerung an die konkrete Existenz des Menschen.* […] *Diese Erinnerung an die Dinge und ihre Nähe, an Herkunft und Zukunft ist gerade deshalb bedeutend, weil die Flucht vor dem Konkreten, vor dem, was sich einem sinnenden und besinnlichen Denken als Sinn zeigt, Menschen immer wieder in die Fernen des Allgemeinen und aus der Heimat in die Fremde treibt und weil das gegenwärtige technische Zeitalter selbst vieles vergessen lässt und zur ‚totalen Gedankenlosigkeit‘ führen könnte.“*[814]

---

811 Heidegger (1955), S. 13

812 Heidegger (1955), S. 20

813 Vgl. Heidegger (1955), S. 23

814 Zaborowski (2014), S. 99f

Als Resümee dieser Überlegung warnt er vor dem Risiko eines Dritten Weltkrieges mit der völligen Vernichtung der Menschheit und vor der grundsätzlichen Gefahr, dass die „*Technik den Menschen auf eine Weise fesseln, behexen, blenden und verblenden könnte, dass eines Tages das rechnende Denken als das einzige in Geltung und Übung bliebe"*[815].

An anderer Stelle äußert sich Heidegger ähnlich wie im obigen Zitat zum Atomzeitalter sehr skeptisch zu der Frage, ob wir Menschen überhaupt in der Lage seien, über die anstehenden Weltveränderungen angemessen nachzudenken:

> *„Dabei ist jedoch das eigentlich Unheimliche nicht dies, dass die Welt zu einer durch und durch technischen wird. Weit unheimlicher bleibt, dass der Mensch für diese Weltveränderung nicht vorbereitet ist, dass wir es nicht vermögen, besinnlich denkend in eine sachgemäße Auseinandersetzung mit dem zu gelangen, was in diesem Zeitalter eigentlich heraufkommt."*[816]

Laut Alfred Denker, der ein Nachwort zu Heideggers Rede über Gelassenheit verfasste, hat Heidegger immer betont, „*dass das Wesen der Physik nichts Physikalisches, dass das Wesen der Technik nichts Technisches und dass das Wesen des Menschen nichts Menschliches sei."* Die Physik kann die Frage „*Was ist Physik*?" nicht mit den Methoden der Physik beantworten. Und: „*Das Problem der unaufhaltsamen Übermacht der Technik* [kann] *nicht technisch gelöst werden"*[817]. Denkers Aussage zum Wesen des Menschen ist sicherlich diskutierbar und kontrovers, in Bezug auf die Physik und Technik allerdings plausibel. Auch die Schlussfolgerung ist klar und nachvollziehbar: Die Technik kann das Problem ihrer eigenen Übermacht nicht lösen.

Heidegger fordert den Willensakt der Selbstbefreiung des Menschen und damit eine „*neue Form des Denkens als die neue Haltung zur Technik*"[818]. Er nennt das „*Gelassenheit zu den Dingen*", also eine Art sachliche Distanz gegenüber dem technischen Objekt, die jedoch explizit keine gedankenlose Distanz ist, sondern eine nachdenkende Distanz, „*eine Haltung der Aufmerksamkeit und Achtsamkeit* […]*, die er ‚die Offenheit für das Geheimnis'*[819] *nennt"*[820]. Nur mit einem besinnlichen Nachdenken, das Heidegger vom rechnenden Denken abgrenzt, kann diese Aufmerksamkeit und Achtsamkeit erreicht werden[821].

---

815 Heidegger (1955), S. 25

816 Denker (2014), S. 54; Originalzitat aus: Martin Heidegger, *Was heißt denken?* hrsg. von Paola-Ludovika Coriando, Frankfurt am Main 2002

817 Zitate in diesem Abschnitt: Denker (2014), S. 55

818 Denker (2014), S. 56

819 Denker zitiert hier Heidegger (1955), S. 24: „*Die Gelassenheit zu den Dingen und die Offenheit für das Geheimnis gehören zusammen. Sie gewähren uns die Möglichkeit, uns auf eine ganz andere Weise in der Welt aufzuhalten. Sie versprechen uns einen neuen Grund und Boden, auf dem wir innerhalb der technischen Welt, und ungefährdet durch sie, stehen und bestehen können. Die Gelassenheit zu den Dingen und die Offenheit für das Geheimnis geben uns den Ausblick auf eine neue Bodenständigkeit. Diese könnte sogar eines Tages geeignet sein, die alte, jetzt rasch hinschwindende Bodenständigkeit in einer gewandelten Gestalt zurückzurufen.*"

820 Denker (2014), S. 65

821 Siehe oben

Drei Konsequenzen lassen sich aus Heideggers Philosophie und insbesondere aus seiner Rede zur Gelassenheit ableiten:

1. Noch stärker als Mitte des letzten Jahrhunderts droht mit der Einführung der KI die absolute Dominanz des „rechnenden Denkens" über das „Nachdenken". Die KI ist hilfreich bei der ersten Kategorie, allerdings zur zweiten Art nicht imstande. Nur der Mensch kann nachdenken.
2. Eine neue Haltung der „nachdenkenden Distanz gegenüber der Technik" ist erforderlich. Dieses Nachdenken muss in der gesamten Gesellschaft erfolgen, nicht nur in den Technikbranchen und Unternehmen, die sich mit KI beschäftigen.
3. Die von Heidegger geforderte „Offenheit für das Geheimnis" bedeutet, dass die Implikationen für die „konkrete Existenz des Menschen" Gegenstand des Nachdenkens über die KI sein müssen.

Aus dieser Perspektive überrascht es nicht, dass Hans Jonas sehr stark von Martin Heidegger beeinflusst wurde, auch wenn die beiden Philosophen als Gegenspieler verstanden werden[822].

11.3.4 Foucaults Kritik der Aufklärung

> *„Nicht die Treue zu doktrinären Elementen ist der Faden, der uns mit der Aufklärung verbinden kann, sondern die ständige Reaktivierung einer Haltung – das heißt eines philosophischen Ethos, das als permanente Kritik unseres historischen Seins beschrieben werden könnte."*[823]
>
> Michel Foucault

Der postmoderne Philosoph Michel Foucault hat sich in zwei Bereichen seiner Arbeit grundsätzlich mit der Aufklärung auseinandergesetzt. Einerseits hat er selbst in Anlehnung an Kants ursprünglichen Beitrag einen Aufsatz mit dem Titel „*Was ist Aufklärung?*" verfasst und darin sein Verständnis der Aufklärung entwickelt. Andererseits hat er sich in mehreren Werken mit Grenzsituationen der modernen Zivilisation, wie z.B. den Umgang mit psychisch Kranken oder Straftätern[824], aufklärungskritisch beschäftigt.

Für Foucault ist die Aufklärung und der damit verbundene Ausgang aus der selbstverschuldeten Unmündigkeit kein singuläres und einmaliges Ereignis, sondern eine Haltung,

---

[822] Vgl. Elm (2021), S. 15: „*Neben weiteren Gemeinsamkeiten von Heidegger und Jonas – z.B. heben beide (a) die* ***Beispiellosigkeit der modernen Technik*** *hervor, sehen sie (b) in ihrer* ***Radikalisierung begründet im dualistischen und im****, so Heidegger,* ***willensmetaphysischen Denken der neuzeitlichen Subjektphilosophie*** [...]*, kritisieren (c) das* ***anthropologisch- instrumentelle Technikverständnis mit seiner Neutralitätsthese technischer Mittel****, betonen (d) neben der* ***Gefahr der ökologischen Katastrophe*** *(e) die* ***Gefährdung des für Seins- bzw. Transzendenz-Ansprüche offenen Menschenwesens****.";* Hervorhebungen DS

[823] Michel Foucault zitiert in Erdmann et al. (1990), S. 33; und im Aufsatz des Autors: Foucault (1990), S. 45

[824] „*Wahnsinn und Gesellschaft: Eine Geschichte des Wahns im Zeitalter der Vernunft*"; „*Überwachen und Strafen: Die Geburt des Gefängnisses*"; „*Sexualität und Wahrheit*"

mit der wir immer wieder zum kritischen Denken herausgefordert sind. Dabei ist er skeptisch, ob wir wirklich jemals mündig werden:

> *„Ich weiß nicht, ob wir jemals mündig werden. Vieles in unserer Erfahrung überzeugt uns, daß das historische Ereignis der Aufklärung uns nicht mündig gemacht hat und daß wir es noch immer nicht sind. Dennoch scheint mir, daß der kritischen Befragung der Gegenwart und unserer selbst, die Kant in einer Reflexion über die Aufklärung formulierte, eine Bedeutung verliehen werden kann.* […]
>
> [Die Aufklärung] *muß als eine Haltung vorgestellt werden, ein Ethos, ein philosophisches Leben, in dem die Kritik dessen, das wir sind, zugleich die historische Analyse der uns gegebenen Grenzen ist und ein Experiment der Möglichkeit ihrer Überschreitung.“*[825]

Grundsätzlich hat sich Foucault in diesem Aufsatz keineswegs von Kant und den Zielen der Aufklärung distanziert. Er hat lediglich betont, dass sie zu keinem Zeitpunkt vollendet war und wir zu allen Zeiten um unsere Mündigkeit im Kontext der jeweiligen Gegenwart kämpfen müssen. Dieser milde und versöhnliche Umgang mit Kants Erbe wurde im Kreise der Anhänger und Gegner Foucaults mit Überraschung aufgenommen[826].

Beim Studium von Foucaults sonstigen Arbeiten (insbesondere der oben genannten Werke[827]) offenbart sich seine Position, wonach in der Moderne, trotz allem Gerede über Humanität und Freiheit, der offene und oftmals erratische Gebrauch von Machtinstrumenten in der vormodernen Zeit durch *„eine heimtückischere und gründlichere Form der Unterdrückung“*[828] ersetzt wurde, in der öffentliche und private Institutionen *„den Menschen überwachen und ‚normalisieren‘“*:

> *„Erstens will Foucault aufzeigen, wie sehr die Überwachung selbst eine Machtausübung ist, die uns oft stärker prägt, als es rohe Gewalt vermag. Zweitens will er uns zeigen, wie eng Wissensansprüche und Machtausübung miteinander verwoben sind, wie sehr jeder vermeintliche Fortschritt in den Humanwissenschaften zu Versuchen geführt hat, Menschen zu kategorisieren und diejenigen zu unterdrücken, die aus der Norm fallen.“*[829]

Genau diese Kategorisierung betreibt die KI (wie u.a. schon von Zuboff, Koenig festgestellt) mit Perfektion, insbesondere in den sozialen Medien und vielen anderen

---

825 Foucault (1990); S. 52f

826 Fleischacker (2018), S. 108: „*These lectures came as a surprise to Foucault's friends and foes alike: the former, because they themselves tended to have an aversion to Kant and felt that figures on the cultural and political left should not look to him for inspiration; and the latter, because they felt that Foucault was misreading or trivializing Kant's point in WE* [Kürzel für: Was ist Aufklärung? Von I. Kant]*, and disingenuous or facetious in claiming to stand in Kant's tradition*“

827 Ich beziehe mich hier insbesondere auf die Analysen von Samuel Fleischacker

828 Fleischacker (2018), S. 108; Übersetzung DS

829 Fleischacker (2018), S. 108f, Übersetzung DS; im Original: “*First, Foucault wants to bring out how much surveillance itself is an exercise of power, shaping who we are often more deeply than brute force can do. Second, he wants to show us how tightly claims to knowledge and exercises of power are interwoven, how much every supposed advance in the human sciences has given rise to attempts to categorize human beings and repress those who fall outside the norm.”*

Internetplattformen. Noch ohne die Möglichkeiten der Digitalisierung zu kennen, beschreibt Foucault die Scheinheiligkeit der Aufklärung:

> *„Foucault bringt die Schattenseiten der Aufklärung ans Licht, die manipulativen, manchmal grausamen Aspekte von Praktiken, die sich als emanzipatorisch darstellen. Und seine Schriften wurden für ihre scharfsinnigen psychologischen und soziologischen Einsichten und für das neue, umfassendere Konzept der Unterdrückung, das sie hervorbringen, gelobt (Foucaults Verständnis der oppressiven Qualität der Überwachung scheint heute immer relevanter zu sein).*“[830]

Zusammenfassend lassen sich die folgenden sieben Punkte zu Foucaults Aufklärungskritik festhalten:

1. Aufklärung ist kein abgeschlossenes Ereignis, sondern eine Haltung, die die Menschen in der Gegenwart und in der Zukunft brauchen, um sich der Mündigkeit zu nähern bzw. sich von der Unmündigkeit zu entfernen.
2. Die Menschen werden in der Moderne nicht weniger unterdrückt, sondern anders, nicht durch zentrale Herrschaftsstrukturen, sondern durch verteilte und dezentrale Strukturen, oftmals „*systematischer und heimtückischer*“[831] als vorher.
3. Überwachung ist ein Macht- und Herrschaftsinstrument, das in der Moderne immer weiter verfeinert wurde.
4. Die Vorgeben von Wissen (im Gegensatz zum Besitz von Wissen) und Machtausübung sind eng miteinander verwoben.
5. Fortschritte in den Humanwissenschaften führen zu ausgeklügelten und verfeinerten „*Kategorisierungen der Menschen*“[832] ...
6. ... und der Unterdrückung derer, die „nicht in die Norm passen“.
7. „Die *Schattenseite der Aufklärung* zeigt sich in *den manipulativen, manchmal grausamen Praktiken, die sich als emanzipatorisch darstellen.“*[833]

Es scheint, dass Shoshana Zuboff drei Jahrzehnte später nahtlos an Foucaults Analyse der Unterdrückung durch Überwachung ansetzt.

---

[830] Fleischacker (2018), S.108; Übersetzung DS; im Original: „*Foucault brings out the seamy side of the Enlightenment, the manipulative, sometimes cruel aspects of practices that represent themselves as emancipatory. And his writings have been hailed for their astute psychological and sociological insight, and for the new, richer conception of oppression they yield (Foucault's understanding of the oppressive quality of surveillance seems ever more relevant today).”*

[831] Fleischacker (2018), S. 108, Übersetzung DS; im Original: „*more insidious and thoroughgoing mode of oppression*“

[832] Ebd, im Original: „*attempts to categorize human beings*“

[833] Wiederholung des obigen Zitats von Fleischacker (2018), S. 109

11.4 KI und Unmündigkeit

An dieser Stelle sei noch einmal wiederholt[834]: Unmündigkeit ist laut Kant das Unvermögen, sich seines Verstandes ohne Leitung eines anderen zu bedienen. Kant hat dies explizit immer im Zusammenhang mit der Freiheit (und Autonomie) verstanden. „*Die Freiheit des Menschen, sich öffentlich ihres Verstandes zu bedienen, ihre Meinung zu vertreten und im Austausch miteinander weiterzuentwickeln, ist die Grundlage der Aufklärung*“[835]. Dafür hat er „*die öffentliche Sphäre neu definiert* [und] *vom Gerichtshof zum Marktplatz weiterentwickelt*“:

> „*War der Mensch zuvor darauf angewiesen, dass ihm in einem öffentlichen Verfahren die Mündigkeit zugesprochen wurde, so erwirbt er sie nun selbst auf dem Forum, dem Marktplatz, in einem wechselseitigen Austausch unter den Bedingungen der Freiheit.*“

Der willens-, handlungs- und gedankenfreie Mensch entwickelt seine Mündigkeit zusammen mit den Mitgliedern seiner Gemeinschaft. Dazu gehört eine weitere wichtige Voraussetzung, die der Erkenntnis:

> „*Mit Hilfe der Freiheit entwickeln die Menschen ihr eigenes Denkvermögen und beginnen, sich selbst und die Welt besser zu verstehen.* ***Kants Mündigkeitsbegriff ist einer, der Erkenntnis unbedingt fordert****. Mündig werde ich, wenn ich mehr und mehr von der Welt erkenne.* […] *Die Aufklärung ist der Prozess, in dem Menschen ihre eigene Freiheit entdecken und behaupten, und schließlich durch diese Freiheit mehr über sich und die Welt erfahren.*“ [Hervorhebung DS]

Für die Prüfung der Hypothese, dass die Künstliche Intelligenz die Mündigkeit des Menschen einschränkt, sind drei Fragen zu untersuchen:

- Trägt die KI dazu bei, dass der Mensch sich seines Verstandes bedienen kann, **ohne Leitung eines anderen**? Anders formuliert: Genießt der Mensch uneingeschränkte Willens-, Handlungs- und Gedankenfreiheit?
- Fördert die KI einen **freien öffentlichen Diskurs**, der dem öffentlichen Gebrauch der Vernunft dient, ohne Bevormundung?
- Unterstützt die KI den Menschen beim **Erkennen der Welt**?

Begegnete man der KI kritiklos und mit blindem Vertrauen, würde man auf alle drei Fragen mit Zustimmung antworten. Bei genauerer Betrachtung kommt man hingegen zu einem anderen Ergebnis.

In vielen Bereichen, in denen die KI eingesetzt wird, verliert der Mensch zunehmend die Gabe, sich selbst zurechtzufinden. Die Beispiele dafür sind vielfältig. Durch die Nutzung von Navigationsgeräten verliert der Mensch mehr und mehr die Fähigkeit, sich mithilfe von Landkarten und Stadtplänen zu orientieren. Ähnliches gilt für viele andere Bereiche,

---

[834] Vgl. Kant (1784b), S. 20

[835] Dieses und folgende Zitate (einschließlich Blockzitate): Villhauer (2009), S. 15

in denen sich KI-gestützte Applikationen oder Plattformen durchsetzen: Ernährungsberatung, Anlageberatung, Partnersuche und Karriereberatung. Das Problem besteht nicht darin, dass diese Unterstützung und das damit verbundene Nudging den Menschen suboptimale Lösungen liefern. Ganz im Gegenteil; für die Gesamtheit der Menschen, die diese Möglichkeiten nutzen, ergibt sich eine objektive Verbesserung des Nutzens. Wenn sich alle Menschen exakt so ernähren, wie es ihnen eine dafür entwickelte KI-Plattform vorschlägt, ergeben sich mit hoher Wahrscheinlichkeit eine höhere Lebenserwartung, geringere Kosten für das Gesundheitssystem und vermutlich auch eine Reduzierung der Umweltbelastung. Wenn sich alle Autofahrer diszipliniert an die Vorgaben ihres KI-gesteuerten und vernetzten Navigationsgerätes halten, stehen weniger Menschen im Stau und es wird auch weniger Energie verbraucht und CO2 erzeugt. Entsprechendes gilt auch für andere bereits existierende und zukünftige Applikationen der Künstlichen Intelligenz. Trotzdem verliert der Mensch an Freiheit, Individualität und Mündigkeit. Wenn Eltern ihren (noch nicht mündigen) Kindern Süßigkeiten und Fast Food vorenthalten, dann tun sie dies zum Wohle ihrer Kinder. Bei erwachsenen Kindern würde man dies als Eingriff in ihre Mündigkeit verurteilen.

Die Entmündigung des Menschen geht allerdings weiter. Ähnlich wie Shoshana Zuboff beklagt Thomas Metzinger in seinem Buch „*Bewusstseinskultur*“[836] die systematische Monetarisierung der menschlichen Aufmerksamkeit in einer „*Aufmerksamkeitsökonomie*“[837] :

> „*Ein Hauptproblem für eine zeitgemäße Bewusstseinskultur ist die neue* ***attention extraction economy*** *[Hervorhebung TM]. Gestützt auf selbstlernende KI und moderne, sich selbst ständig verbessernde Algorithmen, extrahieren sie Aufmerksamkeit aus menschlichen Gehirnen und verwandeln sie in Geld.* [...] *Was in der neuen ‚Aufmerksamkeitsökonomie‘ tatsächlich monetarisiert wird, das ist die Zerstörung von geistiger Autonomie und damit unserer Fähigkeit, die eigene Aufmerksamkeit noch willentlich und selbstbestimmt zu kontrolliere. Aber wir brauchen geistige Autonomie, wenn wir als mündige Bürger unsere Demokratien aufrechterhalten und verteidigen wollen*.“

Die uneingeschränkte Willens-, Handlungs- und Gedankenfreiheit des Menschen wird mit den Methoden der KI unterwandert. Ein selbstbestimmtes Bedienen des eigenen Verstandes ohne Leitung eines anderen wird so immer weniger möglich sein. Es ist unwesentlich, ob es sich beim „Anderen“ um andere Individuen oder andere Technologien handelt.

Einem ähnlichen Paradoxon begegnet man beim öffentlichen Diskurs. Einerseits lässt sich argumentieren, dass noch nie so viele Möglichkeiten bestanden, sich mit Mitbürgern nah und fern auszutauschen zu allen denkbaren Fragen des gesellschaftlichen, politischen, kulturellen und wirtschaftlichen Zusammenlebens. Die dafür geschaffenen

[836] Metzinger (2023)

[837] Dieses und das folgende Blockzitat: Metzinger (2023), S. 70

Diskussionsräume auf den Plattformen der sozialen Medien werden mit KI moderiert. Dort herrscht die Illusion, dass man mit durchschnittlichen und repräsentativen anderen Diskussionsteilnehmern über die Themen diskutiert, die aktuell relevant sind. Das Gegenteil ist der Fall. Dem Teilnehmer werden Themen und Beiträge präsentiert, von denen die KI aus der Auswertung bisheriger Diskussionen weiß, dass man sie offensichtlich in vorherigen Diskussionen verfolgt hat. Es werden inhaltliche Positionen vorgelegt, die andere Diskussionsteilnehmer posten und denen man folgt. Dieser in Bezug auf die Themen- und Teilnehmerauswahl stark eingeschränkte Gedankenaustausch entspricht weder Kants Vorstellungen vom Marktplatz der Ideen noch Habermas‘ Anforderungen an die Diskursethik, die unter anderem einen „*herrschaftsfreien Diskurs*“[838] beinhalten. Der Diskurs in den sozialen Medien bleibt völlig ohne jede Konsequenz für politisches Handeln, Gesetzgebung und das reale Leben der Menschen. Von einem „mündigen“ Bürger kann man in diesem Zusammenhang also nicht sprechen.

Eine weitere Domäne für potenziellen Mündigkeitsverlust ist die Wissenschaft, die ursprünglich am meisten von der Aufklärung profitierte. Die KI kann als Werkzeug in hohem Maße zum Erkenntnisgewinn in den Wissenschaften beitragen. Bedenklich ist allerdings, dass mit den Methoden der KI Korrelationen und Muster erkundet werden, die oftmals die Frage nach kausalen Wirkzusammenhängen offenlässt. Nicht erst seit Aristoteles ist der Mensch auf der Suche nach logischen Zusammenhängen (Deduktion und Induktion) und kausalen Ketten. Er ist auf der Suche nach Hypothesen, die er abduktiv gewinnt (auch mit Hilfe der KI), die er aber immer zu beweisen sucht. Es besteht die Gefahr, dass der Anteil der unbewiesenen wissenschaftlichen Erkenntnisse drastisch zunimmt.

All dies gilt in verstärkter Weise für die generative KI, also zum Beispiel für Large Language Modells wie die Applikation ChatGPT, die Texte produziert, deren Herkunft und Wahrheitsgehalt für den Nutzer und Leser nicht transparent ist. Die Plausibilität derartiger Texte kann nur sehr oberflächlich überprüft werden. Weiterhin wird es bei verstärkter Nutzung derartiger Systeme zukünftig immer schwerer, Texte (und auch Bilder, Videos und sogar Tonaufnahmen) als menschengemacht zu erkennen. Ein Selbstdenken mit eindeutigem Bezug auf menschliche und verifizierbare Quellen wird zunehmend schwierig bis unmöglich.

[838] Begriff, der die Anforderungen an den idealen Diskurs bei Habermas zusammenfasst. Ausführlich dargestellt in Habermas (1981); Wesentliche Elemente des herrschaftsfreien Diskurs sind die Gleichberechtigung der Kommunikationspartner mit gleicher Möglichkeit sich zu äußern und symmetrischer Kommunikation und die Entscheidungsfindung durch den „Zwang des besseren Arguments“

11.5 Zusammenfassung

> „[Die Aufklärung] *muß als eine Haltung vorgestellt werden, ein Ethos, ein philosophisches Leben, in dem die Kritik dessen, das wir sind, zugleich die historische Analyse der uns gegebenen Grenzen ist und ein Experiment der Möglichkeit ihrer Überschreitung.“*[839]
>
> Michel Foucault

Die Analysen und Recherchen in dieser Arbeit dienen zur Klärung der Frage, ob der Menschheit ein selbstverschuldeter Rückfall in die Mündigkeit droht, der dann in letzter Konsequenz viele, nicht zwingend alle, positiven Effekte der Aufklärung negiert.

In diesem Kapitel und in den vier vorherigen Kapiteln konnten diverse Risiken infolge der Einführung und zunehmenden Nutzung Künstlicher Intelligenz identifiziert werden, die in ihrer Gesamtheit wesentliche Sorgen der Aufklärungskritiker des vorherigen Jahrhunderts bestätigen.

Viele der ursprünglichen Argumente der Aufklärungskritik bezogen sich auf Neuerungen in Gesellschaft (Mediengesellschaft), Politik (Faschismus und Totalitarismus), Wirtschaft (Industrieproduktion, Automatisierung) und Technik (Nukleartechnik). Damals war die Künstliche Intelligenz noch gar nicht zu erkennen.

Wie wir festgestellt haben, richtete sich die Kritik gegen bestimmte wiederkehrende Attribute der Entwicklung:

- Die Beschränkung der Vernunft auf den instrumentellen Teil, den Weber die Zweckrationalität nannte, d.h. die Aufgabe der Zwecksetzungskompetenz
- Die zunehmende Ohnmacht des Menschen, die angestoßene Dynamik zu steuern oder gar aufzuhalten
- Die Entwicklung immer perfiderer Instrumente der Überwachung und Machtausübung
- Einschränkungen des ergebnisoffenen öffentlichen Diskurses

Die Argumente der für diese Analyse herangezogenen Kommentatoren und Kritiker – von Heidegger, Horkheimer und Adorno (Frankfurter Schule) über Jonas, Arendt, Foucault und Habermas bis hin zu Zuboff in neuerer Zeit[840] – weisen eine hohe Konsistenz auf, auch wenn sich die gewählten Begriffe und dargestellten Ausprägungen unterscheiden.

Die Künstliche Intelligenz ist kein neues Phänomen, das ein völlig neuartiges Problem schafft. Sie ist vielmehr so etwas wie die Verstärkung einer Entwicklung, die in der modernen Gesellschaft bereits vorher begonnen hat. Der Soziologe Armin Nassehi thematisiert dies in seinem Buch „*Muster. Theorie der digitalen Gesellschaft*“:

---

[839] Foucault (1990), S. 53

[840] Reihenfolge nach Geburtsdaten der Autoren, nicht zwingend nach ihren Veröffentlichungen

> *„Das Bezugsproblem der Digitalisierung ist in der Gesellschaftsstruktur selbst lokalisiert. Während vormoderne Gesellschaften zwar ungeheuer komplexe kulturelle Formen gebildet haben, ist ihre Grundstruktur doch sehr einfach: Alles, buchstäblich alles fügt sich einem Oben-Unten-Schema – soziale Hierarchien, gesellschaftliche Ordnungen, Weltbilder, Taxonomien und auch das Deduktionsprinzip der Logik. Exakt das gilt in der Moderne nicht mehr: Die Verhältnisse werden unübersichtlicher, es etablieren sich unterschiedliche Ordnungsformen nebeneinander, das Differenzierungsprinzip ermöglicht Parallelstrukturen, damit entzieht sich die Struktur der Gesellschaft einer deutlichen Sichtbarkeit."*[841]

Aus Nassehis Sicht bildet die Digitalisierung die zugrundeliegende Gesellschaft ab. Für den Soziologen ergeben sich aus den Möglichkeiten der neuen Technologie ungeahnte Chancen der Erkenntnis für seine Disziplin.

Trotzdem stimmt seine Feststellung, dass die KI in die Gesellschaft insgesamt eingebettet ist. Viele der Gefahren, die in dieser Arbeit thematisiert sind, bestanden schon vorher, dürften sich jedoch mit dieser Technologie drastisch verstärken, und zwar in allen Vertiefungsbereichen: Autonomieverlust des Menschen, Entstehung von Verantwortungslücken bis hin zur Verantwortungslosigkeit, Bedrohung der menschlichen Individualität und Gefährdung der Menschenwürde durch eine „Verdinglichung".

Es ist nicht mehr zu bezweifeln, dass ein umfassender Verlust der Mündigkeit, eine neue selbstverschuldete Entmündigung droht. Eine neue Aufklärung im Sinne des Eingangszitats zu diesem Abschnitt von Michel Foucault ist gefordert: Aufklärung als Haltung.

[841] Nassehi (2019), S. 318

12 Fazit 2: Philosophische Gesamtbeurteilung

In den fünf bisherigen Kapiteln wurden vier kritische Bereiche analysiert, in denen die KI wesentliche Domänen des aufgeklärten Menschen potenziell beeinflusst: seine Autonomie und Freiheit (Kapitel 7), seine Verantwortung bzw. Verantwortlichkeit (Kapitel 8), seine Individualität und sein Subjektsein als Person (Kapitel 9) und seine Menschenwürde (Kapitel 10). Zusammengeführt wurden diese Erkenntnisse anschließend in der Diskussion der Kernthese dieser Arbeit (Kapitel 11), nämlich derjenigen, dass die Menschheit im Rahmen des fortgesetzten und ausgeweiteten KI-Einsatzes Gefahr laufen könnte, wieder in eine selbstverschuldete Unmündigkeit zu geraten und damit zentrale Errungenschaften der Aufklärung wieder aufzugeben.

Korrespondierend mit der gewählten Struktur lassen sich fünf Fokusbereiche der weiteren normativen Begleitung der Weiterentwicklung und des Einsatzes der Künstlichen Intelligenz ableiten:

Beim Einsatz der KI sollten immer die Auswirkungen auf die Autonomie und Freiheit der sie nutzenden oder von ihr beeinflussten Menschen miteinbezogen werden. Der Mensch sollte sich immer darüber im Klaren sein, dass nur er über Autonomie und Urteilskraft verfügt. Die KI ist und bleibt heteronom und birgt die Gefahr, dass sie menschliches Urteilen und Handeln determiniert und damit heteronom macht. Sobald Menschen Tätigkeiten an Maschinen delegieren, die sie (leider) gar zu oft „autonom“ nennen, besteht das Risiko der Einbuße menschlicher Autonomie; im schlimmsten Fall ist dieser Verlust irreversibel. Der Autonomieverlust des Menschen geschieht freiwillig und findet schleichend statt. Jeder einzelne Schritt der Preisgabe eigener Gestaltungs-, Willens- und Handlungsfreiheiten mag unwesentlich und vernachlässigbar erscheinen. Erst in der Summe werden die aufgegebenen Freiheiten spürbar.

Menschen sind für ihr Tun verantwortlich. Verantwortung und Freiheit bedingen sich gegenseitig. Mit dem Einsatz der KI können Bereiche entstehen, in denen die Zuweisung von Verantwortung eingeschränkt oder gar unmöglich wird. Dieses Privileg hatte bisher nur die Natur. Die KI ist und bleibt ein Werkzeug des Menschen. Verantwortlichkeiten müssen in entsprechender Weise ausgestaltet werden. Das leichtfertige Delegieren von Verantwortlichkeiten bei Abläufen, die Entscheidungen ähneln, an eine Technologie, die keine Verantwortung übernehmen kann, ist verantwortungslos. Im Sinne der Verantwortungsethik von Hans Jonas müssen die Nah- und die Fernwirkung räumlich und zeitlich im Blick behalten werden. Es ist ein Diskurs darüber zu führen, ob den zukünftigen Generationen eine Welt hinterlassen wird, in der die Autonomie des Menschen durch den umfassenden Einsatz der KI eingeschränkt ist.

Eine der großen Errungenschaften der letzten Jahrhunderte, nicht erst seit der Aufklärung, sondern schon seit der Renaissance, ist das Verständnis der menschlichen Individualität,

des Individuums oder der Person. Mit den bisweilen subtilen Methoden der KI kann das Einzigartige der einzelnen Person immer mehr eingeschränkt werden. Individualität bedeutet auch, dass der Mensch sich widersprüchlich zum Erwartbaren verhält oder zumindest so verhalten kann, dass sein Verhalten *nicht* vorhersagbar ist. Der Schutz des Menschen als Individuum muss deshalb bei der Ausgestaltung der KI immer bedacht werden.

Die KI kann die Würde des Menschen über verschiedene Hebel einschränken. Einige wurden oben genannt: Freiheitsverlust, Verlust der (Eigen-)Verantwortlichkeit und der Verlust der Individualität. Wichtig ist der Gedanke, dass der Mensch in seinem Subjektsein nicht eingeschränkt werden darf. Die KI ist eine Technik, die selbst niemals Subjekt sein kann und darf und die als Objekt nur mit Objekten umgehen kann. Den Menschen jedoch zum Objekt zu machen, damit er zur KI passt, wäre in der Tat ein gravierender Eingriff in die Menschenwürde. Die Unantastbarkeit der Menschenwürde bedeutet in diesem Sinne auch, dass das Subjektsein des Menschen unantastbar ist.

Der Leitgedanke der Aufklärung ist die „Herausführung des Menschen aus der selbstverschuldeten Unmündigkeit“. Der mündige Mensch denkt selbst, er ist autonom, verantwortlich für sein Tun, er ist eine Person mit Individualität und besitzt als menschliches Subjekt eine Menschenwürde, die nicht nur nach dem deutschen Grundgesetz unantastbar ist. Viele Philosophen der letzten zwei Jahrhunderte beklagen die sukzessive Einschränkung der Mündigkeit des Menschen in der Moderne. Mit den Methoden der KI droht eine weitere Beschleunigung dieser Entwicklung, die wir als Menschen nicht hinnehmen sollten.

13 Praktische Umsetzung der ethischen Leitlinien in Fallstudien

In den folgenden Abschnitten sollen die formulierten ethischen Leitlinien und die damit verbundenen Überlegungen auf einige exemplarische Einsatzfelder der KI angewandt werden. Die aus den Analysen und Betrachtungen dieser Arbeit resultierenden Leitlinien haben den Charakter von „Leitplanken" und nicht zwingend von „Straßensperren". Ein abschließendes Urteil erfordert allerdings nicht nur einen Blick auf die „Leitplanken", sondern im Sinne eines Konsequentialismus auch auf das gesamte Feld der kurz-, mittel- und langfristigen Folgen des Technologieeinsatzes.

Den Einsatz von KI auf dem Schlachtfeld kann man nicht ohne moralphilosophische Betrachtungen des Phänomens Krieg insgesamt diskutieren, unabhängig von den eingesetzten Technologien. Bei der Verwendung von KI in der medizinischen Pflege dürfen der Pflegenotstand und Missbrauch von Patienten durch überlastete Pflegekräfte nicht unberücksichtigt bleiben. Ebensowenig sollte man das sogenannte „autonome Fahren" ohne Deliberation zu menschlichen Fehlern im Straßenverkehr betrachten. Wie von Wilhelm Vossenkuhl angeregt, erfordert das abschließende Urteil in Anlehnung an Kants Modell der Urteilskraft das freie Spiel von Vernunft und Willen[842].

Da viele der einzelnen KI-Anwendungen isoliert nur einen kleinen Bereich der menschlichen Lebenswelt abdecken, könnte man versucht sein zu argumentieren, die einzelnen Autonomieverluste, Einschränkungen von Individualität und Menschenwürde seien nur minimal und nicht existenziell oder umfassend und daher tolerierbar. Deswegen empfiehlt sich die Extrapolation auf das gesamte Bild und die Anwendung der Universalisierungsformel des kategorischen Imperativs: „*Handle nur nach derjenigen Maxime, durch die du zugleich wollen kannst, dass sie ein allgemeines Gesetz werde*."[843]. In jedem Fall sollte auch die erweiterte Form des kategorischen Imperativs nach Jonas mit in Betracht gezogen werden: „*Handle so, dass die Wirkungen deiner Handlung verträglich sind mit der Permanenz echten menschlichen Lebens auf Erden."*[844]

In dieser Arbeit sollen die einzelnen Argumente und Betrachtungsperspektiven aufgezeigt werden, allerdings kann kein finales Urteil erfolgen. Das erfordert einen breiten öffentlichen Diskurs von Beteiligten und (potenziell) Betroffenen mit Blick auf die Auswirkungen in der Nähe und in der Ferne zeitlich und räumlich.

Die Struktur der Vorstellung der Fallstudien ist harmonisiert:

- Vorstellung des Anwendungsbereichs

[842] Vgl. Vossenkuhl (2021), S. 331

[843] Immanuel Kant: AA IV, 421

[844] Jonas (1979), S. 36

- Ziele und erwartete Effekte
- Bewertung im gewählten moralphilosophischen Rahmen
    - Autonomie/Freiheit
    - Verantwortung
    - Individualität/Individuum/Person
    - Menschenwürde
    - Mündigkeit
    - Sonstige moralphilosophische Argumente

Es erfolgt eine Bewertung gemäß den kritischen Fragen in der unten dargestellten Bewertungsmatrix.

