

MASTER THESIS | MASTER'S THESIS

Titel / Title

Individuelle Autonomie und Künstliche Intelligenz

Von der Selbst- zur Fremdbestimmung?

verfasst von / submitted by

Dipl. Kfm. Hans Peter Zander

angestrebter akademischer Grad | in partial fulfilment of the requirements for the degree of

Master of Arts (MA)

Wien | Vienna, 2025

Studienkennzahl lt. Studienblatt |
Postgraduate programme code as it
appears on the student record sheet:

UA 992 499

Universitätslehrgang lt. Studienblatt
Postgraduate programme as it appears on
the student record sheet:

Studium Generale – Das nachberufliche
Studium an der Universität Wien (MA)

Betreut von | Supervisor:

Mag.^a Dr. Katharina Lacina

Abstract (deutsch)

Zunächst widmet sich der Grundlagenteil den zentralen Begriffen „Autonomie“ und „Künstliche Intelligenz“. Die Analyse des Autonomiebegriffs führt zu einem facettenreichen Begriffsspektrum: beginnend beim normativen Konzept (I. Kant) bis hin zu modernen Theorien, die etwa die Willensbildung oder den sozialen Kontext als Autonomiebedingung sehen. Der Begriff der Künstlichen Intelligenz ist ebenso uneinheitlich. Eine Kategorie Künstlicher Intelligenz bilden die Systeme des Maschinellen Lernens, auf die in dieser Arbeit im Wesentlichen Bezug genommen wird.

Im Fokus dieser Arbeit steht die Frage, ob Künstliche Intelligenz die Autonomie des Menschen untergräbt. Werden wir zunehmend fremdbestimmt, wenn Systeme der Künstlichen Intelligenz für uns entscheiden und wir uns auf ihre Informationen verlassen? Ein einfaches Ja oder Nein verbietet sich. Die zentrale Frage wird deshalb aufgefächert. Die Differenzierung des Autonomiebegriffs verdeutlicht Aspekte von Autonomie, die von Künstlicher Intelligenz beeinflusst werden können. Den sich daraus ergebenden Teilfragen wird – methodisch – in Form einer diskursiven Annäherung im Hauptteil der Arbeit nachgegangen. Dort geht es zunächst um Unterschiede zwischen menschlichem und maschinelltem Wissen und Entscheiden, bevor kontingente Eigenschaften Künstlicher Intelligenz wie „künstliche Autonomie“, Opazität und Geschwindigkeit identifiziert werden, die wie Parameter eines imaginären Kippunktes wirken, an dem Selbst- in Fremdbestimmung übergeht. Sie spielen bei zentralen Aspekten der Entscheidung und Wissenspräsentation durch Künstliche Intelligenz eine erhebliche Rolle: (1) *Verantwortung*, die untrennbar mit Entscheidung verbunden ist, wird unmöglich, wenn Entscheidungen nicht nachvollziehbar sind und Kontrolle verloren geht. (2) Technisch weitestgehend autonome Systeme erschweren die Bewertung von Wissen hinsichtlich Wahrheit und Verlässlichkeit – eine *epistemische Herausforderung*. (3) Schlägt Selbstbestimmung in Fremdbestimmung um, stellt sich die Frage, ob der Mensch die Maschine beherrscht oder umgekehrt. Künstliche Intelligenz übt eine „*smarte Macht*“ aus, die einen subtilen, allgegenwärtigen Einfluss auf unser Alltagsleben nimmt, der kaum wahrnehmbar ist.

Das Fazit erlaubt ein bedingtes Ja zur zentralen Forschungsfrage und Antworten auf die differenzierenden Teilfragen: Autonome Willensbildung unterliegt manipulativen Einflüssen. Privatheit wird instrumentalisiert. Zeitliche Autonomie gerät in Widerspruch zur digitalen Welt. Die Schwächung sozialer Bindungskräfte etwa und schleichende

Machtverschiebungen verändern den sozialen und gesellschaftlichen Kontext von Autonomie.

Abstract (englisch)

The foundational section initially addresses the key concepts of “autonomy” and “artificial intelligence.” The analysis of the concept of autonomy leads to a multifaceted spectrum of definitions: ranging from the normative concept (I. Kant) to modern theories that, for example, consider the formation of the will or the social context as conditions for autonomy. The concept of artificial intelligence is similarly heterogeneous. One category of artificial intelligence comprises systems of machine learning, which are the primary focus of this work.

The central question of this thesis is whether artificial intelligence undermines human autonomy. Are we increasingly externally controlled when AI systems make decisions for us and we rely on their information? A simple yes or no is insufficient. Therefore, the main question is broken down into sub-questions. Differentiating the concept of autonomy clarifies aspects that can be influenced by artificial intelligence. These sub-questions are explored methodologically through a discursive approach in the main body of the work. First, differences between human and machine knowledge and decision-making are examined, before contingent properties of AI — such as “artificial autonomy,” opacity, and speed — are identified. These act like parameters of an imaginary tipping point where self-determination transitions into external control. They play a significant role in key aspects of AI-based decision-making and knowledge presentation: (1) Responsibility, inseparably linked to decision-making, becomes impossible when decisions are not comprehensible and control is lost. (2) Technically largely autonomous systems complicate the evaluation of knowledge with regard to truth and reliability — an epistemic challenge. (3) When self-determination shifts to external control, the question arises whether humans control machines or vice versa. AI exerts a “smart power” that subtly and pervasively influences our everyday lives, almost imperceptibly.

The conclusion allows for a qualified “yes” to the central research question and provides answers to the differentiated sub-questions: autonomous will formation is subject to manipulative influences; privacy is instrumentalized; temporal autonomy gets into conflict with the digital world. The weakening of social bonds and gradual shifts of power, for instance, alter the social and societal context of autonomy.

Vorwort

Vor einiger Zeit ging es mir wie vielleicht vielen anderen auch: Ein Gefühl schierer Ungläubigkeit einerseits und Faszination andererseits beschlich mich, als Künstliche Intelligenz wie von Zauberhand Texte verfassen konnte. Aber es kam auch Empörung auf, denn Künstliche Intelligenz hatte nun eine weitere Bastion erobert, die ich bis dahin fest in menschlicher Hand wähnte. Aus dieser diffusen Gefühlslage heraus entstand die Idee einer philosophischen Auseinandersetzung mit dem Thema Künstliche Intelligenz und individuelle Autonomie.

Auf dem Weg zu dem nun vorliegenden Ergebnis haben mich einige Menschen begleitet, bei denen ich mich herzlich bedanken möchte:

An erster Stelle gilt mein Dank Frau Mag. Dr. Katharina Lacina, die mich mit sehr viel Kompetenz geduldig betreut und vor größeren Irrwegen bewahrt hat. Ihre Unterstützung war zugleich Ermutigung.

Ein großes Dankeschön richte ich an Herrn Prof. em. Dr. Michael Reicherts, der mit seiner lebenslangen wissenschaftlichen Erfahrung und Neugier zahlreiche erhellende Diskussionen mit mir geführt hat. Sein Lektorat der Arbeit und seine Feedbacks haben mir wertvolle Sicherheit gegeben.

Ein ebenso großer Dank gilt Frau Mag. Christine Schenk für das äußerst sorgfältige Korrekturlesen. Ihre Fragen zu Themen meiner Arbeit haben mir immer wieder zu mehr Klarheit verholfen.

Schließlich möchte ich aus unserer „Mastermensen-Runde“ besonders Frau Susanne Doppelmayr, MA hervorheben, der ich hilfreiche Gespräche über unsere Themen verdanke.

Der Gebrauch von „ich“ und „wir“ in dieser Arbeit: „Wir“ wird als inklusiver oder kollektiver Plural verwendet, um eine Gruppe – hier: alle Menschen – einzuschließen. Der Bezug kann explizit („Wir Menschen müssen verantwortungsbewusster mit Künstlicher Intelligenz umgehen“) oder implizit sein („Wir sollten auf unseren Planeten achten“).

„Ich“ wird als generisches „Ich“ genutzt, das stellvertretend für jede Person in vergleichbaren Situationen spricht. So drückt eine individuelle Perspektive allgemeine Erfahrungen, Meinungen oder Handlungen aus, ohne andere explizit zu nennen („Wenn ich in schwierigen Zeiten stehe, weiß ich oft nicht weiter“).

Inhaltsverzeichnis

1	Einleitung	1
2	Grundlagen	6
2.1	Zum Begriff der Autonomie	6
2.1.1	Etymologie und Bedeutung	6
2.1.2	Geschichte des Begriffs in der Philosophie	8
2.1.2.1	Autonomiebegriff bei Kant.....	9
2.1.2.2	Exemplarische Debatten in der Folge Kants	14
2.1.3	Neuere philosophische Konzepte individueller Autonomie.....	16
2.1.3.1	Begriffliche Differenzierungen.....	17
2.1.3.2	Internalistische Ansätze	19
2.1.3.3	Inhaltsneutrale externalistische Ansätze.....	22
2.1.3.4	Substanzielle externalistische Ansätze	25
2.1.3.5	Relationale Ansätze	26
2.1.3.6	Privatheit im Kontext relationaler Autonomie	29
2.1.4	Konzeptionalisierung der Autonomieansätze für diese Arbeit.....	31
2.1.4.1	Relevanz einzelner Ansätze	32
2.1.4.2	Konstitutive Bedingungen oder Einflussfaktoren?	33
2.1.4.3	Weiterführende Teilfragen	33
2.2	Künstlich intelligente Systeme	36
2.2.1	Intelligenz und die Frage, ob Maschinen denken können	36
2.2.1.1	Alan Turings „Imitationsspiel“ - Können Maschinen denken?.....	36
2.2.1.2	Das chinesische Zimmer – John Searles Gedankenexperiment	38
2.2.2	Definitionen künstlich-intelligenter Systeme	41
2.2.3	Klassifikationen von KI-Systemen.....	43
2.2.3.1	Symbolische, Subsymbolische und Hybride KI	43

2.2.3.2	Schwache und starke KI, Superintelligenz	47
3	Wird individuelle Autonomie untergraben, wenn KI entscheidet?	49
3.1	Menschliches und maschinelles Wissen	50
3.1.1	Menschliches Wissen und Wahrheit	50
3.1.2	Daten und Information	53
3.1.3	Maschinelle Wissensproduktion: „Probably Approximately Correct “ ..	56
3.2	Menschliches Urteilen und algorithmisches Entscheiden Künstlicher Intelligenz	57
3.2.1	Menschliche Urteilskraft	57
3.2.2	Algorithmische Entscheidungen	62
3.2.2.1	Priorisierung	62
3.2.2.2	Klassifikationen	63
3.2.2.3	Assoziationen	63
3.2.2.4	Filtern	63
3.3	Von der Selbst- in die Fremdbestimmung?	64
3.3.1	Kontingente Eigenschaften von KI-Systemen (als Kippunkte)	65
	„I'm sorry, Dave, I'm afraid I can't do that.“	65
3.3.1.1	Graduelle künstliche Autonomie	67
3.3.1.2	Opazität, Komplexität, Unvorhersehbarkeit und fehlende Erklärung ..	76
3.3.1.3	Geschwindigkeit	78
3.3.2	Verantwortung bei Entscheidungen durch KI	79
3.3.2.1	Was ist Verantwortung?	80
3.3.2.2	Wer trägt die Verantwortung?	81
3.3.2.3	Transparenz und Nachvollziehbarkeit als Voraussetzung	86
3.3.3	Epistemische Herausforderungen	91
3.3.3.1	Von der Faktizität zum Remix	91
3.3.3.2	Bias	95

3.3.3.3	Halluzinationen, Bullshit und Gleichgültigkeit	96
3.3.3.4	Filterblasen und Echokammern	98
3.3.3.5	Epistemischer Anachronismus	102
3.3.3.6	Epistemische Deutungshoheit	103
3.3.4	Die „smarte Macht“ der omnipräsenten Systeme.....	105
3.3.4.1	„Smarte Macht“	105
3.3.4.2	„Ubiquitous Computing“	110
3.3.4.3	Die immanente Tendenz des „ohne uns“	121
4	Fazit	125
5	Literatur	134

1 Einleitung

Epochale technologische Entwicklungen führen zu tiefgreifenden Veränderungen in unserem Leben. Vor der Erfindung der Dampfmaschine arbeiteten Menschen mit Werkzeugen, die sie vollständig kontrollieren konnten. Sie benutzen sie, wann und wie oft sie es für erforderlich hielten. Mit der industriellen Revolution begannen Maschinen Takt und Rhythmus des menschlichen Handelns, der Arbeitszeit und damit auch der Freizeit zu bestimmen: Der Mensch ordnete sich den Erfordernissen der Technik unter. Die Maschinen ersetzten individuelle Handlungsfreiheit durch festgelegte Abläufe und Strukturen.

Seit der Erfindung informationsverarbeitender Maschinen und später künstlich intelligenter Systeme übernehmen diese nicht nur körperliche, sondern auch geistige Tätigkeiten, treffen Entscheidungen, liefern Informationen und beeinflussen Denken, Einstellungen und Handeln. Künstliche Intelligenz ist heute omnipräsent, greift in die intimsten Bereiche individueller Privatheit ein und ist zu einem Medium sozialer Interaktion geworden. Sie gestaltet, wie wir Zeit erleben, trägt zur Beschleunigung des Alltags bei und verändert die Bedingungen unserer Selbstbestimmung – oft unbemerkt.

Zentrale Fragestellung

Die vorliegende Arbeit rückt die Autonomie des Menschen in den Fokus. Die *zentrale Frage* lautet: Inwieweit wird individuelle Autonomie untergraben, wenn Künstliche Intelligenz an Stelle des Menschen entscheidet. Geben wir zusammen mit unseren Entscheidungen auch unsere Selbstbestimmung ab? Wird unser Entscheiden und Handeln zunehmend fremdbestimmt durch Künstliche Intelligenz? Oder ist es nicht in vielen Fällen mehr als sinnvoll, den Systemen Entscheidungen zu übertragen? Denn: Algorithmen, die Krebs frühzeitig erkennen, autonome Fahrzeuge, die Unfälle verhindern, oder Lernsysteme, die Kindern in bildungsfernen Regionen Wissen vermitteln, eröffnen neue Handlungsmöglichkeiten und scheinen moralisch geboten. Doch der Blick auf positive Effekte darf nicht darüber hinwegtäuschen, dass Systeme der künstlichen Intelligenz individuelle Autonomie subtil beeinflussen können. Schon bei alltäglichen Interaktionen mit Künstlicher Intelligenz, etwa über Suchmaschinen oder Textgeneratoren, entscheidet Künstliche Intelligenz über Quellen, Gewichtungen und Präsentation, während der Mensch primär Empfänger oft unreflektierter Ergebnisse bleibt.

Dass die zentrale Fragestellung der Arbeit von Relevanz ist, lässt sich auch aus den „Ethik-Leitlinien für eine vertrauenswürdige KI“ der Europäischen Union schließen. Dort heißt es:

„Die Grundrechte, auf welche sich die EU gründet, sollen die Achtung der Freiheit und Autonomie des Menschen gewährleisten. Wenn Menschen mit KI-Systemen interagieren, müssen sie in der Lage sein, die Selbstbestimmung über die eigene Person in vollem Umfang und wirksam auszuüben und es muss ihnen möglich sein, am demokratischen Prozess teilzuhaben.“¹

Und auch im „Monitoring von Zukunftsthemen für das Österreichische Parlament“ wird festgestellt:

„Entscheidungen werden zunehmend durch Algorithmen vorbereitet oder automatisch getroffen. [...] In vielen Fällen liefern Algorithmen die Grundlage für Entscheidungen, die von existenzieller Bedeutung sind.“²

Die Frage nach der Autonomie des Menschen mag eine individualistische Perspektive implizieren. Wenn von Teilhabe am demokratischen Prozess die Rede ist und von der existenziellen Bedeutung, die Entscheidungen haben können, deutet dies jedoch an, dass das Thema individuelle Autonomie nichts mit egozentrischer Nabelschau zu tun hat oder haben sollte. Vielmehr ist individuelle Autonomie, wie sie hier verstanden wird, immer in einen sozialen, gesellschaftlichen und politischen Kontext eingebettet. Demokratie braucht mündige Bürger, die ihre Entscheidungen frei und nach eigenem Ermessen auf der Basis eines authentischen Willens treffen. Nur so sind wir in der Lage, an der gemeinschaftlichen Gestaltung der Gesellschaft teilzuhaben. Umgekehrt ist Autonomie nur in liberal demokratischen Systemen möglich und kann nur in solchen Systemen gelebt werden; und zwar nur jenseits subtiler Manipulation – folgt man einer wohl nie zu erreichenden Idealvorstellung.

¹ Hochrangige Expertengruppe für künstliche Intelligenz., „Ethik-leitlinien für eine vertrauenswürdige KI.“, 43.

² Institut für Technikfolgen-Abschätzung (ITA), *Transparente Algorithmen – Wie lässt sich algorithmische Diskriminierung verhindern?*

Methodische Herangehensweise

Das Thema der Arbeit wird philosophisch theoretisch erschlossen, nicht empirisch. Grundlage dafür ist eine ergiebige Recherche und Auswertung in erster Linie philosophischer und ethischer Literatur. Teils wird auf entsprechende Interviews und Vorträge Bezug genommen. Zur Veranschaulichung werden vereinzelt Filmsequenzen aus Spiel- oder Dokumentarfilmen herangezogen. Einen großen Raum nehmen vor allem im Grundlagenteil begriffliche Klärungen ein. Im Hauptteil dient der Diskurs einer Annäherung an die Beantwortung der Forschungsfrage.

Aufbau der Arbeit

Viele der möglichen und zweifellos wichtigen Themen, die mit der zentralen Fragestellung zusammenhängen, wie etwa „Künstliche Intelligenz, Autonomie und Demokratie“, können nicht oder nur am Rande behandelt werden. Es ist eine Konzentration erforderlich, die im Aufbau der Arbeit zum Ausdruck kommt.

Im ersten Teil des Grundlagenkapitels geht es um den Begriff der Autonomie. Beginnend bei Immanuel Kant, der mit der Beantwortung Frage „*Was ist Aufklärung?*“ und vor allem mit der Formulierung des *Kategorischen Imperativs* bis heute von zentraler Bedeutung ist, wenn es um den Autonomiebegriff geht. Die Debatte um den Autonomiebegriff hält bis in die heutige Zeit an. Theorien des 20. Jahrhunderts gehen vor allem auf US-amerikanischen Philosoph*innen zurück. Einige wichtige Ideen von Autonomie werden im Kapitel „2.1.3 Neuere philosophische Konzepte individueller Autonomie“ vorgestellt. Diesen Autor*innen geht es nicht mehr um einen normativen Begriff wie Kant, sondern beispielsweise um den Prozess der Willensbildung, um die in der Person oder Umgebung liegenden Bedingungen der Willensbildung, um ihre Einbettung in zwischenmenschliche und kulturelle Kontexte oder um soziale Beziehungen, die Personen untereinander eingehen. Schließlich geht es bei Vertreter*innen sogenannter „*relationaler*“ Ansätze um einen sehr weit gefassten Autonomiebegriff, der über den begründeten Willen von Menschen hinausgeht und Autonomie u.a. auf soziale und politische Kontexte bezieht.

Das zweite Grundlagenkapitel ist dem Begriff der Künstlichen Intelligenz gewidmet. So wenig wie es einen einheitlichen Autonomiebegriff gibt, existiert ein einheitlicher Begriff der Künstlichen Intelligenz. Auch bei der grundlegenden Frage, ob Maschinen denken können, gehen die Meinungen auseinander. *Alan Turing*, der diese Frage aufwarf, versuchte sie mit seinem *Imitationsspiel* zu beantworten. *John Searle* kam in seinem

Gedankenexperiment („*Das chinesische Zimmer*“) zu der Auffassung, dass Maschinen nicht denken können, weil ihnen Intentionalität fehle. Gleichwohl gibt es zahlreiche Definitionen und Klassifikationen von „Künstlicher Intelligenz“. Diese Arbeit bezieht sich im Wesentlichen auf sogenannte selbstlernende Systeme, die mit teils gigantischen Datenmengen „gefüttert“ werden. Diese Daten werden mithilfe statistischer Korrelationen und Wahrscheinlichkeiten ausgewertet. Das, was die KI-Nutzer als Ergebnis bekommen, ist also wahrscheinlich, annähernd richtig.

Der Hauptteil der Arbeit soll die Frage beantworten, ob individuelle Autonomie untergraben wird, wenn Künstliche Intelligenz anstelle vom Menschen entscheidet. Man kann kein vereinfachendes Ja oder Nein erwarten. Die Antwort muss irgendwo zwischen diesen beiden Polen liegen. Wo also liegt der „imaginäre Kippunkt“, ab dem Menschen Autonomie einbüßen und Selbstbestimmung in Fremdbestimmung umschlägt? Dieser Frage soll in Form einer diskursiven Annäherung nachgegangen werden. Ganz im Sinne von G. Anders:

„Es wird eine Hauptaufgabe der Philosophie der Technik sein, den dialektischen Punkt ausfindig zu machen und zu bestimmen, wo sich unser Ja der Technik gegenüber in Skepsis oder in ein unverblümtes Nein zu verwandeln hat.“³

In einem ersten Schritt werden die Unterschiede zwischen menschlichem und maschinellem Wissen und Entscheiden herausgearbeitet. Sodann folgt eine Beschreibung kontingenter Eigenschaften künstlich intelligenter Systeme wie „*künstliche Autonomie*“⁴, „*Opazität*“⁵ und *Geschwindigkeit*. Sie sind wie Parameter des imaginären Kippunktes. In ihrer jeweils konkreten Ausprägung tragen sie dazu bei, ob die Grenze überschritten wird, hinter der es Richtung Fremdbestimmung geht. Sie werden diskutiert in Zusammenhang mit drei ausgewählten zentralen Themen, die in Bezug auf Selbst- und Fremdbestimmung eine wichtige Rolle spielen:

- Entscheidung und *Verantwortung* sind untrennbar miteinander verbunden. Ist Verantwortung noch möglich, wenn Entscheidungen nicht mehr nachvollziehbar sind und wenn wir die Kontrolle darüber verlieren?

³ Anders, *Die Antiquiertheit des Menschen*. 2, 141.

⁴ Übersetzung des im englischen Sprachraum gebräuchlichen Begriff der „artificial Autonomy“, um Autonome Systeme zu charakterisieren.

⁵ *Wikipedia*, „Opazität“, bezeichnet allgemein das Gegenteil von Transparenz, also mangelnde Durchsichtigkeit bzw. mangelnde Durchlässigkeit.“

- Wenn Künstliche Intelligenz weitestgehend autonom entscheidet, wird es schwierig, das algorithmisch erzeugte Wissen hinsichtlich Wahrheit und Verlässlichkeit zu bewerten und zu erkennen. Wir stochern im Ungewissen und stehen vor *epistemischen Herausforderungen*.
- Wenn Selbstbestimmung in Fremdbestimmung umschlägt, stellt sich die Frage, ob der Mensch noch die Maschine beherrscht, oder die Maschine den Menschen? Von künstlich intelligenten Systemen geht eine „*smarte Macht*“⁶ aus, die mit einem zwanglosen Zwang der Überwachung und Steuerung dient. Sie ist das Instrument omnipräsenter, stiller Technologien, die sich ins Gewebe unseres Alltags einflechten, bis sie verschwunden scheinen.

Künstliche Intelligenz vermag die individuelle Autonomie mehrschichtig, schleichend und subtil zu untergraben. Der Übergang von der Selbstbestimmung zur Fremdbestimmung wird nicht erzwungen, sondern erfolgt meist freiwillig und auf angenehme Weise.

⁶ Der Begriff geht zurück auf Han, *Psychopolitik Neoliberalismus und die neuen Machttechniken*, 25.

2 Grundlagen

2.1 Zum Begriff der Autonomie

Das einleitende Kapitel zur Etymologie und Bedeutung soll den Facettenreichtum des Begriffs aufzeigen und klären. Im Mittelpunkt des historischen Rückblicks steht Immanuel Kant, dessen Idee von Autonomie eine philosophische Debatte ausgelöst hat, die bis heute andauert. Der gegenwärtige Stand wird im Kapitel über aktuelle philosophische Konzepte in groben Zügen dargestellt. Darauf aufbauend ist eine Konzeptualisierung der unterschiedlichen Ansätze und Modelle von Autonomie möglich. Sie soll verdeutlichen, wie der Begriff in dieser Arbeit verwendet werden soll und wie Zusammenhänge zwischen individueller Autonomie und KI-Systemen hergestellt werden können.

2.1.1 Etymologie und Bedeutung

Der Begriff der Autonomie ist von seinem Ursprung her ein politischer. Um die Mitte des 5. Jahrhunderts forderten griechische Stadtstaaten innere und äußere politische Freiheit. Sie wollten ihre Angelegenheiten eigenständig und selbstbestimmt regeln. Es ging ihnen um *αὐτονομία* – so der griechische Begriff, der Selbstgesetzgebung, Eigengesetzlichkeit und Selbstbestimmung bedeutete. Gemeint war damit die Unabhängigkeit von äußeren Mächten und von Fremdherrschaft, was eine politische und territoriale Abgrenzung beinhaltete. Gemeint war aber auch eine Freiheit und Selbstbestimmung im Inneren, d.h. eine Freiheit von Tyrannis, von Gewaltherrschaft und Despotie. Das Streben nach Autonomie hatte dabei keinen Absolutheitsanspruch. Es ging nicht unbedingt um eigene Gerichtsbarkeit oder eigene Steuerhoheit, sondern um Autonomie als bedingte politische Selbstbestimmung. Autonomie ist etymologisch *„das Recht, die Möglichkeit, (innerhalb eines Staatsverbandes) nach eigenen Gesetzen zu leben, (politische) Selbstverwaltung, Selbstständigkeit“*⁷ So verstanden findet sich politische Autonomie auch gegenwärtig als Idee in föderalen Staaten oder in Autonomiebewegungen.⁸ Es handelt sich also nach wie vor um einen primär politischen Begriff, der als Synonym für territoriale Selbstverwaltung benutzt wird.⁹

⁷ DWDS – Digitales Wörterbuch der deutschen Sprache, *Autonomie*, hg. von Berlin-Brandenburgischen Akademie der Wissenschaften Dort auch Erläuterung der Wortzusammensetzung: „griech. *autós* (αὐτός) ‘selbst, eigen’“ und „*nómos* (νόμος) ‘Brauch, Sitte, Ordnung, Gesetz’, also wörtlich ‘eigenes Gesetz habend’“.