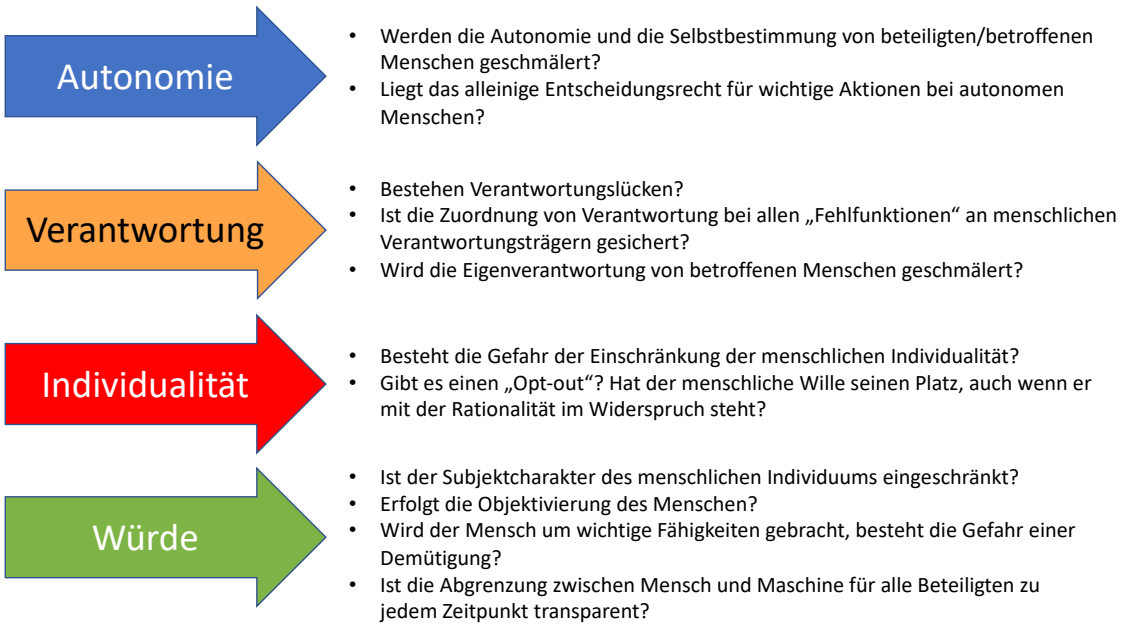

Abbildung 8: Bewertungsrahmen

13.1 Erste Fallstudie: KI in der Pflege

> *"An old lady sits alone in her sheltered accommodation stroking her pet robot seal. She has not had any human visitors for days. A humanoid robot enters the room, delivers a tray of food, and leaves after attempting some conversation about the weather, and encouraging her to eat it all up. The old lady sighs, and reluctantly complies with the robot's suggestions. When she finishes eating, she goes back to stroking the pet robot seal: `At least you give my life some meaning´ she says, as the robot seal blinks at her with its big eyes and makes seal-like sounds in response to her ministrations."*[845]
>
> Amanda Sharkey

Ein wichtiges zukünftiges Anwendungsfeld der KI (und Robotik) liegt im Pflegebereich, spezifisch in der Altenpflege. Aufgrund der demografischen Entwicklung und der Fortschritte in der Medizin nimmt die Nachfrage nach Pflegeleistungen in der Welt insgesamt und insbesondere in Deutschland deutlich zu. In Deutschland lebten 2019 4,1 Millionen pflegebedürftige Menschen, von denen 3,3 Millionen zuhause versorgt wurden und 818.000 stationär/teilstationär untergebracht waren[846]. Nach verschiedenen Modellen wird der Pflegebedarf in allen Pflegegraden in den nächsten Jahrzehnten noch weiter zunehmen. Die Versorgung wird von 1,17 Millionen Pflegekräften (Wert für 2020) in der ambulanten Pflege[847] und in der stationären/teilstationären Pflege zusammen mit vier bis fünf Millionen pflegenden Angehörigen[848] übernommen. Seit Jahren spricht man vom Pflegenotstand, der sich aus einem Mangel an Pflegekräften ergibt und sich mit hoher Wahrscheinlichkeit noch weiter verschärfen wird.

Die sich vergrößernde Lücke zwischen der Nachfrage und dem Angebot an qualifizierten Pflegekräften kann – so die verbreitete Hoffnung – mit modernen Techniken wie Robotik und Künstlicher Intelligenz geschlossen werden. Einerseits können Menschen aufgrund des Einsatzes von Servicerobotern *„länger selbstständig in ihrem vertrauten häuslichen Umfeld leben"*[849]. Andererseits könnten *„sowohl pflegende Angehörige als auch Pflegekräfte im ambulanten und stationären Bereich bei körperlich anstrengenden oder auch repetitiven und zeitraubenden Arbeiten durch robotische Systeme unterstützt werden"*[850].

Die technische Entwicklung für derartige Systeme steht erst am Anfang. Immerhin zeichnet sich ein stark wachsender Markt für Roboter im Gesundheitswesen ab. Bis zum Jahr

---

[845] Zitiert aus Sharkey (2014), S. 63

[846] Quelle: Statistisches Bundesamt, https://www.destatis.de/DE/Themen/Gesellschaft-Umwelt/Gesundheit/Pflege/Tabellen/pflegebeduerftige-pflegestufe.html;jsessionid=27E136C04291D68B7AFFFA32B8F3C294.live721

[847] Quelle: Statistisches Bundesamt, https://www-genesis.destatis.de/genesis/online?sequenz=tabelleErgebnis&selectionname=23621-0001&zeitscheiben=10#abreadcrumb

[848] Deutscher Ethikrat (2020), S. 7

[849] Deutscher Ethikrat (2020), S. 8

[850] Deutscher Ethikrat (2020), S. 8

2025 wird ein Marktvolumen von 50 Mrd. US-Dollar erwartet[851]. In einer ersten Strukturierung unterscheidet Denis Pijetlovic drei Typen[852]:

- Roboter für medizinische Versorgung
- Roboter für die Krankenpflege
- Roboter für die Heimpflege

Im Folgenden soll dagegen eine von Amanda und Noel Sharkey und vom Deutschen Ethikrat aufgegriffene Differenzierung[853] der Robotik in der Pflege hinsichtlich der erbrachten Funktionen verwendet werden:

> *"The three main ways in which robots might be used in elder care are: (1) to assist the elderly and/or their carers in daily tasks; (2) to help monitor their behaviour and health; and (3) to provide companionship."*[854]

Assistenz, Monitoring und Begleitung sind die drei Kategorien des Einsatzes von KI in der Pflege. Assistenzroboter sind Systeme, „*die Pflegende und Gepflegte bei ihren alltäglichen Aufgaben unterstützen"*. Dazu gehören „*Roboter, die bei der Nahrungsaufnahme oder der Körperhygiene unterstützen*, „*intelligente Transportsysteme, die Medikamente oder Wäscheutensilien bereithalten und Hebehilfen*, die bei *körperlich anspruchsvollen Tätigkeiten wie beim Höherlagern von Patienten und beim Transfer aus dem Bett heraus assistieren"*[855]. Auch gehören fahrerlose Kleinfahrzeuge für die Patienten zu dieser Kategorie, ebenso wie Exoskelette[856]. Beim Monitoring, international auch „*electronic care surveillance"* genannt, handelt es sich um Systeme, mit denen regelmäßig medizinische Körperfunktionen erfasst und ausgewertet werden, die z.B. „*Menschen mit Gedächtnisdefiziten an alltägliche Tätigkeiten (Medikamenteneinnahme, Essens- und Flüssigkeitsaufnahme, Toilettengänge etc.) erinnern"*. Für die Unterstützung sozialer Interaktionen können Begleitroboter eingesetzt werden, die emotionale und kommunikative Bedürfnisse befriedigen. Bereitgestellt werden z.B. „*Roboter in Tiergestalt* (*ähneln Robben, Katzen oder Hunden*)", die auf Berührungen und Geräusche mit spezifischen Lauten und Bewegungen reagieren und z.T. auch über Spracherkennung verfügen.

All diese dargestellten Systeme setzen unterschiedliche Technologien der Künstlichen Intelligenz ein; dabei sind sie nicht alle gleichermaßen relevant für die spezifischen Fragestellungen der übergreifenden moralphilosophischen Untersuchung entlang von

---

851 Pijetlovic (2020), S. 62f

852 Pijetlovic (2020), S. 62f

853 Beschreibung und Erläuterung der differenzierten Funktionen zitiert aus: Deutscher Ethikrat (2020), S. 16f

854 Sharkey Sharkey (2012b), S. 27

855 Dieses und die folgenden Zitate im Absatz: Deutscher Ethikrat (2020), S. 16f

856 „*Als ein **Exoskelett** (griechisch von exo „außen" und skeletos „ausgetrockneter Körper") bezeichnet man ein mechanisches Gerüst, welches dem Menschen am Außenkörper angebracht wird. Die Idee dahinter ist, den Körper mit einem Stützkorsett zu unterstützen."*, Quelle: https://www.luttermann.de/leistungen/orthopaedietechnik/exoskelett/

Autonomie, Verantwortung, Individualität und Menschenwürde. So sind die Systeme aus der ersten Kategorie der Assistenzsysteme, die entweder im Hintergrund ablaufen (Transportroboter) oder eindeutig Werkzeuge und Hilfsmittel für das Pflegepersonal sind, für die Belange dieser Arbeit nicht bedeutsam. Anders sieht es bei Monitoring Systemen aus, die immer präzisere und umfänglichere Daten des Patienten erfassen und vielfältig weiterverwenden. Auch die Begleitroboter, die zukünftig komplett andere Formen annehmen können als Kuscheltiere mit Spracherkennung und Reaktion auf Streicheln und Geräusche, sind aus ethischer Sicht von großer Bedeutung.

13.1.1 KI in der Pflege: Autonomie und Freiheit

In Bezug auf die erste Bewertungsdimension „Autonomie und Freiheit" ist es bei der KI-Anwendung in der Pflege besonders wichtig, stringent zu unterscheiden, nämlich zwischen der menschlichen Seite (pflegebedürftige Person und Pfleger) einerseits und technischen Hilfsmitteln (Roboter und KI-Systeme) andererseits . Wie oben im entsprechenden Kapitel dargelegt wurde, besteht ein „*kategorialer Unterschied*"[857], der zu keinem Zeitpunkt „*verwischt werden darf*".

Daraus ergeben sich drei spezifische Schlussfolgerungen: Erstens können nur Patienten, Angehörige mit Vollmacht, Pfleger und Ärzte autonome Entscheidungen treffen. Zweitens muss jeder Einschränkung der Autonomie und Freiheit des Patienten von ihm selbst oder einer bevollmächtigten Person zugestimmt werden. Die Vor- und Nachteile der Einschränkung (aus Sicht des Patientenwohls) müssen sorgfältig abgewogen werden. Bei der Vergitterung eines Bettes zum Schutz vor dem Herausfallen ist dies offensichtlich, beim Einsatz eines Rollators oder eines Rollstuhls, der in regelmäßigen Abständen Informationen über seine Position an einen zentralen Computer übermittelt und ggf. auch das Verlassen eines vorgegebenen Radius verweigert, ist dies erst auf den zweiten Blick klar. Drittens darf dem Patienten im Sinne einer anthropomorphen Verwechselung niemals suggeriert werden, dass er es mit einem Menschen zu tun habe, wenn dies nicht der Fall ist. Dies gilt insbesondere für Patienten mit eingeschränkten kognitiven Fähigkeiten, z.B. Demenzkranken. In Bezug auf den Einsatz von Robotern in Tiergestalt warnt der Ethikrat vor der „*‚Infantilisierung' älterer Menschen*"[858].

Auch hier ist die Autonomie, in anderen Worten die Selbstbestimmung des Patienten, im Fokus. Der Deutsche Ethikrat schreibt dazu:

> *„Selbstbestimmung bezieht sich auf die Fähigkeit der Person, sich im Denken und Handeln an eigenen Überzeugungen, Wünschen und Präferenzen zu orientieren, bzw. auf die Freiheit, selbstgewählte Ziele und Pläne eigenverantwortlich zu verfolgen. Hierzu gehört zunächst*

[857] Vergl. auch: Deutscher Ethikrat (2020), S. 11; In der Schrift des Ethikrats zur Robotik in der Pflege bemüht man hier das Stichwort des „Anthropomorphismus"

[858] Deutscher Ethikrat (2020), S. 19

> *das Recht, über die Technikausstattung einer Einrichtung vor Aufnahme informiert zu werden, und zwar in angemessener Form und gegebenenfalls mit der individuell notwendigen Unterstützung, sodass die Technik mit all ihren für die Person relevanten Konsequenzen ihres Einsatzes verstanden wird."*[859]

Zu gewährleisten ist nicht nur, dass der Patient über die vielfältigen technischen Funktionen informiert wird, sondern auch, dass er bestimmte Funktionen und empfundene Einschränkungen seiner eigenen Autonomie jederzeit ablehnen kann. Letzteres ist in der Tat aus Sicht des Trägers und Betreibers einer Einrichtung problematisch. Zu viele „Opt-Outs" erhöhen die Komplexität und führen zu Kostensteigerungen und Effizienzverlusten.

13.1.2 KI in der Pflege: Verantwortung

Für die Verantwortung gilt die gleiche Abgrenzung wie für die „Autonomie und Freiheit". Nur Personen können Verantwortung übernehmen, niemals Maschinen. Bei der Ausgestaltung von Robotik-Systemen für die Pflege ist es wichtig, dass die Eigenverantwortung des Patienten respektiert und gestärkt wird.

Weiterhin sind die Systeme so zu gestalten, dass alle Parameterveränderungen der medizinischen Betreuung des Patienten in letzter Konsequenz vom Pfleger und vom zuständigen Arzt verantwortlich entschieden werden. Es ist vorstellbar, dass die KI eines Pflegeroboters, der gemäß seiner Monitoringfunktion in Kombination mit vielfältigen Diagnostikfunktionen schon sehr bald auch die Dosierung und Gabe von Medikamenten steuert, und zwar am Arzt und Pfleger vorbei, und so umfassende Veränderungen der Therapie vornimmt. Dies ist explizit auszuschließen.

13.1.3 KI in der Pflege: Respekt des Individuums

Medizinische Einrichtungen wie Krankenhäuser und Pflegeheime sowie ambulante Pflegedienste stehen unter einem hohen Kostendruck und sind deshalb immer wieder gefordert, ihre Prozesse und Abläufe zu standardisieren. Individuelle Dienstleistungen und Betreuungsangebote waren bereits in der Vergangenheit außerhalb von Privatleistungen selten zu erbringen. Auch wenn die KI-Technologie flexibel ausgestaltet werden kann, besteht die Gefahr, dass die Betreuung aus Sicht des Patienten noch stärker vereinheitlicht und standardisiert wird und „passend für die KI und Robotik" gemacht. Damit wird noch weniger Rücksicht auf die individuellen Bedürfnisse des Patienten genommen.

Ein wichtiges Element der Individualität des Menschen ist die Aufrechterhaltung der spezifischen sozialen Kontakte. Das Risiko einer sozialen Isolation aufgrund von altersbedingten Einschränkungen nimmt generell zu. Einerseits können die neuen Technologien

[859] Deutscher Ethikrat, S. 32

einer Vereinsamung entgegensteuern. Andererseits kann die Nutzung der Technologien die eigene Identität und die eigene Außenwirkung negativ beeinflussen, was wiederum zu einer weiteren Reduzierung von sozialen Kontakten führen kann. Auch ist nicht auszuschließen, „dass *sich Bezugspersonen in geringerem Maße verantwortlich oder Besuche als weniger notwendig erachten"*[860], insbesondere wenn die Kombination aus Monitoring- und Assistenzsystemen technisch gut funktioniert und es dem Gepflegten nach dem ersten oberflächlichen Eindruck gut geht.

13.1.4 KI in der Pflege: Menschenwürde

Dreh- und Angelpunkt aller moralphilosophischen, insbesondere menschenwürdeorientierten Betrachtungen zum Einsatz der KI in der Pflege ist – wie schon übergreifend für alle Anwendungen angeregt – die Frage danach, ob der Mensch dadurch sein Subjektsein verliert und noch mehr oder gar ausschließlich zum Objekt wird. Die bisherigen Überlegungen und die gewählten Anwendungsbeispiele zur Autonomie, Verantwortung und Individualität bestätigen diese Perspektive.

Die Beurteilung der Auswirkungen des Einsatzes von Robotik und KI in der Altenpflege ist komplex und fällt von Person zu Person unterschiedlich aus. Hierzu ein Beispiel: Einige Personen empfinden die Tatsache, dass sie überhaupt Unterstützung bei der Körperpflege benötigen, als entwürdigend auch dem Pflegepersonal gegenüber, und begrüßen es daher, dass sie diese Unterstützung durch eine Maschine erfahren. Andere Personen mögen dies genau umgekehrt sehen.

Es besteht ein Paradoxon: Je mehr Sicherheit und Komfort ein Robotersystem bieten kann, desto grösser ist auch die Gefahr einer Verletzung der Menschenwürde[861].

Die britische Wissenschaftlerin Amanda Sharkey hat untersucht, wie sich der Einsatz von Robotern in der Altenpflege auf die Menschenwürde auswirkt, und dabei den in Abschnitt 10.6 vorgestellten Fähigkeitenansatz von Martha Nussbaum verwendet:

> *"The capability approach (CA) is introduced as a different but tangible account of what it means to live worthy of human dignity. It is used here as a framework for the assessment of the possible effects of eldercare robots on human dignity. The CA enables the identification of circumstances in which robots could enhance dignity by expanding the set of capabilities that are accessible to frail older people. At the same time, it is also possible within its framework to identify ways in which robots could have a negative impact, by impeding the access of older people to essential capabilities."*[862]

---

[860] Deutscher Ethikrat (2020), S. 36

[861] Vgl. Sharkey Sharkey (2012a); S. 278: "*It seems almost paradoxical that the more safety the robots provide, the more their use may breach human rights*."; Anmerkung: im Angelsächsischen werden Menschenrechte und Menschenwürde etwas anders voneinander abgegrenzt als im deutschsprachigen Raum

[862] Sharkey (2014), S. 63

Sharkey geht auf die bekannten und von vielen Philosophen nicht gelösten Schwierigkeiten des Umganges mit dem Begriff der Menschenwürde ein, insbesondere aus Sicht einer anwendungsorientierten Praktikerin. Der Fähigkeitenansatz stellt alle Menschen gleich, unabhängig vom Ausmaß ihrer Vorerkrankungen oder Einschränkungen. Die konsequente Nutzung des Ansatzes über ein breites Spektrum von Anwendungen aus den drei Bereichen Assistenz, Monitoring und Begleitung schafft Transparenz über die Vorteile und auch einige Risiken. Freilich stellt sie auch fest, dass der Bewertungsrahmen für den spezifischen Einsatz in der Altenpflege mit der starren Vorgabe der zehn Fähigkeiten eine nötige Flexibilität vermissen lässt und vor allem einige wichtige Aspekte, die Scham und Privatheit betreffen, nicht berücksichtigen. Die größte Herausforderung der neuen Technologien für die Menschenwürde wird auch mit dem Fähigkeitenansatz nicht angemessen erkannt: Die Bedeutung von Interaktion mit echten Menschen in der Altenpflege, wie es auch im Eingangszitat dieses Abschnitts zum Ausdruck kommt. Viele einzelne Technologien in den Bereichen Assistenz, Monitoring und Begleitung sind aus Sicht des Gepflegten, der Pfleger und der Pflegeeinrichtung durchaus berechtigt. Mit dem Fähigkeitenansatz oder anderen Richtlinien gelangt man schnell zu positiven Einschätzungen. Trotzdem führen sie alle zusammen in der Extrapolation zu einem Umfeld, in dem die Gepflegten es immer weniger mit echten Menschen zu tun haben. Roboter reinigen ihre Zimmer und Sanitärbereiche, wechseln die Bettwäsche, bringen ihnen die Speisen und Medikamente und überwachen deren Einnahme, Roboter fahren sie in den Park und zurück und verhindern den „Ausflug“ über gesetzte Grenzen hinaus. Roboter lesen ihnen vor der Nachtruhe Geschichten vor. Die vermeintliche Lückenlosigkeit von Assistenz, Monitoring (oder sollte man „Überwachung“ sagen?) und Begleitung hat möglicherweise und nachvollziehbar zur Folge, dass sich Angehörige und Freunde zurückziehen.

Die isolierte Beurteilung der ethischen Verträglichkeit einzelner Technologien und Funktionalitäten kommt zu positiven oder zumindest neutralen (unauffälligen) Schlüssen. Im Sinne der Verantwortungsethik ist hingegen zusätzlich eine Gesamtsicht auf die kombinierten Auswirkungen aller eingesetzten Technologien auf die Würde des Patienten erforderlich[863]. Erst aus einer derartigen holistischen Perspektive sind die wahren Effekte erkennbar und ergibt sich insgesamt eine kritische Beurteilung.

[863] Vgl. Decker (2002), S. 113

13.2 Zweite Fallstudie: KI in militärischen Waffensystemen

> *"We are now confronted with a new case: death by algorithm. This challenges many of the assumptions engrained in the ethical, legal, religious, and other normative systems of the world."*[864]
>
> Christof Heyns, Ehemaliger Sonderberichterstatter der Vereinten Nationen über außergerichtliche, summarische oder willkürliche Hinrichtungen, 2010-2016

Es überrascht nicht, dass sich neben vielen Branchen der Industrie, den Wissenschaften, den Medien und dem Gesundheitswesen auch das Militär für die Nutzungsmöglichkeiten der Künstlichen Intelligenz interessiert und das schon seit einiger Zeit. In Büchern und Kinofilmen[865] wurden viele Ideen der Nutzung von KI in sogenannten Killerrobotern schon seit Jahrzehnten dargestellt.

In der Realität der Armeen und ihrer Lieferanten weltweit läuft der Einzug der KI deutlich weniger spektakulär und bildgewaltig ab, jedoch nicht weniger transformativ und folgenreich. Sogenannte autonome Waffensysteme („Killer Robots") sind aus mehreren Gründen attraktiv für das Militär[866]. Sie reduzieren das Risiko für Leib und Leben der eigenen Soldaten und verringern potentiell die politischen Kosten eines Krieges. Langfristig kann der Einsatz derartiger Systeme auch finanziell günstiger als der Einsatz menschlicher Soldaten sein. Schließlich benötigen sie kein Gehalt, keine Pension, Lebensversicherung, Unterkunft, Verpflegung oder Gesundheitsversorgung. Auch können sie deutlich leistungsfähiger als Menschen oder von Menschen gesteuerte Systeme sein, insbesondere in Bezug auf Geschwindigkeit, Genauigkeit und Ausdauer.

Militärbündnisse (wie z.B. die NATO), Großmächte, Demokraten und Autokraten überschlagen sich mit Veröffentlichungen, Pressemitteilungen und Interviews, in denen sie die strategische Bedeutung der KI für die Armee der Zukunft beschwören. So schrieb die NATO in ihrem Review im Oktober 2021[867]:

> *„An Artificial Intelligence Strategy for NATO*
>
> *One does not have to look far to see how Artificial Intelligence (AI) – the ability of machines to perform tasks that typically require human intelligence – is transforming the international security environment in which NATO operates. Due to its cross-cutting nature, AI will pose a broad set of international security challenges, affecting both traditional military capabilities and the realm of hybrid threats, and will likewise provide new opportunities to respond to them. AI will have an impact on all of NATO's core tasks of collective defence, crisis management, and cooperative security."*

---

[864] Heyns (2017), S. 47f

[865] Einige Beispiele: Terminator 1-4 (1984, 1991, 2003, 2009), Cyborg 1-3 (1989, 1993, 1994), RoboCop 1-3 (1987, 1993, 1994), Universal Soldier 1-3 (1992, 1999, 2009), Blade Runner (1982), I Robot (2004), Westworld (1973)

[866] Folgende Punkte sehr stark angelehnt an Müller (2016), S. 70 (4/16)

[867] Quelle: https://www.nato.int/docu/review/articles/2021/10/25/an-artificial-intelligence-strategy-for-nato/index.html, gesichtet am 25.8.2022

Dies deckt sich auch mit Bekanntmachungen der Verteidigungsministerien und Armeen einzelner NATO-Mitgliedsländer. Human Rights Watch zitiert eine Verlautbarung der US Air Force, die davon ausgeht, dass die Fähigkeiten der KI („machine capabilities“) bis 2030 eine Qualität erreicht haben werden, so dass Menschen die schwächste Komponente in einer Vielzahl von Systemen und Prozessen werden[868], nach welchen Kriterien auch immer.

Auch die Kontrahenten der NATO haben die Bedeutung der KI für ihre strategischen Interessen erkannt. Der russische Präsident Wladmir Putin erklärte im September 2017, dass derjenige, der „*einen Durchbruch im Bereich Künstliche Intelligenz erreicht*, *damit die Welt beherrschen* könne“[869]. Ohne Zweifel denkt er dabei insbesondere an die Bedeutung dieser Technologie für das Militär. Auch und insbesondere steht China da nicht zurück. Der amerikanische Think Tank Brookings hat dazu im April 2020 einen Bericht veröffentlicht[870]:

> *“As the Chinese People’s Liberation Army (PLA) seeks to become a “world-class military,” its progress in advanced weapons systems continues to provoke intense concern from its neighbors and competitors. The Chinese military and China’s defense industry have been pursuing significant investments in robotics, swarming, and other applications of artificial intelligence (AI) and machine learning (ML). Thus far, advances in weapons systems described or advertised as “autonomous” (自主) or “intelligentized” (智能化) have built upon existing strengths in the research and development of unmanned (无人) systems and missile technology. While difficult to evaluate the sophistication of these emerging capabilities, this initial analysis concentrates on indicators of progress in weapons systems that may possess a range of levels of autonomy.”*

Auch viele weitere Länder mit geopolitischen Ambitionen beschäftigen sich intensiv mit der Entwicklung von KI-basierten Waffensystemen, darunter Israel, Iran und Indien. Letzten Endes müssen sich alle anderen Nationen schon aus Gründen der Selbstverteidigung mit der KI auseinandersetzen.

Henry Kissinger (zusammen mit Eric Schmidt und Daniel Huttenlocher) warnt in seinem Buch „*The Age of AI – And our Human Future*“ vor den Folgen einer unüberlegten Nutzung der KI im militärischen Bereich und fordert eine ethische Begleitung des Prozesses:

> *“Due to the dual-use character of most AI technologies, we have a duty to our society to remain at the forefront of research and development. But this will equally oblige us to understand the limits. If a crisis comes, it will be too late to begin discussing these issues. Once employed in a military conflict, the technology’s speed all but ensures that it will impose*

[868] Zitiert in IHRC (2012), S. 8

[869] Quellen: https://www.heise.de/newsticker/meldung/Putin-Wer-bei-KI-in-Fuehrung-geht-wird-die-Welt-beherrschen-3821332.html, https://www.sueddeutsche.de/digital/kuenstliche-intelligenz-militaer-waffen-kriegsfuehrung-muenchner-sicherheitskonferenz-1.4791986-2, gesichtet am 25.8.2022

[870] Quelle: https://www.brookings.edu/research/ai-weapons-in-chinas-military-innovation/, gesichtet am 25.8.2022

> *results at a pace faster than diplomacy can unfold. A discussion of cyber and AI weapons among major powers must be undertaken, if only to develop a common vocabulary of strategic concepts and some sense of one another's redlines. The will to achieve restraint on the most destructive capabilities must not wait for tragedy to arise. As humanity sets out to compete in the creation of new, evolving, and intelligent weapons, history will not forgive a failure to attempt to set limits.* ***In the era of artificial intelligence, the enduring quest for national advantage must be informed by an ethic of human preservation."***[871] [Hervorhebung DS]

Vor der weiteren Vertiefung soll an dieser Stelle eine Klassifizierung vorgestellt werden, die für Folgediskussionen bedeutsam ist. Das US Department of Defense (DoD) hat schon 2003[872] in einem Papier "Unmanned Effects (UFX): Taking the Human Out of the Loop" drei Kategorien von Waffensystemen unterschieden: „*human-in-the loop*", "*human-on-the-loop*" und "*human-out-of-the-loop*". Bei der ersten Kategorie handelt es sich um ferngesteuerte Waffensysteme. Ein Soldat oder Bediener steuert alle Bewegungen und Aktionen einer ferngesteuerten Waffe, z.B. einer Drohne. Beim zweiten Modell (human-on-the-loop) wird das System und dessen „Verhalten" programmiert und während der Ausführung überwacht. Wie diese Überwachung aussehen kann und welche Möglichkeiten des Eingriffs tatsächlich bestehen, wurde damals offengelassen. Die dritte Kategorie („human-out-of the-loop") sieht keine Rolle des Menschen beim Einsatz mehr vor. Dabei handelt es sich um ein vollautonomes Waffensystem. In ihrer offiziellen Sprache weisen die allermeisten westlichen Armeen die Arbeit an vollautonomen Waffensystemen weit von sich und unterstreichen die „*Verantwortbarkeit als systemtechnisches Prinzip*"[873]:

> *„Der Gedanke der verantworteten Nutzung digitaler Technologien realisiert eine gedankliche Verbindung zwischen dem bewussten Handeln in der Menschenwelt und den automatisch ablaufenden Prozessen in der Welt der Algorithmen."*

Trotzdem bleibt die Tür einen Spalt breit für „human-out-of-the-loop" geöffnet:

> *„In der aktuellen Debatte um die wehrtechnische Nutzung digitaler Technologien ist der Verantwortungsbegriff fundamentaler als das Konzept der meaningful human control, auf das in zahlreichen Dokumenten Bezug genommen wird. Denn auch der Einsatz technisch autonomer Systeme, die nach der Entscheidung, sie einzusetzen, vor der erzielten Wirkung nicht mehr von Menschen kontrollierbar sind,* ***können demnach unter bestimmten, klar abzugrenzenden Bedingungen verantwortbar sein****."*[874] [Hervorhebung DS]

Es drängt sich jedoch der Eindruck auf, dass in Bezug auf die normative Ausgestaltung des KI-Einsatzes beim Militär das vielbeschworene „Kind bereits in den Brunnen gefallen ist".

---

871 Kissinger et al. (2021), S. 176f

872 Amoroso (2020), S.7f

873 W. Koch (2020), S. 37

874 W. Koch (2020), S. 38f

Im letzten Jahrzehnt wurden eine Reihe von weltweiten Initiativen zur Regulierung von „Autonomous Weapons Systems“ (AWS) und Killer Robots[875] angestossen:

- Im Jahr 2013 schloss sich eine Koalition von 72 Nichtregierungsorganisationen (NGOs) aus 31 Ländern zusammen und startete eine „campaign to stop killer robots“.
- 2015 wurde ein offener Brief von mehr als 26.000 Unterzeichnern, darunter 4.000 Forscher aus den Bereichen KI und Robotics, veröffentlicht; sie forderten einen „ban on offensive autonomous weapons beyond meaningful human control“.
- Bei den Vereinten Nationen wurden ab 2014 in diversen Gremien Diskussionen zu den autonomen Waffensystemen geführt.
- 2018 wurde ein Verbot dieser Waffen von lediglich 26 Ländern[876] unterstützt.

Insgesamt scheinen diese Initiativen vor dem Hintergrund anderer aktueller geo- und gesundheitspolitischer Krisen an Momentum zu verlieren.

Die moralphilosophische Diskussionen zu „Autonomen Waffensystemen“ greift einige der zentralen in dieser Arbeit herausgearbeiteten Beurteilungsdimensionen für den Einsatz von KI auf: Autonomie/Freiheit, Verantwortung, Individuum/Individualität, Menschenwürde, in der Regel mit einem sehr starken Fokus auf Autonomie, Verantwortung und Menschenwürde. Diese drei Bereiche sollen deshalb hier – ähnlich wie in der vorherigen Fallstudie – vertieft werden.

Generell ist der Diskurs zum Thema kontrovers. In zahlreichen Publikationen wird nachdrücklich für ein Verbot von AWS argumentiert[877]. Andererseits finden sich auch Gegenpositionen, die sich klar gegen eine Ächtung von AWS aussprechen und deren Entwicklung explizit begrüßen, so zum Beispiel Vincent Müller, der einen Aufsatz mit dem Titel „*Autonomous killer robots are probably good news*“[878] verfasst hat. Sein Argument ist größtenteils konsequentialistisch und von der Überzeugung geleitet, dass mit dem Einsatz derartiger Systeme die Summe der menschlichen Opfer reduziert werden kann. Einige der gegenteiligen Argumente werden in den folgenden Abschnitten aufgegriffen.

875 Sharkey (2019), S. 75f

876 Ägypten, Algerien, Argentinien, Bolivien, Brasilien, Chile, China*, Costa Rica, Djibouti, Ecuador, Ghana, Guatemala, Irak, Kolumbien, Kuba, Mexico, Nicaragua, Österreich, Pakistan, Palästina, Panama, Peru, Uganda, Vatikan, Venezuela, Zimbabwe; * China forderte nur, die Nutzung vollautonomer Waffen zu verbieten, nicht aber deren Entwicklung und Produktion; Quelle: https://www.stopkillerrobots.org/wp-content/uploads/2018/04/KRC_CountryViews_13Apr2018.pdf, gesichtet am 25.8.2022

877 Beispiele: Amoroso (2020), Heyns (2017), IHRC (2014), Sharkey (2019), Sparrow (2007)

878 Müller (2016); ähnliche Argumente auch bei Arkin (2013)

13.2.1 KI in militärischen Waffensystemen: Autonomie

Es finden sich wenige Bereiche, in denen die falsche Verwendung des Autonomiebegriffs, selbst bei Philosophen, einen größeren Schaden angerichtet hat bzw. in Zukunft anrichten wird als bei den sogenannten „Autonomen Waffensystemen". Die Darstellungen des Unterschiedes zwischen der menschlichen Autonomie und der angeblichen Autonomie von Maschinen bleiben wenig überzeugend.

Befürworter derartiger Waffensysteme argumentieren häufig, AWS führten zu einer Verringerung von menschlichen Opfern, insbesondere unter Nichtkombattanten (Zivilisten und anderen geschützten Personen, wie z.B. Sanitäts- und Seelsorgepersonal des Militärs)[879]. Ronald Arkin führt dafür sechs Argumente auf:

- Roboter müssen sich nicht selbst verteidigen, töten daher auch nicht aus Selbstverteidigung oder präventiv.
- Die moderne Sensorik gibt Robotern einen deutlich präziseren Überblick über das Schlachtfeld und bessere Unterscheidung zwischen Kombattanten und Nichtkombattanten bzw. Freunden und Feinden.
- Roboter sind ohne Emotionen und agieren deshalb auch ohne Angst, Ärger, Frust, Rache und Hysterie und begehen daher auch nicht die Fehler, die Menschen in solchen Situationen machen.
- Roboter neigen nicht zu Verzerrungen wie „confirmation bias" oder „scenario fulfillment", die zu falschen Wahrnehmungen der Situation und damit verbundenen Fehlentscheidungen führen (können).
- Elektronische Systeme können mehr Informationen aus verschiedenen Quellen schneller verarbeiten als Menschen, die im modernen Krieg überfordert sind.
- AWS können im „Team mit menschlichen Soldaten" ausgleichend und objektiv wirken.

Das Grundproblem von Arkins Argumentation liegt darin, dass er den kategorialen Unterschied zwischen menschlichen Soldaten und AWS negiert oder zumindest komplett ausblendet. Heyns vertieft den Punkt in seinem Aufsatz von 2017, den er einige Jahre nach seinem Bericht zu AWS und deren Ächtung vor den Vereinten Nationen verfasste:

> *"What if technology develops to the point where it is clear that fully autonomous weapons surpass human targeting, and can potentially save many lives? Would human rights considerations in such a case not mitigate for the use of autonomous weapons, instead of against it? I argue that the rights to life and dignity demand that even under such circumstances, full autonomy in force delivery should not be allowed"*[880].

[879] Vgl. Arkin (2013), S. 320f; einschließlich der folgenden Argumente (übersetzt durch DS)

[880] Heyns (2017), S. 46

Heyns verweist in seiner Gegenargumentation auf den obigen kategorialen Unterschied: Die „Autonomie“ von Robotern ist nicht vergleichbar mit der Autonomie von menschlichen Wesen. Sie sind keine freien moralischen Akteure[881]. Er stellt weiterhin klar heraus, dass AWS niemals in vorhersagbarer Weise auf ihr Umfeld reagieren (können). Dieser Punkt wurde bereits im zweiten Kapitel dieser Arbeit thematisiert. Nach Foerster sind ML-Systeme „nichttriviale Technologien“: synthetisch determiniert, analytisch nicht-determinierbar, vergangenheitsabhängig und nicht vorhersagbar. Daher, so Heyns, entziehen sich solche Systeme per Definition der menschlichen Kontrolle:

> “As such, machine autonomy, beyond a certain point can potentially undermine or limit human autonomy and control over the world.”[882]

Vorherige Innovationen und Technologien gaben dem Krieger Kontrolle über immer mächtigere und wirkungsvollere Waffen. Die sogenannten autonomen Waffensysteme besitzen das Potential, dies umzudrehen und die Identität des Entscheidungsträgers zu tauschen: „*Die Waffe wird der neue Krieger*“[883].

Human Rights Watch (mit Unterstützung von C. Heyns) sieht vier Verstöße gegen das Völkerrecht („International Humanitarian Law“)[884]: Gegen das Differenzierungsgebot („Rule of Distinction“) zwischen Kombattanten und Nichtkombattanten, gegen das Gebot der Verhältnismäßigkeit („Rule of Proportionality“), gegen das Gebot der Abwägung der militärischen Notwendigkeit („Military Necessity“) und gegen die Martens’sche Klausel („Martens Clause“[885]), eine Klausel des Völkerrechts, nach der in allen Bereichen, die im Völkerrecht (noch) nicht explizit geregelt sind, Kombattanten und Zivilpersonen unter dem Schutz von *feststehenden Gebräuchen*, den *Grundsätzen der Menschlichkeit* und *Forderungen des öffentlichen Gewissens* stehen[886].

---

881 Heyns (2017), S. 48

882 Heyns (2017), S. 48

883 Heyns (2017), S. 48; Übersetzung DS

884 IHRC (2012), S. 30-36

885 “***Martens Clause**: Frequently cited as one of the quintessential demonstrations of the humanitarian character of the law of armed conflict (international humanitarian law), the Martens Clause stipulates that in cases not covered by international humanitarian law conventions, neither combatants nor civilians find themselves completely deprived of protection. Instead, in such cases, the conduct of belligerents remains regulated by the principles of the law of nations as they result from the usages of international law, from the laws of humanity, and from the dictates of public conscience.” Quelle: Oxford Bibliographies, https://www.oxfordbibliographies.com/view/document/obo-9780199796953/obo-9780199796953-0101.xml, heruntergeladen am 28.8.2022*

886 Vgl. Abs.9 des Vorspruchs der Mantelkonvention zur Haager Landkiegsordnung (HLKO) von 1899, Abs. 8 des Textes von 1907: „*So lange, bis ein vollständigeres Kriegsgesetzbuch festgestellt werden kann, halten es die hohen vertragschließenden Theile für zweckmäßig, festzusetzen, daß inden Fällen, die ind en von ihnen angenommenen Bestimmungen nicht vorgesehen sind, die Bevölkerungen und Kriegführenden unter dem Schutze und den herrschenden Grundsätzen des Völkerrechts bleiben, wie sie sich aus den unter gesitteten Staaten geltenden Gebräuchen, aus den Gesetzen der Menschlichkeit und aus den Forderungen des öffentlichen Gewissens herausgebildet haben*“

Das Differenzierungsgebot zur Unterscheidung zwischen Kombattanten und Nichtkombattanten, d.h. zwischen Soldaten und Zivilisten, war und ist schon seit jeher ein Gebot, gegen das in den meisten, wenn nicht allen Kriegen verstoßen wurde, aktuell auch in der Auseinandersetzung in der Ukraine, oftmals sogar bewusst und beabsichtigt. Die Nichteinhaltung dieses Gebots erhöht das Grauen des Krieges um Dimensionen. In Bezug auf den Einsatz der KI in sogenannten autonomen Waffensystemen ist zu erwarten, dass Zivilisten noch weniger geschützt werden können als in konventionellen Kriegen. In Zeiten, in denen Soldaten zunehmend keine Uniformen mehr tragen und zu ihrem eigenen Schutz die Nähe zu Zivilisten z.B. auf Marktplätzen oder in öffentlichen Verkehrsmitteln suchen, ist die präzise Unterscheidung zwischen Kombattanten und Nichtkombattanten nicht mehr programmierbar. Noel Sharkey wird bei Human Rights Watch dazu wie folgt zitiert:

> *"Humans understand one another in a way that machines cannot. Cues can be very subtle, and there is an infinite number of circumstances where lethal force is inappropriate."*[887]

Nur ein Mensch mit Emotionen erkennt eine Mutter, die sich um ihre Kinder sorgt und dabei laut schreiend auf den Soldaten zuläuft, von dem sie fürchtet, dass er ihre Kinder erschießen wird[888].