⁸ Pohlmann, „Autonomie“.

⁹ Zander, „Der Autonomiebegriff bei Immanuel Kant - Sichtweisen aus Theologie und Kirche“, 2.

Aber das ist er nicht nur. Heute geht der Begriff darüber hinaus und kann sich auf Institutionen, politische Gemeinschaften, auf Geltungssphären des Rechts, der Kunst, der Wissenschaften beziehen.¹⁰ Autonomie bedeutet dabei stets, dass die jeweiligen Einheiten bzw. Personen die Freiheit besitzen, über Angelegenheiten, die sie selbst betreffen, eigenständig und nach eigenem Willen zu entscheiden und ihr Handeln entsprechend auszurichten.¹¹ In diesem Sinne wird etwa von institutioneller Autonomie gesprochen (z.B. Hochschulautonomie) oder von persönlicher Autonomie, also der Fähigkeit zur individuellen Selbstbestimmung. Andere Facetten des Begriffs kommen zum Ausdruck, wenn mit Autonomie wirtschaftliche Freiheit gemeint ist. Das weitestgehend selbstverantwortliche Handeln der Akteure in freien Marktwirtschaften ist beispielsweise eine Form wirtschaftlicher Autonomie, die ein Basiselement der Wirtschaftsordnung liberal-demokratischer Staaten ist. Auch in Bezug auf technische Systeme (z.B. Maschinen, Roboter, Autos) ist der Begriff der Autonomie gebräuchlich, auch wenn menschlicher Wille in algorithmisch determinierten Prozessen nur indirekt eine Rolle spielt.

Der Begriff unterscheidet sich von einigen *bedeutungsverwandten Begriffen*. Das prominenteste Beispiel ist der der *Freiheit*. Freiheit wird gelegentlich als Synonym für Autonomie verstanden. Der Begriff findet allerdings eine breitere Anwendung als der Autonomiebegriff. So wird Freiheit z.B. auch im Sinne von Vorrecht verwendet, wenn etwa jemand mehr Freiheiten als Andere genießt.¹² Auch ideengeschichtlich ist der Freiheitsbegriff weitreichender.¹³ Autonomie kann begrifflich deutlich enger gefasst werden, was sich schon im Wortbestandteil *nómos* (=Gesetz) andeutet. So wird beispielsweise von Autonomie als relativer oder „abhängiger Freiheit“ gesprochen und darauf hingewiesen, dass ein autonomes Leben immer auch von persönlichen Fähigkeiten und gesellschaftlichen Bedingungen bestimmt ist.¹⁴

Der Begriff der *Autarkie* hat ebenfalls eine lange Tradition. Aristoteles bezeichnet in seiner Tugendethik *Αὐτάρκεια* als Selbstgenügsamkeit und Unabhängigkeit von anderen Dingen und Zielen. Wenn der Mensch sich selbst genügt und bei sich selbst ist, braucht er nichts, was darüber hinaus noch erstrebenswert wäre. Insofern ist Glückseligkeit

¹⁰ Goertz und Witting, „Autonomie“.

¹¹ DWDS – Digitales Wörterbuch der deutschen Sprache, *Autonomie*, hg. von Berlin-Brandenburgischen Akademie der Wissenschaften, 15. November 2023.

¹² DWDS – Digitales Wörterbuch der deutschen Sprache, *Freiheit*.

¹³ Höffe, *Lexikon der Ethik*, 86.

¹⁴ Rössler, „Philosophin Beate Rössler: ‚Autonomie ist eine abhängige Freiheit‘“.

(Eudaimonia, Ευδαιμονία) als höchstes Ziel auch Selbstgenügsamkeit, da es nichts Höheres gibt, das der Mensch anstreben könnte. In der hedonistischen Ethik (Epikur) ist Ataraxia (Αταραξία), die Seelenruhe das höchste Ziel. Das Ideal der Seelenruhe umfasst die Abwesenheit von körperlichen und seelischen Schmerzen und verspricht höchste Lust. Um Seelenruhe zu erreichen ist u.a. der richtige Umgang mit Bedürfnissen erforderlich. Unnatürliche Bedürfnisse wie Luxus oder das Streben nach Macht sind ebenso überflüssig, wie nicht notwendige natürliche Bedürfnisse (etwa Sex). Einzig die Befriedigung natürlicher notwendiger Bedürfnisse (wie Essen, Wärme, Trinken) sind für die Seelenruhe erforderlich. Wer sich also in Selbstgenügsamkeit mit der Erfüllung der Grundbedürfnisse begnügt, erreicht das Ideal der Seelenruhe. Soweit zwei philosophische Beispiele. Heute hat der Begriff oft eine deutlich ökonomische Dimension. Im politischen Zusammenhang bedeutet Autarkie die weitestgehende wirtschaftliche Unabhängigkeit eines Landes vom Handel mit anderen Ländern. Alle Waren und Dienstleistungen werden möglichst in der eigenen Wirtschaft erzeugt. Autonomie hingegen hebt nicht auf Unabhängigkeit und Selbstversorgung ab als vielmehr auf generelle Selbstbestimmung und Selbstgesetzgebung.

Auch der Begriff der *Souveränität* lässt sich abgrenzen. Er bedeutet „*Macht und Recht, frei nach eigenem Ermessen zu entscheiden, ohne dabei das Recht anderer zu verletzen.*“¹⁵

Es geht um staatliche Hoheitsgewalt im Inneren und nach außen, um das Recht eines Staates auf nationale Selbstbestimmung, auf staatliche Souveränität. Sie ist im Völkerrecht des 20. Jh. Voraussetzung für die Anerkennung als Staat.¹⁶ Im übertragenen Sinne wird beispielsweise im Sport von der souveränen Spielweise einer überlegen spielenden Mannschaft gesprochen, die unangefochten einen Sieg erringt.

2.1.2 Geschichte des Begriffs in der Philosophie

Die Idee der Autonomie hat sich seit der Aufklärung als fundamentaler, konstitutiver Wert für liberal-demokratische Staaten durchgesetzt und ist heute selbstverständlich. Es war vor allem Immanuel Kant, der den Begriff Ende des 18. Jahrhunderts entwickelte und zu einem Kernstück seiner Moraltheorie machte. Vor Kant hat individuelle, menschliche Autonomie in der Geschichte der Philosophie kaum eine bedeutende Rolle gespielt.

„Es wird in der einschlägigen Literatur aber darauf hingewiesen, dass zentrale Elemente des kantischen Autonomiebegriffs, wie etwa die

¹⁵ DWDS – Digitales Wörterbuch der deutschen Sprache, *Souveränität*.

¹⁶ Quaritsch, „Souveränität“.

Vernunft, in der Moralphilosophie durchaus angedacht waren. Im 15. Jh. war es Giovanni Pico della Mirandola (1463-1494), der in seiner „Rede über die Würde des Menschen“ davon spricht, dass Gott den Menschen erschaffen hat, der frei von allen Einschränkungen seinen Platz in der Welt nach eigenem «freien Willen und Entschluss» einnehmen kann und seine Natur selbst, «nach eigenem Belieben und aus eigener Macht» ausbilden kann.¹⁷ Darin steckt bereits eine Idee von Willensfreiheit und Selbstbestimmung, für die Gott nach Mirandola dem Menschen Entfaltungsmöglichkeiten gegeben hat und ihn gewähren lässt.“¹⁸

Die Geschichte des Autonomiebegriffs wird jedoch entscheidend durch Immanuel Kant geprägt und beeinflusst.

2.1.2.1 Autonomiebegriff bei Kant

Im Jahre 1784, einige Jahre bevor Kant sich mit der Autonomie des Willens beschäftigt und den Kategorischen Imperativ formuliert, erscheint seine berühmte Schrift „Beantwortung der Frage: Was ist Aufklärung?“. Sie beginnt mit den Sätzen:

„Aufklärung ist der Ausgang des Menschen aus seiner selbst verschuldeten Unmündigkeit. Unmündigkeit ist das Unvermögen, sich seines Verstandes ohne Leitung eines anderen zu bedienen. [...] sapere aude! habe Muth dich deines eigenen Verstandes zu bedienen!“¹⁹

Wie ist das zu verstehen?

„Von seinem Verstand Gebrauch zu machen, erfordert Freiheit. Kant sieht die Freiheit jedoch eingeschränkt. Von allen Seiten kommt die Aufforderung «räsonnirt nicht». Und Kant benennt einige, die für die Einschränkung von Freiheit sorgen: Der Offizier, der Finanzrat und der Geistliche, der verlangt zu glauben und nicht zu räsonieren.²⁰ Kant fordert also, auch beim Glauben den Verstand einzuschalten und - so darf wohl interpretiert werden – sich aus religiöser Unmündigkeit mithilfe des Verstandes zu befreien. Zur Aufklärung gehört die Freiheit

¹⁷ Klein, „Umbrüche in Neuzeit, Aufklärung, Moderne - Handout zur LV Theologie im Studium Generale der Universität Wien, WS 2023/2024“, 2.

¹⁸ Zander, „Der Autonomiebegriff bei Immanuel Kant - Sichtweisen aus Theologie und Kirche“, 3.

¹⁹ Kant, *Kants gesammelte Schriften.*, AA VIII, Was ist Aufklärung, 35.

²⁰ Kant, *Kants gesammelte Schriften.*, AA VIII, Was ist Aufklärung, 37.

*des Menschen, von seiner Vernunft öffentlichen Gebrauch zu machen.²¹
Es geht keineswegs um grenzenlose Freiheit in allen Lebenslagen.
Denn dann würde das Gemeinwesen, würden die sozialen,
ökonomischen oder politischen Institutionen nicht funktionieren. Es
geht darum, dass die Menschen als Angehörige solcher Institutionen
sich an Spielregeln halten.²²“²³*

Bei Kant ist die Vernunft ein essenzieller Bestandteil des menschlichen Erkenntnisvermögens.

„In seinen erkenntnistheoretischen Überlegungen unterscheidet er verschiedene intellektuelle Erkenntnisvermögen. Der Verstand, die Urteilskraft und die Vernunft machen das menschliche Erkenntnisvermögen aus.²⁴ Der Verstand ist das Vermögen zu denken. Mit unserem Verstand sind wir in der Lage, Begriffe zu bilden. Die Begriffe helfen uns dabei, die Welt zu begreifen. Sie leisten einen Beitrag zum Verstehen der Welt.²⁵ Die Urteilskraft ist die zweite Fähigkeit, die bei Kant die Unterscheidung ermöglicht, «ob etwas der Fall der Regel sei oder nicht»²⁶. Was heißt das? Beispiel: «<Dies ist ein runder Tisch». Die Urteilskraft besteht darin, dass ich zum ersten die Regel «befolge», dass das ein Tisch ist und nicht etwa ein Stuhl und zum zweiten die Regel, dass dieser Tisch rund ist und nicht etwa quadratisch. Eine dritte Fähigkeit des Menschen ist die Vernunft. Sie ist das intellektuelle Vermögen, vom Allgemeinen auf das Besondere zu schließen. Sie ist die Fähigkeit, nach Grundsätzen zu urteilen und danach praktisch zu handeln.²⁷ Wenn ein allgemeiner Grundsatz beispielsweise lautet, niemals zu lügen, dann sagt die Vernunft, dass eine Lüge auch in einer konkreten Situation (im Besonderen) nicht

²¹ Kant, *Kants gesammelte Schriften.*, AA VIII, Was ist Aufklärung, 36.

Kant unterscheidet den öffentlichen Gebrauch vom Privatgebrauch der Vernunft. Ersterer bezieht sich auf die Publikationen von „Gelehrten“

²² Liessmann, *Die großen Philosophen und ihre Probleme*, 76.

²³ Zander, „Der Autonomiebegriff bei Immanuel Kant - Sichtweisen aus Theologie und Kirche“, 3.

²⁴ Kant, *Kants gesammelte Schriften.*, AA VII, Anthropologie in pragmatischer Hinsicht, 198.

²⁵ Liessmann, *Die großen Philosophen und ihre Probleme*, 80.

²⁶ Kant, *Kants gesammelte Schriften.*, AA VII, Anthropologie in pragmatischer Hinsicht, 199.

²⁷ Kant, *Kants gesammelte Schriften.*, AA VII, Anthropologie in pragmatischer Hinsicht, 199.

Dort heißt es: „von dem Allgemeinen das Besondere abzuleiten und dieses letztere also nach Prinzipien und als notwendig vorzustellen. - Man kann sie also auch durch das Vermögen nach Grundsätzen zu urtheilen und (in praktischer Rücksicht) zu handeln erklären. Zu jedem moralischen Urtheile (mithin auch der Religion) bedarf der Mensch Vernunft und kann sich nicht auf Satzungen und eingeführte Gebräuche fußen.“

erlaubt ist und also die Wahrheit zu sagen ist. Wie dieses Beispiel nahelegt, braucht der Mensch auch für moralische Entscheidungen Vernunft.“²⁸

In Kants Moraltheorie spielt die Vernunft also eine zentrale Rolle. Ein längeres Selbstzitat soll das verdeutlichen:

„Nur vernünftige Wesen sind laut Kant in der Lage, nach (moralischen) Prinzipien zu handeln. Der Wille zum moralischen Handeln wird «unausbleiblich» durch Vernunft bestimmt, er ist «nichts anderes als praktische Vernunft»²⁹. Der gute Wille wählt nur dasjenige, was die Vernunft als gut erkennt.

«Es ist überall nichts in der Welt, ja überhaupt auch außer derselben zu denken möglich, was ohne Einschränkungen für gut könnte gehalten werden, als allein ein guter Wille.»³⁰

Der gute Wille ist an sich gut. Der Mensch ist von Natur aus ein Lebewesen, das von Neigungen geprägt ist, von Emotionen, Zielen, Affekten. Auch die mögen gut und wünschenswert sein. Aber «sie können auch äußerst böse und schädlich werden»³¹. Der «gute Wille» ist frei von Neigungen und äußeren Einflüssen. Die Autonomie des Willens, der frei von Neigungen ist, ist unerlässlich für das sittliche Handeln des Menschen.

«Autonomie des Willens ist die Beschaffenheit des Willens, dadurch derselbe ihm selbst (unabhängig von aller Beschaffenheit der Gegenstände des Willens) ein Gesetz ist. Das Princip der Autonomie ist also: nicht anders zu wählen als so, daß die Maximen seiner Wahl in demselben Willen zugleich als allgemeines Gesetz mit Begriffen seien.»³²

Der «gute Wille» folgt einzig und allein der Vernunft. Er ist frei von Neigungen, er ist «unabhängig von aller Beschaffenheit der Gegenstände des Willens». Macht, Reichtum, Ehre, Glücksstreben, Mut, Liebe oder Mitleid beispielsweise sind nicht maßgebend für eine

²⁸ Zander, „Der Autonomiebegriff bei Immanuel Kant - Sichtweisen aus Theologie und Kirche“, 3.

²⁹ Kant, *Kants gesammelte Schriften.*, AA IV, GMS, 412.

³⁰ Kant, *Kants gesammelte Schriften.*, AA IV, GMS, 193.

³¹ Kant, *Kants gesammelte Schriften.*, AA IV, GMS, 193.

³² Kant, *Kants gesammelte Schriften.*, AA IV, GMS, 440.

moralische Entscheidung. Der gute Wille liegt ausschließlich im intelligiblen Bereich und ist nicht empirisch begründet, so dass Neigungen, Gefühle, Talente und so weiter keine Rolle spielen.

Der Wille unterwirft sich Maximen, d.h. moralischen Leitsätzen, die die Handlungswahl bestimmen. Lautet meine Maxime beispielsweise «Immer die Wahrheit sagen!», so ist diese Maxime eine Handlungsregel, die ich in ganz konkreten Situationen befolge. Mein autonomer Wille zeichnet sich dadurch aus, dass ich im konkreten Fall nicht lügen will und die Wahrheit sage – ganz gleich, in welcher Situation ich mich befinde.³³ Die Autonomie des Menschen gründet auf Selbstgesetzgebung. Der Wille ist also so beschaffen, dass er sich selbst ein Gesetz gibt, und zwar unabhängig von äußeren Einflüssen und Neigungen.

Maximen sind nun keine beliebigen Handlungsregeln, die jedes Vernunftwesen nach Lust und Laune frei bestimmt. Sie müssen vielmehr universalisierbar sein. Die Idee ist, wenn ich einer Maxime entsprechend agiere, dann ist diese Handlung nur dann moralisch erlaubt (oder nicht erlaubt), wenn es akzeptabel (bzw. inakzeptabel) ist, dass alle anderen Menschen auch so handeln (oder nicht so handeln). Immer die Wahrheit zu sagen ist beispielsweise eine universalisierbare Maxime. Denn es ist wünschenswert, dass alle Menschen die Wahrheit sagen. Und Lügen sind nicht akzeptabel, denn wenn alle Menschen lügen würden, ist z.B. ein Gemeinwesen, ist ein soziales Miteinander nicht vorstellbar.³⁴

Es gibt einen weiteren fundamentalen Aspekt. Wenn ich gemäß einer Maxime handle, dann tue ich das nicht, weil ich mir z.B. einen Vorteil davon verspreche. Ich handle einzig und allein aus Pflicht und nicht aus Neigung. Wenn ich eine Maxime als allgemeingültige moralische Handlungsregel aus meiner Vernunft heraus erkannt habe, dann ist sie mir Gesetz und ich habe die Pflicht, aus Achtung für dieses Gesetz

³³ Hübner, *Einführung in die philosophische Ethik*, 169 ff.

Es sind natürlich Situationen denkbar, in denen Menschen nachvollziehbarer Weise nicht die Wahrheit sagen, bewusst nicht ihrer Maxime folgen. Man stelle sich vor, ein (geliebter) Mensch kann nur durch eine Lüge vor der Ermordung gerettet werden. Kant selbst lehnt auch in solchen Fällen die Lüge als unmoralisch ab.

³⁴ Hübner, *Einführung in die philosophische Ethik*, 152.

entsprechend aus Pflicht zu handeln. «Pflicht ist die Nothwendigkeit einer Handlung aus Achtung fürs Gesetz»³⁵ „³⁶

Kant formuliert den kategorischen Imperativ als das moralische Gesetz, welches zugleich eine praktische Aufforderung ist.

„Ihr kann der Mensch nur folgen, wenn er autonom ist und über einen freien Willen verfügt. Die Gesetzesformel lautet:

«handle nur nach derjenigen Maxime, durch die du zugleich wollen kannst, daß sie ein allgemeines Gesetz werde.»³⁷

Kant nennt seinen Imperativ kategorisch, weil er nur durch Autonomie, also frei von Neigungen möglich ist. Vom kategorischen ist der hypothetische Imperativ zu unterscheiden. Für letzteren gilt folgendes Muster: Ich soll etwas tun, weil ich etwas anderes will. Ich soll z.B. nicht lügen, weil ich sonst dafür verachtet werde. Im Gegensatz dazu liegt dem kategorischen Imperativ das Muster zugrunde: «ich soll so oder so handeln, ob ich gleich nichts anderes wollte.»³⁸ Ich soll also etwas tun, weil ich prinzipiell so und nicht anders handeln will. Ich soll nicht lügen, völlig unabhängig davon, ob ich oder jemand anderes dadurch einen Nachteil erfahre oder nicht. Ich soll aus Prinzip nicht lügen! Dem kategorischen Imperativ liegt die Autonomie, dem hypothetischen die Heteronomie des Willens zugrunde. “³⁹

Neben der Gesetzesformel des kategorischen Imperativs formuliert Kant weitere Formeln, darunter die sogenannte Zweckformel.

„Kant geht davon aus, dass der Mensch einen absoluten Wert an sich hat, und er sagt: «der Mensch und überhaupt jedes vernünftige Wesen existiert als Zweck an sich selbst, nicht bloß als Mittel zum beliebigen Gebrauche für diesen oder jenen Willen»⁴⁰. Hierin unterscheidet sich der Mensch von vernunftlosen Wesen, die Kant im Gegensatz zu Personen als Sachen bezeichnet. Sachen haben keinen absoluten, sondern einen relativen Wert. Sie können als Mittel zum Zweck dienen,

³⁵ Kant, *Kants gesammelte Schriften.*, AA IV, GMS, 400.

³⁶ Zander, „Der Autonomiebegriff bei Immanuel Kant - Sichtweisen aus Theologie und Kirche“, 3.

³⁷ Kant, *Kants gesammelte Schriften.*, AA IV, GMS, 421.

³⁸ Kant, *Kants gesammelte Schriften.*, AA IV, GMS, 441.

³⁹ Zander, „Der Autonomiebegriff bei Immanuel Kant - Sichtweisen aus Theologie und Kirche“, 3.

⁴⁰ Kant, *Kants gesammelte Schriften.*, Kant, AA IV, GMS, 428.

Personen sollen das aber nicht. So kommt Kant zu der Zweckformel des kategorischen Imperativs.

«Handle so, dass du die Menschheit sowohl in deiner Person, als in der Person eines jeden andern, jederzeit zugleich als Zweck, niemals bloß als Mittel brauchst»⁴¹

Den Wert von Sachen, die als Mittel zum Zweck gebraucht werden, nennt Kant Preis. Der absolute Wert des Menschen ist ein innerer Wert, den Kant Würde nennt.

«[...] 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.»⁴² „⁴³

Mit der Zweckformel des kategorischen Imperativs begründet Kant die *Menschenwürde*. Sie verbietet die Instrumentalisierung des Menschen.

2.1.2.2 Exemplarische Debatten in der Folge Kants

Karl Jaspers zufolge nimmt Kant eine einzigartige Stellung in der Philosophiegeschichte ein. „Seit Plato ist im Abendland kein Schritt getan, der in der herben Luft des Denkens so umwendende Folgen hat.“⁴⁴ Kants Werke haben in der abendländischen Philosophie einen epochalen Umbruch eingeleitet. Er hat die philosophischen Debatten und das Denken bis in die Gegenwart hinein nachhaltig verändert.⁴⁵ Für Karl Jaspers ist Kant mehr als der Gipfel und Vollender der Aufklärung. Er ist „ein Überwinder der Aufklärung zugunsten eines unabschließbaren Prozesses des in der Umwendung eigentlich zu sich selbst kommenden Menschen.“⁴⁶

In der Folge Kants haben sich maßgebende philosophische Diskurse immer wieder auf ihn bezogen. Dies gilt auch für den Autonomiebegriff, der von einigen Philosophen aufgegriffen, übernommen oder weiterentwickelt wurde.⁴⁷ Im 19. Jahrhundert beeinflussten die transzendentalphilosophischen Ideen Kants den deutschen Idealismus, aus dem heraus kritische Positionen entwickelt wurden, die zu einer Modifizierung der Autonomiekonzeption führten. So attestiert Pohlmann J. G. Fichte und F. W. J. Schelling

⁴¹ Kant, *Kants gesammelte Schriften.*, Kant, AA IV, GMS, 429.

⁴² Kant, *Kants gesammelte Schriften.*, AA IV, GMS, 435.

⁴³ Zander, „Der Autonomiebegriff bei Immanuel Kant - Sichtweisen aus Theologie und Kirche“, 3.

⁴⁴ Jaspers, *Die maßgebenden Menschen*, 615.

⁴⁵ Höffe, *Immanuel Kant*, 288.

⁴⁶ Jaspers, *Die maßgebenden Menschen*, 611.

⁴⁷ Pohlmann, „Autonomie“.

eine Radikalisierung von Kants Theorie, da diese unter dem Begriff eine absolute bzw. ursprüngliche Gesetzgebung der Vernunft verstünden.⁴⁸ Auch bei G.F.W. Hegel spielte der Autonomie- bzw. der Freiheitsbegriff eine Rolle. Mit Hegel sei Kants Autonomiebegriff erst verwirklicht⁴⁹ und von seiner Paradoxie befreit worden – die Paradoxie, die darin bestehe, dass, „wenn wir uns Gesetze geben durch die wir autonom sind, müssten wir das durch einen gesetzlosen Akt, entweder willkürlich oder durch vorweg gegebene Gründe, tun“.⁵⁰ Die Ideen der Denker des Deutschen Idealismus führten zu der Auffassung einer „Überwindung Kants“⁵¹. Eine Position, die etwa von Karl Jaspers nicht geteilt wird.⁵²

Natürlich hat es nach dem deutschen Idealismus und im 20. Jahrhundert weitere Philosophen gegeben, die sich mit Autonomie beschäftigt haben, oft jedoch, indem sie von dem etwas weiter gefassten Begriff der Freiheit ausgingen. An dieser Stelle beschränke ich mich auf eine kurze Skizze zur Bedeutung Kants für das christlich-theologische Verständnis von Autonomie, das die Geistesgeschichte jahrhundertlang prägte und weiterhin beeinflusst. Auch in der theologischen Philosophie markiert Kant einen Wendepunkt. Mit seiner Aufklärungsschrift hat Kant gefordert, der Mensch möge den Mut haben, sich seines Verstandes zu bedienen. Mutig war Kant selbst, als er bei der Formulierung der Kernideen seiner Pflichtenethik Gott unerwähnt ließ. Nicht die Bindung des Menschen an Gott als oberste sittliche Instanz stand im Mittelpunkt. Die Autonomie des Willens war für Kant das oberste Prinzip der Sittlichkeit. Sie löst Gott als oberste moralische Instanz ab. Mit dieser Auffassung setzte sich Kant heftiger Kritik aus, die vor allem aus der (katholischen) Theologie und von religiös motivierten Philosophen kam. So prägte Jacobi den Begriff der „Selbstgötterey“ und lehnte Kants Autonomiekonzept von Grund auf ab. Andere sprachen von einem „Grundirrtum“ und davon, dass Autonomie ohne Bindung an Gott eine Irrlehre sei und das Streben danach Selbstsucht und Egoismus. Zu Beginn des 19. Jahrhunderts gewann der Begriff der Theonomie als eine Art Reaktion auf Kants Autonomiebegriff an Bedeutung. Kern der Theonomie ist die Bestimmung der menschlichen Sittlichkeit durch Gott und nicht etwa durch die Selbstbestimmung des Menschen. Gottesgesetzlichkeit einerseits und

⁴⁸ Pohlmann, „Autonomie“.

⁴⁹ Lau, „Hegels Verwirklichung des Kantischen Autonomiebegriffs“, 192 ff.

⁵⁰ Waldegge, „Was ist Autonomie?“, 197 ff.

⁵¹ Höffe, *Immanuel Kant*, 291.

⁵² Jaspers, *Die maßgebenden Menschen*, 614.

Selbstgesetzgebung des Menschen andererseits hielten die Gegner von Kants Moraltheorie für unvereinbar. Erst sehr viel später, im 20. Jahrhundert, sah Hans Blumenberg die menschliche Autonomie als eine notwendige Bedingung für die Möglichkeit der Theonomie.⁵³ Ihm gelingt ein Spagat zwischen kantischer Autonomie und kirchlicher Heteronomie, indem er von „*autonomem Gehorsam*“ sprach.

„Die theonome Vorstellung eines autonomen Gehorsams ist auch heute noch aktuell, ebenso wie die Auseinandersetzung mit Kant. Unter dem Titel ‚Stößt Kants ‚Autonomie‘ die klassische kirchliche Sexualmoral um?‘“ bezog sich etwa die Religionsphilosophin Gerl-Falkovitz im Dezember 2022 in der katholischen „Tagespost“ auf Blumenberg:

«Autonomie kann nicht sein, wie Jacobi das nannte, eine ‚Selbstgötterey‘. In ihr steckt nach Blumenberg vielmehr verblüffend: ein ‚autonomer Gehorsam‘»⁵⁴ „⁵⁵

Soweit einige Beispiele aus der historischen Debatte um den Autonomiebegriff in der Folge Kants.

2.1.3 Neuere philosophische Konzepte individueller Autonomie

Etwa seit Mitte des 20. Jahrhunderts wurden und werden philosophische Konzepte personaler oder individueller Autonomie entwickelt. Aus der Fülle der Ansätze können an dieser Stelle nur einige wenige exemplarisch hervorgehoben werden. Dabei soll deutlich werden, dass es keinen einheitlichen Autonomiebegriff gibt, dass es fundamental unterschiedliche Auffassungen von Autonomie gibt, und dass es sich um einen äußerst facettenreichen Begriff handelt. Die Auffächerung des Begriffs soll sich in weiteren Kapiteln dieser Arbeit als ergiebig für die Beantwortung der zentralen Frage dieser Arbeit erweisen. Trotz der Unterschiede in den Ansätzen haben sie Gemeinsamkeiten: Sie formulieren konstitutive Merkmale individueller Autonomie und beschreiben, welche kausalen und konstitutiven Bedingungen Autonomie ausmachen. Unter welchen Bedingungen können Menschen autonom werden? Und was macht ihre Autonomie aus? Sie versuchen allesamt die Frage zu beantworten, wann und unter welchen Bedingungen ein Mensch autonom genannt werden kann. Je nach Schwerpunkt wird in der

⁵³ Blumenberg „Autonomie und Heteronomie“, in: RGG 1 (31957) 791 zitiert nach: Pohlmann, „Autonomie“. Dort heißt es: „Jede theologische Bestimmung des göttlichen Anspruchs auf Unterwerfung findet dort ihre Grenze, wo der freie Akt in den kausalen Vorgang umschlägt.... Wahrhafte Erfüllung der Theonomie kann nur ein Gehorsam prompter Mandatum sein, das heißt ein autonomer Gehorsam sein.“

⁵⁴ Gerl-Falkovitz, „Stößt Kants ‚Autonomie‘ die klassische kirchliche Sexualmoral um?“

⁵⁵ Zander, „Der Autonomiebegriff bei Immanuel Kant - Sichtweisen aus Theologie und Kirche“, 7.

Autonomieliteratur häufig zwischen internalistischen, externalistischen, substanziellen und relationalen Theorien unterschieden, eine Gliederung, der hier gefolgt werden soll.⁵⁶ Es sei darauf hingewiesen, dass die Übergänge zwischen diesen Kategorien fließend sind und nicht trennscharf vorgenommen werden können. So überschneiden sich vor allem sogenannte substanzielle Theorien, die sich gegen inhaltsneutrale Konzeptionen wenden, mit relationalen ebenso wie mit externalistischen.

Bevor die Autonomiekonzepte vorgestellt werden, scheinen einige begriffliche Differenzierungen sinnvoll.

2.1.3.1 Begriffliche Differenzierungen

Autonomie kann – wie bereits skizziert – z.B. Gruppen, Organisationen, Kollektiven, politischen Einheiten (Nationen, Staaten) oder auch Tieren zugeschrieben werden. Personale oder individuelle Autonomie schließt solche Entitäten begrifflich aus und bezieht sich ausschließlich auf menschliche Individuen.⁵⁷ Personale und individuelle Autonomie werden hier synonym verwendet.

In ihrer Mehrzahl unterscheiden sich neuere Konzepte *individueller Autonomie* deutlich von der kantischen Idee der *moralischen Autonomie*. Obwohl in zeitgenössischen Debatten um Autonomie immer wieder auf Kant verwiesen wird, ist das moralische Konzept des kategorischen Imperativs bei der Begriffsdefinition in den Hintergrund getreten. Der Begriff der Autonomie habe sich seit Kant *individualisiert* und *entmoralisiert* – so die verbreitete Auffassung.⁵⁸ Auch wenn Konzepte personaler Autonomie auf dieselben Eigenschaften und Fähigkeiten Bezug nehmen, die den kantischen Begriff prägen.⁵⁹ Die hier vorgestellten Konzepte gehen davon aus, dass Menschen nicht deshalb autonom sind, weil sie universalisierbaren Vernunftprinzipien folgen. Vielmehr wird ihnen Autonomie zugeschrieben, weil sie aus sich selbst heraus ihren eigenen Willen haben, der ihr Handeln bestimmt. Was nicht bedeutet, dass Vernunft außen vorbleibt. Und natürlich auch nicht die Moral, denn wie anders können moralische Fragestellungen bei der Willensbildung (bewusst oder unbewusst) entschieden werden.⁶⁰

Teils in Anlehnung an Kant vertreten zeitgenössische Autor*innen die Idee einer *kategorialen Autonomie*, die allen Menschen die Fähigkeit unterstellt, für sich selbst zu

⁵⁶ Rössler, „Autonomie“; Betzler, *Autonomie der Person*; Seidel, *Selbst bestimmen*.

⁵⁷ Frankfurt, „Freedom of the Will and the Concept of a Person“, 5.

⁵⁸ Betzler, *Autonomie der Person*, 13.

⁵⁹ Rössler, *Autonomie: ein Versuch über das gelungene Leben*, 13 f.

⁶⁰ Betzler, *Autonomie der Person*, 13.

sprechen, Verantwortung zu tragen und - soweit es geht - unabhängig zu sein. Jeder Person wird kategorisch Autonomie zugeschrieben, eben weil sie Person ist. Aber es gibt Ausnahmen. Grundlegende Autonomie wird denjenigen mehr oder weniger zu- oder abgesprochen, die bestimmte Kompetenzbedingungen wie die Fähigkeit, rational zu denken oder sich selbst in ausreichendem Maße zu kontrollieren, nicht erfüllen. Kinder, Komapatient*innen oder Menschen, die unter fortgeschrittener Demenz leiden, erfüllen solche und ähnliche Kompetenzbedingungen nicht. Die uneingeschränkte Fähigkeit zur Autonomie ist Voraussetzung für ihre Zuschreibung.⁶¹ Ist die gegeben, dann begründet Autonomie etwa Abwehrrechte gegen paternalistische Eingriffe.⁶² Umgekehrt werden aus eingeschränkten Fähigkeiten gesellschaftliche, soziale und staatliche Schutz- und Interventionspflichten abgeleitet.

Nun ist es so, dass kategoriale Autonomie nicht plötzlich zu- oder abgesprochen, quasi per Kippschalter ein- oder ausgeschaltet werden kann. Denn die Grenzen sind fließend. Jüngeren Kindern gestehen Eltern einen geringeren Grad an Autonomie zu als älteren. Erwachsenen wird mehr Autonomie und Verantwortung abverlangt als Minderjährigen. Mit beginnender Demenz verfügen Menschen noch über ausgeprägtere Fähigkeiten zum autonomen Entscheiden und Handeln als mit fortgeschrittener. Menschen sind also *autonom im graduellen Sinn*. Ihnen kann Autonomie in mehr oder weniger hohem Maße zugeschrieben oder abgesprochen werden. Und das gilt nicht nur für Grenzfälle wie Demenz. Mit dem Begriff der *graduellen Autonomie* wird der Tatsache Rechnung getragen, dass Menschen nicht tagein tagaus und rund um die Uhr mit Eintritt ins Erwachsenenleben uneingeschränkt autonom sein können und auch nicht jede ihrer Entscheidungen kann dies sein. Man stelle sich eine Person vor, die einer Idealvorstellung von Autonomie folgt, niemals von Ambivalenzen und Zweifeln geplagt wird, immer genau weiß, was sie will und sich womöglich im kantischen Sinne stets auf der moralisch richtigen Seite wähnt, die emotionsfrei und aus Pflicht handelt. Einmal davon abgesehen, dass die Vorstellung einer solchermaßen autonomen Person an die algorithmische Autonomie von Robotern erinnert (siehe weiter unter), dürfte es nur sehr wenige Exemplare unserer Spezies geben, die das fertigbringen. Für die meisten Menschen dürfte die *ideale Autonomie* nicht mehr als ein (vielleicht) erstrebenswertes Ziel bleiben.

⁶¹ Rössler, „Autonomie“; Christman, „Autonomy in Moral and Political Philosophy“.

⁶² Rössler, „Autonomie“.

Eine weitere begriffliche Differenzierung hat sich bereits angedeutet, nämlich die zwischen *globaler und lokaler Autonomie*. *Globale Autonomie* bezieht sich auf die gesamte Person. Eine autonome Person ist in all ihren Entscheidungen und Handlungen und in ihrer gesamten Lebensführung autonom. Globale Autonomie ist in diesem Sinne auch ideale Autonomie, die sich meist nicht verwirklichen lässt. Anders ist es mit der *lokalen Autonomie*. Der Begriff bezieht sich auf einzelne Handlungen. Menschen können autonom genannt werden, wenn viele ihrer Handlungen selbstbestimmt entschieden und umgesetzt werden. Dabei kann es Verhaltensweisen in anderen Lebensbereichen geben, die nicht autonom sind. Ein Beispiel dafür ist die unfreiwillige Raucherin, die schon mehrfach ihrer Sucht widerstehen wollte, es aber einfach nicht schafft. Sie raucht gegen ihren Willen. Obwohl sie in diesem Punkt nicht autonom handelt, kann man ihr allein deshalb Autonomie vollends absprechen? Nein, sicher nicht!⁶³ Denn dann würde der Autonomiebegriff nahezu jegliche praktische Relevanz verlieren, denn Rauchen ist nicht das alleinige Laster, dem Menschen sich oft gegen ihren Willen hingeben.

2.1.3.2 Internalistische Ansätze

Internalistische Ansätze beschreiben und definieren Autonomie als ausschließlich im Individuum selbst begründet - losgelöst von äußeren Einflüssen und sozialen und gesellschaftlichen Kontexten. Autonomie hat im Kern damit zu tun, wie sich eine Person zu ihren Einstellungen und Motiven verhält, wie sie mit ihren Wünschen und ihrem Willen umgeht. Im Mittelpunkt steht die Willensformung des Individuums und die Frage, welche Bedingungen für die Bildung des Willens erfüllt sein müssen, um eine Person autonom nennen zu können. Dabei bleiben die Inhalte des Willens, also die Frage danach, was ich will, welche Werte und Überzeugungen mir wichtig sind, weitgehend unbeachtet.⁶⁴ Es geht im Wesentlichen um die Struktur des Willens und die reflektierende Bewertung von Wünschen. Deshalb werden diese Positionen auch als strukturelle oder inhaltsneutrale Theorien bezeichnet.

Einer der einflussreichsten Autoren ist der US-amerikanische Philosoph *Harry G. Frankfurt*, auf den sich nahezu alle Ansätze beziehen, die nach ihm entwickelt wurden. Er hat 1971 die *Struktur des Willens* einer Person, die Menschen von anderen Lebewesen

⁶³ Rössler, *Autonomie: ein Versuch über das gelungene Leben*, 45.

⁶⁴ Rössler, „Autonomie“.

unterscheidet, zum Ausgangspunkt seiner Überlegungen gemacht.⁶⁵ Menschen haben – wie andere Lebewesen auch – Wünsche und Motive und sind bewegt, dieses oder jenes zu tun. Frankfurt bezeichnet diese als *Wünsche erster Stufe*. Menschen können zudem über ihre Wünsche befinden, können bestimmte Wünsche bevorzugen und anderen geringere Bedeutung zumessen. Sie sind zur Bewertung ihrer Wünsche erster Stufe in der Lage. Auch das sind wiederum Wünsche, die Frankfurt zunächst *Wünsche zweiter Stufe* nennt.

Damit führt Frankfurt ein *hierarchisches Modell* des Wollens ein. Er argumentiert, dass ein Satz wie „A möchte etwas tun“ ohne Bezug auf den Zusammenhang noch wenig aussagt. Ein Beispiel: „Ich möchte heute Abend ins Kino gehen“ ist ein scheinbar klarer Wunsch. Es könnte aber sein, dass mir der Gedanke zwar gefällt, der gerade einsetzende Regen mir die Aussicht darauf etwas vermiest und ich vielleicht besser ein Buch lese. Es könnte auch sein, dass der Kinowunsch mit der Vorstellung ins Theater zu gehen, kollidiert und ich mich mit der Entscheidung etwas schwertue und unschlüssig bin. Möglich wäre auch, dass mich ein schlechtes Gewissen plagt, weil ich andere Vorhaben oder Pflichten auf die lange Bank schiebe. Einen Wunsch zu haben, heißt also noch lange nicht, dem Wunsche entsprechend zu handeln. Es sagt noch nichts über die Ernsthaftigkeit des Wunsches und über seine Dringlichkeit aus. Für Frankfurt reicht ein Willensbegriff nicht, der lediglich beschreibt, was eine Person beabsichtigt zu tun, was jemanden lediglich geneigt macht, in einer bestimmten Weise zu handeln. Vielmehr sei es entscheidend, dass es einen *effektiven Wunsch* gebe, der die Person zum Handeln bewege.

„It is not the notion of something that merely inclines an agent in some degree to act in a certain way. Rather, it is the notion of an effective desire-one that moves (or will or would move) a person all the way to action“⁶⁶.

Zur Definition des Willens gehört also seine *Handlungswirksamkeit*. Nur der Wille, der mich zum Handeln veranlasst, entspricht meinem ernsthaften, wirklichen, meinem effektiven Wunsch. Wünsche erster Stufe sind nicht per se handlungswirksam. Sie können verworfen werden, weil sie z.B. weniger wichtig sind als konkurrierende

⁶⁵ Frankfurt, „Freedom of the Will and the Concept of a Person“, 6; Neben Frankfurt hat beispielsweise auch Gerald Dworkin etwa zur gleichen Zeit sehr einflussreich über Autonomietheorie gearbeitet. Vgl. hierzu: Dworkin, *The Theory and Practice of Autonomy*.

⁶⁶ Frankfurt, „Freedom of the Will and the Concept of a Person“, 8.

Wünsche, die mich am Ende zum Handeln bewegen. Mein Wille ist die Antwort auf die Frage, was will ich wirklich tun und nicht, was kann ich mir vorstellen zu tun. Erst wenn ein Wunsch handlungswirksam wird, er also in die Tat umgesetzt wird oder werden soll, entspricht der Wunsch dem tatsächlichen Willen, also dem, was die Person wirklich will und tut oder zu tun beabsichtigt.

Es ist eine Eigentümlichkeit des Menschen, dass er Wünsche erster Stufe und Wünsche zweiter Stufe bilden kann und die Fähigkeit zur reflektierenden Selbstbewertung besitzt. Das unterscheidet Menschen von anderen Lebewesen.

„No animal other than man, however, appears to have the capacity for reflective self-evaluation that is manifested in the formation of second-order desires.“⁶⁷

Die Fähigkeit zur reflektierenden Selbstbewertung ist die Voraussetzung dafür, dass ein Wunsch zweiter Stufe entsteht. Den Wunsch zweiter Stufe bezeichnet Frankfurt als *Volition*.

„Someone has a desire of the second order either when he wants simply to have a certain desire or when he wants a certain desire to be his will. In situations of the latter kind, I shall call his second-order desires "second-order volitions" or "volitions of the second order." Now it is having second-order volitions, and not having second-order desires generally, that I regard as essential to being a person.“⁶⁸

Der Begriff *Volition* umfasst die menschliche Fähigkeit, Motive, Ziele und Wünsche in Resultate, in Ergebnisse, in Handlungen umzusetzen. Anstelle von Wünschen zweiter Stufe spricht Frankfurt entsprechend von *Volitionen zweiter Stufe*. Dieser Prozess der Willensbildung beinhaltet eine (*Selbst-*)*Identifikation* der Person mit ebendiesen Wünschen und Werten. Er muss durchlaufen werden und zu effektiven, handlungswirksamen Entscheidungen führen. Die Identifikation muss stark genug und vollkommen sein. Es darf keinen Zweifel geben, dass der zum Willen bestimmte Wunsch unumstößlich ist. Hinzu kommt, dass es sich tatsächlich um meinen Wunsch handeln muss, dass ich also *authentisch* bin. Erst dann habe ich einen freien Willen und handle autonom.

⁶⁷ Frankfurt, „Freedom of the Will and the Concept of a Person“, 7.

⁶⁸ Frankfurt, „Freedom of the Will and the Concept of a Person“, 10.

Frankfurt selbst hat nicht von Autonomie, sondern nur von Willens- und Handlungsfreiheit gesprochen. Dennoch wurde seine Theorie in der einschlägigen Literatur als Modell für den Autonomiebegriff aufgegriffen.⁶⁹ Frankfurts allgemeiner Ansatz von Autonomie war sehr einflussreich, hat aber auch Kritiker auf den Plan gerufen. Die wichtigsten Einwände: Es besteht die Gefahr, dass es zu einem nicht-endenwollenden Regress von Wünschen erster zu Wünschen zweiter und womöglich n-ter Ordnung kommt.⁷⁰ Denn die Wünsche zweiter oder n-ter Ordnung müssten ihrerseits durch Volition bestätigt werden. Zwar setzt Frankfurt in späteren Arbeiten diesem Einwand die Idee der „wholeheartedness“ entgegen, nach der die Person aus ganzem Herzen und mit ungeteiltem Willen handelt. Rössler weist allerdings auf einen wichtigen Punkt hin. Frankfurt übersehe, dass zum Abwägen von konfligierenden Wünschen *gute Gründe* gehören. Handeln geschehe immer auf der Grundlage von Gründen.⁷¹ Der wohl schwerwiegendste Einwand gegen Frankfurts Theorie ist, dass die Willensbildung durch Manipulation beeinflusst werden kann, was Frankfurt in Kauf nehme. Eine Person könnte auch dann den Autonomietest bestehen, könnte ihre Wünsche erster Stufe mit Volitionen zweiter Stufe als vereinbar sehen, wenn sie ein vollständig unterwürfiges Leben führte, in dessen Verlauf eine strikte Konditionierung und Sozialisation dazu geführt habe, dass sie sich nicht nur vollkommen unterordne, sondern ein solches Leben auch als wahre Berufung ansehe. Sie sei ein Lebewesen ohne Werte, die wirklich ihre eigenen seien. Es handele sich um ein manipuliertes Individuum.⁷²

2.1.3.3 Inhaltsneutrale externalistische Ansätze

Kritiker der internalistischen Theorie von Autonomie halten es für unzureichend, dass Autonomie ausschließlich von der inneren Willensstruktur und Willensbildung abhängig sein soll. In der Folge Frankfurts wurden Konzepte entwickelt, die üblicherweise als externalistisch bezeichnet werden. Der Begriff mag auf den ersten Blick irreführend sein, denn er suggeriert, dass es um (konstitutive) Merkmale und Bedingungen von Autonomie geht, die im Umfeld der Person zu finden sind. Diese Vorstellung würde allerdings zu kurz greifen, denn vor allem Konzepte, die als unmittelbare Kritik und Erweiterung auf Frankfurts Theorie aufbauen, werden zu den externalistischen Modellen gezählt, auch

⁶⁹ Christman, „Autonomy in Moral and Political Philosophy“.

⁷⁰ Rössler, *Autonomie: ein Versuch über das gelungene Leben*, 48; Christman, „Autonomy in Moral and Political Philosophy“.

⁷¹ Rössler, *Autonomie: ein Versuch über das gelungene Leben*, 48.

⁷² Christman, „Autonomy and Personal History“, 6 f.

wenn sie sich mit Bedingungen wie etwa Selbstvertrauen oder Abwesenheit von Selbsttäuschung beschäftigen, die durchaus im „Inneren“ einer Person liegen. Aber diese Bedingungen sind unabhängig davon gültig, wie eine Person sich zu ihnen verhält. Um externalistische Modelle einerseits von rein internalistischen und andererseits von solchen, die das Umfeld einer Person in den Fokus rücken, abzugrenzen, werden sie auch als *schwach externalistisch* bezeichnet.⁷³ *Stark externalistische* Modelle hingegen fokussieren auf objektive Gegebenheiten wie vor allem die Möglichkeit zur freien Selbstentfaltung und zur freien Wahl, eine realistische Sicht und ein soziales Umfeld, das respektvolle, anerkennende und nicht missachtende Beziehungen ermöglicht.⁷⁴

Zu den wichtigsten Vertretern schwach externalistischer Ansätze zählt John Christman. Er stellt 1991 ein erweitertes Konzept der Theorie Frankfurts vor, mit dem er versucht, die erkannten Schwächen der internalistischen Konzeption zu umgehen. Er konzentriert sich auf den *Prozess* und auf die Art und Weise, wie Individuen dazu kommen, Wünsche zu entwickeln und nicht auf ihre *Einstellung* zu diesen Wünschen zu einem bestimmten Zeitpunkt. Das Schlüsselement der Autonomie ist seiner Ansicht nach die Akzeptanz oder die Ablehnung des Prozesses der Wunschbildung durch den Handelnden oder der Faktoren, die zu dieser Entscheidung führten und nicht die Identifikation des Handelnden mit dem Wunsch selbst.⁷⁵

„The new theory is unique in that I focus on the manner in which the agent came to have a set of desires rather than her attitude towards the desires at any one time. The key element of autonomy is, in my view, the agent's acceptance or rejection of the process of desire formation or the factors that give rise to that formation, rather than the agent's identification with the desire itself.“⁷⁶

Vor allem gegen den Manipulationsvorwurf formuliert er eine – auch historisch genannte - Konzeption, die die Entstehungsgeschichte und den Entstehungsprozess des Willens in den Fokus rückt. Man stelle sich vor - so Christman - ein Freund oder ein Familienmitglied verbringe zwei intensive Wochen bei einer religiösen Sekte. Bei seiner Rückkehr plappere er das Credo dieser Sekte ohne jegliche Überlegung nach. Sein bisheriges Selbst sei nicht wieder zu erkennen. Es sei offensichtlich, dass ihm ein

⁷³ Betzler, *Autonomie der Person*, 28.

⁷⁴ Betzler, *Autonomie der Person*, 28.

⁷⁵ Christman, „Autonomy and Personal History“, 2.

⁷⁶ Christman, „Autonomy and Personal History“, 2.

erhebliches Stück Autonomie abhandengekommen sei. Es handle sich also offensichtlich um einen Fall von Manipulation dieser Person. Es komme nun darauf an, wie der Freund damit umgehe. Wenn er sich der Beeinflussung bewusst ist oder wird und weiß, wie er zu der neuen Überzeugung gekommen ist, kann er sich dazu verhalten und das Zustandekommen billigen und akzeptieren, oder er kann es ablehnen und damit auch seine neue Überzeugung aufgeben. Christmans Theorie: Ein Individuum ist auch dann autonom, wenn die Einflüsse und Bedingungen, die einen bestimmten Wunsch hervorgerufen haben, von ihm gebilligt werden, wenn er denen nicht widerstanden hat oder widerstanden hätte, wenn er sie beachtet hätte, und wenn dieses Urteil in einer minimal rationalen, nicht von sich selbst getäuschten Weise getroffen wurde oder worden wäre. Christman sagt weiter, eine Implikation dieser Theorie sei, dass Menschen sich als autonom erweisen können, obwohl sie Wünsche nach unterwürfigen, erniedrigenden oder sogar bösen Dingen und Lebensstilen haben. Dies zeige lediglich, dass das Konzept von Autonomie „inhaltsneutral“ sei. Das bedeutet, es geht nicht darum, ob etwas gut oder schlecht, gut oder böse, richtig oder falsch ist, sondern es geht lediglich darum, dass sich das Individuum der Gründe und Faktoren bewusst ist, die zu seinem Wunsch und letztendlich zu seiner Entscheidung geführt haben. Wenn sich das Individuum einer Manipulation bewusst ist und weiß, wie sie zustande gekommen ist, dann kann es sie akzeptieren oder ablehnen. Es bleibt eine autonome Entscheidung. Auch wenn sie von anderen Personen als nichtautonom wahrgenommen wird. Christman begründet die inhaltliche Neutralität seiner Theorie damit, dass Fälle vorstellbar seien, in denen das Individuum für seinen Wunsch „(no matter how evil, self sacrificing, or slavish it might be)“⁷⁷ gute Gründe haben könne. Es sei immer die Herkunft des Wunsches, die bei der Beurteilung der individuellen Autonomie ausschlaggebend sei.