Beim Gebot der Verhältnismäßigkeit sieht es ähnlich aus. Die Entscheidung darüber, was in einer einmaligen militärischen Situation verhältnismäßig ist und nicht, kann nur das autonome Subjekt Mensch mit seiner Urteilskraft treffen. Nur ein Mensch kann Mitleid empfinden und entscheiden, dass es ausreicht, den gegnerischen Soldaten mit einem Schuss in die Beine kampfunfähig zu machen. Genauso wie bei der obigen Differenzierung ist dies nicht programmierbar.

Das Gleiche gilt für die Beurteilung der militärischen Notwendigkeit eines Mitteleinsatzes. Hier ergibt sich nach Human Rights Watch noch ein zweites Problem[889]: Aufgrund der reinen Existenz der AWS ergibt sich quasi automatisch das Gebot bzw. eine Verpflichtung ihres Einsatzes mit allen Konsequenzen für die weitere Eskalation: auf den ersten Blick ein minimales Risiko für die eigenen (menschlichen) Kräfte bei gleichzeitiger Optimierung der Zielerreichung um jeden Preis.

Die Martens-Klausel wurde formuliert, um zu verhindern, dass rechtsfreie Räume bestehen und dass immer dann, wenn Dinge nicht geregelt erscheinen[890], Bräuche bzw. Sitten sowie das eigene Gewissen und die Prinzipien der Menschlichkeit entscheiden. Sie ist

---

887 IHRC (2012), S. 31; Originalzitat aus Sharkey (2012c), S. 118

888 IHRC (2012), S. 31f

889 IHRC (2012), S. 35

890 "***Until a more complete code of the laws of war is issued, the High Contracting Parties think it right to declare that in cases not included in the Regulations adopted by them, inhabitants and belligerents remain under the protection and empire of the principles of international law****, as they result from the usages established between civilized nations, from the laws of humanity, and the requirements of the public conscience."* [Hervorhebung DS] Quelle: International Committee of the Red Cross, https://www.icrc.org/en/doc/resources/documents/article/other/57jnhy.htm, gesichtet am 28.8.2022

also ein Aufruf an den autonomen Menschen, nach seinem „inneren Gesetz zu entscheiden“ und seine Vernunft zu gebrauchen. Allerspätestens hier ergibt sich eine unüberwindbare Hürde für die KI.

13.2.2 KI in militärischen Waffensystemen: Verantwortung

> *„Der Mut ist doppelter Art: einmal Mut gegen die persönliche Gefahr, und dann Mut gegen die Verantwortlichkeit, sei es vor dem Richterstuhl irgendeiner äußeren Macht oder der inneren, nämlich des Gewissens.“*[891]
>
> Carl von Clausewitz, aus „Vom Kriege“, 1832

In Kapitel 8 („Kann die KI Verantwortung übernehmen?“) wurde herausgearbeitet, dass mit der Entwicklung und Verbreitung der Künstlichen Intelligenz eine zweifache Verantwortungslücke entstehen kann. Erstens diejenige, die entsteht, wenn Menschen Verantwortung an ein System delegieren, das selbst keine Verantwortung übernehmen kann, dessen Agieren allerdings auch nicht vorhersagbar ist, so dass der Delegierende auch keine Verantwortung übernehmen kann. Zweitens diejenige, die dadurch entsteht, dass die Entwickler und Anwender der KI eine Welt für kommende Generationen hinterlassen, in der die Grenzen der menschlichen Freiräume durch die KI bestimmt sind. Beide Punkte sind in dieser Fallstudie 1:1 anwendbar. Dies soll in diesem Abschnitt vertieft werden.

Das obige Zitat von Clausewitz beschreibt sehr klar, dass der (mutige) Soldat sich in einem zweifachen Spannungsfeld bewegt. Einerseits muss er die persönliche Gefahr für sein eigenes Leibeswohl und Leben (und das seiner Kameraden oder ihm anvertrauten Untergebenen) im Blick haben und andererseits muss er sich immer darüber im Klaren sein, dass er für sein Tun verantwortlich ist. Er trägt diese Verantwortung gegenüber seinen Dienstherren, den Gesetzen und seinem persönlichen Gewissen. Künstliche autonome Waffensysteme oder „Killerroboter“ fürchten nicht um das eigene „Leben“ und Wohlergehen und übernehmen auch keine Verantwortung gegenüber einem Vorgesetzten, einem Gericht oder einem nicht vorhandenen Gewissen.

In seinem Aufsatz zur „*Ethik der wehrtechnischen Digitalisierung*“ fasst der Informatiker und NATO-Berater Wolfgang Koch[892] vier zentrale Aspekte rund um den Begriff der Verantwortung zusammen, vollständig konsistent mit den obigen Ausführungen und Kapitel 8:[893]

---

[891] Quelle: https://www.clausewitz.com/readings/VomKriege1832/_VKwholetext.htm#1-3, gesichtet am 28.8.2022

[892] Prof. Dr. **Wolfgang Koch** lehrt Informatik an der Universität Bonn. Zugleich ist er Chief Scientist des Fraunhofer-Instituts für Kommunikation, Informationsverarbeitung und Ergonomie (FKIE) und leitet dort die Abteilung „Sensordaten- und Informationsfusion“. International engagiert er sich als Fellow des Institute of Electrical and Electronics Engineering (IEEE) und vertritt seine Forschungsgebiete in der NATO Science and Technology Organization (STO).

[893] W. Koch (2020), S. 38

- *„Von Verantwortung zu sprechen, ist nur sinnvoll, wenn sie freiwillig übernommen wird. Verantwortung setzt also Freiheit voraus und die Vorstellung vom Menschen als einer freien Person. Maschinen kann man keine Verantwortung übertragen.*
- *Die Vorstellung vom freien Willen als maßgebliche Ursache von Handlungen impliziert die auch juristisch relevante Vorstellung von der Zurechenbarkeit, die sich als ein wesentliches Kriterium im internationalen Völkerrecht findet.*
- *Impliziert wird ferner die Fähigkeit und innere Bereitschaft, trotz fehlender und widersprüchlicher Regeln „gut" zu handeln. Darin zeigt sich die Einsicht, dass Kasuistik, die Formalisierung verantworteten Handelns, unmöglich ist.*
- *Der in Freiheit verantwortende Wille ist nicht absolut, sondern hängt vom einsichtsfähigen Verstand ab. Das „Wahre" und „Gute" bilden also gemeinsam die gedanklich nicht hintergehbare Grundlage verantwortlichen Handelns."*[894]

Folgt man dieser Argumentation, dann verbietet sich jeder Einsatz von „human-out-of-the-loop" Waffensystemen und es ergeben sich hohe Anforderungen an die „human-on-the-loop" Versionen derselben. In der Tat sieht die praktische Realität ganz anders aus. Die Geschwindigkeit von Abläufen im Kriegsgeschehen und die Menge an systemerzeugten Daten machen einen signifikanten menschlichen Eingriff („meaningful human control") unmöglich. Der Mensch wird zum schwächsten Glied in der Kette, so dass aus dem immer kleineren „on-the-loop" ein „out-of-the loop" wird:

> *"The progression from remote controlled systems to LARs* [Lethal autonomous robotic]*, for its part, is driven by a number of other considerations. Perhaps foremost is the fact that, given the increased pace of warfare, humans have in some respects become the weakest link in the military arsenal and are thus being taken out of the decision-making loop. The reaction time of autonomous systems far exceeds that of human beings, especially if the speed of remote-controlled systems is further slowed down through the inevitable time-lag of global communications. States also have incentives to develop LARs to enable them to continue with operations even if communication links have been broken off behind enemy lines."*[895]

Wir kommen zurück zur Frage nach der moralischen und juristischen Verantwortung für die von AWS oder LAR verursachten Verstöße z.B. gegen Menschen- und Völkerrecht. Dazu liegt eine Reihe von Ideen vor, wie z.B. die Übernahme der Verantwortung durch denjenigen Offizier, der den Einsatz des Systems angeordnet hat, oder des Programmierers bzw. Herstellers des Systems, wenn Programm- oder Produktionsfehler zur „fehlerhaften Operation" geführt haben. Leider erweisen sie sich alle als unangemessen und unpraktisch. So kann der Kommandant keine Verantwortung übernehmen, wenn er die steuernden Programme, die zudem auch noch selbstlernend sind, nicht überblickt und

[894] W. Koch (2020), S. 38

[895] Heyns (2013), S. 10, § 53

versteht. Damit entsteht ein Verantwortungsvakuum[896]. Daraus folgt, dass Probleme, die aus dem Einsatz von AWS erwachsen, in der Regel straffrei bleiben. Länder können die politische Verantwortung übernehmen, werden mutmaßlich immer wieder auf technische Unregelmäßigkeiten verweisen. Ohne menschliche Verantwortung für die Details der Aktionen von Waffensystemen werden die militärischen Operationen willkürlich, wie auch der damit verbundene Verlust von menschlichem Leben[897]. Der Philosoph Peter Asaro aus New York schreibt dazu:

> *"The decision to kill a human can only be legitimate if it is non-arbitrary, and there is no way to guarantee that the use of force is not arbitrary without human control, supervision, and responsibility. It is thus immoral to kill without the involvement of human reason, judgement, and compassion, and it should be illegal."*[898]

Gegner einer Ächtung von AWS bestreiten nicht, dass Verantwortungslücken entstehen werden, d.h., dass sich Situationen ergeben, in denen die Waffensysteme vermeidbare menschliche Opfer verursachen, für die kein Mensch die angemessene Verantwortung übernimmt. Sie vergleichen dies mit seltenen Nebenwirkungen von Medikamenten, mit Todesfolgen oder Verkehrsopfern aufgrund des Zusammenbruchs von Brücken bei oder nach einer Flut oder einem Erdbeben[899]. Beide Beispiele erscheinen nicht angemessen für einen Vergleich. Bei Medikamenten werden Risiken von Nebenwirkungen mit den betroffenen Patienten besprochen oder zumindest auf dem Beipackzettel kommuniziert. Der mündige und autonome Patient wägt erhoffte positive Wirkungen des Medikaments gegen die Gefahr von Nebenwirkungen ab. Für die Opfer autonomer Waffensysteme besteht diese Möglichkeit definitiv nicht. Auch der Vergleich mit dem Risiko, Opfer eines Brückenzusammensturzes nach Überschwemmungen oder Erdbeben überzeugt nicht. Letzteres ist eher vergleichbar mit den Opfern von Naturkatastrophen, wie z.B. dem Risiko, durch einen Blitzschlag oder einen umstürzenden Baum ums Leben zu kommen. Ein AWS ist auf keinen Fall einem Naturphänomen zuzurechnen.

Das zweite Verantwortungsvakuum ist eher langfristig. Es erwächst aus der Tatsache, dass autonome Waffensysteme (AWS/LAR) nach ihrer Einführung wohl dauerhaft zum Arsenal der Armeen dieses Planeten gehören werden und nie wieder abgeschafft werden, ähnlich wie die nuklearen Waffen aus den Zeiten des Kalten Krieges. Schlimmer noch:

---

[896] Heyns (2013), S. 15, Absatz 80: "*The question of legal responsibility could be an overriding issue. If each of the possible candidates for responsibility identified above is ultimately inappropriate or impractical, a* ***responsibility vacuum*** *will emerge, granting impunity for all LAR use. If the nature of a weapon renders responsibility for its consequences impossible, its use should be considered unethical and unlawful as an abhorrent weapon."* [Hervorhebung DS]

[897] Heyns (2017), S. 14: "*One is not allowed to delegate a power that one does not have – and humans do not have the authority to use force without applying their minds. Non-human decision-making, or determinations where there is no meaningful human control regarding the use of lethal force is, according to this argument, inherently "arbitrary", and deaths that result are unlawful deprivations of life."*

[898] Asaro (2012), S. 708

[899] Vgl. Müller (2016), S. 75 (oder 8 von 16)

Nach und nach wird die große Mehrheit der Länder über diese Waffen verfügen wollen und es wird ein Rüstungswettbewerb einsetzen. Henry Kissinger und seine Mitautoren erwarten zusätzliche Herausforderungen für die internationale Sicherheit, unabhängig davon, ob es am Ende sogenannte autonome Waffensysteme geben wird oder nicht:

> *"The AI era risks complicating the riddles of modern strategy further beyond human intention – or perhaps complete human comprehension. Even if nations refrain from the widespread deployment of so-called lethal autonomous weapons – automatic or semiautomatic AI weapons that are trained and authorized to select their own targets and attack without further human authorization – AI holds the prospects of augmenting conventional, nuclear, and cyber capabilities in ways that make security relationships among rivals more challenging to predict and maintain and conflicts more difficult to limit."*[900]

Generell ist zu fürchten, dass die zunehmende Verfügbarkeit von AWS die Einstiegshürden in einen Krieg verkleinern und die Dynamik der Kriegsführung dergestalt verändern, dass zunehmend kriegerische Operationen gestartet werden, die man mit konventionellen Waffen als zu riskant eingeschätzt hätte.

Beide Typen von Verantwortungslücken sind also bedeutsam: Einerseits ist sicherzustellen, dass bei KI-gestützten Waffensystemen immer Personen die Verantwortung für den Einsatz tragen. Andererseits sind Politiker, Diplomaten und Generäle verantwortlich dafür, welche zusätzlichen Herausforderungen sie mit der Einführung dieser Technologien den zukünftigen Generationen hinterlassen. Die Nutzung der KI beim Militär bringt zum einen Vorteile mit sich: höhere Geschwindigkeit, höhere Präzision und besseren Schutz der eigenen Soldaten. Zum anderen bewirkt er eine drastische Erhöhung der Komplexität in militärischen Auseinandersetzungen bei einer geringeren Vorhersagbarkeit der Gesamtdynamik und einem steigenden Risiko der unkontrollierten Eskalation. Das alles muss bedacht und verantwortet werden.

13.2.3 KI in militärischen Waffensystemen: Menschenwürde

Viele würden argumentieren, dass Menschen generell im Krieg ihre Würde verlieren. Anlässlich der Materialschlachten des 1. Weltkrieges wurde der Begriff des Soldaten als „Kanonenfutter" geprägt. Der einzelne Soldat ist nicht mehr Subjekt und Herr seines Schicksals, sondern wird zum Objekt. Eine große Herausforderung ist vor diesem Hintergrund die Diskussion darüber, wie sich die Würde von Kombattanten und Nicht-Kombattanten durch den Einsatz von KI-gestützten sogenannten autonomen Waffensystemen gegenüber der konventionellen Kriegführung noch verschlechtern könnte.

Heyns macht die Verletzung der Menschenwürde – so wie in dieser Arbeit in Kapitel 10 argumentiert – am Verlust des Subjektseins des Menschen fest. Menschen werden als Objekt behandelt, wenn sie wegen ihres instrumentellen Wertes versklavt werden oder

[900] Kissinger et al. (2021), S. 139

als wertloses Einzelwesen in einem Massaker oder einem Terroranschlag ihr Leben verlieren. Er stellt diese beiden Kategorien von Menschen, deren Würde maximal verletzt wird, mit der Kategorie derer gleich, die aufgrund von AWS/LAR ums Leben kommen. Die Personen, die das KI-System zum Ziel der ultimativen Gewalt auswählt, werden zu Objekten, insbesondere wenn sie als Unbeteiligte irrtümlich ins Zielfeld geraten.

> „*Sie haben keine Möglichkeit, an die Menschlichkeit des Gegners zu appellieren oder darauf zu hoffen, dass ihre eigene Menschlichkeit eine Rolle spielt, weil sie sich auf der anderen Seite einer Maschine befinden*."[901]

Heyns stellt heraus, dass auch im konventionellen Krieg viele Menschen keine Chance haben, an die Menschlichkeit des Feindes zu appellieren. Es gibt fraglos immer wieder Ausnahmen, die bei AWS/LAR ausgeschlossen sind. Heyns schreibt dazu weiter:

> *"In a pre-autonomous weapons world, there are two thresholds that force has to meet during armed conflict. The first is that it has to be established that the use of force in particular is legal. The second is that a human being moreover must deem it necessary and warranted in that particular case to use such a force. Autonomous weapons mean the second barrier is removed."*[902]

Dies ist ein sehr starkes Argument: Der Roboter wird im Rahmen seiner ihm vorgegebenen Rahmenbedingungen seinen Auftrag vollständig erfüllen. Der Mensch kann das auch, muss es aber nicht und kann im Einzelfall aus Gründen, die nur ihn als Mensch in seinem Handeln beeinflussen können, davon abweichen. Konsequentialisten mögen argumentieren, dass dies in der Realität selten geschehe und im Verhältnis zur Gesamtzahl der Opfer vernachlässigbar sei.

[901] Heyns (2017), S. 18; Übersetzung DS

[902] Heyns (2017), S. 18

13.3 Erkenntnisse aus den Fallstudien

Die beiden Fallstudien könnten unterschiedlicher kaum sein. Dennoch zeigt die detaillierte Untersuchung, dass die Dimensionen des vorgeschlagenen Bewertungsrahmens insbesondere in Bezug auf Autonomie, Verantwortung und Würde sowie in etwas reduziertem Umfang in Bezug auf Individualität sehr relevant und anwendbar sind.

Abbildung 9: Identifizierte Risiken und Gefährdungen in den Fallstudien

Jede Preisgabe von menschlicher Autonomie und Verantwortlichkeit zugunsten von Maschinen, Robotern oder KI-Systemen muss als solche erkannt werden und ist in Bezug auf die potenzielle Beschädigung oder Verletzung von menschlicher Individualität und Würde kritisch zu reflektieren.

Erschwert wird dieser Diskurs dadurch, dass die Verlagerung oftmals schleichend vonstattengeht und die menschliche Autonomie und Verantwortlichkeit nach und nach reduziert wird. Dies gilt für die sukzessive Einführung von Pflegesystemen, die immer größere Arbeitsumfänge aus dem Arbeitsvolumen eines Pflegers übernehmen und dabei zu einer stetigen Entmenschlichung und Verdinglichung führen und auch für die Schritt-für-Schritt Einführung von AWS.

Welche Probleme sich nach der vollständigen Einführung der KI ergeben, kann oftmals erst im Nachhinein festgestellt werden, was die Durchsetzbarkeit von Korrekturen erschwert. Mittlerweile haben auch amtierende und ehemalige Führungskräfte von Technologiekonzernen diese Problematik erkannt. In einem Interview der Zeitschrift „The Atlantic“ zu den Risiken der Einführung generativer KI entwirft Eric Schmidt, der ehemalige Vorstandsvorsitzende von Google, ein dystopisches Szenario, wonach bestimmte Gruppen die Systeme missbrauchen könnten, um damit rassistisches oder

demokratiefeindliches Gedankengut zu verbreiten. Er fürchtet, dass – ähnlich wie bei den sozialen Medien – diese Problematik viel zu spät erkannt wird:

> „*If you think about the biggest problems in the world, they are all really hard— climate change, human organizations, and so forth. And so, I always want people to be smarter. The reason I picked a dystopian example is because we didn't understand such things when we built up social media 15 years ago. We didn't know what would happen with election interference and crazy people. We didn't understand it and I don't want us to make the same mistakes again.*"[903]

Im selben Artikel spricht ein anderer Unternehmer in Bezug auf die generative KI von einer „*pretty radical uncertainty*". Die Unsicherheit darüber, welche segensreichen oder aber verheerenden Auswirkungen die Technologie haben kann, ist selbst bei denjenigen, die sich am besten damit auskennen sollten, „radikal".

[903] Warzel (2023)

14 Zusammenfassung und Ausblick

> „*Wir müssen uns gegen den Posthumanismus, den Versuch, den Menschen abzuschaffen, wehren. Denn er ist eine Verblendung, die der Selbstvernichtung des Menschen durch seinen digitalen Militärapparat zugutekommt. Wer den Menschen zugunsten des Übermenschen überwinden will, verachtet in Wahrheit das Leben. Doch der einzige Sinn des Lebens ist das Leben selbst. Das gelingende Leben ist der Sinn des Lebens. Die Bedingungen eines gelingenden Lebens werden unter anderem im Nachdenken über das Nachdenken erforscht, das über unseren Denksinn Verbindungen mit der Tatsache aufnimmt, dass wir unserem Leben nicht entrinnen können*."[904]
>
> Markus Gabriel, Schlusswort von „Der Sinn des Denkens"

Diese Arbeit ist zugleich eine Würdigung und eine Entzauberung der Künstlichen Intelligenz sowie eine Warnung an ihre Nutzer.

Die **Würdigung** besteht darin, dass klar dargestellt wurde, welche bahnbrechenden Fortschritte in den knapp siebzig Jahren seit ihrer Begründung und der Prägung des Begriffes „Artificial Intelligence" erreicht werden konnten. Das betrifft die Algorithmen der KI, die flankierenden Technologien und die Vielzahl der Anwendungen in nahezu allen Lebensbereichen des Menschen, im privaten wie im öffentlichen Sektor. Mit der KI verbindet man zu Recht die Hoffnung, die größten Probleme der Menschheit lösen zu können, z.B. im Gesundheitswesen, in Bezug auf den Klimawandel und bei der Modernisierung der Wirtschaft und Gesellschaft und nicht zuletzt in den Wissenschaften. Auch die Erwartungen an die ökonomischen Möglichkeiten der Technologie übertreffen diejenigen anderer Innovationen.

Eine **Entzauberung** ergibt sich aus der Erkenntnis, dass die KI mitnichten wirklich intelligent ist. Die Technologie wird durch menschengemachte Algorithmen gesteuert, die zugegebenermaßen immer komplexer und autoadaptiver geworden sind und so den Eindruck erwecken, sie hätten eine der menschlichen vergleichbaren Intelligenz. Die gewaltigen Leistungssteigerungen der letzten Jahrzehnte gehen auch auf innovative neue Algorithmen etwa für Mustererkennung, Bildanalyse und Sprachverarbeitung zurück, stützen sich aber noch viel stärker auf die deutlich gestiegene Leistungsfähigkeit der Basistechnologien wie Prozessoren, Speichermedien und Sensorik und vor allem auf die exponentiell gestiegene Verfügbarkeit von Daten, die überhaupt erst die vielfältigen Anwendungen der Technologie ermöglichen. Die Künstliche Intelligenz ist und bleibt (synthetisch) deterministisch. Damit wird sie dauerhaft weit von ihrem Vorbild der menschlichen Intelligenz entfernt bleiben – weiter als der Autor dieser Arbeit zu Beginn der Recherchen noch erwartet hätte. Letzteres ergibt sich einerseits aus den klarer erkennbaren fundamentalen Grenzen der Technologie und anderseits aus einem verbesserten Verständnis der menschlichen Intelligenz und des menschlichen Bewusstseins. Es zeichnet sich zudem

[904] Gabriel (2018), S. 318f

ab, wie wenig wir im Jahr 2023 wirklich über das menschliche Bewusstsein, sein Zustandekommen und die Möglichkeiten seines Nachbaus wissen. Mehr offene Fragen als Antworten verbleiben. Vermutlich stoßen wir bei unserer Suche nach einer Lösung fundamentaler Fragen zum menschlichen Bewusstsein auch an kategoriale Grenzen unserer Erkenntnisfähigkeit[905].

Die zentrale Erkenntnis dieser Arbeit hat hingegen den Charakter einer **Warnung**. Diese Warnung ergibt sich nicht aus einer Sorge vor der Gefährlichkeit der Technologie an sich. Alle Warnungen vor der starken KI, der künstlichen allgemeinen Intelligenz (Artificial General Intelligence, AGI), der Intelligenzexplosion und der Singularität erscheinen sehr weit hergeholt, zumindest lassen sich dafür keine Indizien und in der Empirie keine Belege finden. Es ist nicht damit zu rechnen, dass die KI, wie in der Science Fiction häufig dargestellt, ihre eigene Agenda entwickelt und die Herrschaft über die Menschheit übernimmt. Die Künstliche Intelligenz ist und bleibt ein Werkzeug des Menschen. Daran hat sich mit der aktuell breit in den Medien diskutierte Applikation ChatGPT auch nichts geändert. Aus dieser Positionierung wird die Technologie zukünftig nicht heraustreten können, selbst wenn es dazu weiterhin kontroverse Meinungen unter Vertretern verschiedener Disziplinen und der Philosophie gibt[906].

Die tatsächliche Warnung setzt bei der Anwendung dieser Technologie an und letztlich bei der Frage: Was macht die KI mit uns Menschen? Oder noch präziser: Was richten wir mit ihrer Nutzung an? Subjekt bei dieser Betrachtung ist nicht die Technologie, obwohl es einige Stimmen gibt, die sie gern in der Subjektrolle sehen würden, sondern sind wir Menschen selbst. Es konnte sehr klar herausgearbeitet werden, dass erhebliche Gefahren für die menschliche Freiheit, die menschliche Verantwortlichkeit, Individualität und Menschenwürde bestehen. Dadurch, dass die Verantwortung für die Technologie und ihre Anwendung einzig und allein dem Menschen als Subjekt zugewiesen werden kann, sind auch alle Einschränkungen, die daraus für unsere Freiheit, Eigenverantwortung, Individualität, Mündigkeit und Würde erwachsen, selbstverschuldet. Deswegen besteht

[905] Vgl. Siebert (1998), S. 193f

[906] Vgl. Chalmers (2022) und Bender Koller (2020); Chalmers kommt in seiner Untersuchung zur Frage „Could a Large Language Model be Conscious?“ zu einem negativen Ergebnis für derzeitig bekannte Modelle, schließt aber nicht aus, dass sich das innerhalb der nächsten Dekade ändert und fordert weitere Forschung zu dieser Frage: „*My conclusion is that questions about AI consciousness are becoming ever more pressing. Within the next decade, even if we don't have human level artificial general intelligence, we may have systems that are serious candidates for consciousness. Although there are many challenges and objections to consciousness in machine learning systems, meeting those challenges yields a realistic potential research program for conscious AI."*;

Bender und Koller sehen weiterhin die grosse Schwäche von LLMs darin, dass sie die Semantik bzw. Bedeutung (Meaning) der Texte nicht verstehen: „*We argue that the language modeling task, because it only uses form as training data, cannot in principle lead to learning of meaning. We take the term language model to refer to any system trained only on the task of string prediction, whether it operates over characters. Words, or sentences, and sequentially or not. We take (linguistic) meaning to be the relation between a linguistic form and communicative intent.*"

berechtigter Anlass zur Sorge, dass einige wichtige Errungenschaften der Aufklärung wieder preisgegeben werden. Dies wurde allgemein gezeigt und anhand von zwei exemplarischen Fallstudien zum Einsatz der KI in der Pflege und in militärischen Waffensystemen veranschaulicht. Es ist tatsächlich zu befürchten, dass die Menschheit wieder zurückfällt in eine wiederum selbstverschuldete Unmündigkeit und damit die Aufklärung nur eine temporäre Erscheinung war.

Gerade weil die Preisgabe der menschlichen Freiheit, Verantwortlichkeit, Individualität und Würde durch den Einsatz der Künstlichen Intelligenz in kleinen Schritten erfolgt und in vielen Fällen über Generationen hin vonstattengeht, aber letztlich unwiderruflich sein dürfte, ist eine von einem breiten gesellschaftlichen Konsens getragene Begleitung des Einsatzes der Technologie gefordert. Die in diesem Werk herausgearbeiteten Dimensionen stellen einen ersten Ansatz dar. Besonders starke ethische Konflikte werden sich dann ergeben, wenn sich aus einer rein konsequentialistischen Betrachtung signifikante Vorteile der KI-Nutzung ergeben, obwohl die deontologische Ethik zur Vorsicht mahnt. Bei den beiden in dieser Arbeit betrachteten Fallstudien konnte eine derartige Situation festgestellt werden. Der konsequente Einsatz der KI in der Pflege von Alten und Kranken kann aus der Perspektive der Ökonomie und des Lebenserhalts vorteilhaft sein, aber zugleich eine eklatante Einschränkung der Menschenwürde mit sich bringen. Auch beim Einsatz der KI in Waffensystemen ergibt sich aus einer kurzfristigen konsequentialistischen Sichtweise ein klarer Vorteil für den Einsatz der Technologie, der aber – so lautet das Argument dieser Arbeit – beim Einnehmen einer langfristigen Perspektive und auf Basis einer pflichtenethischen Beurteilung keinen Bestand mehr hat.

Spätestens jetzt ist der Zeitpunkt gekommen, dass bei der weiteren Entwicklung der Technologie kurz-, mittel- und langfristige Auswirkungen des Einsatzes von Applikationen der Künstlichen Intelligenz auf die Autonomie und Freiheit sowie menschlicher Verantwortlichkeit, Individualität und Würde mit in Betracht gezogen werden müssen. Neben der weiter forcierten Forschung und Entwicklung der Technologie an sich ist in jedem Anwendungsfall auch eine wissenschaftliche Vertiefung der Auswirkungen auf die Humanität entlang der oben beschriebenen und in dieser Dissertation erarbeiteten Leitlinien erforderlich. Unter dem Arbeitstitel einer „Verantwortlichen KI“ („Responsible AI“) signalisieren viele Unternehmen, die entweder die KI an sich weiterentwickeln oder sie verstärkt in ihren Produkten und Prozessen nutzen wollen, ein gewisses Problembewusstsein. Die dabei verfolgten Ansätze sind allerdings rudimentär, wenig systematisch und oftmals nur „Lippenbekenntnisse“. Eine „Ethik der KI“ sollte eine unerlässliche Komponente der Wirtschafts- und Unternehmensethik sein.

Zusätzlich ist die Frage zu stellen, ob eine Selbstregulierung und -beschränkung der Wirtschaft ausreicht, um Fehlentwicklungen im Sinne dieser Arbeit zu verhindern und aufzuhalten. Es ist zu klären, mit welcher Regulierung und normativen Begleitung von

staatlicher Seite auf regionaler, nationaler und globaler Ebene die Entwicklung angemessen normativ begleitet werden kann. Die bisherigen Ansätze in einigen Regionen (z.B. in der Europäischen Union) greifen zu kurz und sind häufig reaktiv. Nur aus einer disziplinübergreifenden Initiative mit den Natur- und Ingenieurwissenschaften, der Philosophie, den Rechtswissenschaften, der Soziologie und den Politikwissenschaften können effektive und effiziente Eingriffsmöglichkeiten erarbeitet werden. Diese müssen dann zügig über die demokratischen Institutionen ihren Weg in die Gesetzgebung und Regulierung finden.

Aufgrund der globalen Reichweite von Entwicklern und Nutzern der KI und des sich daraus ergebenden weltweiten Wettbewerbs um Technologieführung muss auch dringend ein weltweiter Diskurs in den relevanten Institutionen (z.B. Vereinte Nationen, G7, G20 und WEF) herbeigeführt werden. Die Sorge ist groß, dass „das Kind bereits in den Brunnen gefallen ist“ und viele Entwicklungen mittlerweile unumkehrbar geworden sind. Ähnlich wie in der Fallstudie zum Einsatz der KI in Waffensystemen dargestellt, könnte sich eine Logik ergeben, nach der sich kein Teilnehmer mehr eine Verlangsamung oder einen Ausstieg aus der Technologieentwicklung leisten kann. Analog zur Nukleartechnologie zeichnet sich ein Wettlauf der Nationen („Arms Race“) ab. Die spieltheoretische Logik ist klar: Verhängnisvoller als die unregulierte Weiterentwicklung und -verbreitung der Technologie ist es, selbst nicht dabei zu sein und von allen anderen abgehängt zu werden. Schon im „Gefangenendilemma“ konnte aufgezeigt werden, dass Alleingänge („defection“) vorteilhafter sein können als die gutgläubige Kooperation und der Selbstverzicht mit dem Risiko, dass sich andere nicht an die Vereinbarungen halten (könnten).

Die mittel- bis langfristigen Auswirkungen der Entwicklung auf die Humanität unterliegen einer großen Unsicherheit. Unser Wissen darüber ist nach wie vor beschränkt. Die von Hans Jonas in Bezug auf moderne Technologien beschriebene „*Kluft zwischen der Kraft des Vorherwissens und Macht des Tuns*“[907] vergrößert sich dramatisch. Aus der „*Anerkennung der Unwissenheit*“ leitet Jonas an gleicher Stelle in seiner Ethik der Verantwortung die „*Kehrseite der Pflicht des Wissens*“ ab. Es ist unsere Pflicht, die hier beschriebene Kluft in einem disziplinübergreifenden Kraftakt zu verkleinern oder gar zu schließen. Ansonsten droht der Rückfall des Menschen in die selbstverschuldete Unmündigkeit mit allen Konsequenzen in Bezug auf Freiheit, Individualität und Menschenwürde.

An dieser Stelle soll die aktuelle Diskussion über die Regulierung der KI auf Basis der ersten Erfahrungen mit ChatGPT im Jahr 2023 kommentiert werden. Die KI ist mit dem Erscheinen dieser Applikation endgültig im breiten öffentlichen Diskurs angekommen. Weiterhin gibt es zunehmend Stimmen in Politik, Wissenschaft, Wirtschaft, dort insbesondere aus den Technologieunternehmen, die eine Regulierung fordern, um Gefahren

[907] Jonas (1979), S. 28

für die Humanität abzuwenden sind. Diese Entwicklung ist begrüßenswert und unterstützt auch grundsätzlich die Erkenntnisse und Forderungen dieser Arbeit. Dennoch ist weiterhin Skepsis angebracht.

Es ist zu befürchten, dass die Maßnahmen zu kurz greifen und nur die Probleme in Bezug auf „Data Privacy", Verhinderung von Vorurteilen (biases), Diskriminierungen und Hassrede bzw. Erzeugung von Falschnachrichten (deep fake) angehen. Hierbei handelt es sich tatsächlich um eklatante kurzfristige Herausforderungen, die unbedingt normativ und regulativ zu adressieren sind. Die in dieser Arbeit identifizierten Problemfelder gehen allerdings weit darüber hinaus. Dafür ist bisher noch kein hinreichendes öffentliches Bewusstsein festzustellen.

Ergänzend ist nicht erkennbar, wie die Regulierung konkret auszugestalten ist und weltweit durchgesetzt werden kann, so dass die beklagten oder befürchteten Entwicklungen tatsächlich verhindert werden können. Es verbleibt ein erheblicher Handlungsbedarf für Wissenschaft, Politik und Wirtschaft.

Neue Forschungsfragen für vertiefende Recherchen ergeben sich entlang der vier in dieser Arbeit herausgearbeiteten Dimensionen und der Wahrung der Mündigkeit des Menschen sowie in Bezug auf die wirksame und nachhaltige Regulierung der Technologie:

- Wie kann menschliche Autonomie (und Freiheit) bei gleichzeitiger Nutzung der KI als Werkzeug gewahrt werden? Wie erkennen und verhindern wir Einschränkungen der menschlichen Freiheit im Kleinen und im Großen?
- Wie kann menschliche Verantwortlichkeit gewahrt und die Entstehung von Verantwortungslücken verhindert werden? Wie schaffen wir Verantwortlichkeit für die umfassenden KI-bezogenen Änderungen unserer Lebenswelt für zukünftige Generationen?
- Wie bewahren wir menschliche Individualität in einer Welt, in der mit der KI eine immer systematischere Verhaltenssteuerung erfolgt?
- Wie schützen wir die Menschenwürde und dabei insbesondere den Subjektcharakter des Menschen?
- Wie bewahren wir die Mündigkeit des Menschen gemäß der Forderung von Immanuel Kant: Sapere Aude! Habe Mut, dich deines eigenen Verstandes zu bedienen! Wie sieht eine wirksame Fortsetzung der Aufklärung in diesem Jahrtausend aus?
- Mit welcher Regulierung gelingt die Gradwanderung zwischen einer Nutzung des innovativen Potentials der KI und des Schutzes der Humanität? Welche Ansätze sind weltweit wirksam und nachhaltig? Wie stellt man sicher, dass das gesamte Spektrum der identifizierten Risiken in den Blick genommen wird?

Bei den ersten fünf Fragen ist die grundsätzliche Herausforderung in dieser Arbeit herausgearbeitet worden. Nun geht es um das „Wie“ der Vermeidung der Problematik und übergreifend um die konkreten Instrumente einer wirksamen Regulierung.

Anhänge

Anhang 1: Kurzüberblick zu Aufbau und Funktion des Gehirns

Die grundsätzlichen Erkenntnisse der Hirnforschung, die elementaren Wirkmechanismen und der anatomische Aufbau des Gehirns finden sich in der relevanten Fachliteratur. In diesem Abschnitt sollen nur diejenigen Aspekte und Begrifflichkeiten dargestellt werden, auf die in dieser Arbeit zurückgegriffen oder verwiesen wird.