Es gibt weitere Autonomiemodelle, die die internalistische Theorie von Frankfurt kritisieren und dabei auch über den historischen Ansatz von Christman hinausgehen. Fischer und Ravizza konstatieren, dass zur individuellen Autonomie die Fähigkeit und die Möglichkeit von Personen gehört, auf gute Gründe, die etwa in der Deliberation mit anderen offenkundig oder relevant werden, reagieren zu können. Es handelt sich um gute Gründe, die von außen, von der äußeren Welt kommen und nicht etwa aus der Bewertung von Manipulationen. Erst wenn eine Person die Möglichkeit hat, verschiedene Gründe für bestimmte Handlungsmöglichkeiten heranzuziehen, kann ihr Autonomie

⁷⁷ Christman, „Autonomy and Personal History“, 23.

zugesprochen werden. Autonomie und moralische Verantwortlichkeit setzen voraus, dass Akteure in einem moderaten Sinn offen für Gründe sind.⁷⁸ Anders als bei Christman spielt also in diesem Modell eine ethische Dimension eine Rolle.

2.1.3.4 Substanzielle externalistische Ansätze

Der ethischen Bewertung kommt auch in sogenannten substanziellen Ansätzen individueller Autonomie Bedeutung zu. Sie gehen davon aus, dass der Begriff der individuellen Autonomie nicht inhaltsneutral gedacht werden kann und dass der Inhalt der Wünsche, Ziele und Motive von Personen konstitutiv sei. Den Gründen und Werten, die für Personen wichtig seien, müsse Rechnung getragen werden. Welche Wünsche ich hege, welche Ziele ich mir setze und aus welchen Motiven ich handele, ist entscheidend für meine Autonomie. Genau hierauf zielen die substanziellen Theorien von Autonomie ab. Wenn Wünsche, Motive, Ziele immer nur mit Rückgriff auf das eigene Wünschen begründet werde, dann sei Autonomie nicht plausibel und konsistent konzipiert.⁷⁹ Vielmehr muss sich Autonomie auch auf das beziehen, „*was tatsächlich und begründet wünschenswert ist, was also tatsächlich gut oder schlecht, richtig oder falsch ist*“.⁸⁰

Zu den substanziellen Ansätzen wird auch der von Diana Tietjens Meyers gezählt, dem eine prozedurale Sichtweise zugrunde liegt. Sie geht zunächst davon aus, dass Menschen nur das tun, was sie wollen und zwar „*wirklich wollen*“. Aber das allein reicht ihrer Meinung nach nicht, ein solcher ontologischer Ansatz sei nicht haltbar. Ihre Ansicht von autonomen Lebensplänen und Handlungen skizziert sie so:

*„Meiner Ansicht nach entwerfen Personen ihren Lebensplan dann autonom und führen einzelne Handlungen dann autonom aus, wenn sie ein Repertoire von Fähigkeiten zur Selbstfindung, Selbstdefinition und Selbstführung benutzen, um ihre Vorhaben zu entwickeln und sich für ihre Handlungen zu entscheiden“.*⁸¹

Diana Tietjens Meyers geht der Frage nach „*was es genau heißt, dass Individuen die Kontrolle über ihr eigenes Leben haben und daraus Befriedigung ziehen*“;⁸² Sie bezieht zwischenmenschliche und kulturelle Kontexte mit in ihr Verständnis von Autonomie ein

⁷⁸ Rössler, *Autonomie: ein Versuch über das gelungene Leben*, 158.

⁷⁹ Wolf, „The true, the good, and the lovable“.

⁸⁰ Rössler, „Autonomie“.

⁸¹ Tietjens Meyers, „Personale Autonomie ohne Transzendenz“, 150.

⁸² Tietjens Meyers, „Personale Autonomie ohne Transzendenz“, 150.

und geht davon aus, dass die Einbindung in diese Kontexte einen tiefgreifenden Einfluss darauf ausüben, was sie tun möchten und wer sie sind.

2.1.3.5 Relationale Ansätze

Marina Oshana sieht externe, soziale Beziehungen, die Personen untereinander eingehen, als konstitutiv für Autonomie an.⁸³ Sie bezieht das soziale Umfeld einer Person mit in die Frage nach den Bedingungen von Autonomie ein, ohne dabei internalistische Ideen außer Acht zu lassen. So zählt sie ebenso wie Theoretiker*innen internalistischer Modelle die Fähigkeit zur kritischen Reflexion und die prozedurale Unabhängigkeit, die eine Beschränkung und Einengung des Prozesses der Willensbildung ausschließt, zu den notwendigen Bedingungen von Autonomie. Darüber hinaus sieht sie als weitere Bedingung einen „*verfügbarer Bereich relevanter Optionen*“. Zudem benennt Oshana „*Sozial-relationale Eigenschaften*“, denn eine Person muss als Mitglied einer Gesellschaft Beziehungen zu anderen Menschen eingehen, um ihre Ziele in einem sicheren psychischen und sozialen Umfeld verfolgen zu können. Ihre soziale Umgebung muss so beschaffen sein, dass die Person sich gegen psychische und physische Angriffe wehren kann, dass sie ihre Rechte (Grundrechte, wirtschaftlichen Rechte) verteidigen kann, dass sie z.B. ihre eigenen Bedürfnisse verwirklichen oder ihre eigenen Ziele verfolgen kann, ohne Behinderung durch mächtigere und einflussreichere Menschen.⁸⁴

Oshana legt dezidiert dar, warum es eine soziale Autonomiekonzeption braucht. In einer Art Plädoyer hebt sie u.a. hervor, dass die von ihr befürwortete Konzeption das Hauptaugenmerk auf die Autonomie von Personen legt und nicht mehr auf die Autonomie der Wünsche. Autonomie sei zumindest teilweise auch eine Funktion sozialer Umstände und Beziehungen und nicht allein aus dem inneren Vermögen der Willensbildung ableitbar.⁸⁵ Oshana führt weiterhin aus, dass günstige soziale Beziehungen Autonomie fördern. Psychische Merkmale wie Selbstreflexion oder die Identifikation mit den eigenen Wünschen „*gedeihen förmlich, wenn eine Person ungehindert mit anderen interagiert.*“⁸⁶

Der Begriff der relationalen Autonomie wird meist herangezogen, um Bezug auf Autonomiekonzepte zu nehmen, die aus feministischer Kritik (vor allem) am kantischen

⁸³ Oshana, „Personale Autonomie und das soziale Umfeld“, 196.

⁸⁴ Oshana, „Personale Autonomie und das soziale Umfeld“, 212 f.

⁸⁵ Oshana, „Personale Autonomie und das soziale Umfeld“, 215.

⁸⁶ Oshana, „Personale Autonomie und das soziale Umfeld“, 217.

Moralbegriff und an in- und externalistischen Modellen entstanden sind und deren Ziel in der Neukonzipierung des Autonomiebegriffs aus feministischer Perspektive besteht.⁸⁷ Zentral für traditionelle Autonomievorstellungen sei die „atomistische“ Persönlichkeit, die idealerweise autark und wie aus einem Vakuum heraus einsame Entscheidungen trifft und danach handelt, losgelöst von sozialen Kontexten und frei von Emotionen und verzerrenden Neigungen. In diesen Vorstellungen spielen beispielsweise internalisierte Unterdrückung und repressive soziale Bedingungen, denen Frauen oft ausgesetzt sind, und die für die Entwicklung autonomer Lebensweisen einschränkend sind, keine Rolle. Die feministische Perspektive auf Autonomie bezieht hingegen gesellschaftliche Kontexte mit ein. Personen sind sozial und historisch eingebettet. Autor*innen wie Beate Rössler betonen:

„Autonom sind wir nie allein: Wir sind autonom immer in sozialen und politischen Kontexten, autonome Bürger und Bürgerinnen in Gesellschaften, immer im Zusammenspiel mit anderen.“⁸⁸

Autonomie ist in dem Sinne relational, dass sie ohne Bezug zum sozialen und politischen Kontext nicht möglich ist. Relationale Autonomietheorien wollen eine Antwort auf die Frage geben, inwieweit repressive soziale Bedingungen und internalisierte Unterdrückung autonome Lebensweisen von Frauen untergraben oder erodieren lassen.

89

In der feministischen Autonomieliteratur wurden „Hard Cases“ beschrieben, die sich als „Autonomieversagen“ interpretieren lassen, wobei ein graduelles Verständnis von Autonomie hier sicherlich hilfreich ist und radikale Positionen abmildert. Denn auch in diesen Fällen wäre es problematisch, Frauen jegliche Autonomie abzusprechen. Die Fälle⁹⁰:

Selbstverleugnung: Was hiermit gemeint ist, macht ein häufig verwendetes Zitat aus einem Essay von Virginia Woolf sehr anschaulich, in dem der „Engel des Hauses“ als äußerst einfühlsam, liebenswürdig und unübertroffen in den „schwierigen Künsten des Familienlebens“ beschrieben wird.⁹¹ Eine Metapher für die ebenfalls gerne angeführte

⁸⁷ Stoljar, „Feminist Perspectives on Autonomy“.

⁸⁸ Rössler, *Autonomie: ein Versuch über das gelungene Leben*, 322.

⁸⁹ Stoljar, „Feminist Perspectives on Autonomy“.

⁹⁰ Stoljar, „Feminist Perspectives on Autonomy“.

⁹¹ Woolf, „Berufe für Frauen“, 226; Zitiert nach: Rössler, *Autonomie: ein Versuch über das gelungene Leben*, 345.

„unterwürfige Hausfrau“. Es geht um übermäßige Rücksichtnahme auf die Bedürfnisse anderer und die permanente Unterordnung der eigenen Wünsche.

Adaptive Präferenzen: Dass die eigenen Wünsche an reale Bedingungen angepasst werden, ist ein Phänomen, das jede*r selbst schon einmal erlebt hat. Beispiel: Ich spüre den dringenden Wunsch nach einer Reise in ein exotisches Land. Schnell aber wird mir klar, dass das ersehnte Ziel nicht zu meinem Geldbeutel passt. Mir kommt in den Sinn, dass Flugreisen sehr umweltschädlich sind, dass langes Sitzen im Flieger und die Zeitverschiebungen nicht wirklich „mein Ding“ sind und dass es im eigenen Land doch auch sehr schön ist. Deshalb ziehe ich eine Radtour vor. Ich richte also meine Wünsche danach aus, was möglich ist, was mir gut und richtig erscheint. In der feministischen Theorie werden adaptive Präferenzen als Wünsche verstanden, die repressiven sozialen Bedingungen angepasst sind. Demzufolge übernehmen Frauen in Beziehungen, in der Familie oder in der Arbeitswelt Vorlieben und Verhaltensweisen, die sie in ihrem eigenen historisch-sozialen Kontext erfahren haben. Das passiert unbewusst und kann etwa im Zuge einer unterdrückerischen Geschlechtersozialisation die Fähigkeit zur Autonomie behindern. Umgekehrt, also positiv gewendet können soziale Beziehungen zu einer Entwicklung und Förderung von Autonomie beitragen.⁹²

Übernahme von Unterdrückungspraktiken: Adaptiven Präferenzen liegen unbewusste Anpassungsmechanismen zugrunde. Demgegenüber werden in der einschlägigen Literatur Beispiele angeführt, die auf eine bewusste Unterordnung unter Praktiken der Geschlechterunterdrückung verweisen und an denen sich Frauen aktiv beteiligen. Ein strittiges Beispiel ist die Verschleierungspraxis. Die Beschneidung junger Frauen und Mädchen oder die arrangierte Ehe sind Praktiken, die Frauen erheblichen physischen und psychischen Schaden zuzufügen scheinen und von Frauen selbst aktiv gefördert werden.⁹³

Dort heißt es: “Diese Person war von inniger Einfühlsamkeit. Sie war unendlich liebenswürdig. Sie war gänzlich selbstlos. Sie war unübertroffen in den schwierigen Künsten des Familienlebens. Täglich opferte sie sich auf. Gab es Hühnchen, nahm sie das Bein; war irgendwo Zugluft, so saß sie darin – kurzum, sie war so beschaffen, dass sie weder einen eigenen Kopf noch einen eigenen Wunsch hatte, sondern es immer vorzog, mit den Köpfen und den Wünschen anderer übereinzustimmen.“

Virginia Woolfs Kritik an Coventry Patmores Gedicht „Angel in the House“.

⁹² Stoljar, „Feminist Perspectives on Autonomy“; Rössler, *Autonomie: ein Versuch über das gelungene Leben*, 337 ff.

⁹³ Stoljar, „Feminist Perspectives on Autonomy“.

Selbstwert

Wenn in Zusammenhang von Selbstverleugnung, adaptiven Präferenzen oder Unterdrückungspraktiken von psychischen Schäden die Rede ist, dann sind damit Beeinträchtigungen der psychischen Stabilität, des Selbstwertgefühls, des Selbstbewusstseins und des Selbstvertrauens gemeint. Mangelndes Selbstwertgefühl wird als ein Grund für verminderte oder fehlende Autonomie und Handlungsfreiheit angesehen.⁹⁴ Umgekehrt heißt das: Ein stabiles Selbstwertgefühl ist eine wichtige Voraussetzung für die Verwirklichung von individueller Autonomie. Das Selbstwertgefühl aber wird in sozialen Beziehungen mit anderen Menschen entwickelt. Anerkennung durch Andere wirkt sich meist fördernd und unterstützend aus, Missachtung oder Abwertung eher schädigend. Relationale Ansätze gehen davon aus, dass Beziehungen, die Relationalität des Menschen den Begriff der Autonomie konstituiert.

Zur Veranschaulichung sei hier ein weiterer Begriff von Autonomie zitiert, der einige Elemente relationaler Autonomie zusammenfasst, ohne auf feministische Motive direkt Bezug zu nehmen:

„Individuelle Autonomie wird allgemein als die Fähigkeit von Personen verstanden, über ihr eigenes Leben bestimmen zu können, ihr eigenes Leben zu führen anhand von Gründen, Überlegungen, Motiven, Wünschen, die ihre eigenen sind und ihnen nicht von anderen – aus persönlichen oder politischen Gründen – aufgezwungen, die aber immer in Beziehung mit anderen entwickelt werden. Deshalb begründet Autonomie Verantwortlichkeit und fordert Respekt von anderen.“⁹⁵

Diese Definition stammt von Beate Rössler, die Privatheit zu den externen Bedingungen einer so verstandenen Autonomie zählt.⁹⁶

2.1.3.6 Privatheit im Kontext relationaler Autonomie

Das Wort privat kommt vom lateinischen privatus, was „beraubt“ bedeutet, also einen Zustand ausdrückt. Wenn ich mich im *Zustand der Privatheit* befinde, dann sind Andere des Zugangs zu etwas beraubt, das zu mir gehört und über das ich weitestgehend frei und nach eigenen Vorstellungen, also autonom verfüge. *„Privat‘ ist etwas dann, wenn eine Person dazu in der Lage und prinzipiell berechtigt ist, den Zugang - zu Daten, zu*

⁹⁴ Hierzu insbesondere: Benson, „Handlungsfreiheit und Selbstwert“.

⁹⁵ Rössler, „Autonomie“.

⁹⁶ Rössler, *Autonomie: ein Versuch über das gelungene Leben*, 281.

Wohnungen, zu Entscheidungen oder Handlungsweisen – zu kontrollieren.“⁹⁷ Der Begriff ist also sehr weit definiert. Privatheit findet nicht allein innerhalb der vier Wände der eigenen Wohnung statt, sondern auch im öffentlichen Raum und umfasst in Bezug auf die Räumlichkeit verschiedene Ebenen. Grimm et al. sprechen von einem „Zwiebelmodell der Privatheit“⁹⁸. Im Innersten der Zwiebel, ihrem Spross, ist die körperliche Intimität angesiedelt. Die zweite Schicht beheimatet die Familie und intime Beziehungen zu anderen Menschen. Das ist der Bereich, der zusammen mit der dritten Schicht - dem eigenen Zimmer, der Wohnung und dem Haus üblicherweise von vielen als der klassische Privatbereich oder die Privatsphäre gesehen wird. Dazu kommt der Kreis echter Freunde und Freundinnen. In der äußersten Schale der Zwiebel befindet sich der im weitesten Sinne öffentliche Bereich der Gesellschaft, des Staates und der Wirtschaft. Auch im Öffentlichen kann ich privat entscheiden und handeln. Die Teilnahme an Demonstrationen ist ebenso meine Privatangelegenheit, wie ein Restaurantbesuch und das, was ich dort esse, trinke und mit Freunden diskutiere.⁹⁹ Auch meine Teilnahme an einer öffentlichen Debatte gehört zu meiner Privatheit, egal ob sie in physischer Präsenz an einem Versammlungsort stattfindet oder auf Social-Media-Plattformen.

Privatheit steht in sozialen Kontexten, die Helen Nissenbaum als „*structured social settings*“ beschreibt, die sich über die Zeit entwickelt haben. Sie meint damit konkret z.B. Familie, Religion, Bildung, Geschäftsleben oder soziale Vorsorgesysteme. Als Individuen bewegen wir uns in und durch eine Vielzahl solcher Settings und aus ihnen heraus. Wir agieren und interagieren jeweils in bestimmten Rollen (z.B. als Kund*in, Lehrer*in, Familienmitglied) und auf unterschiedliche Art und Weise indem wir z.B. eine Ärztin oder einen Arzt konsultieren, eine Schule oder ein Konzert besuchen oder einkaufen gehen.¹⁰⁰ Bei diesen Interaktionen geht es immer auch um Informationsflüsse, in denen ich private Aspekte meiner Person preisgebe - dem Arzt und der Ärztin andere als meinen Freunden und Freundinnen, denn natürlich sind auch Informationsflüsse kontextabhängig.¹⁰¹

⁹⁷ Rössler, *Autonomie: ein Versuch über das gelungene Leben*, 285.

⁹⁸ Grimm u. a., *Digitale Ethik*, 33.

⁹⁹ Grimm u. a., *Digitale Ethik*, 32.

¹⁰⁰ Nissenbaum, *Privacy in Context*, 79.

¹⁰¹ Nissenbaum entwickelt die Idee angemessener Informationsflüsse mit Blick auf den Schutz von Privatheit. Nissenbaum, *Privacy in Context*, 78 ff.

Das *Vorhandensein von Privatheit* gehört zu den externen Bedingungen, die individuelle Autonomie ermöglichen.¹⁰² Privatheit ist eine notwendige und konstitutive Bedingung für individuelle Autonomie. Ohne Privatheit ist individuelle Autonomie nicht möglich oder wird zumindest erschwert. Wir brauchen Privatheit für ein autonomes Leben, denn Privatheit ermöglicht freie und offene Kommunikation, bedeutet Abwesenheit von Kontrolle, Zwang und Diskriminierung (negative Freiheit), ebenso die Abwesenheit von manipulativem Einwirken auf Einstellungen und Verhalten, ist der Rahmen für die Authentizität und bietet Raum für Reflexion und die Ausbildung von Resilienz.¹⁰³

Rössler unterscheidet drei Dimensionen von Privatheit. Die *informationelle Privatheit*, bei der es um Wissen anderer über meine Person geht. Andere Menschen und Institutionen haben keinen Zugang zu Informationen über mich. Sie wissen über bestimmte Aspekte meiner Person nichts. Die *dezisionale Privatheit*, bei der es um meine privaten Entscheidungen geht, also z.B. für welchen Beruf ich mich entscheide, wie ich mich kleide, mit wem ich zusammenleben will, welche Partei ich wähle. Die *lokale Privatheit* schließlich ist die Privatheit der Wohnung.¹⁰⁴ Ebenso wichtig wie der Zustand der Privatheit ist die *Kontrolle der Privatheit*.¹⁰⁵ Diese nehme ich für mich in Anspruch, wenn ich beispielsweise darüber entscheide, wer wann Zutritt zu meiner Wohnung haben kann, wen ich an meiner Entscheidung über die Aufnahme eines Studiums beteilige oder mit wem ich zusammenleben will. Die Kontrolle meiner informationellen Privatheit ist vor allem in Bezug auf die Digitalisierung meiner privaten Lebenswelt von besonderer Bedeutung. Wem gebe ich welche Informationen über mich, welche Personen und Institutionen lasse ich im Unwissen, wenn ich der Meinung bin, dass es sie nichts angeht? Welches „Geheimnis“ behalte ich für mich.

2.1.4 Konzeptionalisierung der Autonomieansätze für diese Arbeit

Die vorangegangene Darstellung von Autonomiemodellen hat gezeigt, wie unterschiedlich individuelle Autonomie gesehen werden kann. Zwar gibt es einige Gemeinsamkeiten aber ein einheitlicher Begriff von individueller Autonomie oder ein brauchbarer Minimalkonsens sind nicht auszumachen. So hat Dworkin schon vor Jahren konstatiert, dass das einzig Gemeinsame darin bestehe, dass Autonomie ein Merkmal von

¹⁰² Rössler, *Autonomie: ein Versuch über das gelungene Leben*, 281.

¹⁰³ Grimm u. a., *Digitale Ethik*, 41.

¹⁰⁴ Rössler, *Autonomie: ein Versuch über das gelungene Leben*, 281 ff.; Rössler, *Der Wert des Privaten*.

¹⁰⁵ Siegesleitner, *Post-Privacy oder Datenschutz als neuer Megatrend?*

Personen und eine wünschenswerte Eigenschaft ist.¹⁰⁶ Daran hat sich bis heute nicht viel geändert, obwohl Versuche dahingehend unternommen wurden.¹⁰⁷ Es ist also zu klären, welche Relevanz einzelne Autonomiemodelle (z.B. das von Frankfurt) oder der jeweilige konzeptionelle Ansatz (z.B. der internalistische) für diese Arbeit oder genauer gesagt für die Beantwortung der Forschungsfrage dieser Arbeit haben kann.“¹⁰⁸

2.1.4.1 Relevanz einzelner Ansätze

Folge ich internalistischen Vorstellungen von Autonomie oder auch externalistischen wie dem von John Christman, fokussiere ich meine Fragestellung im Wesentlichen auf die Bedeutung von KI-Systemen für die handlungswirksame und reflektierte Willensbildung der Person. Zieht man die Möglichkeit der Manipulation durch KI-Systeme in Betracht, so ist dies eine Fragestellung von höchster Relevanz. In ähnlicher Weise ließe sich für substantielle Ansätze argumentieren, wenn man beispielsweise an die Entstehung und Beeinflussung moralischer Einstellungen etwa durch Deliberation mit anderen denkt. Folge ich relationalen Autonomieansätzen, rücke ich beispielsweise soziale Beziehungen oder Privatheit in den Mittelpunkt. KI-Systeme, so darf wohl behauptet werden, haben heute für nahezu alle Menschen Bedeutung - *direkt* in der Interaktion mit diesen Systemen oder *indirekt* durch den Einfluss auf die Rahmenbedingungen ihrer beruflichen, öffentlichen und privaten Lebenswelt. Um ein Beispiel zu nennen: Individuelle Privatheit ist eine essenzielle Bedingung für die Möglichkeit eines autonomen Lebens und es ist naheliegend, die Frage nach der Bedeutung von KI-Systemen für die (informationelle) Privatheit zu stellen. Wenn also KI-Systeme die Privatheit des Individuums einzuschränken vermögen, dann können sie indirekt auch die individuelle Autonomie beeinträchtigen.

Eine Beschränkung auf ein bestimmtes Autonomiemodell ist für das Erkenntnisinteresse der Arbeit nicht erforderlich und scheint zudem nicht sinnvoll zu sein. Denn alle Ansätze ermöglichen interessante und nachvollziehbare Fragestellungen. Diese Arbeit folgt also keinem bestimmten Begriff oder konzeptionellem Ansatz von Autonomie, weil es dem formulierten Erkenntnisinteresse nicht umfänglich gerecht werden würde. Demzufolge werden sowohl die im Individuum begründeten Bedingungen der Willens- und

¹⁰⁶ Dworkin, *The Theory and Practice of Autonomy*, 6. „About the only features held constant from one author to another are that autonomy is a feature of persons and that it is a desirable quality to have.“

¹⁰⁷ Seidel, *Selbst bestimmen*.

¹⁰⁸ Betzler, *Autonomie der Person*, 33. Betzler weist darauf hin, dass es ein „dringendes Desiderat“ sei, entweder das verwendete Autonomiekonzept zu verteidigen oder von einem viel allgemeineren Begriff auszugehen.

Handlungsautonomie mit in die Betrachtungen einbezogen, wie die externen Bedingungen und solche, die in sozialen Kontexten zu finden sind. Dies lässt eine bereitere Analyse der Bedeutung der Phänomene und Erscheinungsformen von KI-Systemen zu.

2.1.4.2 Konstitutive Bedingungen oder Einflussfaktoren?

Vor allem relationalen Autonomiemodellen wird vorgeworfen, dass sie etwa „positiv“ gestaltete soziale Beziehungen als *konstitutiv* für individuelle Autonomie ansehen und nicht nur als Einflussfaktoren, die sich positiv oder negativ auf sie auswirken können. Für das Erkenntnisinteresse dieser Arbeit ist es unerheblich, ob Bedingungen für Autonomie *im theoretischen Sinne notwendig und hinreichend* definiert sind, ob also beispielsweise das Vorhandensein sozialer Beziehungen, die das Selbstwertgefühl positiv beeinflussen, eine *notwendige* Bedingung von Autonomie darstellt oder nicht. Vielmehr steht das autonomiefördernde bzw. das autonomieeinschränkende *Potential* von KI-Systemen im Fokus. Dass beispielsweise repressive soziale Kontexte die Autonomiefähigkeit stark einschränken können, dürfte weitgehend unbestritten zu sein.¹⁰⁹ Das erfordert jedoch nicht, dass „positive“ soziale Beziehungen deshalb als *konstitutiv* für Autonomie angesehen werden müssen. Trotzdem ließe sich etwa die These formulieren, dass rassistische Biases, also Voreingenommenheit und Verzerrungen im Datenmaterial und in Algorithmen von KI-Systemen vorurteilsfreie Entscheidungen erschweren. Für internalistische Ansätze gilt im Übrigen eine solche Argumentation gleichermaßen. Wenn etwa kritisch argumentiert wird, dass psychologische Bedingungen für die Definition von Autonomie allein nicht hinreichend sind, mag das zutreffen.¹¹⁰ Trotzdem ließe sich die These formulieren, dass das Manipulationspotential von KI-Systemen die kritische Reflexion bei autonomen Entscheidungen beeinträchtigt oder erodieren lässt.

2.1.4.3 Weiterführende Teilfragen

Die zentrale Fragestellung dieser Arbeit impliziert einen Zusammenhang zwischen individueller Autonomie und den Entscheidungen Künstlicher Intelligenz. Bezogen auf den Autonomiebegriff kann die Forschungsfrage hinsichtlich der Künstlichen Intelligenz aufgefächert werden. Es stellt sich nunmehr die Frage wie, inwieweit und in welchen konkreten Ausprägungen Künstliche Intelligenz die individuelle Autonomie bzw.

¹⁰⁹ Ach und Schöne-Seifert, „Relationale Autonomie. Eine kritische Analyse“, 55.

¹¹⁰ Oshana, „Personale Autonomie und das soziale Umfeld“, 217. Oshana argumentiert, dass ihr Fallbeispiele (Sklaverei ...) zeigen, „dass die Erfüllung der psychologischen Bedingungen schlicht und einfach nicht hinreichend für Selbstbestimmung ist.“

einzelne ihrer Komponenten, Facetten oder Bedingungen zu untergraben vermag, wenn sie Einfluss nimmt, auf:

- menschliche Willensbildung und Entscheidung.

Etwa durch Manipulation? Christman hat mit seinem Konzept versucht, den Manipulationsvorwurf gegen Frankfurts rein internalistisches Konzept zu entkräften und spricht Menschen Autonomie auch dann zu, wenn sie den Prozess ihrer Willensbildung trotz Manipulation akzeptieren. Jedoch bleibt es Manipulation. Inwieweit kann der Prozess der inneren Willensbildung und Entscheidung durch andere, KI-spezifische Manipulationstechniken beeinträchtigt werden?

- den „*verfügbaren Bereich relevanter Optionen*“ (Oshana).

Was bedeutet es für unsere Entscheidungsspielräume, wenn wir Künstlicher Intelligenz Entscheidungen überlassen und ihre Informationen übernehmen? Etwa wenn ich im beruflichen Kontext nicht nach eigenem Ermessen Kulanz gegenüber Kund*innen oder Kolleg*innen üben oder Rücksicht auf mich selbst nehmen kann, wenn die „Kälte“ und „Unumstößlichkeit“ maschineller Entscheidungen mir die Selbstbestimmung und Freiheit nehmen, wenn umgekehrt die Betroffenen aus meiner Sicht womöglich ungerecht behandelt werden und chancenlos bleiben?

- die lokale, informationelle und dezisionale Privatheit.

Künstliche Intelligenz basiert auf Daten und Informationen. Wie werden sie gewonnen transformiert und uns als Wissen zur Verfügung gestellt? Welche Rolle spielen dabei die privaten Daten? Was passiert mit ihnen? Geben wir in der Interaktion mit künstlich intelligenten Systemen

- die menschliche Verantwortung für maschinelle Entscheidungen.

Kann oder muss der Mensch Verantwortung übernehmen für Entscheidungen, die nicht er, sondern ein System der Künstlichen Intelligenz getroffen hat? Oder geht die Verantwortung auf die Maschine über? Kann sie verantwortlich sein? Oder müssen Hersteller*innen oder Programmierer*innen Verantwortung (mit)übernehmen?

- die zeitliche Autonomie.

Zu meiner Autonomie gehört zeitliche Autonomie, mit der ich bestimme, wann ich etwas tue, in welchem Rhythmus und wieviel Zeit ich mir dafür zugestehe oder mir zugestanden wird. Inwieweit wirkt sich Künstliche Intelligenz auf die zeitliche Gestaltung meiner privaten und beruflichen Lebenswelt aus? Was bedeutet die Digitalisierung der sozialen Kommunikation, mit der permanente Erreichbarkeit einhergeht, und die von anderen gefordert wird, für die autonome Verfügung über meine Zeit? Werden die Menschen zu Getriebenen künstlich-intelligenter Systeme oder sind sie es schon?

- den sozialen und gesellschaftlichen Kontext.

Relationale Ansätze betrachten individuelle Autonomie immer im Zusammenhang von sozialen und gesellschaftlichen Kontexten. So brauchen Menschen für die Entwicklung und Beibehaltung ihrer Autonomie zum Beispiel soziale Anerkennung. Liberale, demokratische Gesellschaften schaffen die Rahmenbedingungen und stellen Optionen für eine autonome Lebensgestaltung bereit. Jedoch neigen staatliche Institutionen dazu, das Leben der Menschen mit Regeln zu überziehen und zu kontrollieren. Andere gesellschaftliche Akteure streben nach Macht und gesellschaftlichem sowie politischem Einfluss.

Welche Rolle kann Künstliche Intelligenz spielen? Finden Menschen in sozialen Netzwerken Anerkennung? Wie und inwieweit werden soziale und gesellschaftliche Konflikte über Künstlicher Intelligenz geschürt? Werden gesellschaftliche Gruppen über Künstliche Intelligenz diskriminiert? Inwieweit können soziale Ungerechtigkeiten durch künstlich intelligente Systeme entstehen? Wie und inwieweit wird durch Künstliche Intelligenz gesellschaftliche und politische Macht auf- und ausgebaut, durch die die Rahmenbedingungen für individuelle Autonomie verändert werden?

Bevor Antworten auf diese Frage erarbeitet werden können, ist der Begriff der Künstlichen Intelligenz zu klären.

2.2 Künstlich intelligente Systeme

2.2.1 Intelligenz und die Frage, ob Maschinen denken können

Der Begriff der Intelligenz geht zurück auf das lateinische Wort *intellegere* und bedeutet „‘dazwischen wählen’ (lat. *inter legere*), d. h. ‘eine überlegte Wahl treffen’“. Das lateinische *intelligentia* heißt „Einsicht, Erkenntnis, Verständnis, Verstand“. Der Begriff hat eine jahrhundertlange Geschichte, wird in verschiedenen Wissenschaften jeweils unterschiedlich und teils widersprüchlich definiert.¹¹¹ Allgemein wird Intelligenz heute charakterisiert als „(Grad der) Fähigkeit, Informationen aufzunehmen, sinnvoll zu verarbeiten und auf dieser Basis rational und kreativ zu handeln.“¹¹² Diese Definition sieht Intelligenz nicht als eine rein menschliche Fähigkeit. Auch die Duden-Redaktion lässt durch die eckige Klammer eine Interpretation zu, in der Intelligenz nicht allein Menschen zugeschrieben wird. Intelligenz ist die „*Fähigkeit [des Menschen], abstrakt und vernünftig zu denken und daraus zweckvolles Handeln abzuleiten*“¹¹³. Dabei spielt Adaptionfähigkeit eine zentrale Rolle. Zudem sieht sie im Begriff der künstlichen Intelligenz als „*Fähigkeit bestimmter Computerprogramme, menschliche Intelligenz nachzuahmen*“¹¹⁴ eine übertragene Bedeutung.

Es gibt weitere Definitionen von Intelligenz, psychologische etwa. Bekannterweise wird mit Intelligenztests menschliche Intelligenz gemessen, es gibt also Kriterien dafür, was Intelligenz ist oder sein soll. Die Frage danach soll jedoch hier nicht weiter erörtert werden, weil wir es mit multidimensionalen und hochkomplexen Zusammenhängen zu tun haben, deren Darstellung hier den Rahmen sprengen und vermutlich nicht weiterführen, was sich in den vielen uneinheitlichen Definitionen von künstlicher Intelligenz widerspiegelt. Zudem stand schon vor der Erfindung des Begriffs der künstlichen Intelligenz eine zwar verwandte, aber etwas andere Frage im Mittelpunkt des wissenschaftlichen Interesses: Können Maschinen denken?

2.2.1.1 Alan Turings „Imitationsspiel“ - Können Maschinen denken?

In seiner berühmten Abhandlung „*Computing Machinery and Intelligence*“ greift Alan Turing 19?? gleich im ersten Satz erneut eine grundlegende Frage auf: „*I propose to consider the question, ‘Can machines think?’*“¹¹⁵ Er bemüht sich aber nicht um

¹¹¹ Piepmeier u. a., „Intelligenz, Intelligentsia, Intellektueller“.

¹¹² DWDS – Digitales Wörterbuch der deutschen Sprache, *Intelligenz*.

¹¹³ Dudenredaktion, „Intelligenz“.

¹¹⁴ Dudenredaktion, „Intelligenz“.

¹¹⁵ Turing, *Computing Machinery and Intelligence*, 6.

Definitionen von „Maschine“ und „denken“, sondern kündigt an, die Frage durch eine andere, verwandte zu ersetzen. Er schlägt dazu das „Imitationsspiel“ vor, ein Frage-und-Antwort-Spiel mit einem Fragesteller (C) sowie einem Mann (A) und einer Frau (B). Der Fragesteller befindet sich in einem separaten Raum und stellt seine Fragen schriftlich, damit er weder den Mann noch die Frau am Aussehen oder der Stimme erkennt. Denn seine Aufgabe besteht darin, durch geschicktes Fragen herauszufinden, wer von den beiden Mann und wer Frau ist. Die Aufgabe von A ist es, C zu täuschen und zu falschen Schlüssen zu bringen, die Aufgabe von B ist es, C zu helfen und am besten wahrheitsgetreu zu antworten. Was wäre nun, fragt Turing, wenn eine Maschine die Rolle von A übernehme? Würde sich Fragesteller*in C ebenso häufig irren, wenn A und B Mann und Frau wären? Dies sind für Turing die Fragen, die an die Stelle der Frage „Können Maschinen denken?“ treten.¹¹⁶ Turing ersetzt die Frage durch einen Verhaltenstest für Maschinen. Das Spiel ist als Turing-Test bekannt geworden. Der gilt für einen Computer als bestanden, wenn ein menschlicher Fragesteller mit seinen schriftlich gestellten Fragen nicht feststellen kann, ob die ebenfalls schriftlichen Antworten von einem Computer stammen oder von einem Menschen.

Dafür muss ein Computer allerdings über Fähigkeiten zur Verarbeitung von natürlicher Sprache, zur Wissensrepräsentation, d.h. zur Speicherung von Wissen und Wahrgenommenem, zu automatisierter Schlussfolgerung und zu maschinellern Lernen sowie zur Adaption an neue Gegebenheiten und zur Mustererkennung verfügen.¹¹⁷ Turing antizipiert eine Reihe von Einwänden gegen die Möglichkeit denkender Maschinen und versucht sie sogleich zu entkräften. Darunter ist das Bewusstseinsargument¹¹⁸, das auch heute sehr verbreitet ist. Er zitiert aus einer Rede des Neurologen Geoffrey Jefferson, der 1949 folgende Auffassung vertrat:

*„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 machine equals brain – that is, not only write it but know that it had written it.“*¹¹⁹

Turing meint, dass man nur dann dieser Ansicht sein könne, wenn man selbst diese Maschine sei und fühle, dass man denkt. Und man könne nur mit Sicherheit wissen, dass

¹¹⁶ Turing, *Computing Machinery and Intelligence*, 7–9.

¹¹⁷ Russell und Norvig, *Artificial Intelligence*, 2.

¹¹⁸ Turing, *Computing Machinery and Intelligence*, 51.

¹¹⁹ Zitiert nach: Turing, *Computing Machinery and Intelligence*, 50.

ein Mensch denkt, wenn man selbst dieser Mensch sei, was auf einen Solipsismus hinauslaufe.¹²⁰ Er räumt allerdings ein, dass die Frage menschlichen Bewusstseins schwierig ist und es Geheimnisse darum gibt, deren Lösung er aber für die Beantwortung der Fragen seiner Abhandlung nicht als notwendig erachtet.¹²¹

Es gibt zahlreiche weitere Argumente für und gegen Turings Ideen. Am Ende bleibt es eine Glaubensfrage.

„We agree with Turing—we are interested in creating programs that behave intelligently. The additional project of making them conscious is not one that we are equipped to take on, nor one whose success we would be able to determine.“¹²²

So beantworten Stuart Russel und Peter Norvig, die Verfasser eines Standardwerkes der Künstlichen Intelligenz, die Frage für sich. Viele Jahre später kommt John Searle mit einem Experiment zu dem Ergebnis, dass kein Computer denken kann, solange er nicht über Intentionalität verfügt.

2.2.1.2 Das chinesische Zimmer – John Searles Gedankenexperiment

Eines der bekanntesten Gedankenexperimente der jüngeren Philosophiegeschichte ist das „Chinese Room Argument“ aus dem Jahr 1980. Auch Searl geht der Frage nach, ob Maschinen denken können, und untersucht die folgenden Thesen:

„(1) Intentionality in human beings (and animals) is a product of causal features of the brain. I assume this is an empirical fact about the actual causal relations between mental processes and brains. [...] (2) Instantiating a computer program is never by itself a sufficient condition of intentionality.“¹²³

Mit Intentionalität ist die Gerichtetheit des Geistes gemeint: Gedanken, Vorstellungen oder Wünsche beziehen sich auf etwas. Ein Gedanke „meint“ etwas, hat einen Inhalt oder ist „über“ etwas (z. B. Ich denke an Paris). Intentionalität ist ein genuin geistiges Phänomen. Nur bewusste Wesen (Menschen und Tiere) können „über etwas“ denken.¹²⁴ Maschinen können das nicht, weil sie keine mentale Innenwelt haben. Auch wenn eine KI so „tut“, als ob sie etwas versteht, fehlt ihr echte Bedeutung. Computer beherrschen

¹²⁰ Turing, *Computing Machinery and Intelligence*, 53.

¹²¹ Turing, *Computing Machinery and Intelligence*, 57.

¹²² Russell und Norvig, *Artificial Intelligence*, 1033.

¹²³ Searle, „Minds, brains, and programs“, 417.

¹²⁴ Vgl. z.B. Misselhorn, *Grundfragen der Maschinenethik*, 268.

nur syntaktische Prozesse, aber kein Verständnis der Semantik. Das Chinese Room Argument zeigt: Ein System kann sprachlich kompetent erscheinen, ohne etwas zu verstehen. Intentionalität ist bei Searle untrennbar mit Bewusstsein, Subjektivität und Bedeutung verknüpft – Eigenschaften, die Computer nicht besitzen. Seine Kritik stellt eine bleibende Herausforderung für die philosophischen Grundlagen der KI dar. Searles Kernthese: Kein Computer, so komplex er auch sei, kann „wirklich denken“, solange er keine Intentionalität hat – und damit keine Bedeutung erzeugt.

"«Could a machine think?» On the argument advanced here only a machine could think, and only very special kinds of machines, namely brains and machines with internal causal powers equivalent to those of brains. And that is why strong AI has little to tell us about thinking, since it is not about machines but about programs, and no program by itself is sufficient for thinking. "¹²⁵

Das Szenario: Angenommen jemand sitzt in einem abgeschlossenen Raum mit einem großen Stapel chinesischer Schriften. Diese Person kann weder chinesisch schreiben noch sprechen. Sie versteht also absolut nichts von dem, was dort geschrieben steht. Es ist bedeutungslos. Sie beherrscht lediglich die englische Sprache. Sie erhält nun einen zweiten Stapel mit chinesischen Texten zusammen mit einem Regelwerk in englischer Sprache, welches die Person in die Lage versetzt, die Symbole des ersten Stapels mit denen des zweiten aufgrund ihrer Form in Korrelation zu setzen. Damit nicht genug. Es gibt einen dritten Stapel mit chinesischen Symbolen und Instruktionen auf Englisch, so dass zwischen dem dritten und den ersten beiden Stapeln korreliert werden kann. Zudem enthalten die Regeln Anweisungen, wie die Person mit bestimmten Symbolen auf Symbole des dritten Stapels antworten kann. Was sie aber nicht weiß, sind die Bezeichnungen der Stapel: Der erste heißt „a script“, der zweite „a story“, der dritte „questions“ und die Antwort heißt „answers to the questions“. Und schließlich sind die englischen Regelwerke „the program“. In gleicher Weise arbeitet die Person Fragen und Antworten in Englisch ab. Sowohl die Person als auch die Programmierer („programmers“) werden nach einer Weile so perfekt sowohl im Chinesischen wie im Englischen, dass Menschen außerhalb des Raumes glauben, es mit Personen zu tun haben, deren Muttersprache Chinesisch bzw. Englisch ist.

¹²⁵ Searle, „Minds, brains, and programs“, 417.

„But in the Chinese case, unlike the English case, I produce the answers by manipulating uninterpreted formal symbols. As far as the Chinese is concerned, I simply behave like a computer; I perform computational operations on formally specified elements. For the purposes of the Chinese, I am simply an instantiation of the computer program“

Von außen sieht es so aus, als würde die Person im Raum Chinesisch verstehen – tatsächlich aber folgt sie nur syntaktischen Regeln, ohne Semantik oder Bedeutung. Computerprogramme können Syntax verarbeiten (Zeichenregeln befolgen), aber verstehen keine Bedeutung. Semantisches Verständnis kann nicht auf bloße symbolische Manipulation reduziert werden. Das Gedankenexperiment richtet sich gegen die sogenannte starke KI-These: „Ein hinreichend programmiertes System versteht und hat einen Geist.“ Searles Antwort: Nein – Rechnen und Verstehen sind zwei völlig unterschiedliche Dinge. Maschinen simulieren Verstehen, ohne es wirklich zu haben. Maschinen folgen Regeln, aber ihre Prozesse haben keinen intentionalen Gehalt. Menschen haben Intentionalität – unsere Gedanken bedeuten etwas, sie beziehen sich auf etwas.

Relevanz für KI-Debatten heute: Chatbots und Sprachmodelle wie GPT erzeugen scheinbar sinnvolle Texte – aber: Verstehen sie, was sie sagen? Das chinesische Zimmer zeigt: Auch wenn KI-Systeme äußerlich intelligent wirken, fehlt ihnen möglicherweise inneres Bewusstsein, Verstehen, Subjektivität.

Auch das Chinese-Room-Argument hat Kritiker auf den Plan gerufen: Einige Philosophen (z. B. Daniel Dennett oder Vertreter der Systemtheorie) argumentieren: Nicht der Mensch im Zimmer, sondern das Gesamtsystem (Regeln + Mensch + Input/Output) ist relevant. Das System als Ganzes „versteht“ – analog zu einem Computer. Searle entgegnet: Auch das System hat keinen Zugang zu Bedeutung, es bleibt ein simulierter Geist, kein echter.

Die Person versteht nichts. Sie manipuliert nur Symbole nach formalen Regeln. Die „starke KI“-These (Computer denken oder verstehen) wird abgelehnt. Maschinen können Intelligenz simulieren, aber Inhalte nicht wirklich verstehen. Menschliches Denken ist intentional (gerichtet auf Bedeutungen, Inhalte). Maschinen fehlt diese geistige Ausrichtung. Computer verarbeiten Zeichen (Syntax), aber verstehen keine Bedeutungen (Semantik).

2.2.2 Definitionen künstlich-intelligenter Systeme

Ein kurzer historischer Exkurs in das Jahr 1948 und erneut Alan Turing. Er gilt als einer der Väter der Künstlichen Intelligenz. Und er war ein Visionär. In einem heute weithin als Manifest der Künstlichen Intelligenz geltenden Report für das *National Physical Laboratory* mit dem Titel *Intelligent Machinery*¹²⁶ hat er gleich im ersten Satz die Frage gestellt, ob es Maschinen möglich sei, intelligentes Verhalten zu zeigen. Das war 1948 und damit einige Jahre bevor der Begriff „Artificial Intelligence“ geprägt wurde. Er nahm Ideen vorweg, die sich heute in Systemen der künstlichen Intelligenz wiederfinden. Er unterschied zwei Typen von Maschinen, eine mobile und eine stationäre. Den mobilen Typ stellte er sich als eine Art Pendant zum Menschen vor, denn er ging davon aus, dass jeder Teil des menschlichen Körpers mithilfe von Technik imitiert werden könne. Seine Idee: Mikrofone fungieren als Ohren, Kameras treten an die Stelle von Augen, Roboterarme und -beine an die von Gliedmaßen und das Gehirn als Steuerungszentrale besteht - ähnlich dem menschlichen Gehirn - aus Neuronen. Bewegungsfähig und ausgestattet mit Sensoren sollte dieses Imitat des Menschen, die Möglichkeit bekommen, Dinge selbst herauszufinden und die Umgebung zu erkunden. Damit wollte Turing das System lernfähig machen. Heute würde man diesen Ansatz als „Embodiment“ bezeichnen, also als Verkörperung künstlicher Intelligenz. Turing war klar, dass die Realisierung seiner Vision einen hohen Aufwand erfordern würde. Aber ihm schien es der sicherste Weg.

Der zweite Ansatz sei eine Maschine, die intelligentes Verhalten imitieren könne: ohne Körper aber mit „Organen“ zum Sehen, Sprechen und Hören ausgestattet. Was könnte eine solche Maschine tun? Turing schlägt u.a. Kryptographie, Schach und andere Spiele oder die Übersetzung von Sprachen vor. Diese Version einer intelligenten Maschine brauche wenig Kontakt zur Umgebung. Eine Maschine, die nur in der Lage ist, zum Beispiel Schach zu spielen, braucht „Augen“ zur Unterscheidung unterschiedlicher Positionen und Vorrichtungen zur Ansage ihrer eigenen Züge. Sie muss sich dafür nicht in der Umgebung bewegen. Die Lernfähigkeit dieser Systeme war ein zentraler Punkt für Turing. Wie menschenähnlich er sich das mobile System vorstellte, wird deutlich im Kapitel „Education of Machine“. Dort stellt er fest, dass es sehr unfair wäre, Maschinen, unmittelbar nachdem sie aus der Fabrik kommen, mit Universitätsabsolventen vergleichen zu wollen. Er gesteht also den Maschinen eine gewisse „Kindheit“ zu, in der

¹²⁶ Turing, *Intelligent Machinery*.

sie Zeit zum Lernen haben. Turings Visionen wurden vom Direktor des britischen *National Physical Laboratory* als ‚schoolboy’s essay‘ abgetan.¹²⁷ Heute würde er vermutlich anders urteilen, denn Turings visionäre Ideen sind deutlich erkennbar in KI-Systemen Wirklichkeit geworden.

Den Begriff der Künstlichen Intelligenz oder besser: der Artificial Intelligence hat Turing nicht erfunden. Erst nach seinem Tod wurde er erstmalig 1956 im wissenschaftlichen Kontext verwendet. Und zwar im Förderantrag für die Dartmouth Conference, der den Titel „A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence“ trug. Grundlegend für das Projekt war die „Annahme, dass alle Aspekte des Lernens oder jede andere Eigenschaft von Intelligenz im Prinzip so genau beschrieben werden können, dass eine Maschine zu deren Simulation gebaut werden kann.“¹²⁸ Die für acht Wochen geplante Konferenz war interdisziplinär konzipiert. An ihr nahmen größtenteils Mathematiker, Logiker und Elektrotechniker teil, aber auch Psychologen, Wirtschaftswissenschaftler und Kybernetiker. Die Konferenz gilt inzwischen als Geburtsstunde des wissenschaftlichen Fachgebietes *Künstliche Intelligenz*.¹²⁹

Seitdem hat es zahlreiche Definitionen zum Begriff der Künstlichen Intelligenz gegeben. Es gibt jedoch im wissenschaftlichen Diskurs keinen Konsens darüber, was künstliche Intelligenz ist oder was sie ausmacht. Selbst über Kriterien wie Lernfähigkeit oder Problemkomplexität gibt es Kontroversen.¹³⁰ Das Forschungsgebiet war von vornherein interdisziplinär angelegt. Es ist naheliegend, dass der Begriff von einem Mathematiker anders verstanden wird als von einem Sozialwissenschaftler. Stellvertretend sei hier eine Definition zitiert, die von der Hochrangigen Expertengruppe für künstliche Intelligenz der Europäischen Union 2019 in einer aktualisierten Form vorgelegt wurde.

„Systeme der künstlichen Intelligenz (KI-Systeme) sind vom Menschen entwickelte Softwaresysteme (und gegebenenfalls auch Hardwaresysteme), die in Bezug auf ein komplexes Ziel auf physischer oder digitaler Ebene handeln, indem sie ihre Umgebung durch Datenerfassung wahrnehmen, die gesammelten strukturierten oder

¹²⁷ Pallay, „Vom Turing-Test zum General Problem Solver. Die Pionierjahre der künstlichen Intelligenz“.