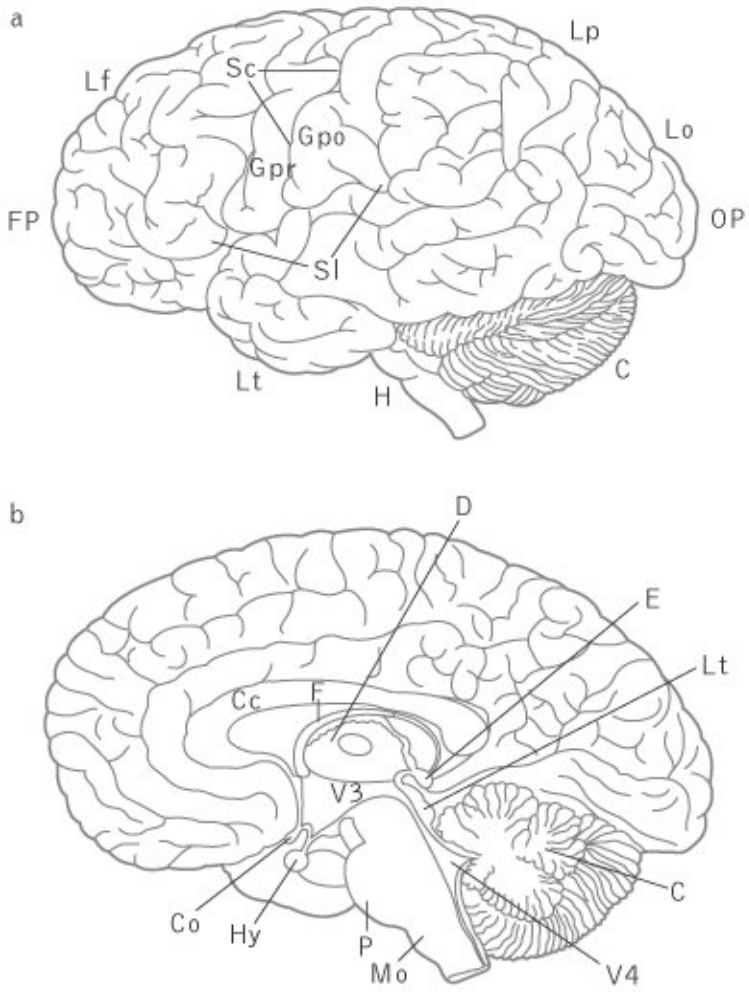

Abbildung 10: Aufbau des menschlichen Gehirns[908]

[908] ***Lateralansicht (oben) und Querschnitt durch das Gehirn des Menschen***. ***C*** *Cerebellum (Kleinhirn),* ***Cc*** *Corpus callosum (Balken),* ***Co*** *Chiasma opticum (Sehnervenkreuzung),* ***D*** *Diencephalon (Zwischenhirn),* ***E*** *Epiphyse,* ***F*** *Fornix,* ***Fp*** *Frontalpol,* ***H*** *Hirnstamm,* ***Lf*** *Lobus frontalis (Frontallappen),* ***Lt*** *lamina quadrigemina,* ***Gpo*** *Gyrus postcentralis (Region der Sensorik),* ***Gpr*** *Gyrus praecentralis (Region der Willkürmotorik),* ***Hy*** *Hypophyse,* ***Lo*** *Lobus occipitalis (Hinterhauptslappen),* ***Lp*** *Lobus parietalis (Scheitellappen),* ***Lt*** *Lobus temporalis (Schläfenlappen),* ***Mo*** *Medulla oblongata (verlängertes Mark), Op Okzipitalpol,* ***P*** *Pons (Brücke),* ***Sc*** *Sulcus centralis,* ***Sl*** *Sulcus lateralis,* ***V3*** *dritter Ventrikel,* ***V4*** *vierter Ventrikel,*; Quelle: Kompaktlexikon der Biologie: https://www.spektrum.de/lexikon/biologie-kompakt/gehirn/4627

Das menschliche Gehirn wiegt etwa drei Pfund. Obwohl es nur ca. zwei Prozent der menschlichen Körpermasse ausmacht, ist es für ein Viertel des Energieverbrauchs über Nährstoffe (Hauptsächlich Glucose) verantwortlich. Die Bestandteile auf der obersten Ebene sind der Hirnstamm, das Zwischenhirn, das Klein-, Mittel- und Großhirn. Letzteres ist in zwei Hälften (auch Hemisphären) geteilt, die über den Corpus Callosum (auch Balken genannt) verbunden sind. Weiße Nervenfasern verbinden die Gehirnhälften miteinander und mit dem Rückenmark und dem Nervensystem. Im Zwischenhirn befindet sich der Thalamus, der wie eine Schalt- und Integrationszentrale für die Sensorik und Motorik funktioniert.

Der auffälligste und wichtigste Teil des Gehirns ist das Großhirn mit seiner faltigen Großhirnrinde, die auch Kortex genannt wird. Die Großhirnrinde mit ihren Furchen und Windungen ist rund 2200 Quadratzentimeter groß, etwa 1,2 bis 4,5 mm dick und besteht aus fünf Schichten. Sie lässt sich in einzelne Felder unterteilen, die mit verschiedenen Funktionen korreliert werden können, also z.B. motorische, auditorische, visuelle und sprachliche Rindenfelder.

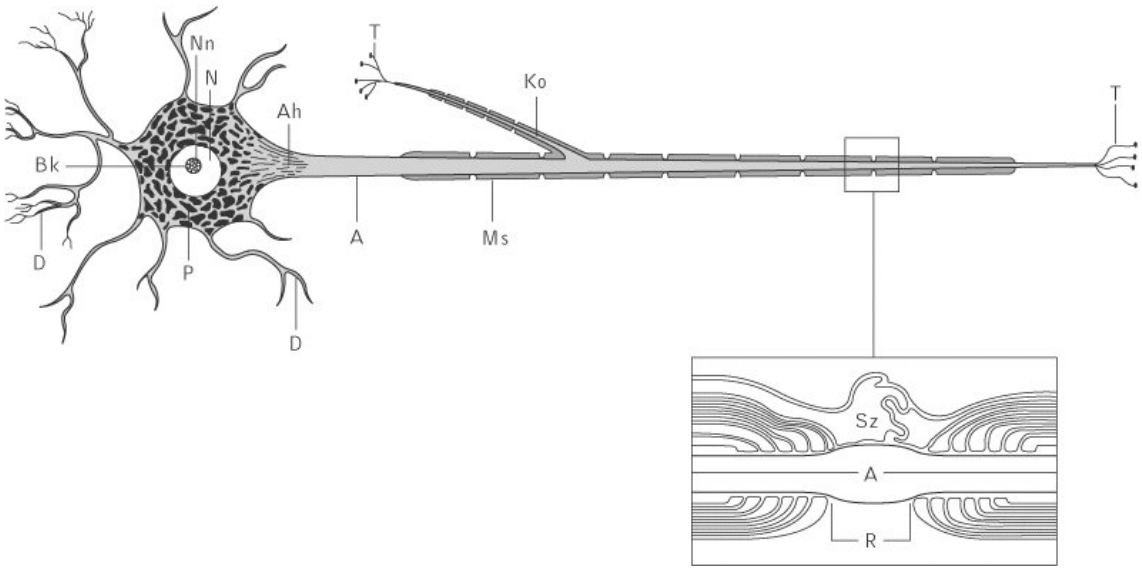

Abbildung 11: Schematischer Bau eines Neurons[909]

Bei den Neuronen handelt es sich um Nervenzellen, die Grundbausteine des Gehirns. Weiterhin finden sich Gliazellen, die Stütz- und Versorgungsfunktionen für die

[909] *Der Ausschnitt zeigt die Feinstruktur eines Ranvier-Schnürrings, und zwar in der unteren Hälfte der Abb. einer Faser des Zentralnervensystems, in der oberen Hälfte einer Faser des peripheren Nervensystems.* ***A*** *Axon,* ***Ah*** *Axonhügel,* ***Bk*** *Barrkörperchen,* ***D*** *Dendriten,* ***Ko*** *Kollaterale,* ***Ms*** *Markscheide,* ***N*** *Nucleus,* ***Nn*** *Nucleolus,* ***P*** *Perikaryon,* ***T*** *Telodendron*; Quelle: Kompaktlexikon der Biologie: https://www.spektrum.de/lexikon/biologie-kompakt/neuron/8159

Nervenzellen wahrnehmen[910]. Eine Vielzahl unterschiedlicher Arten von Nervenzellen können unterschieden werden. Ihre generelle Funktion besteht darin, Erregung aufzunehmen, zu verarbeiten und wieder abzugeben. Neuronen bestehen aus einem Zellkörper (etwa 30 Mikron[911] dick), an dem viele dünne Fortsätze, auch Dendriten genannt, hängen[912]. Die Dendriten nehmen in der Regel die Erregung anderer Nervenzellen auf. Ein am Zellkörper oder einem Hauptdendriten entspringender Fortsatz nennt man Axon. Über dem Axon, von dem es auch mehrere geben kann, werden Erregungen an andere Neuronen weitergegeben. Kommunikation über größere Entfernungen geschieht mit Hilfe von Faserbündeln.

An den Übergabepunkten zu anderen Neuronen sitzen die Synapsen. Jedes einzelne Neuron kann über Synapsen mit tausenden anderen Neuronen verdrahtet sein. Insgesamt gibt es im menschlichen Gehirn etwa eine Million Milliarden Synapsen.

Man unterscheidet zwischen dem präsynaptischen und dem postsynaptischen Teil. Der postsynaptische Teil ist mit einem Dendriten eines anderen Neurons verbunden. Das Neuron ist wegen seiner Membraneigenschaften elektrisch geladen. Bei einer Erregung des Neurons ergibt sich ein elektrischer Impuls entlang des Axons bis in den präsynaptischen Bereich der Synapse und setzt dort aus kleinen Bläschen (Vesikel) Neurotransmitter frei, die sich dann auf der gegenüberliegenden Seite an Rezeptoren oder Kanälen der postsynaptischen Seite anlagern. Die Summe der Anlagerungen an den verschiedenen Dendriten des postsynaptischen Neurons kann auch elektrische Entladungen auslösen (Aktionspotentiale). Die Kommunikation zwischen den Neuronen ergibt sich also aus einer Kombination von elektrischen und chemischen Vorgängen. Die unterschiedlichen Regionen des Gehirns verfügen über verschiedene Neurotransmitter und weitere Substanzen, die aufgrund ihrer Eigenschaften Zeitpunkt, Amplitude und Abfolge neuronaler Entladungen beeinflussen. Die beiden wichtigsten Wirkungen sind Erregung (Exzitation) und Hemmung (Inhibition)[913].

[910] Roth (2009), S. 16f; Edelman (2004), S. 28f

[911] 0,03 mm

[912] Edelman (2004), S. 30

[913] Roth (2009), S. 18

Anhang 2: Das Manifest der Hirnforscher[914]

Elf führende Neurowissenschaftler über Gegenwart und Zukunft der Hirnforschung

Von

Prof. Dr. Hannah Monyer, Ärztliche Direktorin der Abteilung für Klinische Neurobiologie, Universität Heidelberg

Prof. Dr. Frank Rösler, Abteilung Psychologie, Philipps-Universität Marburg

Prof. Dr. Dr. Gerhard Roth, Direktor am Institut für Hirnforschung der Universität Bremen und Rektor des Hanse-Wissenschaftskollegs in Delmenhorst

Prof. Dr. Henning Scheich, Direktor am Leibniz-Institut für Neurobiologie, Magdeburg

Prof. Dr. Wolf Singer, Direktor am Max-Planck-Institut für Hirnforschung, Abteilung Neurophysiologie, Frankfurt am Main

Prof. Dr. Christian E. Elger, Direktor der Klinik für Epileptologie, Universität Bonn

Prof. Dr. Angela D. Friederici, Abteilung Neuropsychologie, Direktorin am Max-Planck-Institut für Kognitions- und Neurowissenschaften Leipzig

Prof. Dr. Christof Koch, California Institute of Technology (Caltech), Computation and Neural Systems, Pasadena

Prof. Dr. Heiko Luhmann, Institut für Physiologie und Pathophysiologie, Johannes-Gutenberg-Universität Mainz

Prof. Dr. Christoph von der Malsburg, Institut für Neuroinformatik, Ruhr-Universität Bochum sowie Computational Vision Lab der University of Southern California, Los Angeles

Prof. Dr. Randolf Menzel, Abteilung Neurobiologie, Freie Universität Berlin

Was wissen und können Hirnforscher heute?

Angesichts des enormen Aufschwungs der Hirnforschung in den vergangenen Jahren entsteht manchmal der Eindruck, unsere Wissenschaft stünde kurz davor, dem Gehirn seine letzten Geheimnisse zu entreißen. Doch hier gilt es zu unterscheiden: Grundsätzlich setzt die neurobiologische Untersuchung des Gehirns auf drei verschiedenen Ebenen an.

- Die oberste erklärt die Funktion größerer Hirnareale, beispielsweise spezielle Aufgaben verschiedener Gebiete der Großhirnrinde, der Amygdala oder der Basalganglien.
- Die mittlere Ebene beschreibt das Geschehen innerhalb von Verbänden von hunderten oder tausenden Zellen.
- Und die unterste Ebene umfasst die Vorgänge auf dem Niveau einzelner Zellen und Moleküle. Bedeutende Fortschritte bei der Erforschung des Gehirns haben wir bislang nur auf der obersten und der untersten Ebene erzielen können, nicht aber auf der mittleren.

Verschiedene Methoden ermöglichen einen Einblick in die oberste Organisationsebene des Gehirns:

[914] Quelle: Gehirn & Geist 6/2004

- Bildgebende Verfahren wie die Positronen-Emissionstomografie (PET) und die funktionelle Magnetresonanztomografie (fMRT), die den Energiebedarf von Hirnregionen messen, besitzen eine gute räumliche Auflösung, bis in den Millimeterbereich. Zeitlich gesehen hinken sie den Vorgängen allerdings mindestens um Sekunden hinterher.
- Die klassische Elektroenzephalografie (EEG) dagegen misst die elektrische Aktivität von Nervenzellverbänden quasi in Echtzeit, gibt aber nicht genau Aufschluss über den Ort des Geschehens.
- Etwas besser – etwa im Zentimeterbereich – liegt die räumliche Auflösung bei der neueren Magnetenzephalografie (MEG), mit der sich die Änderung von Magnetfeldern um elektrisch aktive Neuronenverbände millisekundengenau sichtbar machen lässt.

Drei Ebenen der Erkenntnis

Insbesondere durch die Kombination mehrerer dieser Technologien können wir das Zusammenspiel verschiedener Hirnareale darstellen, das uns kognitive Funktionen wie Sprachverstehen, Bilder erkennen, Tonwahrnehmung, Musikverarbeitung, Handlungsplanung, Gedächtnisprozesse sowie das Erleben von Emotionen ermöglicht.

Damit haben wir eine thematische Aufteilung der obersten Organisationsebene des Gehirns nach Funktionskomplexen gewonnen.

Auch hinsichtlich der untersten neuronalen Organisationsebene hat die Entwicklung völlig neuartiger Methoden wie etwa der Patch-clamp-Technik, der Fluoreszenzmikroskopie oder des Xenopus-Oocyten-Expressionssystems zu einem Erkenntnissprung geführt. Inzwischen wissen wir sehr viel mehr über die Ausstattung der Nervenzellmembran mit Rezeptoren und Ionenkanälen sowie über deren Arbeitsweise, die Funktion von Neurotransmittern, Neuropeptiden und Neurohormonen, den Ablauf intrazellulärer Signalprozesse oder die Entstehung und Fortleitung neuronaler Erregung. Selbst was in einem einzelnen Neuron passiert, können wir mit hoher räumlicher und zeitlicher Auflösung analysieren sowie in Computermodellen simulieren. Dies ist von großer Bedeutung für das grundlegende Verständnis der Arbeitsweise von Sinnesorganen und Nervensystemen sowie für die gezielte Behandlung neurologischer und psychischer Erkrankungen.

Zweifellos wissen wir also heute sehr viel mehr über das Gehirn als noch vor zehn Jahren. Zwischen dem Wissen über die obere und untere Organisationsebene des Gehirns klafft aber nach wie vor eine große Erkenntnislücke. Über die mittlere Ebene – also das Geschehen innerhalb kleinerer und größerer Zellverbände, das letztlich den Prozessen auf der obersten Ebene zugrunde liegt – wissen wir noch erschreckend wenig.

Auch darüber, mit welchen Codes einzelne oder wenige Nervenzellen untereinander kommunizieren (wahrscheinlich benutzen sie gleichzeitig mehrere solcher Codes), existieren allenfalls plausible Vermutungen. Völlig unbekannt ist zudem, was abläuft, wenn hundert Millionen oder gar einige Milliarden Nervenzellen miteinander ”reden“.

Nach welchen Regeln das Gehirn arbeitet; wie es die Welt so abbildet, dass unmittelbare Wahrnehmung und frühere Erfahrung miteinander verschmelzen; wie das innere Tun als ”seine“ Tätigkeit erlebt wird und wie es zukünftige Aktionen plant, all dies verstehen wir nach wie vor nicht einmal in Ansätzen. Mehr noch: Es ist überhaupt nicht klar, wie man dies mit den heutigen Mitteln erforschen könnte. In dieser Hinsicht befinden wir uns gewissermaßen noch auf dem Stand von Jägern und Sammlern.

Die Beschreibung von Aktivitätszentren mit PET oder fMRI und die Zuordnung dieser Areale zu bestimmten Funktionen oder Tätigkeiten hilft hier kaum weiter. Denn dass sich all das im Gehirn an einer bestimmten Stelle abspielt, stellt noch keine Erklärung im eigentlichen Sinne dar. Denn ”wie“ das funktioniert,

darüber sagen diese Methoden nichts, schließlich messen sie nur sehr indirekt, wo in Haufen von hundert Tausenden von Neuronen etwas mehr Energiebedarf besteht. Das ist in etwa so, als versuchte man die Funktionsweise eines Computers zu ergründen, indem man seinen Stromverbrauch misst, während er verschiedene Aufgaben abarbeitet.

Hochdynamische Netzwerke

Vieles spricht dafür, dass neuronale Netzwerke als hochdynamische, nicht-lineare Systeme betrachtet werden müssen. Das bedeutet, sie gehorchen zwar mehr oder weniger einfachen Naturgesetzen, bringen aber aufgrund ihrer Komplexität völlig neue Eigenschaften hervor. Repräsentationen von Inhalten – seien es Wahrnehmungen oder motorische Programme – entsprechen hochkomplexen raumzeitlichen Aktivitätsmustern in diesen neuronalen Netzwerken. Um diesen Signalcode zu entschlüsseln, bedarf es wahrscheinlich paralleler Ableitetechniken, die eine gleichzeitige Messung an vielen Stellendes Gehirns erlauben.

Doch auch wenn viele Geheimnisse noch darauf warten, gelüftet zu werden, hat die Hirnforschung bereits heute einige ganz erstaunliche Erkenntnisse gewonnen. Beispielsweise wissen wir im Wesentlichen, was das Gehirn gut leisten kann und wo es an seine Grenzen stößt. Mit am eindrucksvollsten ist seine enorme Adaptions- und Lernfähigkeit, die – und das ist wohl der überraschendste Punkt – zwar mit dem Alter abnimmt, aber bei weitem nicht so stark wie vermutet. Lange Zeit dachte man, die Hirnentwicklung sei irgendwann in der Jugend abgeschlossen und die neuronalen Netzwerke seien endgültig angelegt. Mittlerweile steht aber fest, dass sich auch im erwachsenen Gehirn zumindest im Kurzstreckenbereich – auf der Ebene einzelner Synapsen – noch neue Verschaltungen bilden können. Außerdem können für bestimmte Aufgaben zusätzliche Hirnregionen rekrutiert werden – etwa beim Erlernen von Fremdsprachen in fortgeschrittenem Alter.

Dank dieser Plastizität kann Hans also durchaus noch lernen, was Hänschen nicht gelernt hat – auch wenn es mit den Jahren deutlich schwerer fällt. Die molekularen und zellulären Faktoren, die der Lern-Plastizität zu Grunde liegen, verstehen wir mittlerweile so gut, dass wir beurteilen können, welche Lernkonzepte – etwa für die Schule – am besten an die Funktionsweise des Gehirns angepasst sind.

Vor allem aus Tierversuchen wissen wir seit einigen Jahren außerdem, dass sich selbst im erwachsenen Gehirn – zumindest an einigen Stellen – noch neue Nervenzellen bilden. Zum jetzigen Zeitpunkt verstehen wir noch nicht, wie sich bei dieser ”Neurogenese“ neue Nervenzellen in alte Verschaltungen einfügen und welche Funktion sie dann über- nehmen. Die Frage, ob sich eine medikamentös induzierte Neurogenese für ursächliche Therapien von neurodegenerativen Erkrankungen einsetzen lässt, können wir daher im Moment noch nicht beantworten.

Die Natur des Geistes

Wir haben herausgefunden, dass im menschlichen Gehirn neuronale Prozesse und bewusst erlebte geistig-psychische Zustände aufs Engste miteinander zusammenhängen und unbewusste Prozesse bewussten in bestimmter Weise vorausgehen. Die Daten, die mit modernen bildgebenden Verfahren gewonnen wurden, weisen darauf hin, dass sämtliche innerpsychischen Prozesse mit neuronalen Vorgängen in bestimmten Hirnarealen einhergehen – zum Beispiel Imagination, Empathie, das Erleben von Empfindungen und das Treffen von Entscheidungen beziehungsweise die absichtsvolle Planung von Handlungen. Auch wenn wir die genauen Details noch nicht kennen, können wir davon ausgehen, dass all diese Prozesse grundsätzlich durch physikochemische Vorgänge beschreibbar sind. Diese näher zu erforschen, ist die Aufgabe der Hirnforschung in den kommenden Jahren und Jahrzehnten.

Geist und Bewusstsein – wie einzigartig sie von uns auch empfunden werden –fügen sich also in das Naturgeschehen ein und übersteigen es nicht. Und: Geist und Bewusstsein sind nicht vom Himmel gefallen,

sondern haben sich in der Evolution der Nervensysteme allmählich herausgebildet. Das ist vielleicht die wichtigste Erkenntnis der modernen Neurowissenschaften.

Was wissen und können Hirnforscher in zehn Jahren?

Was wir in zehn Jahren über den genaueren Zusammenhang von Gehirn und Geist wissen werden, hängt vor allem von der Entwicklung neuer Untersuchungsmethoden ab. Das ”Wo“ im Gehirn, über das uns heute die funktionelle Kernspintomographie Auskunft gibt, sagt uns noch nicht, ”wie“ kognitive Leistungen durch neuronale Mechanismen zu beschreiben sind. Für einen echten Fortschritt in diesem Bereich benötigen wir ein Verfahren, das die Registrierung beider Aspekte in einem ermöglicht.

Wie entstehen Bewusstsein und Ich-Erleben, wie werden rationales und emotionales Handeln miteinander verknüpft, was hat es mit der Vorstellung des ”freien Willens“ auf sich? Die großen Fragen der Neurowissenschaften zu stellen ist heute schon erlaubt – dass sie sich bereits in den nächsten zehn Jahren beantworten lassen, ist allerdings eher unrealistisch. Selbst ob wir sie bis dahin auch nur sinnvoll angehen können, bleibt fraglich. Da-zu müssten wir über die Funktionsweise des Gehirns noch wesentlich mehr wissen.

Sehr wohl aber kann es der Hirnforschung innerhalb der nächsten Dekade gelingen, Erkenntnisse zu erarbeiten, die für Antworten auf diese übergeordneten Fragen entscheidend sein werden. So wollen wir herausfinden, wie Schaltkreise von Hunderten oder Tausenden Neuronen im Verbund des ganzen Gehirns Information codieren, bewerten, speichern und auslesen. Die mittlere Ebene – die Untersuchung der Arbeitsweise von kleineren Bereichen des Nervensystems, von Mikroschaltkreisen – gelangt also zunehmend in den Mittelpunkt der Forschung. Das bisher übliche Verfahren, solche Fragen an Gehirnschnitten zu untersuchen, gehört dann wahrscheinlich der Vergangenheit an, da es nur Momentaufnahmen in einem nicht mehr als Ganzen funktionierenden Schaltwerk darstellen kann. Stattdessen können wir in zehn Jahren wahrscheinlich die räumliche und zeitliche Verteilung von neuronaler Erregung bis auf die Ebene aller beteiligten Neurone in einem Mikroschaltkreis mit bildgebenden Verfahren hoher zeitlicher Auflösung im intakten Nervensystem erfassen. Multiple Photonenmikroskopie, funktionelle Farbstoffe und molekulargenetische Methoden versetzen uns in die Lage, die Regeln des Informationsflusses innerhalb einzelner Neurone und im Verbund von Neuronen zu erkennen.

Voraussetzung für all diese Experimente ist aber, dass die untersuchten Tiere – denn an diesen werden die Versuche vor allem stattfinden – nicht narkotisiert sind und aufgrund schmerzfreier Verfahren ihr natürliches Verhalten zeigen. Nur dann ist es möglich, die Hirnaktivität dieser Tiere beim aktiven Lösen von Aufgaben zu beobachten und dabei die wichtigste Funktion des Gehirns, seine Produktivität und Spontaneität, in die Analyse miteinzubeziehen.

Ganz wesentlich unterstützt wird das Verständnis der Arbeitsweise von Mikroschaltkreisen durch eine detailreiche Modellierung mit Hochleistungsrechnern. Diese Modellierung orientiert sich zukünftig allerdings weniger an den heutigen Konzepten der Informatik und künstlichen Intelligenz als vielmehr an den wirklichen physiologischen Vorgängen. Und zwar nicht nur an denen der unteren Ebene – einzelnen Neuronen mit ihren Ausstattungen an Kanälen und Rezeptoren, ihren wahren Gestalten und ihren plastischen Eigenschaften –, sondern vor allem auch an den neuronalen Prozessen der bisher noch so wenig verstandenen mittleren Ebene, wie sie beim Lernen, beim Erkennen und Planen von Handlungen vorkommen. So wird sich neben der experimentellen Neurobiologie die theoretische Neurobiologie als Forschungsdisziplin durchsetzen, die dann ähnlich wie die theoretische Physik innerhalb der Physik eine große Eigenständigkeit besitzt.

Am Ende der Bemühungen werden die Neurowissenschaften sozusagen das kleine Ein-Mal-Eins des Gehirns verstehen. Daraus lassen sich dann strenge Hypothesen zum Studium übergeordneter Hirnfunktionen ableiten: beispielsweise wie das Gehirn seine zahlreichen Subsysteme so koordiniert, dass kohärente

Wahrnehmungen und koordinierte Aktionen entstehen können. Ohne diesen entscheidenden Zwischenschritt über die „mittlere“ Organisationsebene bleiben die Aussagen über den Zusammenhang zwischen neuronal beobachtbarer Aktivität und kognitiven Leistungen weiterhin spekulativ.

Medizinische Fortschritte

Vor allem was die konkreten Anwendungen angeht, stehen uns in den nächsten zehn Jahren enorme Fortschritte ins Haus. Wahrscheinlich werden wir die wichtigsten molekularbiologischen und genetischen Grundlagen neurodegenerativer Erkrankungen wie Alzheimer oder Parkinson verstehen und diese Leiden schneller erkennen, vielleicht von vornherein verhindern oder zumindest wesentlich besser behandeln können. Ähnliches gilt für einige psychische Krankheiten wie Schizophrenie und Depression. In absehbarer Zeit wird eine neue Generation von Psychopharmaka entwickelt werden, die selektiv und damit hocheffektiv sowie nebenwirkungsarm in bestimmten Hirnregionen an definierten Nervenzellrezeptoren angreift. Dies könnte die Therapie psychischer Störungen revolutionieren – auch wenn von der Entwicklung zum anwendungsfähigen Medikament noch etliche weitere Jahre vergehen werden.

Zudem werden Neuroprothesen wie intelligente Ersatzgliedmaßen oder das künstliche Ohr immer weiter perfektioniert. In zehn Jahren haben wir wahrscheinlich eine künstliche Netzhaut entwickelt, die nicht im Detail programmiert ist, sondern sich nach den Prinzipien des Nervensystems organisiert und lernt. Das wird unseren Blick auf das Sehen, auf die Wahrnehmung, vielleicht auf alle Organisationsprozesse im Gehirn tiefgreifend verändern.

Ebenso werden uns die zu erwartenden weiteren Fortschritte in der Hirnforschung vermehrt in die Lage versetzen, psychische Auffälligkeiten und Fehlentwicklungen, aber auch Verhaltensdispositionen zumindest in ihrer Tendenz vorauszusehen – und ”Gegenmaßnahmen“ zu ergreifen. Solche Eingriffe in das Innenleben, in die Persönlichkeit des Menschen sind allerdings mit vielen ethischen Fragen verbunden, deren Diskussion in den kommenden Jahren intensiviert werden muss.

Was werden Hirnforscher eines Tages wissen und können?

In absehbarer Zeit, also in den nächsten 20 bis 30 Jahren, wird die Hirnforschung den Zusammenhang zwischen neuroelektrischen und neurochemischen Prozessen einerseits und perzeptiven, kognitiven, psychischen und motorischen Leistungen andererseits soweit erklären können, dass Voraussagen über diese Zusammenhänge in beiden Richtungen mit einem hohen Wahrscheinlichkeitsgrad möglich sind. Dies bedeutet, dass man widerspruchsfrei Geist, Bewusstsein, Gefühle, Willensakte und Handlungsfreiheit als natürliche Vorgänge ansehen wird, denn sie beruhen auf biologischen Prozessen.

Eine ”vollständige“ Erklärung der Arbeit des menschlichen Gehirns, das heißt eine durchgängige Entschlüsselung auf der zellulären oder gar molekularen Ebene, erreichen wir dabei dennoch nicht. Insbesondere wird eine vollständige Beschreibung des individuellen Gehirns und damit eine Vorhersage über das Verhalten einer bestimmten Person nur höchst eingeschränkt gelingen. Denn einzelne Gehirne organisieren sich aufgrund genetischer Unterschiede und nicht reproduzierbarer Prägungsvorgänge durch Umwelteinflüsse selbst – und zwar auf sehr unterschiedliche Weise, individuellen Bedürfnissen und einem individuellen Wertesystem folgend. Das macht es generell unmöglich, durch Erfassung von Hirnaktivität auf die daraus resultierenden psychischen Vorgänge eines konkreten Individuums zu schließen.

Im Endeffekt könnte sich eine Situation wie in der Physik ergeben: Die klassische Mechanik hat deskriptive Begriffe für die Makrowelt eingeführt, aber erst mit den aus der Quantenphysik abgeleiteten Begriffen ergab sich die Möglichkeit einer einheitlichen Beschreibung. Auf lange Sicht werden wir entsprechend eine „Theorie des Gehirns“ aufstellen, und die Sprache dieser Theorie wird vermutlich eine andere sein als jene, die wir heute in der Neurowissenschaft kennen. Sie wird auf dem Verständnis der Arbeitsweise von großen

Neuronenverbänden beruhen, den Vorgängen auf der mittleren Ebene. Dann lassen sich auch die schweren Fragen der Erkenntnistheorie angehen: nach dem Bewusstsein, der Ich Erfahrung und dem Verhältnis von erkennendem und zu erkennenden Objekt. Denn in diesem zukünftigen Momentschickt sich unser Gehirn ernsthaft an, sich selbst zu erkennen.

Dann werden die Ergebnisse der Hirnforschung, in dem Maße, in dem sie einer breiteren Bevölkerung bewusst werden, auch zu einer Veränderung unseres Menschenbildes führen. Sie werden dualistische Erklärungsmodelle – die Trennung von Körper und Geist – zunehmend verwischen. Ein weiteres Beispiel: das Verhältnis von angeborenem und erworbenem Wissen. In unserer momentanen Denkweise sind dies zwei unterschiedliche Informationsquellen, die unserem Wahrnehmen, Handeln und Denken zu Grunde liegen. Die Neurowissenschaft der nächsten Jahrzehnte wird aber ihre innige Verflechtung aufzeigen und herausarbeiten, dass auf der mittleren Ebene der Nervennetze eine solche Unterscheidung gar keinen Sinn macht. Was unser Bild von uns Selbst betrifft, stehen uns also in sehr absehbarer Zeit beträchtliche Erschütterungen ins Haus. Geisteswissenschaften und Neurowissenschaften werden in einen intensiven Dialog treten müssen, um gemeinsam ein neues Menschenbild zu entwerfen.

Aller Fortschritt wird aber nicht in einem Triumph des neuronalen Reduktionismus enden. Selbst wenn wir irgendwann einmal sämtliche neuronalen Vorgänge aufgeklärt haben sollten, die dem Mitgefühl beim Menschen, seinem Verliebtsein oder seiner moralischen Verantwortung zugrunde liegen, so bleibt die Eigenständigkeit dieser „Innenperspektive“ dennoch erhalten. Denn auch eine Fuge von Bach verliert nichts von ihrer Faszination, wenn man genau verstanden hat, wie sie aufgebaut ist. Die Hirnforschung wird klar unterscheiden müssen, was sie sagen kann und was außerhalb ihres Zuständigkeitsbereichs liegt, so wie die Musikwissenschaft – um bei diesem Beispiel zu bleiben – zu Bachs Fuge Einiges zu sagen hat, zur Erklärung ihrer einzigartigen Schönheit aber schweigen muss.