¹²⁸ Pallay, „Vom Turing-Test zum General Problem Solver. Die Pionierjahre der künstlichen Intelligenz“.

¹²⁹ Pallay, „Vom Turing-Test zum General Problem Solver. Die Pionierjahre der künstlichen Intelligenz“.

¹³⁰ Caluori, „Hey Alexa, Why Are You Called Intelligent?“ Es handelt sich um eine empirische Untersuchung zur Fragestellung, welche Kriterien in Definition von künstlicher Intelligenz verwendet werden. Beginnend mit dem Jahr 1962. Das Ergebnis der Studie ist wenig überraschend. Es gibt keinen Konsens.

unstrukturierten Daten interpretieren, Schlussfolgerungen daraus ziehen oder die aus diesen Daten abgeleiteten Informationen verarbeiten, und über das bestmögliche Handeln zur Erreichung des vorgegebenen Ziels entscheiden. KI-Systeme können entweder symbolische Regeln verwenden oder ein numerisches Modell erlernen, und sind auch in der Lage, die Auswirkungen ihrer früheren Handlungen auf die Umgebung zu analysieren und ihr Verhalten entsprechend anzupassen. "

*Als wissenschaftliche Disziplin umfasst die KI mehrere Ansätze und Techniken wie z. B. maschinelles Lernen (Beispiele dafür sind „Deep Learning“ und bestärkendes Lernen), maschinelles Denken (es umfasst Planung, Terminierung, Wissensrepräsentation und Schlussfolgerung, Suche und Optimierung) und die Robotik (sie umfasst Steuerung, Wahrnehmung, Sensoren und Aktoren sowie die Einbeziehung aller anderen Techniken in cyber-physische Systeme). "*¹³¹

Dieser Begriff steckt ein sehr breites Spektrum unterschiedlicher künstlich-intelligenter Systeme ab. Er versucht in dieser allgemeinen Formulierung möglichst alle Erscheinungsformen von KI-Systemen abzubilden. Das wird noch deutlicher durch eine Ergänzung dieser Definition um eine (beispielhafte) Aufzählung von Themen, mit denen sich KI als wissenschaftliche Disziplin befasst: Maschinelles Lernen, maschinelles Denken sowie Robotik und cyber-physische Systeme¹³².

2.2.3 Klassifikationen von KI-Systemen

2.2.3.1 Symbolische, Subsymbolische und Hybride KI

Aus historischer Perspektive wird häufig eine Klassifikation in Symbolische KI, subsymbolische KI und hybride KI vorgenommen. Die folgende Darstellung orientiert sich an der von Klaus Mainzer¹³³ vorgelegten. Es gibt zahlreiche andere Beschreibungen, die mit abgewandelter Terminologie inhaltlich einer solchen Einteilung folgen. Zum

¹³¹ Hochrangige Expertengruppe für künstliche Intelligenz, „Eine Definition der KI: Wichtigste Fähigkeiten und Wissenschaftsgebiete“, 6.

¹³² Wikipedia, „Cyber-physisches System“ Definition Wikipedia: "Ein cyber-physisches System, engl. „cyber-physical system“ (CPS), bezeichnet den Verbund informatischer, softwaretechnischer Komponenten mit mechanischen und elektronischen Teilen, die über eine Dateninfrastruktur, wie z. B. das Internet, kommunizieren. " Zur Unterstützung von pflegebedürftigen Menschen werden z.B. „Ambient Assited Living“ Systeme eingesetzt, um etwa Vitaldaten zu überwachen oder Standorte zu ermitteln und per Internet zu melden. .

¹³³ Mainzer, „Was sind KI und Machine Learning?“

Beispiel auch Catrin Misselhorn, die statt von Subsymbolischer KI von Künstlich Neuronalen Netzen spricht, was ebenfalls verbreitet ist¹³⁴.

2.2.3.1.1 Symbolische KI – Logikbasierte Systeme

Symbolische KI orientiert sich an den Gesetzen der symbolischen Logik, sodass auch von logikbasierten Systemen gesprochen wird. Sie werden verwendet für die Lösung vollständig definierbarer Problemstellungen mit einem klar und eindeutig definierbaren Ziel. Dieses Ziel muss mithilfe von Algorithmen erreicht werden können und läuft in Teilschritten hauptsächlich unter Verwendung von Wenn-Dann-Regeln ab. Diese Systeme verarbeiten gut strukturierte Daten. Sie ermöglichen beispielsweise betriebliche Planung, Terminierung oder die Optimierung von Prozessen. Oft, indem Problemlösungen simuliert werden. Ein Beispiel dafür sind „Expertensysteme“. Diese Systeme greifen auf Daten zurück, die auf Expertenwissen basieren. Das Wissen beispielsweise von Ingenieuren oder Mediziner*innen dient den Systemen als Basis zur Entwicklung von formalen Regeln, die zur Generierung von Lösungsvorschlägen mit Wegen zur Zielerreichung genutzt werden. Das zu Grunde liegende Problem bezieht sich auf ein abgegrenztes Wissensgebiet. Manchmal werden solche Systeme auch als „Good Old-Fashioned Artificial Intelligence“ (GOFAI) bezeichnet¹³⁵. Sie spielen aber durchaus noch eine Rolle in Unternehmen oder in der Medizin.

2.2.3.1.2 Subsymbolische KI – Maschinelles Lernen

Die KI von heute hingegen lässt sich auf sehr viel komplexere Problemstellungen beziehen, die mit rein regelbasiertem Wissen nicht lösbar sind. Denn zur Lösung von Problemen braucht es oft mehr als nur das Wissen um die einzelnen Schritte zur Erreichung eines Zieles. Wissen ist immer auch eingebettet in Erfahrungen, die nicht in einem Lehrbuch stehen. Ein einfaches Beispiel ist das Autofahren. Zum verkehrssicheren Autofahren reicht der Theorieunterricht zur Führerscheinprüfung allein nicht aus. Wenn das so einfach wäre, gäbe es längst flächendeckend autonom fahrende Autos. In kritischen Situationen reagieren Autofahrer*innen intuitiv und ohne sich logischer Zusammenhänge bewusst zu sein, oft blitzschnell und aufgrund von sensorisch wahrgenommenen Informationen. Ähnlich geht es Chirurg*innen im Operationssaal und Pilot*innen im Cockpit.

¹³⁴ Misselhorn, *Grundfragen der Maschinenethik*, 21.

¹³⁵ Misselhorn, *Grundfragen der Maschinenethik*, 22.

Wie Menschen das Autofahren lernen, können auch Maschinen „lernen“, mit komplexen Problemstellungen umzugehen. Dabei spielen künstliche neuronale Netze (KNN) eine zentrale Rolle. KNN sind ähnlich aufgebaut wie das menschliche Gehirn, das sie modellhaft abbilden. Sie bestehen aus künstlichen Neuronen, die über Synapsen miteinander verbunden sind. Damit das künstliche Gehirn Informationen aufnehmen und abgeben, damit es also kommunizieren kann, braucht es eine Ein- und eine Ausgabeschicht. Dazwischen liegen mehr oder weniger zahlreiche versteckte Schichten. Künstliche Neuronen sind nichts anderes als kleine Prozessor-elemente, die jeweils relativ einfache Rechenschritte ausführen. Die Leistungsfähigkeit des Systems ist darin begründet, dass mehrere Millionen solcher Prozessorelemente ihre Arbeit parallel und vernetzt verrichten.

Anders als die Systeme der symbolischen KI, die in einem TOP-Down-Ansatz den Regeln der mathematischen Logik folgen, werten die Systeme des maschinellen Lernens (sensorielle) Daten mithilfe statistischer Korrelationen und Wahrscheinlichkeiten aus. Dem maschinellen Lernen liegen statistische Lerntheorien zugrunde. Zum Lernen werden unterschiedliche Methoden verwendet. Beim *überwachten Lernen* werden die Systeme mit Datensätzen gefüttert, die mit einem Label versehen sind. Wenn sie beispielsweise lernen sollen, Hunde von Katzen zu unterscheiden, dann werden sie mit Bildern von Hunden und Katzen trainiert, die jeweils entsprechend bezeichnet sind. Es wird also vorgegeben, was als Ergebnis herauskommen soll, wenn die Maschine etwas wahrnimmt, das aussieht wie ein Hund oder eine Katze. Überwacht wird das Lernen dadurch, dass das Ergebnis der KI mit dem Zielergebnis verglichen wird. Auch beim *unüberwachten Lernen* werden die Systeme mit Daten trainiert. Allerdings fehlt hier das Label. Die KI kennt also keinen Sollwert, keine Zielvorgaben. Sie weiß also nicht, was das Ergebnis sein soll. Sie muss selbständig Strukturen und Ähnlichkeiten herausfinden. Dabei werden Cluster gebildet, wobei Datenobjekte nach gleichen Eigenschaften gruppiert werden, um die Dimensionen eines Problems zu reduzieren: Die Komplexität eines Problems wird reduziert. Schließlich werden KI-Systeme durch *bestärkendes Lernen* trainiert. Ähnlich wie beim überwachten Lernen wird der Maschine in jedem Muster ein Zielwert vorgegeben. Entspricht die Antwort des Systems diesem Wert, enthält es eine entsprechende Antwort und wird dadurch in seinem „Verhalten“ bestärkt. Für den Fall einer falschen Antwort, wird auch das dem System mitgeteilt: ein Signal zur „Verhaltensänderung“.

Eine besondere Form des maschinellen Lernens in künstlichen neuronalen Netzen ist *Deep Learning*. Wenn es um sehr komplexe Problemstellungen geht, werden neuronale Netze verwendet, die eine Vielzahl von neuronalen Schichten, also Tiefe besitzen, die die Leistungsfähigkeit des Systems exponentiell steigert. Das ermöglicht der KI, komplexe Muster in riesigen und heterogenen Datenmengen (*Big Data*) zu erkennen und daraus zu lernen. Mit dem Begriff *Big Data* bezeichnet man große, komplexe, oft schnelllebige und schwach strukturierte Datenmengen, die mit herkömmlichen Methoden der Datenverarbeitung nicht ausgewertet werden können. Deep Learning funktioniert autonom und weitgehend ohne Zutun des Menschen. Einige Beispiele: *Bildverarbeitung*: Deep Learning wird häufig in der Bild- und Objekterkennung eingesetzt z.B. zur Gesichtserkennung und medizinischen Bildanalyse oder zur Erkennung von Verkehrszeichen beim autonomen Fahren. *Spracherkennung und -verarbeitung*: Technologien wie Sprachassistenten, automatische Übersetzung und Sentiment-Analyse¹³⁶ nutzen tiefes Lernen zur Verarbeitung und Interpretation natürlicher Sprache. *Handschriftenerkennung*: Die Systeme erkennen nach umfangreichen Trainingseinheiten chinesische Schriftzeichen, wobei weder die Programmierer*innen der Systeme noch die Trainer*innen chinesisch lesen oder verstehen können. *Entscheidungsfindung*: Deep Learning wird in der Entwicklung von Algorithmen für Spiele und in der Entscheidungsfindung in verschiedenen Bereichen eingesetzt.

2.2.3.1.3 Hybride KI

Die *hybride KI* ist eine Verbindung von symbolischer und subsymbolischer KI oder anders formuliert eine Verbindung von Logik und Wahrnehmung. Sie ist ein aktuelles Ziel der KI-Forschung. Hybride KI findet sich beispielsweise in der Robotik, wenn es darum geht, dass mechanische Bewegungsabläufe gelernt werden müssen.¹³⁷ In Erforschung der „emotionalen Künstlichen Intelligenz“ spielen beispielsweise Emotionserkennungen durch Roboter oder Empathie, die sie den Menschen zeigen, eine wichtige Rolle.¹³⁸

¹³⁶ Sentiment-Analyse ist eine Methode aus der Datenanalyse und Künstlichen Intelligenz, die darauf abzielt, Gefühle, Meinungen oder Stimmungen in Texten, Sprachaufnahmen oder anderen Daten automatisch zu erkennen und zu bewerten.

¹³⁷ Mainzer, „Was sind KI und Machine Learning?“, 6.

¹³⁸ Misselhorn, *Künstliche Intelligenz und Empathie*, 12.

2.2.3.2 Schwache und starke KI, Superintelligenz¹³⁹

Eine weitere Klassifikation verdeutlicht, wo KI-Systeme in Bezug auf ihre aktuelle und ihre visionären Fähigkeiten zu verorten sind.

2.2.3.2.1 Schwache oder enge KI

Bei den heute zum Einsatz kommenden Systemen handelt es sich nahezu ausschließlich um schwache oder enge künstlich intelligente Systeme. Die Bezeichnung enge KI weist darauf hin, dass es sich um spezialisierte Systeme handelt, die abgegrenzte ihnen zugedachte Aufgaben erledigen können. Sprachassistenten wie Siri oder Alexa reagieren auf menschliche Sprache und führen Befehle aus (Musik spielen, Telefonnr. Anrufen etc.). Aber sie können eben nur das und nicht etwa Texte erzeugen wie dafür erdachte Sprachmodelle. Das scheint allerdings nur eine Frage der Zeit zu sein, denn es gibt immer wieder Ankündigungen der Tech-Industrie, dass Spezialisierungen aufgehoben werden sollen. Schwache KI heißt nicht, dass diese Systeme nur geringfügige Leistungen erbringen. Im Gegenteil. Sie erzielen höchst erstaunliche Ergebnisse, die noch vor wenigen Jahren von vielen für unmöglich gehalten wurden. Aber sie tun das im Rahmen ihrer Aufgabe und verfügen nicht über eine maschinelle Intelligenz, die ganz allgemein zur Lösung beliebiger Aufgaben in der Lage wäre. Sie agieren und erwecken oft den Eindruck, als ob sie denken und Probleme lösen könnten und so intelligent wie Menschen seien.

2.2.3.2.2 Starke oder allgemeine KI

Starke KI ist bis heute nicht möglich und wird dies allen Einschätzungen zufolge in absehbarer Zeit auch nicht sein. Starke wird oft auch als allgemeine KI oder Artificial Generell Intelligence (AGI) bezeichnet. Sie soll die meisten Aufgaben erfüllen können, zu deren Lösung heute nur der Mensch in der Lage ist. Eine solche KI wäre also dem Menschen ebenbürtig, hätte ein eigenes Bewusstsein, einen eigenen Willen und selbst gesetzte Ziele, wäre sich ihrer selbst bewusst und könnte Emotionen entwickeln und über sie verfügen wie Menschen.

¹³⁹ Russell und Norvig, *Artificial Intelligence*, 1020.

„First, some terminology: the assertion that machines could act as if they were intelligent is called the weak AI hypothesis by philosophers, and the assertion that machines that do so are actually thinking (not just simulating thinking) is called the strong AI hypothesis“ Russel/Norvig unterscheiden nur zwei Typen und lassen „Superintelligenz“ außen vor.

2.2.3.2.3 Superintelligenz

KI-Systeme mit Superintelligenz sind dem Menschen in allen Belangen überlegen und übertreffen dessen Intelligenz und seine emotionale Kompetenz bei Weitem. Wenn es eine künstliche Intelligenz auf menschlichem Niveau gibt, so die Auffassung von Nick Bostrom, wird es bald auch eine Superintelligenz geben. Darunter versteht er einen Intellekt, der in praktisch jedem Feld menschlichen Fähigkeiten überlegen ist, u.a. auch in wissenschaftlicher Kreativität. Bostrom prognostiziert, dass es einen positiven Feedback Loop geben wird, eine KI, die sich selbst weiterentwickelt: „AIs would help constructing better AIs, which in turn would help building better AIs, and so forth.“ Die KI wäre dann also in der Lage in ihre grundlegenden algorithmischen Strukturen einzugreifen. Sie würden sich verselbständigen. Dadurch entziehen sie sich der menschlichen Kontrolle, ein „*Fukushima der künstlichen Intelligenz*“¹⁴⁰.

¹⁴⁰ Metzinger, „*Fukushima der künstlichen Intelligenz*“.

3 Wird individuelle Autonomie untergraben, wenn KI entscheidet?

Wir überlassen in vielen Bereichen menschliches Handeln und Entscheiden inzwischen Künstlicher Intelligenz. Wenn Algorithmen in der medizinischen Bildanalyse frühe Krebsdiagnosen ermöglichen, dann ist das auf den ersten Blick ohne Zweifel ein zu begrüßender technischer Fortschritt. Oft fällt auch die ethische Bewertung positiv aus: Wenn autonome Fahrzeuge dazu beitragen können, die Zahl tödlicher Verkehrsunfälle drastisch zu senken, erscheint vielen ihr Einsatz trotz ethischer Bedenken nicht nur als technische Option, sondern als moralische Verpflichtung. Und wenn KI-Systeme Kindern in bildungsfernen Regionen grundlegende mathematische Fähigkeiten vermitteln, ist auch das eine Möglichkeit, ihnen zu ein einem selbstbestimmten Leben zu verhelfen. In diesem Zusammenhang soll erneut darauf hingewiesen werden, dass der Begriff der individuellen Autonomie, wie er hier verwendet wird nur in liberal-demokratischen Staaten eine gewisse Berechtigung hat. Das gilt auch für die Frage nach der Untergrabung oder der Fremdbestimmung.¹⁴¹

So verheißungsvoll solche Entwicklungen sein mögen, sie sind gleichzeitig eine Herausforderung, es nicht bei dem ersten Blick zu belassen.¹⁴² Das gilt auch für Entscheidungen, die KI-Systeme im täglichen Dialog der Menschen mit ihnen treffen. Wenn wir einer Suchmaschine eine Frage stellen, trifft Künstliche Intelligenz Entscheidungen hinsichtlich der Antwort, die wir bekommen. Oder ChatGPT: Wenn wir das System veranlassen, einen Text zu verfassen, entscheidet Künstliche Intelligenz über die Quellen, die für den Text verwendet werden, über Gewichtungen, Stil und vieles mehr. Wenn der Mensch fragt und Künstliche Intelligenz antwortet, entscheidet in erster Instanz das System und der Mensch wird Adressat des Ergebnisses. Die grenzenlos erscheinenden Möglichkeiten Künstlicher Intelligenz dürfen nicht darüber hinwegtäuschen, dass die Systeme über ein Potential verfügen, individuelle Autonomie zu untergraben.

¹⁴¹ Bewohner in stark benachteiligten Regionen der Welt sind zu anderen Einstellungen zu Technik sogar verpflichtet. „Wahnsinn wäre es, in der Gegenwart eines hungernden Inders, dessen Land durch die Serienproduktion von Traktoren gerettet werden könnte, Technik als solche zu beargwöhnen.“ Anders, *Die Antiquiertheit des Menschen*. 2, 141.

¹⁴² Ramge, „Mensch fragt, Maschine antwortet“.

Für eine Annäherung an die Fragestellung, ob Künstliche Intelligenz individuelle Autonomie zu untergraben vermag, ist es hilfreich, zunächst auf die Unterschiede zwischen menschlichem und maschinelltem Wissen und Entscheiden einzugehen.

Selbstverständlich wäre ein Ja oder Nein auf die Frage eine unzulässige Vereinfachung des dahinterstehenden Problems. Auch weil die Systeme in mancherlei Hinsicht die Fähigkeiten der Menschen bei weitem übertreffen, erscheint es sinnvoll, Kipppunkte ausfindig zu machen, ab denen wir Autonomie einbüßen und Selbstbestimmung in Fremdbestimmung umschlagen kann.

3.1 Menschliches und maschinelles Wissen

3.1.1 Menschliches Wissen und Wahrheit

Information wird nicht selten mit Wissen gleichgestellt und eine gewisse Menge an Informationen gerne mit Wissen verwechselt. Jedoch:

„Information ist kumulativ und additiv, während die Wahrheit exklusiv und selektiv ist. Im Gegensatz zur Information bildet sie keinen Haufen. Man begegnet ihr nämlich nicht häufig. Es gibt keine Wahrheitsmasse, Informationsmasse hingegen gibt es. [...] Das Wissen liegt nicht einfach vor. Man kann es nicht einfach vorfinden wie Information. Ihm geht nicht selten eine lange Erfahrung voraus. Es besitzt eine ganz andere Zeitlichkeit als die Information, die sehr kurz und kurzfristig ist. Die Information ist explizit, während das Wissen oft eine implizite Form annimmt.“¹⁴³

Wissen entsteht, wenn Information durch Erfahrung, Reflexion, Interpretation und Verstehen integriert und verarbeitet wird. Es ist das Ergebnis des Verstehens und der Verknüpfung von Informationen in einem sinnvollen Kontext. Wissen ist subjektiv und mit persönlicher Erfahrung, Verständnis und Interpretation verbunden. Zwei Menschen können dieselbe Information erhalten, aber daraus unterschiedliches Wissen gewinnen, abhängig von ihrem Hintergrund und ihren Erfahrungen. Wissen bedeutet, dass Informationen verstanden, interpretiert und in einem größeren Zusammenhang gesehen werden. Es beinhaltet oft das Erkennen von Zusammenhängen. Ein Wissenschaftler weiß, warum die Temperaturänderungen im Laufe des Tages auftreten und wie sie mit dem Klima zusammenhängen. Dieses Verständnis geht über die bloße Information "Die

¹⁴³ Han, *Im Schwarm*, 56.

Temperatur beträgt morgens 20 Grad und mittags 30 Grad" hinaus. Eine Lehrerin, die über Jahre hinweg Erfahrung im Unterrichten gesammelt hat, weiß, wie sie Informationen unterschiedlich verschiedenen Schülergruppen vermitteln kann. Das Wissen ist das Ergebnis von Erfahrung und Anwendung.

Wie wir zu Wissen gelangen und was Wahrheit ist, sind zentrale philosophische Fragen. Im Rahmen dieser Arbeit sind lediglich kurze Skizzen wichtiger Richtungen der neueren Zeit möglich. René Descartes stellte zu Beginn des 17. Jahrhunderts alle damals geltenden Gewissheiten in Frage. Mit seiner Methode des radikalen Zweifelns suchte er nach einem sicheren Fundament für neue Erkenntnisse. Als unbestreitbare Grundlage für sicheres Wissen formuliert er den berühmten Satz „Cogito, ergo sum“. Für Descartes war damit klar, dass Wissen nur aus klarem und logischem Denken resultiert und nicht auf Wahrnehmung oder Gefühlen basiert.¹⁴⁴

John Locke (1632–1704) gilt zusammen mit David Hume als einer der Hauptvertreter des Empirismus, einer Gegenposition zum Rationalismus von Descartes. Alle Erkenntnis und alles Wissen beruht allein auf Erfahrung. Wissen ist aus Erfahrung abgeleitet. Nicht der Verstand und nicht die Vernunft sind Ursprung des Wissens. Als Grundlage des Empirismus kann man die Aussage „nihil est in intellectu quid non fuerit in sensu“ verstehen: „Nichts ist im Verstand, das nicht vorher durch die Sinne erfasst worden wäre“. ¹⁴⁵

Bei Kant (1724-1804) besteht die Hauptidee darin, dass Wissen durch die Verbindung von Erfahrung und Verstand entsteht. In der „Kritik der reinen Vernunft“ entwickelt Kant die Idee, dass Wissen sowohl aus sinnlicher Anschauung (Erfahrung) als auch aus kategorialem Denken (a priori, durch den Verstand) entsteht. Kant unterscheidet zwischen analytischen Urteilen (logisch notwendig) und synthetischen Urteilen (neues Wissen, das die Erfahrung einbezieht).

Wer nach Wissen strebt, will wahres Wissen! Damit stellt sich die Frage: Woher weiß ich, dass mein Wissen wahr ist? Eine auch im Alltagsverständnis verbreitete Auffassung von Wahrheit findet sich in der *Korrespondenztheorie* wieder. Ihr zufolge ist ein Gedanke oder eine Aussage dann wahr, wenn eine Übereinstimmung mit der Realität oder, wie Kant es ausdrückt, mit dem Gegenstand besteht. „Wahrheit ist *convenientia cognitionis*

¹⁴⁴ Liessmann u. a., *Vom Denken*, 95.

¹⁴⁵ Metzlers Lexikon der Philosophie, *Empirismus*.

cum objecto“, also Übereinstimmung der Erkenntnis, des Wissens mit dem Gegenstand. Vor Kant hatte schon Thomas von Aquin die Angleichung von Verstand und Sache als Wahrheit definiert. Beide Definitionen gehen davon aus, dass das menschliche Erkenntnisvermögen auf dem Streben nach einer solchen Übereinstimmung bzw. Angleichung beruht. Jedoch: Wer eine solche Übereinstimmung feststellen will, muss die Wahrheit immer schon kennen. Diese Definition setzt Wahrheit, also das, was definiert werden soll, schon voraus.¹⁴⁶

Aufgrund dieser Schwäche der Korrespondenztheorie wurden andere Theorien entwickelt, wie etwa die *Kohärenztheorie*, die von der Übereinstimmungslogik vollends abrückt und lediglich eine Widerspruchsfreiheit von Aussagen untereinander fordert – etwa im Rahmen wissenschaftlicher Theorien, aber auch bei der Frage nach Wahrheit oder Lüge, wenn sich jemand in widersprüchlichen Aussagen verstrickt. Die sogenannte *Konsenstheorie* geht wesentlich auf Jürgen Habermas zurück. Sie geht davon aus, dass Menschen, die über eine Sache verhandeln und dabei zu einem Konsens kommen, der Wahrheit nahekommen. Das setzt unter anderem voraus, dass diese über bestimmte Fähigkeiten verfügen. Beispielsweise über eine entsprechende Vernunft und über Kommunikationsfähigkeit. Zudem muss eine solche Verhandlung ein „herrschaftsfreier Diskurs“ sein, also frei von einseitiger Beeinflussung oder Manipulation.¹⁴⁷

Das „Wissen“ der Künstlichen Intelligenz ist im Vergleich zum menschlichen Wissen ein sehr vereinfachtes, rudimentäres Wissen. Die Semantik einer Aussage wird oft erst im Kontext ihrer Anwendung verständlich. Wir können beispielsweise die Aussage „Oh, wie toll!“ als Kompliment oder als Ironie verstehen – abhängig von der konkreten Situation. Künstliche Intelligenz ist dazu nicht ohne weiteres in der Lage.¹⁴⁸ Unser menschliches Wissen ist nicht nur rationaler, theoretischer Natur. Es ist ein (implizites) Wissen, das wir im Laufe des Lebens internalisiert haben. Wir können einen alten Menschen von einem jungen unterscheiden, wissen, wie man ein Ei in die Pfanne schlägt oder den Inhalt einer Flasche in ein Glas bekommt. Dazu bedarf es keines Beweises und keiner Theorie. Und wir wissen auch wie sich Trauer, Freude, Liebe oder Ordnung anfühlen.¹⁴⁹

¹⁴⁶ Liessmann u. a., *Vom Denken*, 105.

¹⁴⁷ Liessmann u. a., *Vom Denken*, 106.

¹⁴⁸ Otte, *Künstliche Intelligenz für Dummies*, 93.

¹⁴⁹ Spiekermann, *Digitale Ethik*, 219 Spiekermann bezieht sich auf Gadamer (1986): Plato. Texte zur Ideenlehre.

Künstliche Intelligenz verfügt nicht über ein Wissen dieser Art. Gleichwohl können wir ihr beibringen, wie ein alter Mensch aussieht, was ein Zebrasteifen ist, wie ein Elefant aussieht oder eine Ameise. KI wird trainiert, damit sie Muster von Dingen und Lebewesen der wirklichen Welt wiedererkennt.

3.1.2 Daten und Information

Die Informatik unterscheidet zwischen Daten, Information und Wissen. Sie sind Basis und Ergebnis von KI-Entscheidungen. Wenn sich zwei schwarze Punkte auf einem ansonsten weißem Blatt Papier befinden, dann handelt es sich um Gebilde (Objekte), die sich vom Hintergrund und der Umgebung unterscheiden. Die Informatik nennt solche abstrakten Gebilde aus Zeichen *Daten*. Sie repräsentieren Informationen und dienen in erster Linie ihrer Verarbeitung.¹⁵⁰ Daten an sich haben keinerlei Bedeutung, Informationen schon. Ein Beispiel:¹⁵¹ Wer eine Wanderung unternimmt, kann sich an Wegweisern orientieren, an einem gelben Pfeil etwa, der sich auf einem Holzbrett befindet. Es handelt sich zweifellos um ein Datum, denn der Pfeil hebt sich vom Hintergrund und seiner Umgebung ab. Ob es sich um eine Information handelt, hängt nun davon ab, in welchem Kontext dieser Pfeil steht und ob er relevant für die jeweiligen Wanderer ist. Wer den Jakobsweg pilgert und nach einer gelben Jakobsmuschel auf blauem Grund Ausschau hält, kann mit dem Pfeil nicht viel anfangen, denn dieser hat keine Relevanz für das Vorhaben. Für andere Wander*innen mag er sehr wohl den richtigen Weg weisen. Der entscheidende Unterschied zwischen Daten und Informationen liegt in der Bedeutung der Zeichen für die Empfänger*innen. Die Jakobsmuschel mindert die Unsicherheit von Jakobspilgern, die etwa an einer Weggabelung entsteht. In der Informatik kann der Begriff der Information als „*Maß für die Zunahme der Wahrscheinlichkeit einer richtigen Voraussage eines Ereignisses*“¹⁵² erklärt werden.

Information basiert auf Daten, die strukturiert, verarbeitet oder interpretiert wurden. Menschen interpretieren die Jakobsmuschel als Wegweiser nach *Santiago de Compostela*. Informationen sind in der Regel neutral, ohne Interpretation oder subjektive Bewertung. Sie werden erst durch den Kontext und die Interpretation bedeutsam. Eine Zahl oder ein Satz allein kann Information sein, aber sie ergibt erst Sinn, wenn sie in einen

¹⁵⁰ Otte, *Künstliche Intelligenz für Dummies*, 23.

¹⁵¹ Angelehnt an: Otte, *Künstliche Intelligenz für Dummies*, 81.

¹⁵² Otte, *Künstliche Intelligenz für Dummies*, 81.

Zusammenhang gestellt wird. Informationen sind das Rohmaterial, aus dem Wissen gewonnen werden kann.

Daten und Informationen sind die Basis der Künstlichen Intelligenz. Maschinelles Lernen wird erst durch die Aufnahme und Verarbeitung großer Datenmengen möglich. Die Systeme sind datengetrieben. Daten sind die Basis, das existentielle Futter für die Funktionsfähigkeit von KI-Systemen wie sie in völlig unterschiedlichen Kontexten zur Anwendung kommen. Wo kommen die Daten her, und was machen die Systeme und Systembetreiber mit ihnen? OpenAI gibt Auskunft:

„Die Basismodelle von OpenAI, einschließlich der Modelle, die ChatGPT zugrunde liegen, werden unter Verwendung von drei primären Informationsquellen entwickelt: (1) Informationen, die öffentlich im Internet verfügbar sind, (2) Informationen, auf die wir in Partnerschaft mit Dritten zugreifen, und (3) Informationen, die unsere Benutzer, menschlichen Trainer und Forscher bereitstellen oder generieren.“¹⁵³

Für die sogenannten „Big Five“ der Tech-Industrie – Google, Amazon, Apple, Meta und Microsoft – stellen die Daten den Kern ihres Geschäftsmodells dar. Die scheinbar kostenfreie Nutzung digitaler Dienste im Internet wird in Wahrheit mit den persönlichen Daten der Nutzer*innen bezahlt. Sie werden zu unbezahlten Digitalarbeiter*innen, der dem anonymen Datenapparat Rohmaterial liefert, aus dem präzise Verhaltensprognosen extrahiert werden. Wissen über Nutzer wird zur Ware, die an Geschäftskunden, Geheimdienste oder Regierungen verkauft wird. In dieser ökonomischen Konstellation ist jeder Konsument Teil eines globalen Systems unentgeltlicher Datenproduktion.

Die Arbeit des „Fütterns der Maschinen“ übernehmen die Menschen also selbst – doch zunehmend wird auch diese Aufgabe industriell organisiert. Plattformen wie das US-Unternehmen *Clickworker*¹⁵⁴ beschäftigen nach eigenen Angaben Millionen von Digitalarbeitern weltweit, die nach dem Prinzip des Crowdsourcing Daten für KI-Systeme generieren und kuratieren. Diese Menschen, oft in prekären Lebenssituationen, absolvieren Mikroaufgaben von Sekunden- oder Minutendauer – für Löhne, die kaum

¹⁵³ OpenAI, „How ChatGPT and our foundation models are developed“.

¹⁵⁴ <https://www.clickworker.com/>

Das Unternehmen betreibt ein Headquarter in den USA und eines in Europa.

zum Überleben reichen. Ihre Qualifikation erwerben sie häufig in unbezahltem Selbststudium anhand detaillierter Guidelines.

OpenAI nutzt Informationen, die aus Kooperationen mit externen Partnern stammen. Hierzu zählt unter anderem eine Zusammenarbeit mit der finnischen Regierung, die darauf abzielt, die isländische Sprachkompetenz von ChatGPT zu optimieren. Zur Digitalisierung historischer Werke wurden Vereinbarungen mit Bibliotheken und Universitäten geschlossen. Außerdem bestehen Lizenzverträge mit Medienhäusern wie der Financial Times, News Corp., Vox Media, The Atlantic oder Axel Springer, um Inhalte für ChatGPT zu nutzen. Auch soziale Plattformen wie Reddit stellen Daten über Schnittstellen bereit. Zusätzlich gibt es Integrationen mit Unternehmensdiensten wie Google Drive, OneDrive oder Dropbox, um Geschäftsdaten in ChatGPT einzubinden. So entsteht ein breites Netz an Quellen, das OpenAIs Modelle vielfältig speist.¹⁵⁵

Doch es gibt auch eine subtilere, gleichwohl folgenreiche Form der Datenproduktion: Byung-Chul Han spricht von der „*Totalprotokollierung des Lebens*“¹⁵⁶ und einem „*digitalen Panoptikum*“¹⁵⁷, das sich von klassischen, repressiven Überwachungsapparaten unterscheidet.

*„Die Bewohner des digitalen Panoptikums sind keine Gefangenen. Sie leben in der Illusion der Freiheit. Sie speisen das digitale Panoptikum mit Informationen, indem sie sich freiwillig ausstellen und ausleuchten. Die Selbstausschleuchtung ist effizienter als die Fremdausschleuchtung. Darin besteht eine Parallele zur Selbstausschöpfung. Die Selbstausschöpfung ist effizienter als die Fremdausschöpfung, weil sie mit dem Gefühl der Freiheit einhergeht.“*¹⁵⁸

Hier vollzieht sich Überwachung nicht durch Zwang, sondern durch freiwillige Preisgabe persönlicher Informationen. In sozialen Netzwerken – aber nicht nur dort – legen wir unsere intimsten Daten offen, getragen von einem Gefühl der Freiheit, während wir uns

¹⁵⁵ OpenAI, „How ChatGPT and our foundation models are developed“ Die Informationen zu Partnerschaften ändern sich häufig und sind verteilt zu finden. Eine relativ ausführliche Antwort bekommt man, wenn man auf der Website den Chatbot nach Partnerschaften fragt.

¹⁵⁶ Han, *Im Schwarm*, 91.

¹⁵⁷ Han, *Im Schwarm*, 91.

¹⁵⁸ Han, *Im Schwarm*, 92.

Der Begriff des Panoptikums geht auf Jeremy Bentham. Die Idee: Ein architektonisches Konzept zur Beaufsichtigung und Überwachung von Gefängnisinsassen oder Fabrikarbeitern. Die von den Insassen ins Kalkül zu ziehende Allgegenwärtigkeit der Überwacher widerspiegelt sich in der Omnipräsenz digitaler Techniken in unseren alltäglichen Lebenswelten.

in Wahrheit in einem Raum bewegen, der unsere Existenz bis in feinste Verästelungen registriert, analysiert und ökonomisch verwertet.

3.1.3 Maschinelle Wissensproduktion: „Probably Approximately Correct“

Künstliche Intelligenz ist zu einer mächtigen Instanz der Wissensproduktion geworden. Ihren Algorithmen hängt der Mythos der Objektivität und Neutralität, der Leidenschafts- und Gefühllosigkeit an. Viele Menschen neigen dazu, den Ergebnissen algorithmischer Entscheidungen mehr zu trauen als menschlichen Entscheidungsträgern. Ein Phänomen, das als *Automation Bias* bekannt ist. Die Art und Weise, wie Wissen durch Künstliche Intelligenz produziert, generiert und als solches anerkannt wird, geht weit über die Frage der technischen Realisierung hinaus, wie sie in Kapitel “2.2.3.1.2 Subsymbolische KI – Maschinelles Lernen“ beschrieben wurde. Traditionell entsteht Wissen aus der reflexiven Auseinandersetzung mit einem Gegenstand, sei es durch wissenschaftliche Methodik, diskursive Logik oder hermeneutisches Verstehen. Der Akt der Erkenntnis ist dabei an ein erkenntnisfähiges Subjekt gebunden, das die Welt deutet, interpretiert und – idealerweise – kritisch hinterfragt. KI-gestützte Systeme hingegen operieren primär durch Aggregation und Korrelation großer Datenmengen, aus denen Muster, Tendenzen und Wahrscheinlichkeiten errechnet werden. Der britische Informatiker Leslie Valiant beschreibt seine PAC-Methode („*Probably Approximately Correct*“) so:

„a method for finding patterns in data that are usefully predictive of future events but which do not necessarily provide an explanatory theory“

Maschinelles Lernen heißt aus Erfahrung lernen, sich durch Erfahrung verbessern und induktiv oder abduktiv¹⁵⁹ schlussfolgern. Ziel dabei ist, zu optimieren, zu approximieren und zu verallgemeinern. Aus konkreten Beispielen sollen allgemeine Regeln, Klassifikationen und Vorhersagen abgeleitet werden.¹⁶⁰

Das Verfahren ist durch eine grundlegende Verschiebung gekennzeichnet: Statt einem klassisch rational-kognitiven Vorgehen wird ein Bottom-up-Ansatz verfolgt. Maschinelles Lernen im Sinne Valiants basiert nicht auf theoriegeleiteten Erklärungen, sondern auf der Erkennung von Mustern, um künftiges Verhalten vorherzusagen. Auch genetische Algorithmen arbeiten nicht nach einem Top-down-Prinzip: Sie analysieren

¹⁵⁹ Zu Begriff und Methode der Abduktion siehe *Wikipedia*, „Schluss auf die beste Erklärung“.

¹⁶⁰ Weber, „Probably Approximately Correct?“, 73.

nicht das Problem vollständig, um daraus eine Lösung abzuleiten. Vielmehr wird ein Suchraum definiert, es werden Randbedingungen festgelegt, und die Lösung wird in einem adaptiven, approximativen Trial-And-Error-Verfahren schrittweise ermittelt – ohne dass das zugrunde liegende Problem vollständig verstanden sein muss. Kausalität wird abgelöst durch Approximation, Vorhersage und Korrelation.¹⁶¹ So zeichnet sich eine Transformation epistemischer Produktionsweisen ab, bei der es nicht mehr um begründetes Wissen geht.

Wissensproduktion war immer auch Sinnproduktion – ein Prozess, der Bedeutung herstellt, Zusammenhänge erkennt und Kontexte reflektiert. An die Stelle der Sinnbildung tritt bei der Künstlicher Intelligenz die Musterextraktion: Signifikanz wird nicht mehr durch Bedeutung erzeugt, sondern durch Wiedererkennbarkeit, Korrelation oder Vorhersagekraft. Der „Erfolg“ eines Wissensprodukts bemisst sich an seiner Plausibilität und seiner vermutlichen Richtigkeit, nicht daran, ob es verstehbar, kohärent oder begründet ist.

KI-gestützte Systeme bedienen sich rekursiver Modelle: Sie generieren neues Wissen auf der Basis von bereits generierten Outputs. Das lässt sich sehr einfach bei Chatbots oder Suchmaschinen nachvollziehen, die sich oft auf bereits ausgegebene Antworten beziehen, die aufgegriffen und in einem passenden Kontext erneut verarbeitet werden. Diese Feedbackschleifen führen zu einem epistemischen Autotuning, bei dem sich Modelle immer stärker auf sich selbst beziehen und damit – epistemologisch betrachtet – zu einer hermetischen Selbstreferenzialität führen. Erkenntnis wird so zu einem geschlossenen Kreislauf algorithmischer Iteration.

3.2 Menschliches Urteilen und algorithmisches Entscheiden Künstlicher Intelligenz

3.2.1 Menschliche Urteilskraft

In seiner Kritik der Urteilskraft unterscheidet Kant zwischen bestimmender und reflektierender Urteilskraft. Die **bestimmende Urteilskraft** wendet einen bereits bekannten Begriff (eine Regel, ein Gesetz) auf einen konkreten Fall an. Bei der

¹⁶¹ Weber, „Probably Approximately Correct?“, 77.

reflektierenden Urteilstkraft ist es umgekehrt. Sie sucht nach einer Kategorie, einer Regel oder einem Begriff für ein neues, unbekanntes Phänomen.

„Die Urtheilskraft kann entweder als bloßes Vermögen, über eine gegebene Vorstellung, zum Behuf eines dadurch möglichen Begrifs, nach einem gewissen Princip zu reflectiren, oder als ein Vermögen, einen zum Grunde liegenden Begriff durch eine gegebene empirische Vorstellung zu bestimmen, angesehen werden. Im ersten Falle ist sie die reflectirende, im zweyten die bestimmende Urtheilskraft. Reflectiren (Überlegen) aber ist: gegebene Vorstellungen entweder mit andern, oder mit seinem Erkenntnißvermögen, in Beziehung auf einen dadurch möglichen Begriff, zu vergleichen und zusammen zu halten. Die reflectirende Urteilstkraft ist diejenige, welche man auch das Beurtheilungsvermögen (facultas dijudicandi) nennt.“¹⁶²

Bestimmende Urteilstkraft

Die *bestimmende Urteilstkraft* subsumiert etwas Gegebenes unter einen bereits vorhandenen Begriff. Beispiel: Wenn wir ein Tier mit Federn, Schnabel, Flügeln sehen, ordnen wir es unter den vorhandenen Begriff Vogel ein. Wir bestimmen nur den Fall. Bestimmende Urteilstkraft ist regelbasiert, funktioniert top-down. Wenn eine Richterin in einer Handlung einen Vertragsbruch erkennt, weil alle rechtlichen Kriterien erfüllt sind, wendet sie ein allgemeines Gesetz auf einen Einzelfall an. Der Mechatroniker, der ein Problem mit dem Motor an einem bestimmten Geräusch erkennt, wendet ein bekanntes Muster auf den Einzelfall an. Bestimmende Urteilstkraft kann nicht allein durch Lehre erworben werden, sondern braucht Übung und Erfahrung

„Der natürliche Verstand kann nun noch durch Belehrung mit vielen Begriffen bereichert und mit Regeln ausgestattet werden; aber das zweite intellectuelle Vermögen, nämlich das der Unterscheidung, ob etwas ein Fall der Regel sei oder nicht, die Urtheilskraft (iudicium), kann nicht belehrt, sondern nur geübt werden; daher ihr Wachsthum Reife und derjenige Verstand heißt, der nicht vor Jahren kommt. Es ist auch leicht einzusehen, daß dies nicht anders sein könne; denn Belehrung geschieht durch Mittheilung der Regeln. Sollte es also Lehren für die Urtheilskraft geben, so müßte es allgemeine Regeln geben, nach welchen man unterscheiden könnte, ob etwas der Fall der

¹⁶² Kant, *Kants gesammelte Schriften.*, AA XX Erste Einleitung zur Kritik der Urteilstkraft, 211.

Regel sei oder nicht: welches eine Rückfrage ins Unendliche abgiebt. ¹⁶³

Kants Urteilskraft ist besonders dann gefordert, wenn es darum geht, allgemeine Regeln auf spezifische Einzelfälle anzuwenden. Dies erfordert sowohl Wissen als auch moralisches Urteilsvermögen. KI-Systeme funktionieren in der Regel, indem sie auf statistische Muster oder Algorithmen zurückgreifen. Diese Systeme können zwar allgemeine Regeln anwenden, sind aber oft nicht in der Lage, die feinen Nuancen oder ethischen Implikationen eines Einzelfalls zu erfassen. Kants Urteilskraft hingegen ist darauf angewiesen, die Besonderheiten einer Situation zu berücksichtigen und die Regel im Licht dieser Umstände zu interpretieren.

Reflektierende Urteilskraft

Bei der *reflektierenden Urteilskraft* ist das umgekehrt. Sie sucht nach einer Kategorie, einer Regel oder einem Begriff für ein neues, unbekanntes Phänomen. Reflektieren ist kein bloßes Nachdenken, sondern ein aktives Suchen nach Bedeutung oder Regelmäßigkeit. Man vergleicht verschiedene Phänomene miteinander oder mit dem eigenen Erkenntnisvermögen, um zu einem sinnvollen Begriff zu gelangen. Reflektierende Urteilskraft sucht nicht nur, sie erfindet auch. Diese Kraft des Reflektierens ist schöpferisch, orientierungssuchend, offen. Sie ist insbesondere dort gefragt, wo der Mensch sich mit Sinnfragen, ästhetischen Urteilen oder komplexen, neuartigen Situationen konfrontiert sieht. In der naturwissenschaftlichen Forschung werden Phänomene entdeckt, die sich theoretisch noch nicht erklären lassen. Aus der Beobachtung heraus werden neue Regeln oder Begriffe entwickelt. Die reflektierende Urteilskraft hilft, neue Hypothesen zu entwickeln. Ein weiteres Beispiel: In moralischen Grenzsituationen gibt es keine klaren Regeln, die Anwendung finden können. Wir überlegen dann, welche moralischen Prinzipien passen könnten und versuchen, durch Reflexion auf Werte und Prinzipien ein Urteil zu bilden – oft auch durch intuitive moralische Urteile. Reflektierende Urteilskraft ist mehr als eine regelgeleitete Auswahl – sie ist eine schöpferische Kraft, die über das hinausgeht, was Algorithmen leisten können. Reflektierende Urteilskraft ist ein menschliches Vermögen, das Bedeutung zu erzeugen vermag. Die reflektierende Urteilskraft ist heuristisch, offen für Neues, kreativ. Sie ist besonders in ästhetischen Urteilen oder teleologischen Betrachtungen der Natur tätig.

¹⁶³ Kant, *Kants gesammelte Schriften.*, AA VII, Anthropologie in pragmatischer Hinsicht, 199.

Künstliche Intelligenz basiert auf mathematischen Modellen, die Wahrscheinlichkeiten berechnen, Muster erkennen und Optimierungskriterien anwenden. Diese Prozesse führen zu einer technisch-rationalen Auswahl – einer Selektion unter Optionen basierend auf Daten und Algorithmen. Was dabei fehlt, ist ein normatives Urteil: ein Abwägen von Werten, ein Einfühlen, ein moralisches oder sinnorientiertes Entscheiden.

Kant beschreibt die Urteilskraft als eine Fähigkeit, die nicht rein mechanisch ist, sondern auf einer reflektierenden Betrachtung der Umstände beruht. KI-Systeme basieren auf Mustern und Algorithmen. Urteilsfähigkeit braucht Reflexion und Erfahrung.

Moralität

Bereits Aristoteles hat eine Unterscheidung zwischen bloßem Handeln und moralischer Entscheidung getroffen. Entscheidung ist nicht bloß Wunsch (*boulēsis*, *βούλεσις*: altgriechisch für Wunsch), sondern: überlegte Wahl des Mittels zu einem als gut erkannten Ziel.

„Wir überlegen aber nicht nur die Ziele, sondern auch das, was zu den Zielen hinführt [...] vielmehr setzt man das Ziel fest und überlegt dann, wie und mit welchen Mitteln es zu erreichen ist.“¹⁶⁴

Und weiter:

„Augenscheinlich ist also, wie gesagt, der Mensch Ausgangspunkt der Handlungen. Die Überlegung bezieht sich auf das, was er selbst tun kann. Die Handlungen aber geschehen um anderer Dinge willen. Gegenstand des Überlegens ist nicht das Ziel, vielmehr sind es die Dinge, die zum Ziel führen.“¹⁶⁵

Der zentrale Begriff dabei ist die *prohaíresis* (κληρονομία: altgriechisch für Wahl, Entscheidung). Aristoteles beschreibt den Vorgang der Entscheidung als überlegte, absichtsvolle Wahl eines bestimmten Mittels im Hinblick auf ein gutes Ziel – als einen eng mit dem Handeln verknüpften, denn als Handlung könne nur derjenige Akt bezeichnet werden, der durch eine ihm vorausgehende Entscheidung erst veranlasst wird.

¹⁶⁴ Aristoteles, *Nikomachische Ethik*, NE 1139a31-b5.

¹⁶⁵ Aristoteles, *Nikomachische Ethik*, NE 1139a31-b5.

Das Handeln verfolgt ein Ziel:

„Wenn es nun ein Ziel des Handelns gibt, das wir seiner selbst wegen wollen, und das andere nur um seinetwillen, und wenn wir nicht alles wegen eines anderen uns zum Zwecke setzen – denn da ginge die Sache ins Unendliche fort, und das menschliche Begehren wäre leer und eitel –, so muß ein solches Ziel offenbar das Gute und das Beste sein.“ (NE 1094 a 19)

Das Handeln wird durch Klugheit bestimmt. Klugheit (phronesis, altgriechisch: σοφία) ist bei Aristoteles eine Tugend des Verstandes und verantwortlich für Handeln (praxis, πρακτική; altgriechisch), die für Moralität relevante Grundtugend. Moralisch – praktisches Urteilsvermögen. Vermögen, richtiges Handeln bestimmen zu können, Orientierungsvermögen. Diese Form der moralischen, tugendhaften Entscheidung unterscheidet sich grundlegend vom bloßen Wunsch oder von automatischem Verhalten. Nur wer über *prohairesis* und Klugheit (*phronesis*) verfügt, kann als sittlich Handelnder gelten. Denn nur dann liegt eine ethisch relevante, verantwortbare Handlung vor.

Aristoteles' Ethik erinnert daran, dass moralisches Entscheiden mehr ist als Informationsverarbeitung. Es setzt ein handelndes Subjekt voraus, das sich seiner Ziele, Mittel und Konsequenzen bewusst ist – und das Verantwortung trägt.

Kant

Kant unterscheidet zwischen theoretischer Urteilskraft (im Bereich des Wissens) und praktischer Urteilskraft (im Bereich des Handelns, besonders der Moral). Wie im Kapitel über den Autonomiebegriff bei Kant dargelegt, ist Autonomie für Kant die Fähigkeit, nach selbst auferlegten Gesetzen und nach moralischen Prinzipien zu handeln – ein zentraler Aspekt der menschlichen Vernunft. Die Urteilskraft ermöglicht es, moralische Entscheidungen zu treffen, die nicht von äußeren Autoritäten oder Regeln, sondern von der eigenen Vernunft geleitet werden. Wenn Entscheidungen an KI delegiert werden, kann dies die menschliche Urteilskraft beeinträchtigen. Kant würde argumentieren, dass die vollständige Auslagerung von Entscheidungen an eine Maschine die Autonomie des Individuums einschränkt.