Anhang 3: Bewusstseinstheorien

Tabelle 5: Eine Auswahl von Bewusstseinstheorien nach Seth Bayne[915]

Theory	Primary claim
Higher-order theory (HOT)	Consciousness depends on meta-representations of lower-order mental states
Self-organizing meta- representational theory	Consciousness is the brain's (meta-representational) theory about itself
Attended intermediate representation theory	Consciousness depends on the attentional amplification of intermediate-level representations
Global workspace theories (GWTs)	Consciousness depends on ignition and broadcast within a neuronal global workspace where fronto-parietal cortical regions play a central, hub-like role
Integrated information theory (IIT)	Consciousness is identical to the cause–effect structure of a physical substrate that specifies a maximum of irreducible integrated information
Information closure theory	Consciousness depends on non-trivial information closure with respect to an environment at particular coarse-grained scales
Dynamic core theory	Consciousness depends on a functional cluster of neural activity combining high levels of dynamical integration and differentiation
Neural Darwinism	Consciousness depends on re-entrant interactions reflecting a history of value-dependent learning events shaped by selectionist principles
Local recurrency	Consciousness depends on local recurrent or re-entrant cortical processing and promotes learning
Predictive processing	Perception depends on predictive inference of the causes of sensory signals; provides a framework for systematically mapping neural mechanisms to aspects of consciousness
Neuro-representationalism	Consciousness depends on multilevel neurally encoded predictive representations
Active inference	Although views vary, in one version consciousness depends on temporally and counterfactually deep inference about self-generated actions
Beast machine theory	Consciousness is grounded in allostatic control-oriented predictive inference
Neural subjective frame	Consciousness depends on neural maps of the bodily state providing a first-person perspective
Self comes to mind theory	Consciousness depends on interactions between homeostatic routines and multilevel interoceptive maps, with affect and feeling at the core
Attention schema theory	Consciousness depends on a neurally encoded model of the control of attention
Multiple drafts model	Consciousness depends on multiple (potentially inconsistent) representations rather than a single, unified representation that is available to a central system
Sensorimotor theory	Consciousness depends on mastery of the laws governing sensorimotor contingencies
Unlimited associative learning	Consciousness depends on a form of learning which enables an organism to link motivational value with stimuli or actions that are novel, compound and non-reflex inducing
Dendritic integration theory	Consciousness depends on integration of top-down and bottom-up signalling at a cellular level
Electromagnetic field theory	Consciousness is identical to physically integrated, and causally active, information encoded in the brain's global electromagnetic field
Orchestrated objective reduction	Consciousness depends on quantum computations within microtubules inside neurons

[915] Übernommen aus Seth Bayne (2022), S. 441, Table 1

Literaturverzeichnis

Achtner, Wolfgang (2006): „Einleitung“. In: Achtner, W. et al. (Hrsg.): *Künstliche Intelligenz und menschliche Person*. 2006. Marburg: N.G. Elwert, S. 1-12
(zit. Achtner (2006), Seite)

Adorno, Theodor W. (1951): *Minima Moralia. Reflexionen aus dem beschädigten Leben*. 12. Auflage 2019. Frankfurt: Suhrkamp
(zit. Adorno (1951), Seite)

Adorno, Theodor W.; **Rabinbach**, Anson G. (1975): „Culture Industry Reconsidered“. In: *New German Critique.* Autumn 1975No. 6. Durham, North Carolina: Duke University Press. S. 12-19
(zitiert Adorno Rabinbach (1975), Seite)

Alvesson, Mats; **Sandberg**, Jörgen (2011): “Generating research questions through problematization”. In: *Academy of Management Review*. 2011. Vol. 36. No. 2, S. 247-271
(zit. Alvesson Sandberg (2011), Seite)

Alvesson, Mats; **Sandberg**, Jörgen (2020): “The problematizing review: A counterpoint to Elsbach and Van Knippenberg’s argument for integrative reviews”. In: *Journal of Management Studies*. 2020. 57 (6), 1290-1304
(zit. Alvesson Sandberg (2020), Seite)

Amoroso, Daniele (2020): Autonomous Weapon Systems and International Law. A Study on Human-Machine Interactions in Ethically and Legally Sensitive Domains. 2020. Napoli: Edizioni Scientificiche Italiane NOMOS
(zit. Amoroso (2020), Seite)

Arendt, Hannah (1951): *The origins of totalitarianism*. 1951. United Kingdom: Penguin Random House
(zit. Arendt (1951), Seite)

Arendt, Hannah (1958): *The human condition*. 2018. edition with foreword by Danielle Allen and introduction by Margaret Canovan. Chicago: The University of Chicago Press
(zit. Arendt (1958), Seite)

Arendt, Hannah (1967): *Vita Activa. Oder vom tätigen Leben*. 18. Taschenbuchauflage 2016. München/Berlin: Piper
(zit. Arendt (1967), Seite)

Arendt, Hannah (1970): *Macht und Gewalt*. 28. Taschenbuchauflage 2021. München/Berlin: Piper
(zit. Arendt (1970), Seite)

Arendt, Hannah (1986): *Elemente und Ursprünge totaler Herrschaft*. 20. Auflage 2017. München/Berlin: Piper
(zit. Arendt (1986), Seite)

Arkin, Ronald C. (2013): “Legal Autonomous Systems and the Plight of the Non-combatant”. In: *The Political Economy of Robots. Prospects for Prosperity and Peace in the Automated 21st Century*. Ryan Higgins (Hrsg.). 2018. Cham, Schweiz: Palgrave Macmillan. S. 317-326
(zit. Arkin (2013), Seite)

Asaro, Peter (2012): „On banning autonomous weapon systems.: human rights, automation, and the dehumanization of lethal decision-making”. In: *International Review of the Red Cross*. Vol. 94, Nr. 886, Summer 2012. S. 687-709
(zit. Asaro (2012), Seite)

Augstein, Jakob (Hrsg., 2017): *Reclaim Autonomy. Selbstermächtigung in der digitalen Weltordnung.* 2017. Berlin: Suhrkamp
(zit. Augstein (2017), Seite)

Augustinus (2009): *Confessiones. Bekenntnisse. Lateinisch/Deutsch*. Kurt Flasch, Burkhard Mojsisch (Hrsg.). 2009. Stuttgart: Reclam
(zit. Augustinus (2009), Seite)

Barinaga, Marcia (1990): "The mind reveled? Some neuroscientists think that recently discovered oscillations of electrical potential at 40 hertz hold the key the brain assembles sense impressions into a single object". In: Science. August 24th 1990. Vol. 249, Issue 4971. S. 856-858
(zit. Barinaga (1990), Seite)

Bartneck, Christoph; **Lütge**, Christoph; **Wagner**, Alan; **Welsh**, Sean (2019): *Ethik in KI und Robotik***.** 2019. München: Carl Hanser
(zit. Bartneck Lütge (2019), Seite)

Barz, Wolfgang (2006): „Naturalisierung der Intentionalität. Ein philosophischer Holzweg". In: *Deutsche Zeitschrift für Philosophie*, Berlin 54 (2006), S. 189-200
(zit. Barz (2006), Seite)

Bayertz, Kurt (1995): „Eine kurze Geschichte der Herkunft der Verantwortung". In: *Verantwortung. Prinzip oder Problem?* Kurt Bayertz (Hrsg.). 1995. Darmstadt: Wissenschaftliche Buchgesellschaft: S. 3-71
(zit. Bayertz (1995), Seite)

Becchi, Paolo (2013): *Das Prinzip Menschenwürde – eine Abhandlung*. 2013. Berlin: Duncker & Humblot
(zit. Becchi (2013), Seite)

Beck, Hanno (2014): *Behavioral Economics. Eine Einführung*. 2014. Wiesbaden: Springer Gabler
(zit. Beck (2014), Seite)

Beckermann, Ansgar (2008): *Das Leib-Seele-Problem. Eine Einführung in die Philosophie des Geistes*. 2. durchgesehene Auflage 2011. Paderborn: Ferdinand Schöningh
(zit. Beckermann (2008), Seite)

Beiersdörfer, Kurt (2003): *Was ist Denken?* 2003. Paderborn: Ferdinand Schöningh
(zit. Beiersdörfer (2003), Seite)

Bendel, Oliver (Herausgeber, 2019): *Handbuch Maschinenethik.* 2019. Wiesbaden: Springer
(zit. Bendel (2019), Seite)

Bender, Emily M.; **Koller**, Alexander (2020): "Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data". *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. July 2020. S. 5185-5198
(zit. Bender Koller (2020), Seite)

Bhagoji, Arjun Nitin; **He**, Warren; **Li**, Bo; **Song**, Dawn (2018): *Practical Black-box Attacks on Deep Neural Networks using Efficient Query Mechanisms*. In: Proceedings of the Computer Vision – ECCV 2018. https://link.springer.com/conference/eccv
(zit. Bhagoji et al. (2018), Seite)

Bieri, Peter (Hrsg., 1981): *Analytische Philosophie des Geistes*. 4. Neu ausgestattete Auflage 2007. Weinheim und Basel: Beltz Verlag
(zit. Bieri (1981), Seite)

Block, Ned (1995): „Eine Verwirrung über eine Funktion des Bewusstseins“. In: *Bewusstsein*. Hrsg. Metzinger, Thomas. 5. Auflage 2005. Paderborn: mentis. S. 523-581
(zit. Block (1995), Seite)

Boddington, Paula (2017): *Towards a Code of Ethics for Artificial Intelligence*. 2017. Cham, Switzerland: Springer International Publishing
(zit. Boddington (2017), Seite)

Böhler, Dietrich; **Herrmann**, Bernadette (Hrsg., 2015): „Hans Jonas. Das Prinzip Verantwortung. Grundlegung“. Band I/2 Erster Teilband von *Kritische Gesamtausgabe der Werke von Hans Jonas*. 2015. Freiburg: Rombach
(zit. Böhler Herrmann (2015), Seite)

Böhler, Dietrich; **Herrmann**, Bernadette (Hrsg., 2017): „Hans Jonas. Das Prinzip Verantwortung. Tragweite und Aktualität einer Zukunftsethik“. Band I/2 Zweiter Teilband von *Kritische Gesamtausgabe der Werke von Hans Jonas*. 2017. Freiburg: Rombach
(zit. Böhler Herrmann (2017), Seite)

Böhme, Gernot (2009): Der mündige Mensch. Denkmodelle der Philosophie, Geschichte, Medizin und Rechtswissenschaft. 2009. Darmstadt: WBG
(zit. Böhme (2009), Seite)

Bongardt, Michael; **Burckhart**, Holger; **Gordon**, John-Stewart; **Nielsen-Sikora**, Jürgen (Hg., 2021): *Hans Jonas – Handbuch. Leben – Werk – Wirkung*. 2021. Berlin: J.B.Metzler Verlag
(zit. Bongardt et al. (2021), Seite)

Bostrom, Nick (2014): *Superintelligenz. Szenarien einer kommenden Revolution*. 3. Auflage 2018. Berlin: Suhrkamp
(zit. Bostrom (2014), Seite)

Bostrom, Nick (2018): *Die Zukunft der Menschheit. Aufsätze*. 1. Auflage 2018. Berlin: Suhrkamp
(zit. Bostrom (2018), Seite)

Brandt, Horst D. (Hg., 1999): *Immanuel Kant. Was ist Aufklärung? Ausgewählte kleine Schriften*. Nachdruck 2019. Hamburg: Felix Meiner
(zit. Brandt (1999), Seite)

Brendel, Elke (2006): „Mit Gödel zum Antimechanismus? Versuch einer Bestandsaufnahme der Debatte um die Implikationen der Gödelschen Unvollständigkeitstheoreme für die KI“. In: Achtner, W. et al. (Hrsg.): *Künstliche Intelligenz und menschliche Person*. 2006. Marburg: N.G. Elwert, S. 39-53
(zit. Brendel (2006), Seite)

Brown, Richard E. (2016): „Hebb and Cattell: The Genesis of the Theory of Fluid and Crystallized Intelligence”. In: *Frontiers in Human Neuroscience*. December 2016, Vol. 10, Article 606. https://www.frontiersin.org/articles/10.3389/fnhum.2016.00606/full
(zit. Brown (2016), Seite)

Brüntrup, Godehard (2003): „Zur Kritik des Funktionalismus“. In: *Ist der Geist berechenbar. Philosophische Reflexionen*. Hrsg.: Köhler, Wolfgang R.; Mutschler, Hans-Dieter. 2003. Darmstadt: Wissenschaftliche Buchgesellschaft. S. 58ff
(zit. Brüntrup (2003), Seite)

Brüntrup, Godehard (2018): *Philosophie des Geistes. Eine Einführung in das Leib-Seele-Problem*. 2018. Stuttgart: Kohlhammer
(zit. Brüntrup (2018), Seite)

Buddeberg, Eva (2011): Verantwortung im Diskurs – Grundlinien einer rekonstruktiv-hermeneutischen Konzeption moralischer Verantwortung im Anschluss an Hans Jonas, Karl-Otto Apel Emmanuel Lévinas. 2011. Berlin/Boston: de Gruyter
(zit. Buddeberg (2011), Seite)

Bunge, Mario (1977): „State and Events". In: W. E. Hartnett (Hrsg.). *Systems: Approaches, Theories, Applications*. 1977. Boston: Reidel
(zit. Bunge (1977), Seite)

Bunge, Mario (1980): *The mind-body problem. A psychobiological approach.* 1980. Oxford, New York, Sydney, Paris, Frankfurt: Pergamon Press
(zit. Bunge (1980), Seite)

Bunge, Mario (1984): *Das Leib-Seele-Problem. Ein psychobiologischer Versuch.* 1984, 1. Deutsche Ausgabe. Tübingen: J.C.B. Mohr
(zit. Bunge (1984), Seite)

Bunge, Mario (1993): **"**Realism and antirealism in social science*".* In: *Theory and Decision*. 1993. 35, S. 207-235. https://doi.org/10.1007/BF01075199
(zit. Bunge (1993), Seite)

Bunge, Mario (2010): *Matter and Mind. A Philosophical Inquiry*. 2010. Heidelberg: Springer
(zit. Bunge (2010), Seite)

Bunge, Mario (1999): *The philosophy-sociology connection*. 1999. New Brunswick (USA), London (UK): Transaction Publishers
(zit. Bunge (1999), Seite)

Bunge, Mario; **Ardila**, Ruben (1987): *Philosophy of Psychology*. 1987. Heidelberg: Springer
(zit. Bunge Ardila (1987), Seite)

Bunge, Mario; **Ardila**, Ruben (1990): *Philosophie der Psychologie*. 1990. Tübingen: Mohr
(zit. Bunge Ardila (1990), Seite)

Bunge, Mario; **Llinás**, Rudolfo (1978): „The mind-body-problem in the light of contemporary neurobiology". In: *16th World Congress of Philosophy, Section Papers.* 1978. S. 131-133
(zit. Bunge Llinás (1978), Seite)

Burge, Tyler (1995): „Zwei Arten von Bewußtsein". In: *Bewusstsein*. Hrsg. Metzinger, Thomas. 5. Auflage 2005. Paderborn: mentis. S. 583-594
(zit. Burge (1995), Seite)

Burge, Tyler (2007): *Foundations of Mind*. 2009. New York: Oxford University Press
(zit. Burge (2007), Seite)

Busch, Christoph; **De Franceschi**, Alberto (Hrsg., 2021): *Algorithmic Regulation and Personalized Law. A Handbook.* 2021. München: C.H. Beck
(zit. Busch De Franceschi (2021), Seite)

Buxmann, Peter; **Schmidt**, Holger (2019): „Grundlagen der Künstlichen Intelligenz und des Maschinellen Lernens*"*.In: *Künstliche Intelligenz. Mit Algorithmen zum wirtschaftlichen Erfolg.* Herausgegeben von Buxmann Schmidt. 2019. Berlin: Springer. S. 3-20
(zit. Buxmann Schmidt (2019), Seite)

Callaghan, Victor; **Miller**, James; **Yampolskiy**, Roman; **Armstrong**, Stuart (Hrsg., 2017): *The technological singularity. Managing the journey*. Berlin: Springer
(zit. Callaghan et al. (2017), Seite)

Cao, Y., **Li**, S., **Liu**, Y., **Yan**, Z., **Dai**, Y., **Yu**, P. S., & **Sun**, L. (2023): *A comprehensive survey of ai-generated content (aigc): A history of generative ai from gan to chatgpt*. arXiv preprint

arXiv:2303.04226
(zit. Cao et al. (2023), Seite)

Cassirer, Ernst (1918): *Kants Leben und Lehre*. Nachdruck von 1994. Darmstadt: Wissenschaftliche Buchgesellschaft
(zit. Cassirer (1918), Seite)

Cassirer, Ernst (1923): „Zur Einführung“ In: *Immanuel Kant. Was ist Aufklärung? Ausgewählte kleine Schriften*. Hrsg. Brandt, Horst. Nachdruck 2019. Hamburg: Felix Meiner. S. 3-19
(zit. Cassirer (1923), Seite)

Cattell, Raymond B. (1941): “Some theoretical issues in adult intelligence testing”. *Psychol. Bulletin*. 1941. 38, S. 592
(zit. Cattell (1941), Seite)

Cattell, Raymond B. (1943): “The measurement of adult intelligence”. *Psychol. Bulletin*. 1943. 3. S. 153-193
(zit. Cattell (1943), Seite)

Cattell, Raymond B. (1963): “Theory of fluid and crystallized intelligence.: a critical experiment”. In: *Journal of educational psychology*. 1963. Vol 54, No. 1, S. 1-22
(zit. Cattell (1963), Seite)

Chalmers, David J. (1996): *The conscious mind. In search of a Fundamental Theory*. 1996. New York: Oxford University Press
(zit. Chalmers (1996), Seite)

Chalmers, David J. (2000): “What is a Neural Correlate of Consciousness?”. In Thomas Metzinger (Herausgeber): *Neural Correlates of Consciousness: Empirical and Conceptual Questions*. Cambridge MA: MIT Press. S. 17–39
zit. Chalmers (2000), Seite)

Chalmers, David (2010): “The Singularity. A philosophical Approach”. In: *Journal of Consciousness Studies*. 2010. Bd. 17, S. 7-65
(zit. Chalmers (2010), Seite)

Chalmers, David (2022): *Could a Large language Model be Conscious?* November 28, 2022. Transcripted talk given at NeurIPS conference in New Orleans. https://philpapers.org/archive/CHACAL-3.pdf
(zit. Chalmers (2022), Seite)

Christaller, Thomas; **Wehner**, Josef (Herausgeber, 2003): *Autonome Maschinen*. 2003. Wiesbaden: Westdeutscher Verlag
(zit. Christaller Wehner (2003), Seite)

Christian, Brian (2020): *The Alignment Problem. Machine Learning and Human Values.* 2020. New York: W.W. Norton&Company
(zit. Christian (2020), Seite)

Christley, Ron (2008): “Philosophical Foundations of Artificial Consciousness”. In: *Artificial Intelligence in Medicine*. 2008. Vol. 44. Elsevier. S. 119-137
(zit. Christley (2008), Seite)

Smith **Churchland**, Patricia (1995): „Die Neurobiologie des Bewusstseins“. In: *Bewusstsein*. Hrsg. Metzinger, Thomas. 5. Auflage 2005. Paderborn: mentis. S. 463ff
(zit. Churchland (1995), Seite)

Crane, Tim (2007): Intentionalität als Merkmal des Geistigen. Sechs Essays zur Philosophie des Geistes. 2007. Frankfurt am Main: Fischer
(zit. Crane (2007), Seite)

Cruse, Holk; **Dean**, Jeffrey, **Ritter**, Helge (1998): *Die Entdeckung der Intelligenz. Oder können Ameisen denken?* Taschenbuchausgabe von 2001. München: dtv
(zit. Cruse Dean Ritter (1998), Seite)

Cruse, Holk; **Dean**, Jeffrey, **Ritter**, Helge (1999): „Was ist Intelligenz?" In: *Intelligenz zwischen Mensch und Maschine. Von der Hirnforschung zur künstlichen Intelligenz*. Herausgeber: Wellmann, Karl-Heinz; Thimm, Utz. 1999. Münster: LIT. S. 92ff
(zit. Cruse Dean Ritter (1999), Seite)

Crush, Rick; **Smith Churchland**, Patricia (1995): „Lücken im Penrose-Parkett". In: Metzinger, Thomas (Hrsg.): *Bewusstsein. Beiträge aus der Gegenwartsphilosophie*. 2. durchgesehene Auflage 1996. Paderborn: Ferdinand Schöningh. S. 221 - 249
(zit. Crush Churchland (1995), Seite)

Damasio, Antonio (2017): *Im Anfang war das Gefühl. Der biologische Ursprung menschlicher Kultur*. 2017. München: Siedler
(zit. Damasio (2017), Seite)

Dannemann, Rüdiger (2015): „Nachwort". In: Rüdiger Dannemann (Hrsg.). *Georg Lukács – Werksauswahl in Einzelbänden*. Band 3. 2015. Bielefeld: Aisthesis Verlag. S. 177 - 218
(zit. Dannemann (2015), Seite)

Decher, Friedhelm (2015): *Handbuch der Philosophie des Geistes*. 2015. Darmstadt: Wissenschaftliche Buchgesellschaft
(zit. Decher (2015), Seite)

Decker, Michael (2002): „Robotik. Perspektiven für menschliches Handeln in der zukünftigen Gesellschaft". In: *Technikfolgenabschätzung – Theorie und Praxis*. Nr.2, 11. Jg., Juli 2002
(zit. Decker (2002), Seite)

Decker, Michael (2008): „Caregiving robots and ethical reflection: the perspective of interdisciplinary technology assessment". In: *AI&Society* 22 315 – 330
(zit. Decker (2008), Seite)

Denker, Alfred (2014): "Heimat, Technik und Gelassenheit auf Heideggers Denkweg. Eine Spurensuche". In: Martin Heidegger: *Gelassenheit. Heideggers Meßkircher Rede von 1955*. 2. Auflage 2015. München: Verlag Karl Alber. S. 41 - 70
(zit. Denker (2014), Seite)

Dennett, Daniel C. (1998): *Brainchildren: essays on designing minds*. 1998. Cambridge, Massachusetts: A Bradford book. The MIT Press
(zit. Dennett (1998), Seite)

Dennett, Daniel C. (2006): „Qualia eliminieren". In: *Grundkurs Philosophie des Geistes. Band1: Phänomenales Bewusstsein*. Hrsg.: Metzinger, Thomas. Auflage 2009. Paderborn: Mentis. S. 205-249
(zit. Dennett (2006), Seite)

Dennett, Daniel C. (2017): „Die Singularität – Ein moderne Legende?" In: „*Was sollen wir von Künstlicher Intelligenz halten?*". Herausgegeben von John Brockman. 2017. Frankfurt am Main: Fischer. S. 123 - 127
(zit. Dennett (2017), Seite)

Dennett, Daniel C. (2018): *Von den Bakterien zu Bach und zurück. Die Evolution des Geistes*. 2. Auflage 2018. Berlin: Suhrkamp
(zit. Dennett (2018), Seite)

Descartes, René (1637): *Discours de la Méthode. Bericht über die Methode*. Französisch / deutsche Ausgabe. 2022. Stuttgart: Reclam
(zit. Descartes (1637), Seite)

Descartes, René (2009): *Meditationen*. neu übersetzte Ausgabe von 2009. Hamburg: Meiner Verlag
(zit. Descartes (2009), Seite)

Deutscher Ethikrat (Hrsg., 2020): *Robotik für gute Pflege. Stellungnahme.* 2020. Berlin: Deutscher Ethikrat
(zit. Deutscher Ethikrat (2020), Seite)

Deutscher Ethikrat (Hrsg., 2023): *Mensch und Maschine – Herausforderungen durch Künstliche Intelligenz. Stellungnahme.* Nicht lektorierte Fassung vom 20.3.2023. Berlin: Deutscher Ethikrat
(zit. Deutscher Ethikrat (2023), Seite)

Domingos, Pedro (2015): The master algorithm. How the quest for the ultimate learning machine will remake our world. 2017. UK: Penguin Random House
(zit. Domingos (2015), Seite)

Dreier, Horst (2013): „Artikel 1 I. Menschenwürde" In: H. Dreier (Hrsg.). *Grundgesetz Kommentar*. Band I. 3. Auflage. 2013. Tübingen: Mohr-Siebeck. S. 154 - 261
(zit. Dreier (2013), Seite)

Dretske, Fred (1980): „The intentionality of cognitive states". In *Midwest Studies in Philosophy*. Vol. 5, 1980. *Studies in Epistemology*. S. 281 – 284
(zit. Dretske (1980), Seite)

Du Bois-Reymond, Emil (1872): *Über die Grenzen des Naturerkennens.* Ein Vortrag in der zweiten öffentlichen Sitzung der 45. Versammlung deutscher Naturforscher und Ärzte zu Leipzig. 2. Auflage 1872. Leipzig: Verlag von Veit & Comp.
(zit. Du Bois-Reymond (1872), Seite)

Du Sautoy, Marcus (2019): *The creativity code. Art and Innovation in the age of AI*. 2019. Cambridge, Massachusetts: Harvard University Press
(zit. Du Sautoy (2019), Seite)

Edelman, Gerald M. (1989): *The remembered present. A biological theory of consciousness*. 1989. New York: Basic Books
(zit. Edelman (1989), Seite)

Edelman, Gerald M., **Tononi**, Giulio (2000): *Gehirn und Geist. Wie aus Materie Bewusstsein entsteht*. Deutsche Ausgabe 2002. München: Beck
(zit. Edelman Tononi (2000), Seite)

Edelman, Gerald M. (2004): *Das Licht des Geistes. Wie Bewusstsein entsteht*. Taschenbuchausgabe 2007. Hamburg: Rowohlt Taschenbuch Verlag
(zit. Edelman (2004), Seite)

Elm, Ralf (2021): „Hans Jonas und Martin Heidegger. Nähe und Distanz in Ontologie und Verantwortungsdenken". Langfassung des Artikels „Heidegger" in Michael Bongardt u.a. (Hrsg.): *Hans Jonas-Handbuch*. https://www.hans-jonas-zentrum.de/files/cto_layout/downloads/Langtext-Jonas-und-Heidegger_Ralf-Elm-2020.pdf
(zit. Elm (2021), Seite)

Erdmann, Eva; **Forst**, Rainer; **Honneth**, Axel (Hrsg., 1990): *Ethos der Moderne. Foucaults Kritik der Aufklärung*. 1990. Frankfurt/Main: Campus Verlag
(zit. Erdmann et al. (1990), Seite)

Ertel, Wolfgang (2008): *Grundkurs Künstliche Intelligenz. Eine praxisorientierte Einführung*. 4. Auflage 2016. Wiesbaden: Springer Vieweg
(zit. Ertel (2008), Seite)

Europäische Kommission (2019): *Ethik Leitlinien für vertrauenswürdige KI. Hochrangige Expertengruppe für Künstliche Intelligenz.* Abschlussbericht in deutscher Sprache. 2019
(zit. EU Ethik-Leitlinien für eine vertrauenswürdige KI (2019), Seite)

Ewert, Dirk (2006): „Die Gödelschen Theoreme und die Frage nach der Künstlichen Intelligenz in theologischer Sicht". In: Achtner, W. et al. (Hrsg.): *Künstliche Intelligenz und menschliche Person*. 2006. Marburg: N.G. Elwert, S. 55-76
(zit. Ewert (2006), Seite)

Fahrenberg, Jochen (2008): „Gehirn und Bewusstsein. Neuro-philosophische Kontroversen". In: S. Gauggel und M. Herrmann (Hrsg.). *Handbuch der Neuro- und Biopsychologie*. 2008. Göttingen: Hogrefe
(zit. Fahrenberg (2008), Seite)

Falkenburg, Brigitte (2012): *Mythos Determinismus. Wieviel erklärt uns die Hirnforschung*? 1. Auflage 2012. Berlin Heidelberg: Springer Spektrum
(zit. Falkenburg (2012), Seite)

Feil, Ernst (1982): „Autonomie und Heteronomie nach Kant". In: *Freiburger Zeitschrift für Philosophie und Theologie*. Band 29 (1982), Heft 3. S. 389-441
(zit. Feil (1982), Seite)

Fichte, Johann Gottlieb (1796): *Grundlage des Naturrechts nach Prinzipien der Wissenschaftslehre.* Herausgegeben von Manfred Zahn. Digitaldruckausgabe von 1991. Hamburg: Meiner
(zit. Fichte (1796), Seite)

Fichte, Johann Gottlieb (1798/99): *Wissenschaftslehre nova methodo*. Kollegnachschift K. Chr. Fr. Krause. Herausgeber: Erich Fuchs. 2., verb. Ausgabe 1994. Hamburg: Meiner
(zit. Fichte (1798/99), Seite)

Fischer, Klaus (2003): „Drei Grundirrtümer der Maschinentheorie des Bewußtseins". In: *Ist der Geist berechenbar? Philosophische Reflexionen*. Herausgeber: Köhler, Wolfgang R.; Mutschler, Hans-Dieter. 2003. Darmstadt: Wissenschaftliche Buchgesellschaft. S. 33ff
(zit. Fischer (2003), Seite)

Fitzi, Gregor; **Matsuzaki**, Hironori (2013): Menschenwürde und Roboter. In: *Menschenwürde und Medizin. Ein interdisziplinäres Handbuch*. Herausgeber: Joerden, Jan. 2013. Berlin: Duncker und Humblot. S. 919-931
(zit. Fitzi Matsuzaki (2013), Seite)

Flashar, Hellmut (2013): *Aristoteles. Lehrer des Abendlandes*. 3. Auflage 2014. München: C.H. Beck
(zit. Flashar (2013), Seite)

Fleischacker, Samuel (2013): *What is Enlightenment? Kant's questions*. 2013. Milton Park, New York: Routledge
(zit. Fleischacker (2013), Seite)

Fleischacker, Samuel (2018): "Kant in the Dialectic of Enlightenment". In: *Aufklärungs-Kritik und Aufklärungs-Mythen. Horkheimer und Adorno in philosophischer Perspektive.* Sonja Lavaert, Winfried Schröder (Hrsg.). 2018. Berlin/Boston: De Gruyter, S. 123 – 142
(zit. Fleischacker (2018), Seite)

Ford, Martin (2019): Die Intelligenz der Maschinen. Mit Koryphäen der Künstlichen Intelligenz im Gespräch. Innovationen, Chancen und Konsequenzen für die Zukunft der Gesellschaft. Erste deutsche Auflage. 2019. Frechen: mitp
(zit. Ford (2019), Seite)

Foerster, Heinz von (1993a): „Zukunft der Wahrnehmung: Wahrnehmung der Zukunft". In: *Wissen und Gewissen.* Siegfried J. Schmidt (Hrsg.). 10. Auflage 2019. Frankfurt am Main: Suhrkamp, S. 194-210
(zit. Foerster (1993a), Seite)

Foerster, Heinz von (1993b): „Prinzipien der Selbstorganisation im sozialen und betriebswirtschaftlichen Bereich". In: *Wissen und Gewissen.* Siegfried J. Schmidt (Hrsg.). 10. Auflage 2019. Frankfurt am Main: Suhrkamp, S. 233-268
(zit. Foerster (1993b), Seite)

Foerster, Heinz von (1995): „Entdecken oder Erfinden. Wie läßt sich Verstehen verstehen?". In: *Einführung in den Konstruktivismus*. München/Zürich: Piper, S. 60-67
(zit. Foerster (1995), Seite)

Försterling, Wolfram (2016): Aufklärung oder Unmündigkeit. Wieweit strahlt das Licht der Vernunft? Die Werte der Aufklärung im 21. Jahrhundert. 2016.Baden-Baden: Nomos
(zit. Försterling (2016), Seite)

Foucault, Michel (1990): „Was ist Aufklärung". In: *Ethos der Moderne. Foucaults Kritik der Aufklärung.* Erdmann, Eva; Forst, Rainer; Honneth, Axel (Hg.). 1990. Frankfurt: Campus Verlag. S. 35-54
(zit. Foucault (1990), Seite)

Frické, Martin (2015): „Big Data and its Epistemology". In: *Journal of the associataion for information science and technology*. 2015. 66(4). S. 651-661
(zit. Frické (2015), Seite)

Fromm, Erich (1945): *Die Furcht vor der Freiheit*. 25. Auflage 2021. München: dtv
(zit. Fromm (1945), Seite)

Funke, Joachim (2022): „Was ist Intelligenz? Die psychologische Sicht". In: *Künstliche Intelligenz. Macht der Maschinen und Algorithmen zwischen Utopie und Realität*. Hrsg.: Alfred Krabbe, Herrmann Michael Niemann, Thomas von Woedtke. 2022. Leipzig: Evangelische Verlagsanstalt. S. 87-110
(zit. Funke (2022), Seite)

Gabriel, Markus (2015): *Ich ist nicht Gehirn. Philosophie des Geistes für das 21. Jahrhundert*. 2. Auflage 2018. Berlin: Ullstein
(zit. Gabriel (2015), Seite)

Gabriel, Markus (2018): *Der Sinn des Denkens*. 2018. Berlin: Ullstein
(zit. Gabriel (2018), Seite)

Galloway, Scott (2017): *The Four. Die geheime DNA von Amazon, Apple, Facebook und Google*. 4. Auflage 2019. Kulmbach: Börsenmedien/Plassen Verlag
(zit. Galloway (2017), Seite)

Gardner, Howard (2002): *Intelligenzen. Die Vielfalt des menschlichen Geistes.* 2002. Stuttgart: Klett-Cotta
(zit. Gardner (2002), Seite)

Gessmann, Martin (2009): *Philosophisches Wörterbuch*. 23., vollständig neu bearbeitete Auflage. 2009. Stuttgart: Kröner
(zit. Gessmann (2009), Seite)

Geyer, Christian (2004, Hrsg.): *Hirnforschung und Willensfreiheit. Zur Deutung der neuesten Experimente*. 9. Auflage 2016. Frankfurt: Suhrkamp
(zit. Geyer (2004), Seite)

Gisbertz, Philipp (2017): „Menschenwürde in der angloamerikanischen Rechtsphilosophie. Ein Vergleich zur kontinentaleuropäischen Begriffsbildung". In: *Studien zur Rechtsphilosophie und Rechtstheorie*. Hrsg.: Ralf Dreier, Robert Alexy, Martin Borowski. Band 70. 2017. Baden-Baden: Nomos
(zit. Gisbertz (2017), Seite)

Gödel, Kurt (1931): „Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme". In: *Monatshefte für Mathematik und Physik* 38. 1931. S. 173–198
(zit. Gödel (1931), Seite)

Gödel, Kurt (1932): „Über Vollständigkeit und Widerspruchsfreiheit". In: *Kurt Gödel. Collected Works*. Vol. I, Publications 1929-1936. Feferman et al. (Hrsg.). 1986. Oxford: Oxford University Press, S. 234-236
(zit. Gödel (1932), Seite)

Good, Irving John (1965): „Speculations Concerning the Forst Ultraintelligent Machines". In: *Advances in Computers*, Vol. 6 (1965) 31ff
(zit. Good (1965), Seite)

Görz, Günther; **Schneeberger**, Josef; **Schmid**, Ute (Hrsg., 2003): *Handbuch der Künstlichen Intelligenz*. 5., überarbeitete und aktualisierte Auflage 2014. München: Oldenbourg
(zit. Görz Schneeberger Schmid (2003), Seite)

Görz, Günther, **Schmid**, Ute; **Braun**, Tanya (Hrsg., 2021): *Handbuch der Künstlichen Intelligenz*. 6. Neuauflage 2021. Berlin: de Gruyter Oldenbourg
(zit. Görz et al. (2021), Seite)

Gregory, Richard (1994): "Seeing Intelligence". In: *What is intelligence?* Edited by Jean Khalfa. 1994. Cambridge: Cambridge University Press, S. 13-26
(zit. Gregory (1994), Seite)

Grunwald, Armin (Herausgeber, 2013): *Handbuch Technikethik*. 2013. Stuttgart: J.B. Metzler
(zit. Grunwald (2013), Seite)

Guilford, J.P. (1950): „Creativity". In: *American Psychologist*. 5(9). 1950. S. 444-454
(zit. Guilford (1950), Seite)

Guilford, J.P. (1984): „Varieties of divergent production". In: *Creative Behavior*. Vol. 18, Issue 1. March 1984. S. 1-10
(zit. Guilford (1984), Seite)

Günther, Gotthard (1957): *Das Bewusstsein der Maschinen. Eine Metaphysik der Kybernetik.* Neuauflage von 2021 auf Basis der 2. Auflage. Frankfurt am Main: Klostermann
(zit. Günther (1957), Seite)

Haagen, Christian (2021): Verantwortung für Künstliche Intelligenz. Ethische Aspekte und zivilrechtliche Anforderungen bei der Herstellung von KI-Systemen. 2021. Baden-Baden: Nomos
(zit. Haagen (2021), Seite)

Habermas, Jürgen (1981): *Theorie des kommunikativen Handelns*. 2 Bde.. 11. Auflage 2019. Frankfurt/Main: Suhrkamp
(zit. Habermas (1981), Seite)

Habermas, Jürgen (1988): *Der philosophische Diskurs der Moderne. Zwölf Vorlesungen*. 13. Auflage 2019. Frankfurt am Main: Suhrkamp
(zit. Habermas (1988), Seite)

Habermas, Jürgen (1990): *Die Moderne – ein unvollendetes Projekt.* 3. Auflage 1994. Leipzig: Reclam
(zit. Habermas (1990), Seite)

Haenlein, Michael; **Kaplan**, Andreas (2019): *A brief history of Artificial Intelligence: On the past, present, and future of artificial intelligence.* California Management Review. 2019. Vol. 61(4). S. 5-14
(zit. Haenelein Kaplan (2019), Seite)

Halbig, Christoph; **Henning**, Tim (Herausgeber, 2012): *Die neue Kritik der instrumentellen Vernunft.* 2012. Berlin: Suhrkamp
(zit. Halbig Henning (2012), Seite)

Harrach, Sebastian (2014): Neugierige Strukturvorschläge im maschinellen Lernen. Eine technikphilosophische Verortung. 2014. Bielefeld : Transcript Verlag
(zit. Harrach (2014), Seite)

Heidbrink, Ludger; **Langbehn**, Claus; **Loh**, Janina (2017): *Handbuch Verantwortung*. 2017 Wiesbaden: Springer VS
(zit. Heidbrink Langbehn Loh (2017), Seite)

Heidegger, Martin (1955): *Gelassenheit. Heideggers Meßkircher Rede von 1955*. 2. Auflage 2015. München: Verlag Karl Alber
(zit. Heidegger (1955), Seite)

Heinemann, Lars (2011): „Normativität bei Max Weber. Zum Spannungsverhältnis von Wertfreiheit und Verstehen*"*. In: Ahrens, Beer et al. (Hg.) *Normativität: Über die Hintergründe sozialwissenschaftlicher Theoriebildung. https://www.researchgate.net/publication/285770509_Normativitat_bei_Max_Weber_Zum_Spannungsverhaltnis_von_Wertfreiheit_und_Verstehen*
(zit. Heinemann (2011), Seite)

Helbing, Dirk (Herausgeber, 2019): Towards Digital Enlightenment. Essays on the Dark and Light Sides of the Digital Revolution. 2019. Cham, Switzerland: Springer
(zit. Helbing (2019), Seite)

Hering, Steffen; **Schultz**, Nora; **Galert**, Thorsten (2018): „Menschenwürde im Angesicht neuer Technologien". Ethik Med 30, 375–383 (2018). https://doi.org/10.1007/s00481-018-0511-y
(zit. Hering et al. (2018), Seite)

Heyns, Christof (2013): *Report of the Special Rapporteur on extrajudicial, summary or arbitrary executions*. United Nations Human Rights Council, Twenty-third session, Agenda item 3, 9 April 2013 (A/HRC/23/47)
(zit. Heyns (2013), Seite)

Heyns, Christof (2017): „Autonomous weapons in armed conflict and the right to a dignified life: an African perspective". In: *South African Journal on Human Rights*, Vol. 33, 2017-Issue 1. S. 46-71
(zit. Heyns (2017), Seite)

Hilgendorf, Eric (2012): „Können Roboter schuldhaft handeln? Zur Übertragbarkeit unseres normativen Grundvokabulars auf Maschinen". In: *Jenseits von Mensch und Maschine. Ethische und rechtliche Fragen zum Umgang mit Robotern, Künstlicher Intelligenz und Cyborgs*. Herausgeber: Hilgendorf, Eric; Beck, Susanne. 2012. Baden-Baden: Nomos
(zit. Hilgendorf (2012), Seite)

Hilgendorf, Eric (2013): „Menschenwürde und die Idee des Posthumanen". In: *Menschenwürde und Medizin. Ein interdisziplinäres Handbuch*. Herausgeber: Joerden, Jan. 2013. Berlin: Duncker und Humblot. S. 1047ff
(zit. Hilgendorf (2013), Seite)

Hilgendorf, Eric (2017): „Autonomes Fahren im Dilemma. Überlegungen zur moralischen und rechtlichen Behandlung von selbsttätigen Kollisionsvermeidesystemen“, in: *Autonome Systeme und neue Mobilität*, Tagungsband, Eric Hilgendorf (Hrsg.), Nomos, Baden-Baden, 2017, S. 143ff
(zit. Hilgendorf (2017), Seite)

Historisches Wörterbuch der Philosophie (1976). Herausgeber: Ritter, Joachim; Gründer, Karlfried. Darmstadt: Wissenschaftliche Buchgesellschaft
(zit. Historisches Wörterbuch der philosophischen Begriffe (1976), Seite)

Höffe, Otfried; **Forschner**, Maximilian; **Horn**, Christoph; **Vossenkuhl**, Wilhelm (1977): *Lexikon der Ethik*. 8. überarbeitete und ergänzte Ausgabe. 2023. München: C.H. Beck
(zit. Höffe et al. (1977), Seite)

Hoffmann, Dirk W. (2011): Theoretische Informatik. 2011. 2., aktualisierte Auflage. Carl Hanser Fachbuchverlag, München
(zit. Hoffmann (2011), Seite)

Hofstetter, Yvonne (2016): *Das Ende der Demokratie. Wie die künstliche Intelligenz die Politik übernimmt und uns entmündigt*. 2. Auflage 2018. München: C. Bertelsmann/Penguin
(zit. Hofstetter (2016), Seite)

Homeister, Matthias (2005): *Quantum Computing verstehen. Grundlagen – Anwendungen – Perspektiven.* 5. Auflage. 2018. Wiesbaden: Springer Vieweg
(zit. Homeister (2005), Seite)

Horkheimer, Max (1946): “Reason Against itself: Some Remarks on Enlightenment”. In: *Theory, Culture & Society*. 1993;10(2):79-88. doi:10.1177/026327693010002004
(zit. Horkheimer (1946), Seite)

Horkheimer, Max (1947): *Zur Kritik der instrumentellen Vernunft*. Dt. Ausgabe 2007. Frankfurt am Main: Fischer Taschenbuch Verlag
(zit. Horkheimer (1947), Seite)

Horkeimer, Max, **Adorno**, Theodor W. (1944): *Dialektik der Aufklärung. Philosophische Fragmente*. 23. Auflage. 2017. Frankfurt: Fischer Taschenbuch Verlag
(zit. Horkheimer Adorno (1944), Seite)

Hubig, Christoph (1990): „Verantwortung in Wissenschaft und Technik. Fragen und Probleme“ In: Hubig, Christoph (Hrsg.): *Verantwortung in Wissenschaft und Technik* : Kolloquium an der Technischen Universität Berlin, WS 1987/88. Berlin : Univ.-Bibliothek der Technischen Univ., 1990 (TUB-Dokumentation, Kongresse und Tagungen. 54). S. 1-10. http://dx.doi.org/10.18419/opus-7685
(zit. Hubig (1990), Seite)

Hubig, Christoph (1995): „Verantwortung und Hochtechnologie“. In: *Verantwortung. Prinzip oder Problem?* Kurt Bayertz (Hrsg.). 1995. Darmstadt: Wissenschaftliche Buchgesellschaft: S. 98-142
(zit. Hubig (1995), Seite)

Hume, David (1739): *A treatise of human nature: Being an attempt to introduce the experimental method of reasoning*. 1739. Originalausgabe. London: John Noon, White-Hart, near Mercer’s-Chapel in Cheapside
(zit. Hume (1739), Seite)

Husserl, Edmund (1900): *Logische Untersuchungen. Erster Band. Prolegomena zur reinen Logik*. 7. Auflage. 1993. Tübingen: Max Niemeyer Verlag
(zit. Husserl (1900), Seite)

IHRC (International Human Rights Clinic, Human rights Program at Harvard Law School, 2012): *Losing Humanity. The Case against Killer Robots*. 2012. Human Rights Watch (Herausgeber). https://www.hrw.org/sites/default/files/reports/arms1112_ForUpload.pdf, heruntergeladen am 28.8.2022
(zit. IHRC (2012), Seite)

IHRC (International Human Rights Clinic, Human rights Program at Harvard Law School, 2014): *Shaking the foundations. The human rights implications of killer robots.* 2014. Human Rights Watch (Herausgeber)
(zit. IHRC (2014), Seite)

Ilkou, Elenio; **Koutraki**, Maria (2020): "Symbolic Vs Sub-symbolic AI Methods: Friends or Enemies?". In: *Proceedings of the CIKM 2020 Workshops, October 19-20, Galway, Ireland.* https://ceur-ws.org/Vol-2699/paper06.pdf
(zit. Ilkou Koutraki (2020), Seite)

Jackson, Frank Cameron; (1982): "Epiphenomenal Qualia". In: *The Philosophical Quarterly*, Bd. 32, No. 127 (April 1982), S. 127ff
(zit. Jackson (1982), Seite)

Jackson, F. (2003): *Mind and Illusion*. Royal Institute of Philosophy Supplement, 53, S.251-271
(zit. Jackson (2003), Seite)

Jacob, Pierre (2023): "Intentionality", *The Stanford Encyclopedia of Philosophy* (Spring 2023 Edition), Edward N. Zalta & Uri Nodelman (eds.), URL: https://plato.stanford.edu/archives/spr2023/entries/intentionality/
(zit. Jacob (2023), Absatz)

Jaedtke, Kathleen (2019): „Kann Künstliche Intelligenz kreativ sein?". In: *Bigdata-Insider*. 2019. https://www.bigdata-insider.de/kann-kuenstliche-intelligenz-kreativ-sein-a-824939/
(zit. Jaedtke (2019))

Jonas, Hans (1979): *Das Prinzip Verantwortung. Versuch einer Ethik für die technologische Zivilisation*. Erste deutsche Auflage 1984. Berlin: Suhrkamp
(zit. Jonas (1979), Seite)

Jonas, Hans (1981): *Macht oder Ohnmacht der Subjektivität*. Erste Auflage 1981. Frankfurt am Main: Insel Verlag
(zit. Jonas (1981), Seite)

Jones, Nicola (2014): *The Learning Machines*. In: Nature Vol. 505: S. 146 – 148
(zit. Jones (2014), Seite)

de Jonquières, Guy. (2017): *The world turned upside down: the decline of the rules-based international system and the rise of authoritarian nationalism*. In: International Politics 54, S. 552–560 (2017). https://doi.org/10.1057/s41311-017-0049-5
(de Jonquières (2017), Seite)

Kaminski, Andreas (2012): *Lernende Maschinen: naturalisiert, transklassisch, nichttrivial? Ein Analysemodell ihrer informellen Wirkungsweise*. In: Andreas Kaminski & Andreas Gelhard (Herausgeber). „Zur Philosophie der informellen Technisierung". 2014. Darmstadt: Wissenschaftliche Buchgesellschaft. S. 58-81
(zit. Kaminski (2012), Seite)

Kaminski, Andreas (2020): *Gründe geben. Maschinelles Lernen als Problem der Moralfähigkeit von Entscheidungen*. In: Klaus Wiegerling, Michael Nerurkar und Christian Wadephul (Hg.): „Datafizierung und Big Data: Ethisch, anthropologische und wissenschaftstheoretische Perspektiven". Bielefeld: Springer. S. 151-174
(zit. Kaminski (2020), Seite)

Kaminski, Andreas; **Gelhard**, Andreas (2014): *Zur Philosophie der informellen Technisierung*. 2014. Darmstadt: Wissenschaftliche Buchgesellschaft
(zit. Kaminski Gelhard (2014), Seite)

Kant, Immanuel (1781): *Kritik der reinen Vernunft. 1. Ausgabe 1781.* In: Werke (Akademieausgabe), Band IV. https://korpora.zim.uni-duisburg-essen.de/kant/verzeichnisse-gesamt.html
(zit. Kant (1781), IV AA)

Kant, Immanuel (1784a): „Idee zu einer allgemeinen Geschichte in weltbürgerlichen Absicht" In: *Immanuel Kant. Was ist Aufklärung? Ausgewählte kleine Schriften*. Hrsg. Brandt, Horst. Nachdruck 2019. Hamburg: Felix Meiner. S. 3-19
(zit. Kant (1784a), Seite)

Kant, Immanuel (1784b): „Beantwortung der Frage: Was ist Aufklärung?" In: *Immanuel Kant. Was ist Aufklärung? Ausgewählte kleine Schriften*. Hrsg. Brandt, Horst. Nachdruck 2019. Hamburg: Felix Meiner. S. 20-27
(zit. Kant (1784b), Seite)

Kant, Immanuel (1785a): *Grundlegung der Metaphysik der Sitten.* In: Werke (Akademieausgabe), Band IV. Berlin: de Gruyter
(zit. Kant (1785a), Seite)

Kant, Immanuel (1785b): „Grundlegung der Metaphysik der Sitten*"*. In: *Immanuel Kant. Kritik der praktischen Vernunft. Grundlegung zur Metaphysik der Sitten.* Werkausgabe Band VII. Wilhelm Weischedel (Hrsg.). Berlin: Suhrkamp, S. 7-102
(zit. Kant (1785b), Seite)

Kant, Immanuel (1786): „Was heißt: sich im Denken orientieren?" In: *Immanuel Kant. Was ist Aufklärung? Ausgewählte kleine Schriften*. Hrsg. Brandt, Horst. Nachdruck 2019. Hamburg: Felix Meiner. S. 45-61
(zit. Kant (1786), Seite)

Kant, Immanuel (1787a): *Kritik der reinen Vernunft. 2. Ausgabe 1787.* In: Werke (Akademieausgabe), Band III. https://korpora.zim.uni-duisburg-essen.de/kant/verzeichnisse-gesamt.html
(zit. Kant (1787a), III AA)

Kant, Immanuel (1787b): „Kritik der reinen Vernunft*"*. In: *Immanuel Kant. Kritik der praktischen Vernunft.* Werkausgabe Band IV. Wilhelm Weischedel (Hrsg.). Berlin: Suhrkamp
(zit. Kant (1787b), Seite)

Kant, Immanuel (1788): „Kritik der praktischen Vernunft*"*. In: *Immanuel Kant. Kritik der praktischen Vernunft. Grundlegung zur Metaphysik der Sitten.* Werkausgabe Band VII. Wilhelm Weischedel (Hrsg.). Berlin: Suhrkamp, S. 103-302
(zit. Kant (1788), Seite)

Kant, Immanuel (1790): „Kritik der Urteilskraft*"*. In: *Immanuel Kant. Kritik der Urteilskraft.* Werkausgabe Band X. Wilhelm Weischedel (Hrsg.). Berlin: Suhrkamp, S. 69-456
(zit. Kant (1790), Seite)

Kant, Immanuel (1797): „Die Metaphysik der Sitten*"*. In: *Immanuel Kant. Die Metaphysik der Sitten.* Werkausgabe Band VIII. Wilhelm Weischedel (Hrsg.). Berlin: Suhrkamp, S. 303634
(zit. Kant (1797), Seite)

Kant, Immanuel (1798): „Anthropologie in pragmatischer Hinsicht*"*. In: *Immanuel Kant. Schriften zur Anthropologie, Geschichtsphilosophie, Politik und Pädagogik 2.* Werkausgabe Band XII. Wilhelm Weischedel (Hrsg.). Berlin: Suhrkamp, S. 395-690
(zit. Kant (1798), Seite)

Kant Lexikon: siehe Willaschek

Keil, Geert (2007): Willensfreiheit. 2007. Berlin: Walter de Gruyter
(zit. Keil (2007), Seite)

Kemmerling, Andreas (2003): „Was ist menschlicher Geist? Neue Wissenschaft und alte Begriffe". In: *Ist der Geist berechenbar? Philosophische Reflexionen*. Hrsg.: Köhler, Wolfgang R.; Mutschler, Hans-Dieter. 2003. Darmstadt: Wissenschaftliche Buchgesellschaft. S. 168-187
(zit. Kemmerling (2003), Seite)

Kersting, Kristian; **Meyer**, Ulrich (2017): *From Big Data to Big Artificial Intelligence? Algorithmic Challenges and Opportunities of Big Data*. Künstliche Intelligenz 32, 3–8 (2018). https://doi.org/10.1007/s13218-017-0523-7
(zit. Kersting Meyer (2017), Seite)

Keysers, Christian; **Gazzola**, Valeria (2014): *Hebbian learning and predictive mirror neurons for actions, sensations and emotions*. In: Philosophical Transactions B Royal Society. 2014 June 5. 369 (1644): 2013075
(zit. Keyser Gazzola (2014))

Kiggins, Ryan (Hg., 2018): *The Political Economy of Robots. Prospects for Prosperity and Peace in the Automated 21st Century*. 2018. Cham, Schweiz: Palgrave Macmillan
(zit. Kiggins (2018), Seite)

Kim, Jaegwon (2005): *Physicalism, or something near enough*. 2005. Princeton & Oxford: Princeton University Press
(zit. Kim (2005), Seite)

Kim, Jaegwon (2007): "Emergenz: Zentrale Gedanken und Kernprobleme". In: *Grundkurs Philosophie des Geistes. Band 2: Das Leib-Seele-Problem*. 2. Auflage 2013. Paderborn: mentis. S. 297-318
(zit. Kim (2007), Seite)

Kissinger, Henry A; **Schmidt**, Eric; **Huttenlocher**, Daniel: *The age of AI and our human future*. 2021. London: John Murray (Publishers)
(zit. Kissinger et al. (2021), Seite)

Kitchin, Rob (2014): "Big Data, new epistemologies and paradigm shifts". In: *Big Data & Society*. April-June 2014. S. 1-12
(zit. Kitchin (2014), Seite)

Koch, Christof (2004): *The Quest for Consciousness. A neurobiological Approach.* 2004. Englewood, Colorado: Roberts & Company Publishers
(zit. C. Koch (2004), Seite)

Koch, Christof (2020): Bewusstsein. Warum es verbreitet ist, aber nicht digitalisiert werden kann. 1. Dt. Auflage 2020. Berlin: Springer
(zit. C. Koch (2020), Seite)

Koch, Wolfgang (2020): „Zur Ethik der wehrtechnischen Digitalisierung. Informations- und ingenieurwissenschaftliche Aspekte". In: *Ethische Herausforderungen digitalen Wandels in bewaffneten Konflikten*. **Rogg**, Matthias; **Scheidt**, Sophie; **Schubert**, Hartwig von (Hrsg.). 2020. Hamburg: German Institute for Defense and Strategic Studies. S. 17-54
(zit. W. Koch (2020), Seite)

Koenig, Gaspard (2019): *Das Ende des Individuums. Reise eines Philosophen in die Welt der Künstlichen Intelligenz*. 1. deutsche Auflage 2021. Berlin: Galiani Verlag
(zit. Koenig (2019), Seite)

Köhler, Wolfgang R.; **Mutschler**, Hans-Dieter (Hrsg., 2003): Ist der Geist berechenbar? Philosophische Reflexionen. 2003. Darmstadt: Wissenschaftliche Buchgesellschaft
(zit. Köhler Mutschler (2003), Seite)

Krause, Ulf von (2021): *Künstliche Intelligenz im Militär. Chancen und Risiken für die Sicherheitspolitik*. 2021. Wiesbaden: Springer VS
(zit. Krause (2021), Seite)

Kurzweil, Ray (2005): *The singularity is near. When humans transcend biology*. 2006. New York: Penguin
(zit. Kurzweil (2005), Seite)

Landgrebe, Jobst; **Smith**, Barry (2022): *Why Machines will never rule the world. Artificial Intelligence without fear.* 2022. New York: Routledge
(zit. Landgrebe Smith (2022), Seite)

Lavaert, Sonja; **Schröder**, Winfried (Hrsg. 2018): *Aufklärungs-Kritik und Aufklärungs-Mythen*. *Horkheimer und Adorno in philosophischer Perspektive.* 2018. Berlin/Boston: De Gruyter
(zit. Lavaert Schröder (2018), Seite)

Legg, Shane; **Hutter**, Markus (2007): "A collection of Definitions of Intelligence". In: *Frontiers in Artificial Intelligence and Applications,* Bd. 157 (2007) S. 17-24; https://arxiv.org/abs/0706.3639; S. 1 - 12
(zit. Legg Hutter (2007), Seite)

Leibniz, Gottfried Wilhelm (1724): *Hauptschriften zur Grundlegung der Philosophie. Teil II.* Übersetzt durch Artur Buchenau, herausgegeben durch Ernst Cassirer. 1996. Hamburg: Meiner
(zit. Leibniz (1925), Seite)

Lexikon der Mathematik in sechs Bänden. 2001. Berlin Heidelberg: Spektrum Akademischer Verlag
(zit. Lexikon der Mathematik (2001), Seite)

LeCun, Yann; **Bengio**, Yoshua; **Hinton**, Geoffrey: "Deep Learning". In: *Nature*, Vol. 521. 28 May 2015. S. 436-444
(zit. LeCun Bengio Hinton (2015), Seite)

Lenk, Hans (1971): *Philosophie im technologischen Zeitalter*. 1971. Stuttgart: Kohlhammer
(zit. Lenk (1971), Seite)

Lenk, H.; **Maring**, M. (1993): "Verantwortung – Normatives Interpretationskonstrukt und empirische Beschreibung". In: L.H. Eckensberger und U. Gähde (Hrsg.). *Ethische Norm und empirische Hypothese*. Frankfurt a. M.: Suhrkamp: S. 222-243
(zit. Lenk Maring (1993), Seite)

Libet, Benjamin (2005): Mind Time. *Wie das Gehirn Bewusstsein produziert*. 2005. Frankfurt a.M.: Suhrkamp
(zit. Libet (2005), Seite)

List, Christian (2019): *Why free will is real*. 2019. Cambridge, Massachusetts: Harvard University Press
(zit. List (2019), Seite)

List, Christian (2021): *Warum der freie Wille existiert*. 2021. Karlsruhe, Leipzig: wbg
(zit. List (2021a), Seite)

List, Christian (2021): "Group Agency and Artificial Intelligence". In: *Philosophy & Technology*. August 2021 (34). S. 1213-1241. Online: http://philsci-archive.pitt.edu/19406/.
(zit. List (2021b), Seite)

Ludlow, Peter; **Nagasawa**, Yujin; **Stoljar**, Daniel (Hrsg., 2004): *There is something about Mary. Essays on Frank Jackson's knowledge argument*. 2004. Cambridge MA: The MIT Press
(zit. Ludlow Nagasawa Stoljar (2004), Seite)

Lucas, John R. (1961): "Minds, Machines, and Gödel*"*. In: *Philosophy*, Vol. 36, No. 137 (Apr. – Jul., 1961), S. 112-127: Cambridge University Press on behalf of Royal Institute of Philosophy
(zit. Lucas (1961), Seite)

Lucas, John R. (1996): „Minds, Machines, and Gödel: A retrospect". In: P. J. R. Millican & A. Clark (eds.). *Etica E Politica*. 1996. Clarendon Press. S. 1
(zit. Lucas (1996), Seite)

Luhmann, Niklas (1990): *Die Wissenschaft der Gesellschaft*. 1990. Frankfurt: Suhrkamp
(zit. Luhmann (1992), Seite)

Lukács, Georg (1923): „Die Verdinglichung und das Bewußtsein des Proletariats". In: Rüdiger Dannemann (Hrsg.). *Georg Lukács – Werksauswahl in Einzelbänden*. Band 3. 2015. Bielefeld: Aisthesis Verlag.
(zit. Lucas (1923), Seite)

Lutz, Bernd (Hrsg., 2015): *Metzler Philosophen-Lexikon. Von den Vorsokratikern bis zu den Neuen Philosophen.* 2015. Sonderausgabe und zugleich dritte, aktualisierte und erweiterte Auflage. Stuttgart: J.B. Metzler
(zit. Metzler Philosophen-Lexikon (2015), Seite)

Lutz-Bachmann, Matthias (2019): *Autonomie, I. Philosophisch*, Version 22.10.2019, 17:30 Uhr, in: Staatslexikon8 online, URL: https://www.staatslexikon-online.de/Lexikon/Autonomie (abgerufen: 16.05.2021)
(zit. Lutz-Bachmann (2019)

Lyytinen, Kalle; **Yoo**, Youngijn (2002): "Issues and Challenges in ubiquitous computing". In: *Communications of the ACM*. December 2002 / Vol. 45, No. 12
(zit. Lyytinen Yoo (2002), Seite)

Marcuse, Herbert (1941): *Vernunft und Revolution. Hegel und die Entstehung der Gesellschaftstheorie*. Erste deutsche Auflage. 2020. Berlin: Suhrkamp
(zit. Marcuse (1941), Seite)

Mainzer, Klaus (2014): *Die Berechnung der Welt. Von der Weltformel zu Big Data*. 2014. München: C.H. Beck
(zit. Mainzer (2014), Seite)

Mainzer, Klaus (2015): *Künstliche Intelligenz. Wann übernehmen die Maschinen*. 2. Auflage 2018. Berlin: Springer
(zit. Mainzer (2015), Seite)

Mainzer, Klaus (2020): *Quantencomputer. Von der Quantenwelt zur Künstlichen Intelligenz.* 2020. Berlin: Springer
(zit. Mainzer (2015), Seite)

Mainzer, Klaus; **Kahle**, Reinhard (2022): *Grenzen der KI theoretisch - praktisch, ethisch.* 2022. Berlin: Springer
(zit. Mainzer Kahle (2022), Seite)

Manhart, Klaus (2022): „Eine kleine Geschichte der Künstlichen Intelligenz. KI und Machine Learning“. In: Computerwoche. https://www.computerwoche.de/a/eine-kleine-geschichte-der-kuenstlichen-intelligenz,3330537,6
(zit. Manhart (2022))

Margalit, Avishai (2012): *Politik der Würde. Über Achtung und Verachtung*. Zweite deutsche Auflage. 2018. Berlin: Suhrkamp
(zit. Margalit (2012), Seite)

Marx, Karl (1894): *Das Kapital. Kritik der politischen Ökonomie. Dritter Band*. Herausgegeben von der internationalen Marx-Engels-Stiftung. 2004. Berlin: Akademie Verlag
(zit. Marx (1894), Seite)

McCarthy, John (2000): “Free Will – Even for robots”. In: *Journal of Experimental & Theoretical Artificial Intelligence*. January 2000, S. 341-352
(zit. McCarthy (2000), Seite)

McCulloch, W.S., **Pitts**, W.H.: (1943): *A logical calculus of the ideas immanent in nervous activity,* Bulletin of mathematical biophysiscs, Vol. 5, 1943, S. 115-133
(zit. McCulloch Pitts (1943), Seite)

McCulloch, Warren (1965): *Embodiments of Mind*. New paperback edition 1988. Massachusetts: MIT Press
(zit. McCulloch Pitts (1965), Seite)

McKinsey (2023): *What is generative AI*. https://www.mckinsey.com/featured-insights/mckinsey-explainers/what-is-generative-ai#
(zit. McKinsey (2023), Seite)

Merkel, Reinhard (2006): *Willensfreiheit und rechtliche Schuld. Eine strafrechtsphilosophische Untersuchung*. 2. Auflage 2014. Baden-Baden: Nomos
(zit. Merkel (2006), Seite)

Metzinger, Thomas (Herausgeber, 1995): *Bewusstsein. Beiträge aus der Gegenwartsphilosophie*. 2. durchgesehene Auflage 1996. Paderborn: Ferdinand Schöningh
(zit. Metzinger (1995), Seite)

Metzinger, Thomas (Herausgeber, 2006): *Grundkurs Philosophie des Geistes. Band 1: Phänomenales Bewußtsein*. 2006. Paderborn: mentis
(zit. Metzinger Band 1 (2006), Seite)

Metzinger, Thomas (Herausgeber, 2007): *Grundkurs Philosophie des Geistes. Band 2: Das Leib-Seele-Problem*. 2. Auflage 2013. Paderborn: mentis
(zit. Metzinger Band 2 (2007), Seite)

Metzinger, Thomas (2009): Der EGO Tunnel. Eine neue Philosophie des Selbst: Von der Hirnforschung zur Bewusstseinsethik. 6. Auflage 2017. München: Piper
(zit. Metzinger (2009), Seite)

Metzinger, Thomas (Herausgeber, 2010): *Grundkurs Philosophie des Geistes. Band 3: Intentionalität und mentale Repräsentation*. 2010. Paderborn: mentis
(zit. Metzinger Band 3 (2010), Seite)

Metzinger, Thomas (2023): *Bewusstseinskultur. Spiritualität, intellektuelle Redlichkeit und die planetarische Krise*. 4. Auflage 2023. Berlin: Berlin Verlag
(zit. Metzinger (2023), Seite)

Metzler Philosophen-Lexikon: siehe Lutz, Bernd (Hrsg., 2015)
(zit. Metzler Philosophen-Lexikon (2015), Seite)

Mirandola, Pico de la (1486): „Rede über die Würde des Menschen“. In: *Texte zur Menschenwürde*. Franz Josef Wetz (Hrsg.). 2. Auflage 2019. Stuttgart: Reclam. S. 82-85
(zit. Mirandola (1486), Seite)

Misselhorn, Catrin (2018): *Grundfragen der Maschinenethik*. 2. durchgesehene Auflage 2018. Stuttgart: Philipp Reclam
(zit. Misselhorn (2018), Seite)

Mitchell, Melanie (2019): *Artificial Intelligence. A Guide for Thinking Humans*. 2019. New York: Farrar, Strauss and Giroux
(zit. Mitchell (2019), Seite)

Mohr, Georg (2001): „Der Begriff der Person bei Kant, Fichte und Hegel“. In: *Person. Philosophiegeschichte - Theoretische Philosophie – Praktische Philosophie*. Dieter Sturma (Hrsg.). 2001. Paderborn: mentis. S. 103-142
(zit. Mohr (2001), Seite)

Müller, Jean Moritz (2015): „Künstliche Intelligenz“ In: Demmerling Stekeler-Weithofer: *Sprachphilosphie.* 2015. Berlin/Boston: DeGruyter
Online: Müller, Jean. (2015). Künstliche Intelligenz. 10.1515/wsk.15.0.kunstlicheintelligenz.
(zit. Müller (2015), Seite)

Müller, Hans-Peter; **Sigmund**, Steffen (Hg., 2020): *Max Weber Handbuch. Leben – Werk – Wirkung.* 2. Aktualisierte und erweiterte Auflage 2020. Berlin: J.B. Metzler
(zit. Müller Sigmund (2020), Seite)

Müller, Vincent C. (2016): “Autonomous killer robots are probably good news”. In Ezio Di Nucci and Filippo Santoni de Sio (Hg.), *Drones and responsibility: Legal, philosophical and socio- technical perspectives on the use of remotely controlled weapons* (London: Ashgate), 67-81. http://www.ashgate.com/isbn/9781472456724 DOI: 10.4324/9781315578187-4
(zit. Müller (2016), Seite)

Nagel, Thomas (1974): *What is it like to be a bat? Wie ist es, eine Fledermaus zu sein?* 1974. Stuttgart: Philipp Reclam
(zit. Nagel (1974), Seite)

Nassehi, Armin (2019): *Muster. Theorie der digitalen Gesellschaft*. 2019. München: C.H. Beck
(zit. Nassehi (2019), Seite)

Nelkin, Norton (1993): “What is consciousness?” In: *Philosophy of Science*. Bd. 60, Nr. 3 (September 1993), S. 419ff
(zit. Nelkin (1993), Seite)

Nemitz, Paul; **Pfeffer**, Matthias (2020): *Prinzip Mensch. Macht, Freiheit und Demokratie im Zeitalter der Künstlichen Intelligenz*. 2020. Bonn: Dietz
(zit. Nemitz Pfeffer (2020), Seite)

Nida-Rümelin, Julian (2001): *Strukturelle Rationalität. Ein philosophischer Essay über praktische Vernunft*. 2001. Stuttgart: Philipp Reclam
(zit. Nida-Rümelin (2001), Seite)

Nida-Rümelin, Julian (2005): *Über menschliche Freiheit*. 2005. Stuttgart: Philipp Reclam
(zit. Nida-Rümelin (2005, Seite)

Nida-Rümelin, Julian (2011): *Verantwortung*. 2011. Stuttgart: Philipp Reclam
(zit. Nida-Rümelin (2011), Seite)

Nida-Rümelin, Julian (2020): *Eine Theorie praktischer Vernunft*. 2020. Berlin/Boston: Walter de Gruyter
(zit. Nida-Rümelin (2020), Seite)

Nida-Rümelin, Julian; **Battaglia**, Fiorella (2019): „Mensch, Maschine und Verantwortung“. In: Bendel. Oliver (Hrsg.). *Handbuch Maschinenethik*. 2019. Wiesbaden: Springer, S. 57-72
(zit. Nida-Rümelin Battaglia (2019), Seite)

Nida-Rümelin, Julian; **Weidenfeld**, Nathalie (2018): *Digitaler Humanismus. Eine Ethik für das Zeitalter der Künstlichen Intelligenz*. 2. Auflage 2018. München: Piper
(zit. Nida-Rümelin Weidenfeld (2018), Seite)

Nielsen-Sikora, Jürgen (2017): *Hans Jonas. Für Freiheit und Verantwortung*. 2017. Darmstadt: WBG
(zit. Nielsen-Sikora (2017), Seite)

Nietzsche, Friedrich (2010): *Also sprach Zarathustra.* 7. Auflage 2018. München: C.H. Beck
(zit. Nietzsche (2010), Seite)

Nilsson, Nils J. (2014): *Die Suche nach Künstlicher Intelligenz. Eine Geschichte von Ideen und Erfolgen*. 2014. Berlin: Akademische Verlagsgesellschaft
(zit. Nilsson (2014), Seite)

Nimtz, Christian (2009): “Physisches” und Multi-Realisierbarkeit, oder: zwei Probleme für den Physikalismus gelöst“. In: *Physikalismus, Willensfreiheit, Künstliche Intelligenz*. Backmann, Marius; Michel, Jan G. (Hrsg.). 2009. Paderborn: mentis. S. 23ff
(zit. Nimtz (2009), Seite)

Nussbaum, Martha (2015): *Fähigkeiten schaffen. Neue Wege zur Verbesserung menschlicher Lebensqualität*. 2. Auflage 2019. Freiburg: Karl Alber
(zit. Nussbaum (2015), Seite)

Obama, Barack (2009): *President Barack Obama’s Inaugural Address*. 20.1.2009. https://obamawhitehouse.archives.gov/blog/2009/01/21/president-barack-obamas-inaugural-address (zuletzt eingesehen am 5.8.2021)
(zit. Obama (2009))

O’Neill, Onara (1989): *Constructions of reason. Explorations of Kant’s practical philosophy*. 1989. Cambridge: Cambridge University Press
(zit. O‘Neill (1989), Seite)

Ottmann, Henning (2012): Geschichte des politischen Denkens. Das 20. Jahrhundert. Von der kritischen Theorie bis zur Globalisierung. 2012. Stuttgart: J.B. Metzler
(zit. Ottmann (2012), Seite)

Ouchchy, Leila; **Coin** Allen; **Dubljevic** (2020): “AI in the headlines: the portrayal of the ethical issues of artificial intelligence in the media”. In: *AI&Society*. 35, 927–936 (2020). https://doi.org/10.1007/s00146-020-00965-5
(zit. Ouchchy et al. (2020))

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C.L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L.E., Simens, M., Askell, A., Welinder, P., Christiano, P.F., Leike, J., & Lowe, R.J. (2022): *Training language models to follow instructions with human feedback*. ArXiv, abs/2203.02155.
(zit. Ouyang et al. (2022), Seite)

Pauen, Michael et al. (Herausgeber, 2005): *Bewusstsein. Philosophie, Neurowissenschaften, Ethik*. 2005. Paderborn: Wilhelm Fink
(zit. Pauen (2005), Seite)

Paulick, Christian (2018): „Macht“ [online]. *socialnet Lexikon*. Bonn: socialnet, 17.09.2018 [Zugriff am: 01.05.2022]. Verfügbar unter: https://www.socialnet.de/lexikon/Macht
(zit. Paulick (2018))

Peirce, Charles Sanders (1934): *Collected papers of Charles Sanders Peirce***.** Vol. V (Pragmatism and Pragmaticism), Vol. VI (Scientific Metaphysics). Edited by Charles Hartshorne and Paul Weiss. Fourth printing 1974. Cambridge, Massachusetts: The Belknap Press of Harvard University Press
(zit. Peirce (1934), Seite)

Peirce, Charles Sanders (1973): *Vorlesungen über Pragmatismus*. neu herausgegeben von Elisabeth Walter. 1991. Hamburg: Meiner
(zit. Peirce (1973), Seite)

Penrose, Roger (1994a): *Schatten des Geistes. Wege zu einer neuen Physik des Bewusstseins.* 1. deutsche Ausgabe. 1995. Heidelberg, Berlin, Oxford: Spektrum Akademischer Verlag
(zit. Penrose (1994a), Seite)

Penrose, Roger (1994b): "Mathematical Intelligence". In: *What is intelligence?*. Edited by Jean Khalfa. 1994. Cambridge: Cambridge University Press, S. 107-136
(zit. Penrose (1994b), Seite)

Pentland, Alex (2014): "The death of individuality: What really governs your actions?". *New Scientist*. 5 April 2014. S. 30-31
(zit. Pentland (2014), Seite)

Pfordten, Dietmar von der (2016): *Menschenwürde*. München: C.H. Beck
(zit. Pfordten (2016), Seite)

Philosophisches Wörterbuch (2009). Herausgeber: Martin Gessmann. 23., vollständig bearbeitete Auflage. Stuttgart: Kröner
(zit. Philosophisches Wörterbuch (2009), Seite)

Pijetlovic, Denis(2020): *Das Potential der Pflege-Robotik. Eine systemische Erkundungsforschung.* 2020. Wiesbaden: Springer Gabler
(zit. Pijetlovic (2020), Seite)

Pinker, Steven (2008): "The Stupidity of Dignity". In: *The New Republic*. May 28, 2008
(zit. Pinker (2008))

Pinker, Steven (2018): *Enlightenment now. The case for reason, science, humanism, and progress*. 2018. New York: Viking/Penguin Random House
(zit. Pinker (2018), Seite)

Platon (1958): *Der Staat (Politaia)*. Übersetzt und herausgegeben von Karl Vretska. Bibliographisch ergänzte Ausgabe 2015. Stuttgart: Reclam
(zit. Platon (1958), Seite)

Popitz, Heinrich (1986): *Phänomene der Macht*. 2. Auflage 1992. Tübingen: Mohr Siebeck
(zit. Popitz (1986), Seite)

Porat, Ariel; **Strahilevitz**, Jacob Lior (2021):" The Concept of Personalized Law". In: Busch, De Franceschi (Hrsg.): *Algorithmic Regulation and Personalized Law. A Handbook*. 2021. München: C.H. Beck
(zit. Porat Strahilevitz (2021), Seite)

Prinz, Wolfgang (2004): „Worüber dürfen Hirnforscher reden – und in welcher Weise?*"*. In: Geyer, Christian (Hrsg.): *Hirnforschung und Willensfreiheit. Zur Deutung der neuesten Experimente*. 9. Auflage 2016. Frankfurt: Suhrkamp
(zit. Prinz (2004), Seite)

Prinz, Wolfgang (2013): *Selbst im Spiegel. Die soziale Konstruktion von Subjektivität*. 2013. Berlin: Suhrkamp, S. 20-26
(zit. Prinz (2013), Seite)

Putnam, Hilary (1960): "Minds and Machines". In: Sideny Hook (Hrsg.), *Dimensions of Minds*. New York, USA: New York University Press. S. 138-164
(zit. Putnam (1960), Seite)

Putnam, Hilary (1967): "Psychological Predicates". in *Art, Mind, and Religion*. W. Capitan and D. Merrill (Hrsg.). S. 37-48
(zit. Putnam (1967), Seite)

Ramage, Magnus; **Shipp**, Karen (2009): *Systems Thinkers*. 2. Auflage 2020. London: Springer Nature
(zit. Ramage Shipp (2009), Seite)

Razumnikowa, Olga M. (2013): „Divergent versus convergent thinking". In: Carayannis, E.G. (Hrsg.): *Encyclopedia of Creativity, Innovation and Entrepreneurship*. 2013. New York: Springer, S. 546-552
(zit. Razumnikowa (2013), Seite)

Reihlen, Markus; **Habersang**, Stefanie; **Nikolova**, Natalia (2022): „Realist Inquiry". In C. Neesham et al., *Handbook of Philosophy of Management*. 2022. Germany: Springer
(zit. Reihlen et al. (2022), Seite)

Ravenscroft, Ian (2008): *Philosophie des Geistes*. Eine Einführung. 2008. Stuttgart: Philipp Reclam
(zit. Ravenscroft (2008), Seite)

Rheinberger, Hans-Jörg (2007): „Wie werden aus Spuren Daten, und wie verhalten sich Daten zu Fakten?" In: Gugerli, David u.a. (Hrsg.): *Nach Feierabend: Zürcher Jahrbuch für Wissengeschichte*, Nr. 3. 2007. Zürich-Berlin: Diaphanes, S. 117-125
(zit. Rheinberger (2007), Seite)

Rhodes, Mel (1961): „An Analysis of Creativity". In: *The Phi Delta Kappa*. Apr. 1961, Vol. 42, No. 7. S. 305-310
(zit. Rhodes (1961), Seite)

Röd, Wolfgang (1989): „Dritter Teil. Traditionalistische Strömungen". In: Poggi, Stefani; Röd, Wolfgang (Hrsg.). *Die Philosophie der Neuzeit 4. Positivismus, Sozialismus im 19. Jahrhundert*. Sonderauflage 2021. München: Beck. S. 249-304
(zit. Röd (1989), Seite)

Rogg, Matthias; **Scheidt**, Sophie; **Schubert**, Hartwig von (Hrsg., 2020): *Ethische Herausforderungen digitalen Wandels in bewaffneten Konflikten*. 2020. Hamburg: German Institute for Defense and Strategic Studies
(zit. Rogg et al. (2020), Seite)

Rost, Detlef H. (2013): *Handbuch Intelligenz*. 2013. Weinheim, Basel: Beltz
(zit. Rost (2013), Seite)

Roth, Gerhard (2004a): „Worüber dürfen Hirnforscher reden – und in welcher Weise?". In: Geyer, Christian (Hrsg.): *Hirnforschung und Willensfreiheit. Zur Deutung der neuesten Experimente*. 9. Auflage 2016. Frankfurt: Suhrkamp, S. 66-85
(zit. Roth (2004a), Seite)

Roth, Gerhard (2004b): „Wir sind determiniert. Die Hirnforschung befreit von Illusionen". In: Geyer, Christian (Hrsg.): *Hirnforschung und Willensfreiheit. Zur Deutung der neuesten Experimente*. 9. Auflage 2016. Frankfurt: Suhrkamp, S. 218-222
(zit. Roth (2004b), Seite)

Roth, Gerhard (2009): *Aus Sicht des Gehirns*. Auflage 2015. Frankfurt: Suhrkamp
(zit. Roth (2009), Seite)

Rothaar, Markus (2013): „Menschenwürde qua Autonomie und Anerkennung: Kant und Fichte“. In: *Menschenwürde und Medizin: ein interdisziplinäres Handbuch*. Herausgeber: Joerden, Jan. Berlin: Duncker und Humblot. S. 73ff
(zit. Rothaar (2013), Seite)

Rott, Hans (2009): „Die Freiheit in den Zeiten neurowissenschaftlichen Fortschritts“. In: Mühling, Markus (Hrsg.): *Gezwungene Freiheit? Personale Freiheit im pluralistischen Europa*. 2009. Göttingen: Vandenhoek & Ruprecht, S. 117-134
(zit. Rott (2009), Seite)

Runciman, David (2018): *How democracy ends*. 2018. London, UK: Profile Books
(zit. Runciman (2018), Seite)

Runco, Mark A.; **Jaeger**, Garrett J. (2012): „The Standard Definition of Creativity. Comments and Corrections”. In: *Creativity Research Journal*. 24(I), 2012. S. 92 – 96
(zit. Runco Jaeger (2012), Seite)

Russell, Bertrand (1945): *Philosophie des Abendlandes*. Sonderausgabe 2012. Zürich: Europa Verlag
(zit. Russell (1945), Seite)

Russell, Stuart; **Norvig**, Peter (2004): *Künstliche Intelligenz. Ein moderner Ansatz*. 3 aktualisierte Auflage 2012. München: Pearson Deutschland
(zit. Russell Norvig (2004), Seite)

Russell, Stuart; **Norvig**, Peter (2010): *Artificial Intelligence. A modern approach.* Third edition. 2010. Harlow, UK: Pearson Education Limited
(zit. Russell Norvig (2010), Seite)

Sadin, Éric (2023): *Die Stimme gegen den Chat erheben.* In: FAZ 17.2.2023 Nr. 41, S. 14
(zit. Sadin (2023))

Safranski, Rüdiger (2021): *Einzeln Sein. Eine philosophische Herausforderung*. 2021. München: Carl Hanser
(zit. Safranski (2021), Seite)

Sartre, Jean-Paul (1997): *Die Transzendenz des Ego. Philosophische Essays 1931 – 1939.* 2. Auflage 2010. Reinbek bei Hamburg: Rowohlt Taschenbuch Verlag
(zit. Sartre (1997), Seite)

Schäfer, Erich (1994): *Grenzen der Künstlichen Intelligenz. John R. Searles Philosophie des Geistes*. Stuttgart Berlin Köln: W. Kohlhammer
(zit. Schäfer (1994), Seite)

Schmid Noerr, Gunzelin (2019): „Zum Begriff der Aufklärung in der Dialektik der Aufklärung“. In: *Zur Kritik der regressiven Vernunft. Beiträge zur „Dialektik der Aufklärung“*. Gunzelin Schmid Noerr, Eva Maria Ziege (Hrsg.). 2019. Wiesbaden: Springer VS
(zit. Schmid Noerr (2019), Seite)

Schmid Noerr, Gunzelin, **Ziege**, Eva Maria (2019): *Zur Kritik der regressiven Vernunft. Beiträge zur „Dialektik der Aufklärung“*. 2019. Wiesbaden: Springer VS
(zit. Schmid Noerr Ziege (2019), Seite)

Scholz, Erhard (2006): „Die Gödelschen Unvollständigkeitssätze und das Hilbertsche Programm einer „finiten“ Beweistheorie“. In: Achtner, W. et al. (Hrsg.): *Künstliche Intelligenz und menschliche Person*. 2006. Marburg: N.G. Elwert, S. 15-38
(zit. Scholz (2006), Seite)

Schulz, Lorenz (2001): *Normiertes Misstrauen. Der Verdacht im Strafrecht*. 2001. Frankfurt am Main: Vittorio Klostermann
(zit. Schulz (2001), Seite)

Schwarting, Detlef (2019): *Künstliche Intelligenz, Eine moralphilosophische Analyse*, Masterarbeit, Fakultät für Philosophie, Wissenschaftstheorie und Religionswissenschaft, LMU, September 2019, München
(zit. Schwarting (2019), Seite)

Searle, John R. (1980): „Minds, Brains, and Programs". In: *Behavioral and Brain Sciences 3 (3)*.1980. S. 417-457
(zit. Searle (1980), Seite)

Searle, John R. (1992): *The rediscovery of the mind*. 1992. Cambridge Massachusetts: MIT Press
(zit. Searle (1992), Seite)

Searle, John R. (1993): „The Problem of Consciousness". In: *Social Research, Bd*. 60, Nr. 1. 1993. S. 3-15
(zit. Searle (1993-1), Seite)

Searle, John R. (1993): *Die Wiederentdeckung des Geistes*. 1. Taschenbuchausgabe 1996. München: Artemis/Suhrkamp
(zit. Searle (1993-2), Seite)

Searle, John R. (2001): *Geist, Sprache und Gesellschaft. Philosophie in der wirklichen Welt*. 2001. Frankfurt am Main: Suhrkamp
(zit. Searle (2001), Seite)

Searle, John R. (2005): *Consciousness: What we still don't know*. Review of "The quest of Consciousness" by Christof Koch. In: The New York Review of Books. January 13 2005
(zit. Searle (2005))

Searle, John R. (2006): *Geist. Eine Einführung*. 2006. Frankfurt am Main: Suhrkamp
(zit. Searle (2006), Seite)

Searle, John R. (2013): „Theory of mind and Darwin's legacy". In: *PNAS (Proceedings of the National Academy of Sciences of the United States of America)* Vol. 110, Suppl. 2, 2013. S. 10343-10348
(zit. Searle (2013-1), Seite)

Searle, John R. (2013): "Can Information Theory Explain Consciousness? Review of 'Consciousness: Confessions of a Romantic Reductionist' by Christof Koch". In: *The New York Review of Books*. January 10, 2013
(zit. Searle (2013-2))

Sen, Amartya (2010): *Die Idee der Gerechtigkeit*. 2010. München: Beck
(zit. Sen (2010))

Seth, Anil K.; **Bayne**, Tim (2022): "Theories of Consciousness*"*. In: *Nature Reviews – Neuroscience*. Vol. 23. July 2022. S. 439-452
(zit. Seth Bayne (2022))

Schmidt, Andreas T.; **Engelen**, Bart (2020): „The ethics of nudging: An overview". In: Philosophy *Compass*. 2020. Wileyonlinelibrary.com/journal/phc3
(zit. Schmidt Engelen (2020))

Schneiders, Werner (1997): *Das Zeitalter der Aufklärung*. 5. Auflage. 2014. München: C:H. Beck
(zit. Schneiders (1997)

Schirrmacher, Frank (Herausgeber, 2015): *Technologischer Totalitarismus. Eine Debatte.* 2. Auflage 2018. Berlin: Suhrkamp
(zit. Schirrmacher (2015), Seite)

Shanahan, Murray (2021): *Die technologische Singularität*. 2021. Berlin: Matthes & Seitz Berlin
(zit. Shanahan (2021)

Sharkey, Noel; **Sharkey**, Amanda (2012a): „The rights and wrongs of Robot Care". In: Lin., Patrick; Abney, Keith, Bekey, George (2012). *Robot Ethics. The ethical and social implications of robotics*. 2012. Cambridge, MA, London: MIT Press. S. 267-282
(zit. Sharkey Sharkey (2012a), Seite)

Sharkey, Amanda; **Sharkey**, Noel (2012b): "Granny and the robots: ethical issues in robot care for the elderly". In: *Ethics and Information Technology*. 2012. 14(1), S. 27-40
(zit. Sharkey Sharkey (2012b), Seite)

Sharkey, Amanda (2014): "Robots and human dignity: a consideration of the effects of robot care on the dignity of older people". In: *Ethics and Information Technology*. 2014. 16(1), S. 63-75
(zit. Sharkey (2014), Seite)

Sharkey, Noel (2012): "Killing made easy: From Joysticks to Politics". In: Lin, Abney, Bekey (Hrsg.): *Robot Ethics: The ethical and social implications of robotics.* 2012. Cambridge, Massachusetts: The MIT Press. S. 111-128
(zit. Sharkey (2012c), Seite)

Sharkey, Amanda (2019): "Autonomous weapon systems, killer robots and human dignity". In: *Ethics and Information Technology*. 2019. 21(1), S. 75-87
(zit. Sharkey (2019), Seite)

Shatz, Carla (1992): *The developing brain*. Sci. Am. 267, S. 60-67
(zit. Shatz (1992), Seite)

Siebert, Carsten (1999): *Qualia. Das Phänomenale als Problem philosophischer und empirischer Bewusstseinstheorien*. Dissertation. Philosophische Fakultät I. HU Berlin. 1999. Online: https://edoc.hu-berlin.de/handle/18452/15088
(zit. Siebert (1999), Seite)

Siedentop, Larry (2014): *Inventing the Individual. The Origins of Western Liberalism*. 2014. Milton Keynes Great Britain: Penguin Random House
(zit. Siedentop (2014), Seite)

Siedentop, Larry (2015): *Die Erfindung des Individuums. Der Liberalismus und die westliche Welt*. 2015. Stuttgart: Klett-Cotta
(zit. Siedentop (2015), Seite)

Singer, Wolf (2004): „Verschaltungen legen uns fest: Wir sollten aufhören, von Freiheit zu sprechen". In: Geyer, Christian (Hrsg.): *Hirnforschung und Willensfreiheit. Zur Deutung der neuesten Experimente*. 9. Auflage 2016. Frankfurt: Suhrkamp, S. 30-65
(zit. Singer (2004), Seite)

Singer, Tassilo (2019): Dehumanisierung der Kriegsführung. Herausforderungen für das Völkerrecht und die Frage nach der Notwendigkeit menschlicher Kontrolle. 2019. Berlin: Springer
(zit. Singer (2019), Seite)

Skinner, B.F. (1971: *Jenseits von Freiheit und Würde*. 2018. Hamburg: Rowohlt
(zit. Singer (1971), Seite)

Snell, Bruno (1975): Die Entdeckung des Geistes. Studien zur Entstehung des europäischen Denkens bei den Griechen. 2009. Göttingen: Vandenhoeck & Ruprecht
(zit. Snell (1975), Seite)

Soldati, Gianfranco (2016): „Intentionale Inexistenz und Bewusstsein". In: *Studia Philosophica. Intentionalität und Subjektivität*. Schweizerische Zeitschrift für Philosophie.75, 2016. Basel: Schwabe. S. 83-100
(zit. Soldati (2016), Seite)

Sparrow, Robert (2007): „Killer Robots". In: *Journal of Applied Philosophy*. Vol. 24. No. 1, 2007. S. 62-77
(zit. Sparrow (2007), Seite)

Spearman, C. (1904): "'General intelligence,' objectively determined and measured". In: *The American Journal of Psychology*. 1904 15(2). S. 201–293. https://doi.org/10.2307/1412107
(zit. Spearman (1904), Seite)

Standing Committee of the One Hundred Year Study of Artificial Intelligence (2021): *Gathering Strength, Gathering Storms: The One Hundred Year Study on Artificial Intelligence (AI100) 2021 Study Panel Report*. Technical report. Stanford University, Stanford
(zit. Standing Committee of the One Hundred Year Study of Artificial Intelligence (2021), Seite)

Stekeler, Pirmin (2014): *Hegels Phänomenologie des Geistes. Ein dialogischer Kommentar*. Band 1: Gewissheit und Vernunft. 2012. Hamburg: Meiner
(zit. Stekeler (2014), Seite)

Stoecker, Ralf (2019): *Theorie und Praxis der Menschenwürde*. 2019. Paderborn: mentis
(zit. Stoecker (2019), Seite)

Sturma, Dieter (2003): „Autonomie. Über Personen, Künstliche Intelligenz und Robotik". In: Christaller, Thomas; Wehner, Joseph (Hrsg.): *Autonome Maschinen*. 2003. Wiesbaden: Westdeutscher Verlag. S. 38-55
(zit. Sturma (2003), Seite)

Szanto, Thomas (2012): *Bewusstsein, Intentionalität und mentale Repräsentation.* 2012. Berlin/Boston: De Gruyter
(zit. Szanto (2012), Seite)

Taylor, Charles (1989): *Sources of the Self. The making of Modern Identity*. 1989. Cambridge: Cambridge University Press
(zit. Taylor (1989), Seite)

Tegmark, Max (2017): *Leben 3.0. Mensch sein im Zeitalter Künstlicher Intelligenz.* 1. dt. Auflage 2019. Berlin: Ullstein
(zit. Tegmark (2017), Seite)

Thaler, Richard H.; **Sunstein**, Cass (2008): *Nudge*. 2008. New Haven: Yale University Press
(zit. Thaler Sunstein (2008), Seite)

Thaler, Richard H.; **Sunstein**, Cass (2009): *Nudge. Wie man kluge Entscheidungen anstößt*. 2009. Berlin: Econ Verlag/Ullstein Taschenbuch
(zit. Thaler Sunstein (2009), Seite)

Thurstone, L. L. (1946). „Theories of intelligence". In: *Scientific Monthly*. 1946 62. New York, S. 101-112.
(zit. Thurstone (1946), Seite)

Tononi, Giulio (2004): "An information integration theory of consciousness". In: *BMC Neuroscience*. 2004. https://bmcneurosci.biomedcentral.com/articles/10.1186/1471-2202-5-42
(zit. Tononi (2004), Seite)

Turing, Alan (1950): „Computing Machinery and Intelligence“. In *Mind*, Bd. 59, Nr. 236 (October 1950), S. 433ff
(zit. Turing (1950), Seite)

Vallor, Shannon (2016): *Technology and the Virtues. A philosophical guide to a future worth wanting.* Paperback edition. 2018. New York: Oxford University Press
(zit. Vallor (2016), Seite)

Villhauer, Bernd (2009): “Mündigkeit und Unmündigkeit nach Kants Schrift „Beantwortung der Frage. Was ist Aufklärung?“. In: *Der mündige Mensch. Denkmodelle der Philosophie, Geschichte, Medizin und Rechtswissenschaft*. Böhme, Gernot (Hrsg.). 2009. Darmstadt: Wissenschaftliche Buchgesellschaft. S. 13-23
(zit. Villhauer (2009), Seite)

Vinge, Vernor (1993): “The coming technological singularity: How to survive in the post-human era”. In: *The technological singularity. Managing the journey.* Callaghan et al. (Hrsg.). 2017. Berlin: Springer. S. 245-256
(zit. Vinge (1993), Seite)

Vossenkuhl, Wilhelm (2006): *Die Möglichkeit des Guten. Ethik im 21. Jahrhundert.* 2006. München: C.H. Beck
(zit. Vossenkuhl (2006), Seite)

Vossenkuhl, Wilhelm (2017): „The Practice of Following Rules“. In: *Wittgenstein-Studien 8*, 2017, S. 137-158
(zit. Vossenkuhl (2017), Seite)

Vossenkuhl, Wilhelm (2021): *Was gilt. Über den Zusammenhang zwischen dem, was ist, und dem, was sein soll*. 2021. Hamburg: Meiner
(zit. Vossenkuhl (2021), Seite)

Wadephul, Christian (2016): „*Führt Big Data zur abduktiven Wende in Wissenschaften?“* In: *Berliner Debatte Initial*, Bd. 27, S. 2016-4
(zit. Wadephul (2016), Seite)

Wagner, Hans (1992): *Die Würde des Menschen*. 1992. Würzburg: Königshausen und Neumann
(zit. Wagner (1992), Seite)

Wagner, Johanna (2020): *Künstliche Intelligenzen als moralisch verantwortliche Akteure*. 2020. Paderborn: Brill mentis
(zit. Wagner (2020), Seite)

Walde, Bettina (2002): *Metaphysik des Bewusstseins. Ein naturalistischer Erklärungsansatz*. 2002. Paderborn: mentis
(zit. Walde (2002), Seite)

Walde, Bettina (2006): *Willensfreiheit und Hirnforschung. Das Freiheitsmodell des egistemischen Libertarismus*. 2006. Paderborn: mentis
(zit. Walde (2006), Seite)

Warzel, Charlie (2023): „What have humans just unleashed? Call it tech’s optical-illusion era: Not even the experts know exactly what will come next in the AI revolution.”. In: *The Atlantic*. March 16 2023
(zit. Warzel (2023))

Weber, Max (1919): *Wissenschaft als Beruf*. 2015. Stuttgart: Reclam
(zit. Weber (1919a), Seite)

Weber, Max (1919): *Politik als Beruf*. 2017. Stuttgart: Reclam
(zit. Weber (1919b), Seite)

Weber, Max (1920): *Gesammelte Aufsätze zur Religionssoziologie*. 9. Auflage 1988. Tübingen: Mohr
(zit. Weber (1920), Seite)

Weber, Max (1922): Wirtschaft und Gesellschaft. Hg.: Alexander Ulfig. 2005. Zwei Teile in einem Band. Frankfurt a. M.: Zeitausendeins
(zit. Weber (1922), Seite)

Wellmann, Karl-Heinz; **Thimm**, Utz (Herausgeber, 1999): *Intelligenz zwischen Mensch und Maschine. Von der Hirnforschung zur künstlichen Intelligenz.* 1999. Münster: LIT
(zit. Wellmann Thimm (1999), Seite)

Wetz, Franz Josef (Hrsg., 2019): *Texte zur Menschenwürde*. 2. Auflage 2019. Stuttgart: Reclam
(zit. Wetz (2019), Seite)

Wilhelms, Günter (2017): "Systemverantwortung". In: *Handbuch Verantwortung*. Heidbrink, Ludger; Langbehn, Claus; Loh, Janina (Hrsg.). 2017. Wiesbaden: Springer VS
(zit. Wilhelms (2017), Seite)

Wiener, Norbert (1952): *Mensch und Menschmaschine*. Ausgabe von 2022. Frankfurt am Main: Klostermann
(zit. Wiener (1952), Seite)

Willaschek, Marcus; et al. (Hrsg., 2015): *Kant Lexikon. Band 1, 2 & 3.* 2021. Berlin/Boston: Walter de Gruyter
(zit. Willaschek (2015), Seite)

Willaschek, Marcus; et al. (Hrsg., 2017): *Kant Lexikon. Studienausgabe*. 2017. Berlin/Boston: Walter de Gruyter
(zit. Willaschek (2017), Seite)

Wolff, Johanna (2015): „Eine Annäherung an das Nudge-Konzept nach Richard H. Thaler und Cass R. Sunstein aus rechtswissenschaftlicher Sicht". In: *RW Rechtswissenschaft*. Jahrgang 6 (2015). Heft 2. S. 194-222
(zit. Wolff (2015), Seite)

Wolff, Michael (2013): „Kants Auflösung des Leib-Seele-Problems". In: Dina Emundts & Sally Sedgwick: *Internationales Jahrbuch des Deutschen Idealismus. Bewusstsein*. 11/2013. Berlin/Boston: Walter de Gruyter, S. 49-76
(zit. Wolff (2013))

Wörterbuch der philosophischen Begriffe (2013). Hrsg.: Regenbogen, Arnim; Meyer, Uwe. 2013. Hamburg: Felix Meiner
(zit. Wörterbuch der philosophischen Begriffe (2013), Seite)

Yeung, Karen (2017): „'Hypernudge': Big Data as a mode of regulation by design". In: *Information, Communication & Society*. 2017, 20 (1). S. 118-–136. DOI: 10.1080/1369118X.2016.1186713
(zit. Yeung (2017), Seite)

Zaborowski, Holger (2014): "Unterwegs zur Gelassenheit. Überlegungen zur Bedeutung von Heideggers Denken". In: Martin Heidegger: *Gelassenheit. Heideggers Meßkircher Rede von 1955*. 2. Auflage 2015. München: Verlag Karl Alber. S. 71-104
(zit. Zaborowski (2014), Seite)

Zahavi, Dan (2008): „Intentionalität und Bewusstsein (V. Logische Untersuchungen, §§ 1-21, Beilage der VI. Untersuchung)". In: Mayer, Verena (Hrsg.): *Edmund Husserl. Logische*

Untersuchungen. 2008. Berlin: Akademie Verlag
(zit. Zahavi (2008), Seite)

Zipp, Jan Sebastian; **Vey**, Karin (2018): „Das kreative System – Überlegungen zur künstlichen Kreativität“. In: *Informatik Spektrum*. 41_1_2018. 2018, S. 27-37
(zit. Zipp Vey (2018), Seite)

Zuboff, Shoshana (2018a): *The age of surveillance capitalism. Paperback edition* 2019. London: Profile books
(zit. Zuboff (2018a), Seite)

Zuboff, Shoshana (2018b): *Das Zeitalter des Überwachungskapitalismus.* 2018. Frankfurt/New York: Campus
(zit. Zuboff (2018b), Seite)

Personenverzeichnis / Biografien

Die biografischen Angaben zu den zitierten Autoren dienen der Einordnung der entsprechenden Personen in Bezug auf Disziplinen, akademische Spezialisierung und Rollen. Sie entstammen, wenn nicht anders angegeben, öffentlich zugänglichen Quellen einschließlich Netz-Ressourcen.

Adorno, Theodor W. (Wiesengrund) (*1903 in Frankfurt am Main, †1969 in Visp, Kt. Wallis), „*dt. Philosoph, Kulturkritiker und Musiktheoretiker, Studium in Frankfurt am Main und in Wien (Komposition); nach der Emigration, zuerst nach England (1934), dann nach Amerika (1938), Mitarbeiter des Instituts für Sozialforschung in New York; ab 1949 Professor in Frankfurt a. Main. Adorno ist neben M. Horkheimer der wichtigste Vertreter der Kritischen Theorie in ihrer ersten Generation.*“ Quelle: Philosophisches Wörterbuch (2009), 23. Auflage, S. 7

Arendt, Hannah (*1906 in Hannover; †1975 in New York), „*dt.-amerik. Philosophin und politische Denkerin; studierte u.a. bei M. Heidegger, K. Jaspers und R. Bultmann. Nach der Emigration lebte sie zunächst in Frankreich, ab 1941 in den USA. Dort Lehrtätigkeit an verschieden Universitäten, ab 1968 Prof. an der New School for Social Research in New York. Im Zentrum von A.s Werk steht die Frage nach dem Wesen der Politik, die sie in kritischer Anlehnung an die Daseinsanlyse Heideggers ausarbeitet und unter der Voraussetzung einer radikalen Gefährdung des Politischen in der Moderne erörtert.*“ Quelle: Philosophisches Wörterbuch (2009), 23. Auflage, S. 50

Bayes, Thomas (* um 1701 in London; †1761 in Tunbridge Wells) war ein englischer Mathematiker, Statistiker, Philosoph und presbyterianischer Pfarrer. Nach ihm ist der Satz von Bayes benannt, der in der Wahrscheinlichkeitsrechnung große Bedeutung hat.

Beckermann, Ansgar (*1945 in Hamburg) war Professor der Geistesphilosophie an der Universität Bielefeld.

Bengio, Yoshua (*1964 in Paris) ist ein kanadischer Informatiker. Er wurde bekannt für seine Forschung zu künstlichen neuronalen Netzen und Deep Learning, für die er als einer der Pioniere mit Geoffrey Hinton und Yann LeCun gilt.

Bieri, Peter (*1944 in Bern; †2023 in Berlin) war ein Schweizer Philosoph und Schriftsteller.

Block, Ned Joel (*1942 in Chicago) ist ein US-amerikanischer Philosoph und Professor an der New York University.

Boden, Margaret Ann (*1936 in London) ist Professorin für Psychologie und Philosophie an der University of Sussex. In ihren Publikationen und Arbeiten beschäftigt sie sich immer wieder mit Fragen der Künstlichen Intelligenz.

Bois-Reymond, Emil Heinrich du (*1818 in Berlin; †1896 ebenda) war ein deutscher Physiologe und theoretischer Mediziner, der als Begründer der experimentellen Elektrophysiologie und Mitbegründer des Faches Physiologie als naturwissenschaftlicher Disziplin gilt.

Brendel, Elke (*1962 in Frankfurt am Main) ist eine deutsche Philosophin und Professorin für Logik und Grundlagenforschung an der Universität Bonn.

Brentano, Franz Clemens Honoratus Hermann Josef (*1838 Boppard am Rhein; †1917 in Zürich) war ein deutscher Philosoph und Psychologe.

Brüntrup, Godehard (*1957 in Fulda) Godehard Brüntrup SJ ist ein deutscher Philosoph und Jesuit und seit 2003 Professor für Philosophie an der Hochschule für Philosophie München mit den Schwerpunkten Metaphysik, Philosophie des Geistes und Sprachphilosophie.

Bunge, Mario Augusto (*1919 in Buenos Aires; † 2020 in Montreal, Québec, Kanada) war ein argentinischer Philosoph und Physiker.

Burge, Charles Tyler (*1946) ist ein amerikanischer Philosoph und Professor an der University of California, Los Angeles.

Cassirer, Ernst (*1874 in Breslau; †1945 in New York), „*war Schüler von H. Cohen, gehörte der Marburger Schule des Neukantianismus an und gilt heute als ein Pionier der Kulturphilosophie. 1919-33 Prof. in Hamburg, emigrierte nach England und lehrte 1933-35 in Oxford, dann in Göteborg, ab 1941 in den USA an der Yale University, 1944-45 an der Columbia University in New York*.“ Quelle: Philosophisches Wörterbuch (2009), 23. Auflage, S. 125

Cattell, Raymond Bernard (*1905 in West Bromwich im ehemaligen Staffordshire, England; † 1998 in Honolulu) war ein britisch-US-amerikanischer Persönlichkeitspsychologe.

Chalmers, David (*1966 in Sydney, Australien) ist ein australischer Philosoph. Seine Hauptarbeitsgebiete liegen im Bereich der Sprachphilosophie und der Philosophie des Geistes.

Church, Alonzo (*1903 in Washington, D.C.; †1995 in Hudson, Ohio) war ein US-amerikanischer Mathematiker, Logiker und Philosoph und einer der Begründer der theoretischen Informatik.

Churchland, Paul M. (*1942) ist ein an der University of California in San Diego lehrender kanadischer Philosoph. Er ist Ehemann der Philosophin Patricia Churchland. Sein Hauptarbeitsgebiet liegt in der Philosophie des Geistes und der Neurophilosophie.

Crick, Francis Harry Compton (*1916 in Northampton, England; †2004 in San Diego, USA) war ein britischer Physiker und Molekularbiologe. Er erhielt 1962 zusammen mit James Watson und Maurice Wilkins den Medizin-Nobelpreis für die Entdeckung der Molekularstruktur der Desoxyribonukleinsäure (DNA). Später wandte er sich den Neurowissenschaften und der Theorie des Bewusstseins zu.

Cruse, Holk (*1942 in Stuttgart) ist ein deutscher Biologe, der auf dem Gebiet der Biokybernetik arbeitet.

Davidson, Donald Herbert (*1917 in Springfield, Massachusetts; † 2003 in Berkeley, Kalifornien) war ein US-amerikanischer analytischer Philosoph und ein Schüler von Willard Van Orman Quine.

Dennett, Daniel Clement (*1942 in Boston) ist ein US-amerikanischer Philosoph und gilt als einer der führenden Vertreter in der Philosophie des Geistes. Er ist Professor für Philosophie und Direktor des Zentrums für Kognitionswissenschaft an der Tufts University.

Descartes, René (latinisiert Renatus Cartesius; *1596 in La Haye en Touraine; †1650 in Stockholm) war ein französischer Philosoph, Mathematiker und Naturwissenschaftler, *„ ‘Vater der neueren Philosophie‘ genannt, denn er begründete den von der Souveränität der Vernunft überzeugten modernen Rationalismus.“* Quelle: Philosophisches Wörterbuch (2009), 23. Auflage, S. 158

Dretske, Fred (*1932 in Illinois, USA; †2013) war ein amerikanischer Philosoph der Erkenntnistheorie und der Philosophie des Geistes

Edelman, Gerald Maurice (*1929 in New York City; †2014 in La Jolla, Kalifornien) war ein US-amerikanischer Mediziner, Biochemiker und Molekularbiologe (Immunologie, Neurowissenschaft). 1972 erhielt er gemeinsam mit Rodney R. Porter den Nobelpreis für Physiologie oder Medizin für seine Entdeckungen im Bereich der chemischen Struktur von Antikörpern. Neben seinen Forschungen zur Immunologie schuf Edelman, der sich ab etwa 1972 den Neurowissenschaften zuwandte, Theorien zum menschlichen Bewusstsein und die sogenannte Neural Group Selection Theory, die Entwicklungsprozesse im Gehirn beschreibt.

Falkenburg, Brigitte (*1953 in Nürnberg) ist eine deutsche Physikerin und Philosophin. Von 1997 bis 2019 war sie Professorin für Theoretische Philosophie mit Schwerpunkt Philosophie der Wissenschaft und Technik an der TU Dortmund.

Fodor, Jerry Alan (*1935 in New York City; †2017 ebd.) war ein amerikanischer Philosoph und Kognitionswissenschaftler. Er lehrte an der Rutgers University in New Jersey.

Foerster, Heinz von (*1911 als Heinz von Förster in Wien; †2002 in Pescadero, Kalifornien) war ein österreichischer Physiker, Kybernetiker und Philosoph.

Gabriel, Markus (*1980 in Remagen) ist ein deutscher Philosoph. Er lehrt seit 2009 als Professor an der Universität Bonn.

Gardner, Howard Earl (*1943 in Scranton, Pennsylvania) ist ein US-amerikanischer Erziehungswissenschaftler. Er ist Professor für Kognition und Pädagogik an der Harvard Graduate School of Education und außerordentlicher Professor für Psychologie an der Harvard University.

Gödel, Kurt Friedrich (*1906; †1978) war ein österreichischer und später amerikanischer Mathematiker, Philosoph und einer der bedeutendsten Logiker des 20. Jahrhunderts.

Good, Irving John (*1916; †2009) war ein britischer Mathematiker, der im zweiten Weltkrieg als Kryptologe mit Alan Turing im Bletchey Park Projekt zusammengearbeitet hat. Die Zusammenarbeit mit Turing setze er nach dem Krieg fort, insbesondere in Bezug auf die Entwicklung von Computern und Bayesian Statistics. Er war Professor an der University of Manchester und später an der Virginia Tech.

Gregory, Richard L. (*1923 in London; †2010 in Bristol) war ein britischer Psychologe und Neurowissenschaftler.

Guilford, Joy Paul (*1897 in Marquette, Nebraska; † 1987 in Los Angeles) war ein faktorenanalytisch arbeitender Persönlichkeits- und Intelligenzforscher.

Harari, Yuval Noah (*1976 in Kiryat Ata, Bezirk Haifa) ist ein israelischer Historiker.

Hebb, Donald Olding (*1904 in Chester, Nova Scotia, Kanada; †1985 ebenda) war ein kanadischer kognitiver Psychobiologe und Professor für Psychologie an der McGill-Universität in Montreal, Kanada.

Heyns, Christof (*1959; †2021) war ein südafrikanischer Rechtswissenschaftler. Er war Professor für Menschenrechtsfragen, Direktor des Instituts für Internationales und vergleichendes Recht in Afrika an der Universität Pretoria sowie Sonderberichterstatter der Vereinten Nationen über außergerichtliche, summarische oder willkürliche Hinrichtungen.

Hinton, Geoffrey E. (*1947 in Wimbledon, Großbritannien) ist ein britischer Informatiker und Kognitionspsychologe, der ebenfalls wie Yann LeCun und Yoshua Bengio, seine Mitpreisträger des Turing-Awards von 2018 vor allem für seine Beiträge zur Theorie künstlicher neuronaler Netze bekannt ist.

Hume, David (*1711 in Edinburgh; †1776 ebenda) war ein schottischer Philosoph, Ökonom und Historiker.

Husserl, Edmund Gustav Albrecht (*1859 in Proßnitz in Mähren, Kaisertum Österreich; †1938 in Freiburg im Breisgau) war ein österreichisch-deutscher Philosoph und Mathematiker und Begründer der philosophischen Strömung der Phänomenologie.

Horkheimer, Max (*1895 in Stuttgart, †1973 in Nürnberg), „*dt. Philosoph und Soziologe Studium der Philosophie in Frankfurt am Main, München und Freiburg, Promotion in Frankfurt 1922, Habilitation 1925 (mit einer Arbeit über I. Kants Kritik der Urteilskraft) 1930-33 Prof. für Philosophie daselbst, ab 1933 Direktor des von ihm gegründeten Instituts für Sozialforschung, das er im amerikanischen Exil in New York, später in Los Angeles weiterführte. Nach seiner Rückkehr 1947 lehrte er bis zu seiner Emeritierung in Frankfurt a. M. Philosophie und war 1951-53 Rektor der dortigen Universität.*“ Quelle: Philosophisches Wörterbuch (2009), 23. Auflage, S. 321

Jackson, Frank Cameron (*1943) ist ein australischer Philosoph. Er studierte an der Universität Melbourne Mathematik und Philosophie und ist heute Professor für Philosophie an der Australian National University. Seine Hauptarbeitsgebiete sind die Philosophie des Geistes und die Metaphysik.

Jäger, Adolf Otto (*1920; †2002) war Professor an der Freien Universität Berlin. Er formulierte das Berliner Intelligenzstrukturmodell.

Jonas, Hans (*1903 in Mönchengladbach; †1993 in New Rochelle) war ein deutsch-amerikanischer Philosoph, der von 1955 bis 1976 als Professor an der New School for Social Research in New York lehrte.

Kanitscheider, Bernulf (*1939 in Hamburg; † 2017) war ein Philosoph und Wissenschaftstheoretiker.

Kim, Jaegwon (*1934; †2019) war ein amerikanischer Philosoph koreanischer Abstammung, der seine Hauptarbeitsbereiche in der Philosophie des Geistes und der Erkenntnistheorie hatte. Er war zuletzt Professor für Philosophie an der Brown University.

Koch, Christof (*1956 in Kansas City, USA) ist ein deutsch-amerikanischer Neurowissenschaftler.

Koenig, Gaspard (*1982) ist ein französischer Essayist und Philosoph.

Kurzweil, Ray (*1948 in Queens, New York) ist ein US-amerikanischer Autor, Erfinder, Futurist und Leiter der technischen Entwicklung (Director of Engineering) bei Google LLC.

LeCun, Yann (*1960 in Soisy-sous-Montmorency) ist ein französischer Informatiker und Träger des Turing Awards 2018. Er wurde bekannt mit der Entwicklung der Convolutional Neural Networks (CNN) bei Bell Labs und seine gemeinsamen Arbeiten mit Yoshua Bengio und Geoffrey Hinton auf dem Gebiet der künstlichen neuronalen Netzwerke.

Leibniz, Gottfried Wilhelm (*1646 in Leipzig; †1716 in Hannover) war ein deutscher Philosoph, Mathematiker, Jurist, Historiker und politischer Berater der frühen Aufklärung.

Lewis, David Kellogg (*1941 in Oberlin, Ohio; † 2001 in Princeton, New Jersey) war ein US-amerikanischer Philosoph.

Libet, Benjamin (*1916 in Chicago, Illinois; † 2007 in Davis, Kalifornien) war ein US-amerikanischer Physiologe.

Lovelace, Ada (geborene Hon. Augusta Ada Byron; *1815 in London; †1852 ebenda), war eine britische Mathematikerin.

Lucas, John Randolph (*1929; †2020 in Somerset) war ein britischer Philosoph.

Mainzer, Klaus (*1947 in Opladen) ist ein deutscher Philosoph und Wissenschaftstheoretiker.

Margalit, Avishai (*1939 in Afula, Palästina) ist ein israelischer Philosoph. Er befasste sich im Schwerpunkt mit der Philosophie der Sprache, dem Thema der praktischen Vernunft und der Sozial- und politischen Philosophie.

McCarthy, John (*1927 in Boston, Massachusetts; † 2011 in Palo Alto, Kalifornien) war ein US-amerikanischer Logiker, Informatiker und Autor. Er gilt als der Erfinder der Programmiersprache LISP.

McCulloch, Warren Sturgis (*1898 in Orange, New Jersey; †1969 in Cambridge, Massachusetts) war ein amerikanischer Neurophysiologe und Kybernetiker. Er ist bekannt für frühe Arbeiten über die Funktionsweise des Gehirns und neuronale Netze.

Metzinger, Thomas (*1958 in Frankfurt am Main) ist ein deutscher Philosoph und Seniorprofessor für theoretische Philosophie an der Universität Mainz. „*Er studierte an der Johann Wolfgang Goethe-Universität in Frankfurt/Main Philosophie, Ethnologie und Theologie. Dort promovierter er 1985 über das Leib-Seele-Problem. 1992 erlangte er die Habilitation an der Justus-Liebig-Universität in Gießen. Im Jahr 2000 wurde er Professor für Philosophie der Kognitionswissenschaft an der Universität Osnabrück. Allerdings wechselte er kurze Zeit später an die Universität Mainz. Zu seinen Forschungsschwerpunkten gehören u.a.: Analytische Philosophie des Geistes; Wissenschaftstheorie und philosophische Probleme der Neuro- und Kognitionswissenschaften und der Psychologie; Wissenschaftstheorie und philosophische Probleme der Künstliche-Intelligenz-Forschung; Geschichte des Leib-Seele-Problems nach dem Zweiten Weltkrieg; neuere Theorien des Geistes, insbesondere des phänomenalen Bewusstseins und der mentalen Repräsentation.“* (Quelle: https://www.giordano-bruno-stiftung.de/beirat/metzinger-thomas (Download vom 29.7.2022))

Minsky, Marvin Lee (*1927 in New York; †2016 in Boston, Massachusetts) war ein amerikanischer Forscher auf dem Gebiet der künstlichen Intelligenz (KI).

Nagel, Thomas (*1937 in Belgrad) ist ein US-amerikanischer Philosoph. Er lehrt an der New York University School of Law und bearbeitet ein weites Themenspektrum. Er lehrte unter anderem an der University of California in Berkeley und an der Princeton University.

Nussbaum, Martha (*1947), ist Professorin für Rechtswissenschaften und Ethik an der University of Chicago. Sie gilt als eine der einflussreichsten Philosophinnen der Gegenwart.

O'Neill, Onara (*1941 in Aughafatten, Nordirland) ist eine britische Philosophin und Politikerin. Sie lehrte als Professorin an der Cambridge University.

Pearle, Judea (*1936 in Tel Aviv) ist ein israelisch-amerikanischer Informatiker und Philosoph mit Schwerpunkt in Fragen der KI.

Peirce, Charles Santiago Sanders (*1839 in Cambridge, Massachusetts; †1914 in Milford, Pennsylvania) war ein US-amerikanischer Mathematiker, Philosoph, Logiker und Semiotiker.

Penrose, Roger (*1931 in Colchester, Essex) ist ein britischer Mathematiker und theoretischer Physiker. Ihm wurde 2020 für seine Arbeiten an Schwarzen Löchern und die Allgemeine Relativitätstheorie der Nobelpreis für Physik zur Hälfte zuerkannt.

Pentland, Alex (*1951 in USA) ist Professor am MIT und der Direktor des dortigen Human Dynamics Laboratory sowie der Forschungsinitiative MIT Connection Science.

Pinker, Steven (*1954 in Montreal) ist ein US-amerikanisch-kanadischer Experimentalpsychologe, Kognitionswissenschaftler, Linguist, Beiratsmitglied der deutschen Partei der Humanisten und populärwissenschaftlicher Autor. Prof. an der Harvard University.

Pitts, Walter (*1923 in Detroit, Michigan; †1969 in Cambridge, Massachusetts) war ein amerikanischer Logiker, der auf dem Gebiet der kognitiven Psychologie arbeitete.

Popitz, Heinrich (*1925 in Berlin; †2002 in Freiburg im Breisgau) war ein deutscher Soziologe, der vor dem Hintergrund der Philosophischen Anthropologie bedeutende Beiträge zur Allgemeinen Soziologie leistete. Popitz publizierte insbesondere zu elementaren Begriffen wie Soziale Norm, Soziale Rolle oder Macht und Gewalt.

Prinz, Wolfgang (*1942 in Ebern, Unterfranken) ist ein deutscher Psychologe und Kognitionswissenschaftler.

Putnam, Hilary Whitehall (*1926 in Chicago, Illinois; †2016 in Arlington, Massachusetts) war ein amerikanischer Philosoph. Er gilt als eine der Schlüsselfiguren der Sprachphilosophie und der Philosophie des Geistes im 20. Jahrhundert.

Ritter, Helge (*1958 in Naila) ist ein deutscher Neuroinformatiker und Professor an der Universität Bielefeld.

Rost, Detlef H. (*1945 in Olsberg) ist ein Marburger Psychologe, der vor allem durch das von ihm initiierte Marburger Hochbegabtenprojekt (MHP) bekannt wurde.

Roth, Gerhard (*1942 in Marburg; †2023) war ein deutscher Biologe und Hirnforscher. Er war Direktor am Institut für Hirnforschung in Bremen. Mit Buchpublikationen beteiligte er sich an aktuellen neurobiologischen und philosophischen Streitfragen.

Russell, Stuart Jonathan (*1962 in Portsmouth) ist ein britischer Wissenschaftler auf dem Gebiet der künstlichen Intelligenz. Er ist Professor für Informatik an der Universität von Berkeley, Kalifornien.

Sautoy, Marcus Peter Francis du (*1965 in London) ist ein britischer Mathematiker und Autor von populärwissenschaftlichen Büchern.

Searle, John Rogers (*1932 in Denver, Colorado) ist ein amerikanischer Philosoph. Seine Hauptarbeitsgebiete sind die Sprachphilosophie, die Philosophie des Geistes, Sozialontologie sowie Teile der Metaphysik. Searle war Professor für Philosophie an der University of California, Berkeley.

Sen, Amartya (*1933) ist emeritierter Professor für Philosophie und Professor für Ökonomie an der Harvard Universität. 1998 erhielt er den Nobelpreis für Ökonomie.

Shannon, Claude Elwood (*1916 in Petoskey, Michigan; †2001 in Medford, Massachusetts) war ein US-amerikanischer Mathematiker und Elektrotechniker. Er gilt als Begründer der Informationstheorie.

Sharkey, Amanda war bis zu ihrer Pensionierung Associate Professor (Senior Lecturer) im Depart-ment of Computer Science an der Universität Sheffield, UK; Noel E. Sharkey (*1948) ist ein briti-scher Informatiker und Professor für künstliche Intelligenz und Robotik ebenfalls an der Universität Sheffield.

Sherrington, Sir Charles Scott (*1857 in London; † 1952 in Eastbourne, Sussex) war ein britischer Neurophysiologe. Für seine Entdeckungen auf dem Gebiet der Funktionen der Neuronen erhielt er 1932 gemeinsam mit Edgar Douglas Adrian den Nobelpreis für Medizin.

Simon, Herbert Alexander (*1916; † 2001) war ein US-amerikanischer Sozialwissenschaftler. Im Jahr 1978 erhielt er den Alfred-Nobel-Gedächtnispreis für Wirtschaftswissenschaften.

Singer, Wolf Joachim (*1943 in München) ist ein deutscher Neurophysiologe und Hirnforscher.

Skinner, B. F. (* 1904; † 1990), war ein amerikanischer Psychologe und Verhaltensforscher, der als einer der führenden Vertreter des Behaviorismus gilt.

Snell, Bruno (*1896 in Hildesheim; †1986 in Hamburg) war ein klassischer Philologe, Hochschullehrer, Universitätsdekan und Rektor.

Soldati, Gianfranco (*1959 in Locarno) ist ein Schweizer Philosoph und Professor für Philosophie.

Spearman, Charles Edward (*1863 in London; † 1945 ebenda) war ein britischer Psychologe, der unter anderem durch seine 1904 publizierte Zweifaktorentheorie der Intelligenz bekannt wurde.

Sternberg, Robert J. (*1949 in New Jersey) ist ein US-amerikanischer Psychologe.

Sturma, Dieter (*1953 in Minden) ist ein deutscher Philosoph. Sturma ist seit 2007 Professor für Philosophie unter besonderer Berücksichtigung der Ethik in den Biowissenschaften an der Universität Bonn.

Tegmark, Max Erik (*1967 in Stockholm) ist ein schwedisch-US-amerikanischer Kosmologe und Wissenschaftsphilosoph.

Thaler, Richard H. (*1945 in East Orange, New Jersey) ist ein US-amerikanischer Wirtschaftswissenschaftler und Professor an der Booth School of Business der University of Chicago. Er gilt als einer der weltweit führenden Verhaltensökonomen.2017 wurde Thaler mit dem Nobelpreis für Wirt-schaftswissenschaften ausgezeichnet.

Thurstone, Louis Leon (*1887 in Chicago; †1955 in Chapel Hill, North Carolina) war ein US-amerikanischer Ingenieur und Psychologe.

Tononoi, Giulio (*1960 in Trient, Trentino-Südtirol) ist ein italienischer Facharzt für Psychiatrie und Neurowissenschaftler. Er hat eine Professur für Psychiatrie an der Universität Madison-Wisconsin und leitet das dortige Center for Sleep and Consciousness.

Turing, Alan Mathison (*1912 in London; †1954 in Wilmslow, Cheshire) war ein britischer Logiker, Mathematiker, Kryptoanalytiker und Informatiker. Er gilt heute als einer der einflussreichsten Theoretiker der frühen Computerentwicklung und Informatik.

Vinge, Vernor Steffen (*1944) ist ein amerikanischer Mathematiker, Informatiker und Science-Fiction-Autor.

Wagner, Hans (*1917 in Plattling, Niederbayern; † 2000 in Bonn) war ein deutscher Philosoph. Er lehrte und forschte als Professor in Würzburg (1953–1961) und Bonn (1961–1982) mit einer Unter-brechung durch eine Gastprofessur an der Yale University.

Walde, Bettina (*1972 in Schwäbisch Hall), Privatdozentin am Institut für Philosophie der LMU München

Wang, Hao (*1921 in Jinan, Provinz Shandong, China; †1995 in New York City) war ein chinesisch-US-amerikanischer Logiker, Mathematiker und Philosoph.

Wapnik, Wladimir Naumowitsch (*1936) ist ein russisch-amerikanischer Mathematiker und Hauptentwickler der Support-Vektor-Maschine (mit Alexei Jakowlewitsch Tscherwonenkis) sowie einer zugehörigen statistischen Lerntheorie (auch Vapnik-Chervonenkis-Theorie genannt).

Weber, Maximilian „Max“ Carl Emil (*1864 in Erfurt; †1920 in München) war ein deutscher Soziologe und Nationalökonom.

Weizenbaum, Joseph (*1923 in Berlin; † 2008 in Berlin) war Mathematiker und Informatiker.

Wolff, Michael (*1942 in Solingen) ist ein deutscher Philosoph. Von 1982 bis 2007 lehrte er als Professor für Philosophie an der Universität Bielefeld.

Zadeh, Lotfi A. (eigentlich Lotfali Askar-Zadeh, *1921 in Baku, Aserbaidschan; †2017 in Berkeley, Kalifornien) war ein US-amerikanischer Mathematiker, Informatiker, Elektroingenieur.

Zuboff, Shoshana ist eine amerikanische Ökonomin. Sie studierte Philosophie an der University of Chicago und promovierte in Sozialpsychologie an der Harvard University. Ab 1981 war sie Professorin an der Harvard Business School.

Glossar

Äquivok

„... gleichlautend, aber bedeutungsverschieden, also mehrdeutig. Die Verwendung äquivoker Begriffe nennt man Äquivokation; sie kann zu logischen Fehlschlüssen führen.“[916]

Bayes-Netze

„Bayes-Netze *stellen eine sehr vielseitig einsetzbare Struktur dar, die in vielen Bereichen der Informatik Anwendung finden. Bayes-Netze dienen der kompakten Speicherung und Verarbeitung unsicheren Wissens. Neue Informationen an einer Stelle des Netzes, in Form von Wahrscheinlichkeiten, wirken sich unter Umständen auf das ganze Netz aus. Darüber hinaus existieren Algorithmen, die es Bayes-Netzen ermöglichen selbstständig dazuzulernen.“*[917]

Behaviorismus

„Behaviorismus*, methodologische Richtung, nach der auf innere Zustände von anderen Personen nur geschlossen werden kann über das Verhalten, ihre Verhaltensäußerungen und Verhaltensdispositionen*[918]

„In seiner gröbsten Version besagt der Behaviorismus, dass der Geist einfach das Verhalten des Körpers ist. Es gibt nichts über das Verhalten des Körpers hinaus, das für das Mentale konstitutiv ist.“[919]

Bindungsproblem

„Wie können verschiedene Kortexareale und Modalitäten synchron und zusammenhängend agieren (und simultan Bewegung, Farbe, räumliche Ausrichtung und so weiter verarbeiten), obwohl sie räumlich voneinander getrennt sind und es kein übergeordnetes und exekutives Areal gibt? “[920]

Computationalismus
auch:
Computer-
funktionalismus

„Als ***Computationalismus, Komputationale Theorie des Geistes*** *(eng. computational theory of mind, CTM) oder auch als Computertheorie des Geistes wird eine in der Kognitionswissenschaft bzw. in den Neurowissenschaften verbreitete Theorie bezeichnet, wonach die kognitiven Fähigkeiten und das Bewusstsein des menschliche Geistes - im Sinn des englischen Begriffs „Mind“ - auf Berechnungsvorgängen beruhen, die vom Gehirn vollzogen werden.“*[921]

Degeneration

„***Degeneration****, 1) allg. Entartung, Zerfall, Rückbildung. 2)in der Molekularbiologie die Eigenschaft des genetischen Codes, dass die meisten Aminosäuren von mehreren Codons codiert werden. Dadurch lässt sich erklären, dass für die 20 biogenen Aminosäuren und drei Stopp-Codons 64 Tripletts vorhanden sind. Glycin, die einfachste Aminosäure, wird durch die vier Basentripletts GGU, GGA, GGC und GGG codiert.“*[922]

---

916 Wörterbuch der philosophischen Begriffe (2013), S. 60

917 Quelle: Marc Wagner, „Bayes-Netze – Eine Einführung“, 2000, Vortrag, Universität Erlangen,

918 Wörterbuch der philosophischen Begriffe (2013), S. 98

919 Searle (2006), S. 58

920 Edelman (2004), S. 149

921 Quelle: „anthrowiki.at“, 15.6.2019

922 Quelle: https://www.spektrum.de/lexikon/biologie-kompakt/degeneration/2914

Entelechie

„*Entelechie, gr., zusammengesetzt aus en >in<, telos >Ziel< und echein >haben<, das, was sein Ziel in sich selbst hat; seit Aristoteles die Form, die sich im Stoff verwirklicht, bes., die im Organismus liegende Kraft, die ihm von innen her zur Selbstentwicklung und -vollendung bringt. Dementsprechend bezeichnet Aristoteles die Seele als die erste Entelechie eines organischen, lebensfähigen Körpers. ...“*[923]

Epiphänomen

„ *Die Begleiterscheinung“*[924]

Epiphänomenalismus

„*Als Epiphänomen bezeichnet man ganz allgemein etwas, das zwar kausal verursacht wurde, aber selber keine bedeutsame kausale Wirkung hat.* ***Epiphänomenalismus*** *betrachtet etwa die Gedanken eines Menschen als Produkte von körperlichen Vorgängen, wobei weder die Gedanken auf den Körper zurückwirken noch zwischen den Gedanken selbst ursächliche Zusammenhänge bestehen. Der Epiphänomenalismus unterscheidet demnach die Bereiche Geist und Körper aber erlaubt keine Reduktion des Einen auf das Andere, und nimmt dabei an, dass Gehirnprozesse das Eigentliche sind und die geistigen Vorgänge ein reines Epiphänomen. In der Psychologie spielt der Epiphänomenalismus daher insofern eine Rolle, als er als eine spezielle Form des Dualismus betrachtet werden kann, der die Probleme des interaktionistischen Dualismus (Descartes) vermeidet, den mentalen Erlebnissen kausale Wirksamkeit als Ursache für folgende Ereignisse abspricht. Der Epiphänomenalismus postuliert daher, dass die Materie kausal auf den Geist wirkt, jedoch umgekehrt der Geist auf die Materie keinerlei Einfluss hat.“*[925]

Expertensystem

„*1. Begriff: in der Künstlichen Intelligenz (KI) wird ein Programm oder ein Softwaresystem als Expertensystem bezeichnet, wenn es in der Lage ist, Lösungen für Probleme aus einem begrenzten Fachgebiet (Wissensdomäne) zu liefern, die von der Qualität her denen eines menschlichen Experten vergleichbar sind oder diese sogar übertreffen (Expertenwissen).*

Bes. bewährt als Expertensysteme haben sich wissensbasierte Systeme; deshalb werden beide Begriffe oft synonym verwendet.

2. Bestandteile (Regelfall): Wissensbasis, Inferenzmaschine, Wissenserwerbskomponente, Dialogkomponenten und Erklärungskomponente.

3. Klassifikation nach Aufgabenstellung:

(1) Diagnosesysteme, die auf der Basis teils gegebener, teils zu suchender Symptome Fälle klassifizieren;

---

[923] Wörterbuch der philosophischen Begriffe (2013), S. 184

[924] Wörterbuch der philosophischen Begriffe (2013), S. 190

[925] Stangl, W. (2019). Stichwort: 'Epiphänomenalismus'. Online Lexikon für Psychologie und Pädagogik. WWW: https://lexikon.stangl.eu/21653/epiphaenomenalismus/ (2019-06-21)

(2) Beratungssysteme, die im Dialog mit dem Menschen eine auf den vorliegenden Fall bezogene Handlungsempfehlung geben;

(3) Konfigurationssysteme, die auf der Basis von Selektionsvorgängen unter Berücksichtigung von Unverträglichkeiten und Benutzerwünschen komplexe Gebilde zusammenstellen;

(4) Planungssysteme, die einen Ausgangszustand durch eine Folge von Aktionen in einen Endzustand überführen.“[926]

Funktionalismus

„*Der Funktionalismus besagt: einzelne, konkrete Gehirnzustände, also Token-Gehirnzustände, werden dadurch zu mentalen Zuständen, dass sie eine bestimmte Art von Funktion im Gesamtverhalten des Organismus erfüllen. Beispiel: Zu sagen, dass Jones glaubt, es regnet, heißt zu sagen, dass in ihm ein bestimmtes Ereignis, ein bestimmter Zustand oder Prozess stattfindet, der von bestimmten Arten äußerer Reizung verursacht wird – zum Beispiel nimmt er wahr, dass es regnet. Kurz gesagt, mentale Zustände werden als Zustände definiert, die bestimmte Funktionen haben.“*[927]

Fuzzy Logic

„*vage Logik, unscharfe Logik; Bereich der Logik, der die semantische Interpretation von Aussagen ermöglicht, die nicht als eindeutig wahr oder falsch eingestuft werden können (z.B. „Peter ist groß.”). Diskrete Wahrheitswerte (wahr und falsch bzw. 1 und 0) werden durch einen stetigen Bereich (i.d.R. Intervall von 0 bis 1) ersetzt. Für Werte aus diesem Bereich werden aussagenlogische Operationen definiert.“*[928]

Geltung

„*Geltung ist ein normativer Anspruch, der den Maßstäben, welche die Menschen im Denken u. Handeln beachten sollen, inhaltlich u. zeitlich Dauer u. Zuverlässigkeit gibt. Im Erkennen sind dies die Maßstäbe der Wahrheit u. ihrer Rechtfertigung, im Handeln sind es die Maßstäbe des Guten u. der Moral, in den Beziehungen der Menschen untereinander sind es die Würde, Freiheit, Gleichheit, Recht u. Gesetz, Treue u. Sympathie, in der Politik sind es die Menschenwürde, das Gemeinwohl, die Verfassung, u. die innere und äußere Sicherheit, in der Wirtschaft sind es die Werte der Arbeit, des Geldes u. der Produkte.*“[929]

Homonymie

„*Zu gr. Homonymia „Gleichnamigkeit“, ein Verhältnis zwischen zwei Wörtern gleicher Aussprache oder Schreibung, aber unterschiedlicher wortgeschichtlicher Herkunft und deutlich verschiedener Bedeutung.“*[930]

Individuum

„*Individuum (Mz. Individuen), lat., Lehnübers. Von gr. atomon ’Unteilbares‘; 1. Das Einzelwesen. Ursprgl fällt der Begriff I. zusammen mit*

[926] Gabler Wirtschaftslexikon (Online): https://wirtschaftslexikon.gabler.de/definition/expertensystem-35743

[927] Searle (2006), S. 71f

[928] Gabler Wirtschaftslexikon; Online: https://wirtschaftslexikon.gabler.de/definition/fuzzy-logic-34198

[929] Höffe et al. (1977), S. 97f

[930] Wörterbuch der philosophischen Begriffe (2013), S. 295

dem des Atoms. In der Scholastik war er auf die menschliche Persönlichkeit eingeschränkt, im 16. Jh. bereits erhält er 2. die Bedeutung ‚besondere Person', und heute bez. er meist den ‚Einzelmenschen' im Verhältnis bzw. Gegensatz zur Gemeinschaft (Individualismus)."[931]

Intentionale Zustände

„*Intentionale Zustände sind Zustände mit „semantischen Eigenschaften". Ein **erstes Beispiel** für eine semantische Eigenschaft ist das Besitzen eines Inhalts: Geistige Zustände, aber auch die Sätze einer Sprache und die in ihnen vorkommenden Ausdrücke, haben einen Inhalt. Physikalische Zustände haben nach allgemeinem Verständnis keinen Inhalt, denn für sich allein genommen bedeuten sie nichts.*

*Das **zweite Beispiel** für eine semantische Eigenschaft ist die Bezugnahme (Philosophen sprechen hier manchmal auch von „Referentialität" oder davon, dass ein geistiger Zustand – ganz ähnlich wie ein sprachlicher Ausdruck – „auf etwas referiert"). Das bedeutet dann, dass sie von etwas handeln, weil sie Bezug auf Einzeldinge oder Arten von Gegenständen in der Welt nehmen. Gedanken oder andere intentionale Zustände beziehen sich also immer auf etwas. Ein physikalischer Zustand dagegen verweist nicht über sich selbst hinaus, er überschreitet sozusagen sein einfaches Dasein in keiner Weise.*

*Das **dritte wichtige Beispiel** für eine semantische Eigenschaft ist der Besitz von „Korrektheitsbedingungen". Wenn die Beziehung zwischen dem intentionalen Zustand und seinem und seinem Gegenstand korrekt ist, dann müssen dafür nämlich bestimmte Bedingungen erfüllt sein. Das kann zum Beispiel heißen, dass er Erfüllungsbedingungen oder Wahrheitsbedingungen besitzt.*

Das also ist es, was wir – jedenfalls unter einer ersten Annäherung – meinen, wenn wir sagen, dass intentionale Zustände semantische Eigenschaften besitzen: Sie haben Inhalte, sie referieren auf etwas und sie können Erfüllungs- oder Wahrheitsbedingungen besitzen."[932]

Intentionalität

„*In einem allgemeinen Verständnis bezeichnet **Intentionalität** die Zielgerichtetheit des Handelns oder der Gefühle. Als philosophischer Terminus wurde er von Brentano zur Charakterisierung der psychischen Phänomene eingeführt. In seiner Psychologie vom empirischen Standpunkt zeigt Brentano auf, dass den psychischen Phänomenen wie Denken, Lieben und Hassen eine intentionale Struktur eigen ist. Zur näheren Charakterisierung führt er den Begriff der »mentalen Inexistenz« an. Er erläutert dies als eine Beziehung auf einen Inhalt, ein Gerichtetsein auf ein Objekt oder auch als immanente Gegenständlichkeit. Brentano verweist in diesem Zusammenhang darauf, dass das »Etwas-als-etwas-Vorstellen« der Eindeutigkeit des Begriffs entsprechen müsse, indem das*

[931] Wörterbuch der philosophischen Begriffe (2013), S. 314

[932] Zitiert aus Metzinger Bd. 3 (2010), S. 33

Etwas als Reales i.S. eines obersten Gattungsbegriffs für Dingliches aufzufassen sei.“[933]

Konnektionismus

„***Konnektionismus*** *m [von latein. conectere = verknüpfen], Englisch: connectionism, eine Schule in der kognitiven Psychologie und im Bereich der künstlichen Intelligenz (KI), die sich von dem bis ca. 1980 dominierenden Paradigma abgesetzt hat, demgemäß (natürliche und künstliche) Kognition wesentlich auf der sequentiellen Manipulation von Symbolen nach Art eines Computerprogramms beruht. Eine Wiege des Konnektionismus war das 1982 von Geoffrey Hinton, James McClelland und David Rumelhart in San Diego gegründete PDP-Projekt (Abk. für E parallel distributed processing: parallel-verteilte Informationsverarbeitung). Es werden Netzwerke von einfachen Verarbeitungseinheiten (Englisch: units; Einheit) betrachtet, die in massiv paralleler Weise prozessieren. Das Netzwerk als Ganzes verwandelt Eingaben (Englisch: inputs) in Ausgaben (Englisch: outputs), ohne dass den vielen räumlich und zeitlich verteilten Zwischenschritten ein benennbarer kognitiver Gehalt (eine Bedeutung) zugeordnet werden kann; man spricht deshalb auch von subsymbolischer Verarbeitung.“*[934]

Kybernetik

„***Kybernetik*** *(gr. kybernetikë (technë) ‘Steuermannskunst‘), von „N. Wiener (Cybernetics or control and communication in the animal and the machine, 1948) begründete und von ihm so benannte Wissenschaft von den kybernetischen Systemen, d.s. dynamische Systeme, die innerhalb eines bestimmten Stabilitätsbereichs über eine Folge von Systemzuständen durch Rückkopplung einem Gleichgewichtszustand zustreben. Allgemeine Merkmale solcher Systeme sind z.B. die Regelung und Informationsverarbeitung sowie die Selbstorganisation und Selbstreproduktion. Gegenstand der K. ist die mathematische Beschreibung und modellhafte Erfassung der Struktur und Funktion solcher Systeme (zur genaueren Bestimmung kybernetischer Systeme). Die Entwicklung der K. erfolgt innerhalb verschiedener Einzeldisziplinen, z.B. in der Spieltheorie. Unterschieden werden kann zwischen der theoretischen oder allgemeinen K., die sich formal mit der Struktur und dem Verhalten von Systemen beschäftigt, und der angewandten K., die sich mit der Anwendung der kybernetischen Methoden und Erkenntnisse auf unterschiedlichsten Gebieten beschäftigt. Ein Beispiel ist die ökonomische K., die sich mit den kybernetischen Merkmalen von Systemen innerhalb der Wirtschaft befasst. Im engeren Sinn wird der Begriff der K. auch verwendet als Sammelbezeichnung für die theoretischen Grundlagen der elektronischen Daten- und Informationsverarbeitung, Nachrichtentechnik und Automatentechnik."*[935]

---

[933] Zitiert aus Metzler Lexikon der Philosophie; Online: https://www.spektrum.de/lexikon/philosophie/intentionalitaet/993

[934] Lexikon der Neurowissenschaft, im Netz unter Spektrum.de

[935] Wörterbuch der philosophischen Begriffe (2013), S. 370f

Materialismus

„***Materialismus**, Neubildung von Materie, die Weltanschauung, nach der es keine andere Wirklichkeit gibt als die Materie, so dass auch Seele, Geist und Denken als Kräfte oder Bewegungen der Materie aufgefasst werden. […] Im 20. Jh. besonders einflussreich hat sich die These vom Parallelismus von Psychischem und Physischem erwiesen, so z.B. in der Identitätstheorie. […] Nach dieser Position sind Bewusstseinsphänomene mit Gehirnprozessen identisch; danach sind Aussagen über Bewusstseinsphänomene inadäquat, solange sie nicht in ihrer Struktur ihrer Gehirnprozesse aufgeklärt sind.“*[936] Die Begriffe grundsätzlich unterschiedlichen Materialismus und Physikalismus werden oft austauschbar verwendet.

Morphologie

„***Morphologie** w [von *morpho- , griech. logos = Kunde; Adj. morphologisch], Formenlehre, Ge-staltlehre, eine Disziplin der Biologie, die sich mit der Körper-Gestalt, dem Aufbau und den Lageverhältnissen der Organe (bei Einzellern der Organelle) von Lebewesen befaßt (Bauplan, Typus).*“[937]

Neuromorphe Hardware

„*Klassische Computer basieren auf einer sogenannten Von-Neumann-Architektur, in der Prozessorkerne sequenziell Befehle ausführen und dabei die Daten im zentralen Speicher bearbeiten. Das heißt, die Rechenleistung der Computersysteme ist abhängig von der Datenübertragungsrate zwischen Prozessor und Speicher. Man spricht hier vom »Von-Neumann-Flaschenhals«. Mit zunehmend anspruchsvolleren Anwendungen haben sich deshalb Hochleistungsrechner mit Multi-Core-Architekturen durchgesetzt, die Berechnungen hochgradig parallelisiert ausführen können. Tatsächlich aber sind die Möglichkeiten Berechnungen zu parallelisieren durch den Zugriff auf gemeinsame Speicherressourcen immer zu einem gewissen Grad limitiert. Neueste Fortschritte im Bereich Deep Learning fordern diese Einschränkungen besonders heraus, weil die hochgradig parallelisierte Struktur tiefer Neuronaler Netze ganz spezifisch verteilte Speicherzugriffsmuster erfordert. Solche Zugriffsmuster können mit herkömmlicher Computertechnologie kaum effizient abgebildet werden. **Neuromorphe Hardware** geht diese Herausforderung an und hilft dabei, Geräten und Systemen künstlichen Intelligenz (KI) zur verleihen.*

Neuromorphe Hardware basiert auf spezialisierten Rechnerarchitekturen, die die Struktur (Morphologie) Neuronaler Netze (NN) von Grund auf widerspiegeln: Dedizierte Verarbeitungseinheiten bilden direkt in der Hardware die Funktionsweise von Neuronen nach, zwischen denen ein physisches Verbindungsnetz (Bus-System) für den schnellen Austausch von Informationen sorgt. Dieses Konzept ist prinzipiell vom menschlichen Gehirn inspiriert, wo biologische Neuronen und Synapsen in ähnlicher Weise zusammenarbeiten. Spezialisierte neuromorphe

[936] Wörterbuch der philosophischen Begriffe (2013), S. 400

[937] Quelle: Lexikon der Biologie, https://www.spektrum.de/lexikon/biologie/morphologie/44060

Einheiten sind zwar weniger flexibel als klassische Mehrzweckprozessoren (CPUs), dafür aber außerordentlich leistungsfähig und energieeffizient im Einsatz für Training und Inferenz von tiefen Neuronalen Netzen (Deep Neural Networks, DNNs).“[938]

Ontologie

„Der Zweig der Philosophie, der sich mit dem Wesen der Realität beschäftigt. Herausragende Themen sind die Allgemeinbegriffe Ding, Änderung, Raum, Gesetz, Verursachung, Leben, Geist, Gesellschaft.“[939]

Physikalismus

*„**Physikalismus**, eine Variante des logischen Empirismus, ausgehend von der Annahme, dass alle Resultate der Erfahrungswissenschaften in der Sprache einer Einheitswissenschaft formulierbar sind, wie sie beispielhaft für die Sprache der Physik entwickelt worden sei (zuerst in: O. Neurath, Empirische Soziologie, 1931); auch allgemein: eine Wissenschaftsauffassung, nach der alle Aussagen über natürliche Sachverhalte aus den Gesetzen der Physik deduzierbar sind.“*[940]

Man unterscheidet zwischen **reduktiven Physikalismus** (z.B. Funktionalismus) und **nichtreduktiven Physikalismus** (z.B. Monismus, Emergenz- und Supervenienztheorien)

Die grundsätzlich unterschiedlichen Begriffe Materialismus und Physikalismus werden oft austauschbar verwendet.

Qualia

*„Singular **Quale**, ein vermutlich zuerst von C.I. Lewis (Mind and the World Order, 1929) verwendeter Terminus, der bei ihm sinnverwandt den qualitativen Charakter eines Erfahrungsinhaltes meint: etwa den spezifischen Charakter einer bestimmten Rotwahrnehmung. In N. Goodmans Erkenntnistheorie spielen Qualia die Rolle einer letzten Fundierung aller Erkenntnis (The Structure of Appearance, 1951).*

Am wichtigsten ist der Terminus heute in der modernen Philosophie des Geistes. Qualia werden dort als Eigenschaften des persönlichen Erlebens aufgefasst: Sehe ich etwa auf ein rotes Objekt, so erlebe ich die Röte auf eine ganz bestimmte Weise. In Anlehnung an Th. Nagel kann man sagen, dass es „irgendwie ist, etwas Rotes zu sehen“ (What is it like to be a Bat? 1974). Diese besonderen Qualitäten des Erlebens werden häufig eine Reihe von sekundären Eigenschaften zugeschrieben. Sie gelten als „unanalysierbar“ oder „einfach“, „privat“ oder „subjektiv“ (ich kann nicht wissen, wie es für andere ist, Rot zu sehen) und „unaussprechlich“ (ich kann einem Blinden nicht erklären, wie es ist, Rot zu sehen). Qualia werden als eines der großen Probleme der Philosophie des Geistes angesehen, weil sie sich dem Zugriff einer objektiv-materialistischen Wissenschaft prinzipiell zu entziehen scheinen.“[941]

---

[938] Wörtlich zitiert aus: https://www.iis.fraunhofer.de/de/ff/kom/ki/neuromorphic.html

[939] Bunge (1984), S. 280

[940] Wörterbuch der philosophischen Begriffe (2013), S. 501

[941] Wörterbuch der philosophischen Begriffe (2013), S. 537

Repräsentation

*„**Repräsentation**, ‚Vergegenwärtigung', ‚Vertretung', allg. die Stellvertretung durch eine anderen, oder durch etwas anderes; in der Psychologie die Lehre, nach der die Vorstellungen die mit ihnen gemeinten Gegenstände darstellen, das Nichtgegenwärtige vergegenwärtigen oder durch ein Symbol **repräsentieren** (vertreten)"*[942]

Supervenienz

***„Supervenienz** bezeichnet die Abhängigkeitsbeziehung zwischen zwei Gegenstandsbereichen (Eigenschaften) oder Vokabularen (Prädikaten). Der Begriff wird vor allem in der Philosophie des Geistes (mentale und physikalische Eigenschaften), in der Moralphilosophie (moralische und nichtmoralische Eigenschaften) oder in der Philosophie der Kunst (ästhetische und nichtästhetische Eigenschaften) verwendet. Die S. soll eine Abhängigkeitsbeziehung zwischen den »hinzukommenden«, supervenienten Eigenschaften von den unterliegenden Basiseigenschaften ausdrücken. Die Grundidee ist dabei, dass sich eine Entität hinsichtlich einer supervenienten Eigenschaft nur dann ändern kann, wenn sich auch hinsichtlich der Basiseigenschaften eine Änderung vollzieht. Umgekehrt formuliert besagt die Supervenienzthese, dass zwei Entitäten, die identische Basiseigenschaften aufweisen, auch hinsichtlich der supervenienten Eigenschaften identisch sind (z.B. zwei Situationen, die hinsichtlich sämtlicher nichtmoralischer Eigenschaften identisch sind, sind auch hinsichtlich ihrer moralischen Eigenschaften identisch). Diese Abhängigkeit der supervenienten Eigenschaften von den Basiseigenschaften gilt nicht in der umgekehrten Richtung: Es ist möglich, dass zwei hinsichtlich der Basiseigenschaften unterscheidbare Entitäten hinsichtlich der supervenienten Eigenschaften identisch sind (im Kontext der Philosophie des Geistes besagt dies, dass ein und dieselbe mentale Eigenschaft von Ereignissen mit unterschiedlichen physikalischen Eigenschaften aufgewiesen werden kann).*

Besonders für den nicht-reduktiven Physikalismus ist der Begriff der S. von großer Bedeutung, um die angenommene einseitige Abhängigkeit z.B. mentaler Eigenschaften von physikalischen Eigenschaften hinreichend spezifizieren zu können. Dabei sind verschiedene Spielarten der S. zu beachten, die mit ihrer Grundidee jeweils logisch kompatibel sind, aber zu jeweils unterschiedlichen Bestimmungen des postulierten Abhängigkeitsverhältnisses führen. So lautet beispielsweise die Definition der schwachen S., dass es in unserer Welt w für zwei Objekte oder Wesen x und y nicht möglich ist, in Bezug auf ihre Basiseigenschaften ununterscheidbar zu sein und gleichzeitig in irgendeiner Hinsicht in ihren supervenierenden Eigenschaften zu differieren. Tatsächlich erweist sich die schwache S. zur Spezifizierung des Abhängigkeitsverhältnisses als unzureichend, da eine andere Welt w' denkbar ist, deren Basiseigenschaften mit denen aus w identisch sind, aber beispielsweise mentale

[942] Wörterbuch der philosophischen Begriffe (2013), 567

Eigenschaften unabhängig von physikalischen Eigenschaften realisiert werden. Dies verdeutlicht, dass der Proponent des nicht-reduktiven Physikalismus zeigen muss, welche Spielart der S. (weitere Kandidaten wären hier z.B. die globale oder lokale S.) das Abhängigkeitsverhältnis z.B. zwischen mentalen und physischen Eigenschaften tatsächlich zureichend im Sinne eines nicht-reduktiven Physikalismus determiniert."[943]

Support-Vector-Machine

„Die Support Vector Machine (SVM) *ist eine mathematische Methode, die im Umfeld des maschinellen Lernens zum Einsatz kommt. Sie gestattet das Klassifizieren von Objekten und ist vielfältig nutzbar. Unterstützt werden die lineare und die nicht-lineare Objektklassifizierung. Typische Anwendungsbereiche sind die Bild-, Text- oder Handschrifterkennung. Eine der zentralen Aufgaben des maschinellen Lernens ist die Klassifizierung von Daten. Ziel der Klassifizierung ist es, anhand vorhandener Daten und Datenzuordnungen zu entscheiden, welcher Klasse ein neues Datenobjekt zugeordnet werden kann. Zunächst wird eine Datenbasis mit Trainingsobjekten benötigt, deren Klassenzuordnung bekannt ist. Auf Basis dieser Daten versuchen die verschiedenen Algorithmen des maschinellen Lernens Trennlinien oder Trennflächen zu finden, mit denen sich neue Objekt den richtigen Klassen zuordnen lassen. Eine Support Vector Machine kann die Objektklassen mithilfe von Trennungsebenen einteilen. Diese Ebenen werden so gewählt, dass zwischen verschiedenen Klassen ein möglichst großer Bereich frei von Objekten bleibt. Die Trennungsfläche mit dem größten Objekt-freien Bereich gilt als optimale Lösung.*

Eine SVM unterstützt sowohl die lineare als auch die nicht-lineare Trennbarkeit von Objekten. In realen Problemstellungen sind die Datenobjekte in den meisten Fällen nicht durch rein lineare Grenzen zu klassifizieren. Für nicht-lineare Klassengrenzen verwendet die Support Vector Machine den Kernel-Trick. Hierbei werden die Trennungsvektoren in eine zusätzliche Dimension (Hyperebene) überführt. Je höher die Anzahl der Dimensionen ist, desto komplexere Trennungsflächen lassen sich mit den Trennungsvektoren mehrerer Hyperebenen realisieren. Die Umwandlung der linearen Trennungsflächen der verschiedenen Hyperebenen in nicht-lineare Trennungsflächen findet bei der Rücktransformation der Dimensionen statt. Es können sogar nicht zusammenhängende Trennungsflächen entstehen.

Damit die Hoch- und Rücktransformation der verschiedenen Dimensionen rechnerisch nicht zu aufwendig wird, verwendet die SVM sogenannte Kernelfunktionen zur Beschreibung der Trennflächen. Der

[943] Metzler Lexikon der Philosophie, Online Version, https://www.spektrum.de/lexikon/philosophie/supervenienz/1971, 5.7.2019

Begriff Kernel-Trick leitet sich aus der Verwendung dieser Kernelfunktionen ab."[944]

Turingmaschine

„Eine Turingmaschine besteht aus einem Prozessor und einem (potenziell) unbegrenztem Band, das in Felder unterteilt ist. Die Elementaroperationen eines Turing-Programms besagen, dass der Prozessor nacheinander (sequenziell) das Band im Arbeitsfeld mit endlich vielen Symbolen bedrucken, löschen, nach links und rechts um ein Feld verschieben oder stoppen kann. Turingmaschinen sind ideale mathematische Maschinen, da sie unbegrenzt steigerbare Speicherkapazitäten voraussetzen, die man sich als unbegrenzt verlängerbares Rechenband vorstellen kann. So lässt sich z.B. ein Programm für einen Zählprozess angeben, das ausgehend von einem leeren Feld auf den nachfolgenden Feldern zum Inhalt des vorherigen Feldes jeweils die Einheit 1 hinzufügt."[945]

Urteilskraft

„***Urteilskraft**, bezeichnet grundsätzlich das Vermögen, über das abstrakte Wissen von etwas hinaus in der Lage zu sein, Dinge einzuordnen. In der scholastischen Tradition wird sie vis aestimativa genannt, bei G. W. Leibniz ars indicandi, bei A.G. Baumgarten dient sie zur Definition des Geschmacks als „Kraft der Seele, von einer klar empfundenen Vollkommenheit oder Unvollkommenheit zu urtheilen". Nachhaltige Prägung erhält der Begriff durch I. Kant, der U. definiert als das „Vermögen, unter Regeln zu subsumieren; d.h. zu unterscheiden, ob etwas unter einer gegebenen Regel … stehe oder nicht". Da die U. in der Anwendung von Regeln auf Fälle besteht, kann es für ihren Gebrauch nicht selbst wieder Regeln geben, sonst würde ein Regress der Regelbegründung drohen. Sie kann folglich nicht erlernt werden, sondern ist eine Gabe und ein Talent. Kant setzt U. mit „Mutterwitz" gleich, wer ihn nicht hat, ist dumm. Die bestimmende U. findet zu gegebenen Verstandesbegriffen die passenden Anschauungen und Besonderheiten, die reflektierende U. sucht umgekehrt zu vorliegenden Einzelfällen, die exemplarisch erscheinen, den passenden Oberbegriff. In seiner Kritik der Urteilskraft von 1790 expliziert Kant Funktion und Rolle der U. als Voraussetzung für das Erkennen der Natur („teleologische Urteilskraft") und in der Beurteilung ästhethischer Gegenstände („ästethische Urteilskraft")."*[946]

---

944 Zitiert aus BIGDATA INSIDER „*Was ist eine Support Vector Machine?*"; https://www.bigdata-insider.de/was-ist-eine-support-vector-machine-a-880134/

945 Mainzer (2014), S. 73

946 Philosophisches Wörterbuch (2006), S. 733

Stichwort- und Namensverzeichnis