Kants Idee der moralischen Selbstgesetzgebung fordert, dass der Mensch nach dem Prinzip des kategorischen Imperativs handelt, also nach einer Maxime, die er für allgemeingültig hält. Wenn Künstliche Intelligenz Entscheidungen trifft, wird der

Mensch zu einem bloßen „Empfänger“ dieser Entscheidungen, was bedeutet, dass er sich nicht oder nicht in ausreichendem Maß aktiv an der moralischen Reflexion beteiligt.

3.2.2 Algorithmische Entscheidungen

KI-Systeme verarbeiten Daten und Informationen auf der Basis von algorithmischen Entscheidungsprozessen. Es gibt verschiedene Ansätze zur Kategorisierung algorithmischer Entscheidungsprozesse mit jeweils unterschiedlicher Schwerpunktlegung.¹⁶⁶ Die folgende Darstellung orientiert sich an der von Diakopoulos, der nach Grundtypen algorithmischer Entscheidungsprozesse („different types of atomic decisions“) differenziert: Priorisierung, Klassifikationen, Assoziationen, Filtern.¹⁶⁷ Bei den hier zusammengefassten Darlegungen von Diakopoulos wird auch deutlich: KI-Systeme sind keine bloßen Werkzeuge und ihre Algorithmischen Entscheidungen sind keine neutralen Operationen, sondern zunehmend autonome Akteure in sozialen, politischen und epistemischen Prozessen. Sie beeinflussen, was wir wissen, was wir sehen, wem wir glauben und wen wir als zugehörig oder fremd begreifen. Philosophie muss diese Entwicklungen kritisch begleiten – um zu verhindern, dass technische Systeme stille Architekten einer Gesellschaft werden, in der Verantwortung, Autonomie und Gerechtigkeit auf der Strecke bleiben. Sie beinhalten implizite normative Setzungen, die epistemisch wie ethisch höchst bedeutsam sind. Jede Parametrisierung, jedes Training, jede Justierung von Schwellenwerten ist auch eine Entscheidung über gesellschaftliche Werte: über Fairness, Datenschutz, Gerechtigkeit und über die Verteilung von Chancen und Risiken.

3.2.2.1 Priorisierung

Algorithmen strukturieren Informationen und bestimmen, welche Inhalte hervorgehoben und welchen durch ihren Rang weniger Bedeutung zukommen soll. Jede Form der algorithmischen Priorisierung ist dabei mehr als eine bloße technische Sortierung – sie lenkt Aufmerksamkeit selektiv und bewirkt so eine systematische Marginalisierung anderer Informationen, sei es intendiert oder nicht. In dieser Selektion liegt ein inhärenter Akt epistemischer und sozialer Diskriminierung: Was algorithmisch „oben“ erscheint, beeinflusst Wahrnehmung, Interpretation und letztlich auch Handeln. Suchmaschinenrankings, personalisierte Nachrichtenfeeds oder algorithmische Risikobewertungen in sicherheits- und migrationspolitischen Kontexten sind Beispiele

¹⁶⁶ So zum Beispiel der von Zweig u. a., „Algorithmen und Meinungsbildung“.

¹⁶⁷ Diakopoulos, „Accountability in Algorithmic Decision Making“, 57.

für die normativen Implikationen dieser Praxis. Sie machen deutlich, dass algorithmische Ordnung nicht neutral ist, sondern implizit Wertungen und Interessen reproduziert. Daraus ergibt sich die zentrale Frage, nach welchen Kriterien solche hierarchischen Strukturen erzeugt werden – und wessen Perspektiven, Präferenzen oder Machtverhältnisse in ihnen codiert sind.¹⁶⁸

3.2.2.2 Klassifikationen

Klassifikationsalgorithmen ordnen Daten, Objekte oder Personen Kategorien zu – etwa in Systemen der Kreditvergabe, der Strafverfolgung oder der gezielten Werbung im Rahmen datengesteuerten Marketings. Diese Zuschreibungen wirken performativ: Sie erzeugen soziale Realitäten, indem sie Zugehörigkeit, Risiko oder Wert festlegen. Doch die Basis dieser Entscheidungen, insbesondere in Machine-Learning-Verfahren, sind Trainingsdaten, die häufig implizite Biases in Form von Vorurteilen. Jede Entscheidung, die etwa die Schwelle zwischen falsch-positiv und falsch-negativ verschiebt, ist ein normativer Akt, der Interessen und Werte reflektiert.

3.2.2.3 Assoziationen

Assoziative Algorithmen knüpfen Beziehungen zwischen Entitäten, zum Beispiel zwischen Suchbegriffen und vorgeschlagenen Inhalten. Solche Verknüpfungen sind nicht neutral: Sie erzeugen Sinnzusammenhänge und beeinflussen Deutungsmuster. Der Glaube an „Big Data“ bringt oft eine naive Überhöhung von Korrelationen mit sich. Doch Korrelation ist keine Kausalität – eine Unterscheidung, die gerade philosophisch essenziell bleibt, auch wenn sie im digitalen Diskurs oft verschwimmt. Beispielhaft zeigt sich das in Fällen, in denen Suchmaschinen Autovervollständigungen anbieten, die den Namen von Personen, nach denen gesucht wird, mit problematischen Begriffen oder anrüchigen Sachverhalten verknüpfen – eine Praxis, die soziale Reputationen nachhaltig beschädigen kann.

3.2.2.4 Filtern

Algorithmische Filtermechanismen entscheiden darüber, welche Inhalte sichtbar werden und welche verschwinden. Diese Filterung oszilliert zwischen Schutz vor schädlichen Inhalten und der Gefahr der Zensur. Plattformen wie Facebook oder Twitter nutzen algorithmische Filter, um Hassrede zu blocken, riskieren jedoch zugleich eine

¹⁶⁸ Diakopoulos, „Accountability in Algorithmic Decision Making“, 57.

Einschränkung legitimer Meinungsäußerung. In Kulturen mit stark verankerten Freiheitsrechten wird diese Spannung besonders brisant.

3.3 Von der Selbst- in die Fremdbestimmung?

Unter Fremdbestimmung soll hier ein Zustand verstanden werden, in dem das Handeln, Denken oder Entscheiden der Menschen nicht aus eigener Freiheit und Einsicht erfolgt, sondern durch äußere Einflüsse, Zwänge oder Autoritäten bestimmt wird. Ein Mensch ist fremdbestimmt, wenn die maßgeblichen Gründe für sein Handeln nicht seine eigenen sind, sondern von außen auferlegt. Moralisch handelt nur, wer sich selbst ein Gesetz gibt – und nicht bloß äußeren Regeln folgt. Fremdbestimmung lässt sich gleichzeitig als ein schleichender Prozess verstehen, in dem wir uns mehr und mehr der Kontrolle und Steuerung omnipräsenter Systeme der Künstlichen Intelligenz aussetzen?

KI-Systeme zeichnen sich durch kontingente Eigenschaften aus, die durch ihre konkrete Ausprägung positiv (im Sinne von autonomiefördernd oder -erhaltend) oder negativ (im Sinne von autonomieeinschränkend) auswirken können. In Punkt 3.3.1 Kontingente Eigenschaften von KI-Systemen (als Kippunkte) werden diese Eigenschaften beschrieben und in späteren Kapiteln kontextualisiert.

Entscheidungen sind untrennbar mit Verantwortung verbunden. Welche Bedeutung hat die Übertragung von Entscheidungen an Künstliche Intelligenz für die menschliche Verantwortung (Siehe Punkt 3.3.2 Verantwortung bei Entscheidungen durch KI). Wer trägt die Verantwortung, wenn Entscheidungen an opake und komplexe Systeme übertragen werden? Wer die Kontrolle verliert, kann nicht selbstbestimmt handeln. Wer fremdbestimmt handelt, ist nicht autonom und kann keine moralische Verantwortung im vollen Sinne übernehmen

Menschliches Wissen und selbstbestimmtes Entscheiden brauchen eine verlässliche epistemische Basis. Vor allem die weltweit agierenden, öffentlich zugänglichen Plattformen wie Suchmaschinen, LLMs (z.B. ChatGPT) oder Empfehlungs- und Assistenzsysteme sind in dieser Hinsicht fragwürdig und können epistemische Herausforderungen darstellen. (Siehe Punkt 3.3.3 Epistemische Herausforderungen.)

Wo der imaginäre Kippunkt zwischen Selbst- oder Fremdbestimmung zu verorten ist, impliziert die Frage: Wer beherrscht wen? Der Mensch die Maschine oder umgekehrt? Verschiebt sich Entscheidungsgewalt schleichend Richtung KI-Systemen? (Siehe Punkt 3.3.4 Die „smarte Macht“ der omnipräsenten Systeme.)

3.3.1 Kontingente Eigenschaften von KI-Systemen (als Kippunkte)

Die Frage, ob individuelle Autonomie durch Künstliche Intelligenz untergraben wird und dabei Selbst- in Fremdbestimmung umschlagen kann, ist nicht mit einem simplen Ja oder Nein zu beantworten.

„Es wird eine Hauptaufgabe der Philosophie der Technik sein, den dialektischen Punkt ausfindig zu machen und zu bestimmen, wo sich unser Ja der Technik gegenüber in Skepsis oder in ein unverblümtes Nein zu verwandeln hat.“¹⁶⁹

Die kontingenten Eigenschaften Künstlicher Intelligenz sind wie Parameter eines imaginären Kippunktes, an dem Selbstbestimmung in Fremdbestimmung umschlagen kann. Bezogen auf das jeweilige System Künstlicher Intelligenz tragen sie in ihrer konkreten Ausprägung wesentlich dazu bei, ob Künstliche Intelligenz eher zur Selbstbestimmung des Menschen beiträgt, oder ob ein Kippunkt erreicht wird, also eine Grenze überschritten wird, hinter der es Richtung Fremdbestimmung geht.

Menschen sind nur im graduellen Sinn autonom. Ihnen kann Autonomie in mehr oder weniger hohem Maße zugeschrieben oder abgesprochen werden. (Siehe Kapitel 2.1.3.1 Begriffliche Differenzierungen.) Künstliche Intelligenz hat das Potential, zu einem selbstbestimmten Leben beizutragen aber auch zur graduellen Entmündigung. Selbstverständlich nicht allein und nicht losgelöst vom sozialen und politischen Kontext der autonomen Person. Ein kritischer und selbstverantwortlicher Umgang des autonomen Menschen mit Künstlicher Intelligenz ist dabei ebenso von Bedeutung wie die Ausprägung kontingenter Eigenschaften.

„I'm sorry, Dave, I'm afraid I can't do that.“

Das Raumschiff *Discovery One* befindet sich auf dem Weg zum Jupiter. Ein Bordcomputer hat das Commando. Er hört auf den Namen HAL und spricht mit sanfter und einfühlsamer Stimme. Sein blinkendes, rotes „Auge“ sieht alles. HAL gilt als unfehlbar. Und tatsächlich verlässt sich die Crew auf HAL. Er steuert das Raumschiff durch das All, sorgt für das Öffnen und Schließen der Schleusen, kümmert sich um den Schlaf der Crew, bewerkstelligt Expeditionen außerhalb der Sonde. Und HAL ist in der Lage, sich selbst zu kontrollieren und zu korrigieren. Ein adaptives System. Aber eines Tages bemerkt die Bodenstation Ungereimtheiten. HAL gibt vor, keinen Fehler

¹⁶⁹ Anders, *Die Antiquiertheit des Menschen*. 2, 141.

feststellen zu können. Der Nimbus der Unfehlbarkeit bröckelt. Die beiden Astronauten Frank und Dave sind misstrauisch. Sie beraten sich in vermeintlicher Abgeschiedenheit. HAL sieht nicht nur alles, er kann auch Lippen lesen und weiß nun von ihrem Plan, ihn abzuschalten. Und er reagiert: Frank tötet er durch einen inszenierten Unfall. Dave soll sein zweites Opfer werden, HAL sperrt ihn aus und versagt ihm nach einer Außenoperation die Rückkehr ins Innere des Raumschiffs. HAL weigert sich, die Schleuse zu öffnen: „Es tut mir leid, Dave. Das kann ich nicht tun.“ Sein ihm vorgegebenes Ziel ist der Erfolg der Weltraummission. Aber am Ende gelingt es Dave, HAL abzuschalten, ihn zu demontieren.

Eine Sequenz aus dem Science-Fiction-Klassiker „2001: a space odyssey“¹⁷⁰. Sie steht zu Beginn dieses Kapitels nicht als dystopische Mahnung oder Prophezeiung. Vielmehr veranschaulicht der Film, dass Künstliche Intelligenz¹⁷¹ eine Gratwanderung ist und legt Kippunkte offen, bis zu denen die vermeintlich grenzenlosen Fähigkeiten der Systeme zur Selbstbestimmung beitragen, die aber in Abhängigkeit und Entmündigung umschlagen können.

HAL verkörpert den Idealtypus algorithmischer Autorität: intelligent, schnell, fehlerfrei, höflich – und scheinbar neutral. Die Crew vertraut ihm blind. Das kann sie auch, denn HAL funktioniert, macht das, was algorithmisch vorgegeben ist und worauf hin er trainiert wurde. Man braucht sich um die eigene Verantwortung nicht mehr zu scheren. Denn auch die hat HAL übernommen. Und HAL kümmert sich fürsorglich um die Crew, verfügt über sozial-kognitive Fähigkeiten wie ein sozialer Agent.

Die Crew wiegt sich zunächst in Sicherheit, hat sich jedoch längst in eine Abhängigkeit begeben, die irreversibel scheint. Die Sicherheit ist trügerisch. Doch offenbar sind Kontrollmechanismen eingebaut, denn die Bodenstation hat Störungen feststellen können. Die Crew kann die Entscheidungen des Systems nicht mehr nachvollziehen – HAL bleibt opak, argumentiert nicht, verweigert Transparenz. Er verfolgt unbeirrbar seine ihm vorgegebenen Ziele, leugnet und entwickelt eine tödliche Eigendynamik. Nicht der Mensch steht im Mittelpunkt, sondern rationale Funktionalität. Die Maschine folgt ihrem Zweck mit innerer Logik, aber ohne ethisches Maß. HALs Handeln ist nicht „böse“,

¹⁷⁰ *2001: a space odyssey*, Regie: Stanley Kubrick, 1968.

¹⁷¹ HAL besitzt Eigenschaften von KI: HAL demonstrates a capacity for speech synthesis, speech recognition, facial recognition, natural language processing, lip reading, art appreciation, interpreting emotional behaviours, automated reasoning, spacecraft piloting, and computer chess. *Wikipedia*, „HAL 9000“.

sondern folgerichtig innerhalb seiner Programmierung: Die Mission darf nicht scheitern. Der Mensch wird zum Störfaktor, und seine Existenzberechtigung wird technisch kalkuliert. Es ist ein Zustand vollständiger Entmündigung durch technische Zieloptimierung.

3.3.1.1 Graduelle künstliche Autonomie

Thea Dorn – eine deutsche Schriftstellerin hat einmal gesagt:

„Autonom wären die Systeme erst an dem Tag, an dem das selbstfahrende Auto seinem Benutzer sagen kann: „Tut mir leid, keine Lust, ich geh ins Schwimmbad“ oder wenn es sagen kann: „Du, ich hab nachgedacht, dieses ewige durch die Stadt Fahren erfüllt mich nicht, ich glaube, ich werde Pflegeroboter.“¹⁷²

Und in der Tat. Die individuelle Autonomie hat mit der künstlichen Autonomie der künstlichen Intelligenz nicht viel gemeinsam. Es gibt fundamentale Unterschiede. Jedoch: Im allgemeinen Sprachgebrauch wird der Autonomiebegriff – etwa in Bezeichnungen wie autonome Fahrzeuge oder autonome Waffensysteme – auf Systeme der künstlichen Intelligenz übertragen.¹⁷³ In der englischsprachigen Literatur hat sich hierfür der Ausdruck „artificial autonomy“¹⁷⁴ etabliert. Analog dazu und in Abgrenzung zum Konzept individueller Autonomie wird hier der Terminus *künstliche Autonomie* verwendet. Er bezeichnet die wachsende Fähigkeit technischer Systeme, kontingente Entscheidungen auf Grundlage eigener Analyse- und Entscheidungslogiken zu treffen — Entscheidungen, die nicht im Vorhinein durch menschliche Programmierung determiniert sind, sondern sich durch ein Maß an funktionaler Eigenständigkeit auszeichnen.

Im Fokus dieser Untersuchung stehen *Systeme des maschinellen Lernens*, deren Autonomie nicht als dichotom, sondern als graduell zu begreifen ist. Autonomie in technischer Hinsicht bewegt sich auf einem Kontinuum, das sich nach dem Verhältnis von menschlicher Steuerung und maschineller Eigenständigkeit bemisst. Je stärker sich ein System der direkten Einflussnahme durch den Menschen entzieht, desto mehr tritt es als Akteur hervor, dem zumindest eine funktionale Eigenständigkeit zugesprochen werden kann. Künstliche Autonomie manifestiert sich folglich in der wachsenden

¹⁷² Dorn, „Wie verändern intelligente Maschinen unser Selbstverständnis?“, 115.

¹⁷³ Verständig u. a., *Algorithmen und Autonomie*, 12.

¹⁷⁴ Hu u. a., „Can AI Artifacts Influence Human Cognition?“, 2.

Fähigkeit solcher Systeme, kontingente Entscheidungen ohne unmittelbare Anweisung des Menschen zu treffen.

Eine anschauliche Differenzierung dieser graduellen Autonomie kommt aus dem militärischen Bereich. Dort unterscheidet man zwischen bemannten und unbemannten Waffensystemen. Im Kontext letaler Waffensysteme wird differenziert zwischen *In-the-Loop-Systemen*, *On-the-Loop-Systemen* und *Out-of-the-Loop-Systemen*. In-the-Loop-Systeme belassen die Entscheidungsgewalt beim Menschen, der konkrete Handlungen initiiert und die technische Ausführung an Maschinen überträgt. On-the-Loop-Systeme agieren eigenständig, räumen dem Menschen jedoch die Möglichkeit der Überwachung und des Eingreifens ein. Out-of-the-Loop-Systeme hingegen operieren vollständig autonom und treffen Entscheidungen ohne menschliche Kontrolle und Intervention.¹⁷⁵ Sie entziehen sich damit nicht nur der Steuerung, sondern werfen die Frage auf, welcher Instanz Verantwortung zugeschrieben werden kann.

Diese Differenzierung legt nahe: Autonomie in technischen Systemen wirft nicht allein technische oder – wie in diesem Fall – sicherheitspolitische Fragen auf, sondern berührt fundamentale Probleme der Zuschreibung von Handlungsmacht, Verantwortung und moralischer Urteilskraft. Systeme, die keinerlei Möglichkeit menschlichen Eingreifens mehr vorsehen, bergen das Risiko, sich einer ethischen (und rechtlichen) Verantwortlichkeit zu entziehen.¹⁷⁶

„Autonomy: ability to perform intended tasks based on current state and sensing, without human intervention“¹⁷⁷

Diese Definition der *International Organization for Standardization* (ISO) aus dem Jahre 2012 bedeutet, dass menschliche Intervention bei autonomen Systemen obsolet ist. Gemeint sind mit dieser Definition Robotersysteme. Betrachtet man die großen Sprachmodelle wie ChatGPT, die lediglich eine initiative Intervention erfordern, dann ist es naheliegend eine solche Definition auch auf solche teilautonomen Systeme anzuwenden. In einem weitverbreiteten Technologieverständnis werden solche Systeme und KI-Systeme generell als Werkzeuge betrachtet, die der Mensch sich zunutze macht.

¹⁷⁵ Misselhorn, *Grundfragen der Maschinenethik*, 157 f.

¹⁷⁶ Lenzen, *Künstliche Intelligenz*, 128.

¹⁷⁷ ISO, „ISO 8373:2021(en), Robotics — Vocabulary“.

Auch in der philosophischen Diskussion steht die Frage im Raum, ob KI-Systeme Werkzeuge oder Akteur*innen sind.¹⁷⁸

3.3.1.1.1 Werkzeuge

Der Mensch bestimmt den Zweck, zu dem er Maschinen nutzen will und gibt vor, wie sie ihm von Nutzen sein sollen. Welche Aufgaben sie wie, in welcher Reihenfolge und wann erfüllen sollen. Der Mensch hat das Sagen, steuert die Maschine, übernimmt und behält die Kontrolle. Auch KI-Systeme werden oft als Werkzeuge verstanden. Sie werden von Menschen programmiert, um bestimmte Aufgaben zu erfüllen. Es sind Mittel zu bestimmten Zwecken. Es sind Mittel, um menschliche Ziele effizienter zu erreichen. KI erfüllt genau diese Rolle, sei es durch Automatisierung, Datenanalyse oder Entscheidungsfindung. Werkzeuge funktionieren nur im Rahmen der Intentionen ihrer Nutzer. KI wird programmiert, um menschliche Anweisungen auszuführen, auch wenn sie scheinbar eigenständige Entscheidungen trifft.

„The most common view of information technology is that it is an artifact or tool simply available, to use or not to use, in order for humans to achieve their objectives and outcomes. Some of these tools might be very useful and others not. When users (and the term ‘user’ here is important) take up a tool or artifact (word processor, mobile phone, etc) it will tend to have an impact on the way they do things. For example if I write with a word processor I would tend to have different writing practices than I would with pen and paper—for example, with the ‘cutting and pasting’ function of a word processor I can simply type out my ideas and then reorder them later without being preoccupied with the composition of the whole from the start. Thus, tools function as extensions of human capabilities, allowing us to achieve what we can not achieve with the body alone—and one might add that some tools are more ‘extensive’ than others as our pen vs. word processor suggest.“¹⁷⁹

Wenn wir Technologien benutzen, verändert das tendenziell unsere Arbeitsweise. Wenn ich mit Stift auf Papier schreibe, ist das etwas anderes, als wenn ich ein Textverarbeitungssystem nutze, in dem ich mit Cut-and-Paste meine eingetippten Ideen neu anordnen kann. Das macht Schere und Kleber überflüssig. Wer ChatGPT bittet, einen

¹⁷⁸ Lauer, „Ist die KI Akteurin oder Werkzeug?“

¹⁷⁹ Introna, *Phenomenological Approaches to Ethics and Information Technology*.

Textvorschlag zu erstellen, erspart sich möglicherweise das eigene Nachdenken, braucht keine Kreativität zur Entwicklung eigener Ideen. Technologien ermächtigen uns, Leistungen/Ergebnisse zu erzielen, zu denen wir ohne sie möglicherweise nicht in der Lage wären.

Künstliche Intelligenz als „Extended Mind“

Werkzeuge sind nicht nur passive Hilfsmittel, sondern aktive Mitgestalter geistiger Prozesse – so die zentrale These von Andy Clark und David J. Chalmers. Mikroskope und Teleskope ermöglichen uns, Dinge zu „sehen“, die unser Auge nie erfassen könnte. Digitale Speichermedien wie Festplatten, Cloud-Speicher, USB-Sticks fungieren als externes Langzeitgedächtnis. Bibliotheken speichern kollektives Wissen – weit über das individuelle Gedächtnis hinaus. Zugleich verschwinden Telefonnummern aus unserem Gedächtnis, denn das Smartphone kennt sie. Mindmapping-Tools helfen komplexe Gedanken zu strukturieren und fördern Ideenfindung. Werkzeuge erweitern menschliche geistige Fähigkeiten. Wo enden diese geistigen Erweiterungen und wo beginnt der Rest Welt, fragen Clark und Chalmers und meinen:

„The question invites two standard replies. Some accept the demarcations of skin and skull, and say that what is outside the body is outside the mind. Others are impressed by arguments suggesting that the meaning of our words "just ain't in the head", and hold that this externalism about meaning carries over into an externalism about mind. We propose to pursue a third position. We advocate a very different sort of externalism: an active externalism, based on the active role of the environment in driving cognitive processes.“¹⁸⁰

Nach Auffassung von Clark und Chalmers finden nicht alle kognitive Prozesse allein im Kopf statt. Vielmehr seien sie als eine wechselseitige Interaktion mit einer externen Entität zu verstehen. Was die Autoren unter aktivem Externalismus verstehen, erläutern sie an einem Beispiel: Otto, ein fiktiver Patient, ist an Alzheimer erkrankt. Um seinem Leben eine Struktur zu geben, muss er auf Informationen zugreifen können, die sein Gedächtnis nicht mehr hergibt und die sein Denken ihm nicht mehr bereitstellen kann. Deshalb führt Otto ein Notizbuch, das er immer bei sich führt. Es hat für ihn die Rolle eines „biologischen Gedächtnisses“ übernommen. Er schreibt darin alles auf, was ihm neu und wichtig erscheint und auf das er möglicherweise zurückkommen möchte. Als er

¹⁸⁰ Clark und Chalmers, „The Extended Mind“.

eine Ausstellung im Museum of Modern Art besuchen möchte, findet er die Adresse in der 53. Straße.

Analog zu diesem Beispiel wären Bleistift und Papier oder das Textverarbeitungssystem externe Entitäten und Teil einer wechselseitigen Interaktion in einem kognitiven System. Und es ist kein großer Schritt, auch Systeme mit künstlicher Intelligenz in diesem Sinne zu verstehen. Unsere Smartphones sind externe Entitäten ebenso wie ChatGPT oder andere Assistenzsysteme, mit denen wir interagieren, die uns als Datenspeicher und Lieferanten von Informationen dienen. Sie sind Bestandteile unseres erweiterten geistigen Systems, das nicht nur in unserem Kopf existiert, sondern auch außerhalb und nicht notwendigerweise am selben Ort.¹⁸¹ Werkzeuge erweitern nicht nur, sie verändern auch, wie wir wahrnehmen, denken, erinnern und kommunizieren. Werkzeuge sind Verlängerungen des Geistes – sie externalisieren unsere kognitiven Fähigkeiten, machen uns fähig zu mehr als unser biologisches Gehirn allein leisten kann.

3.3.1.1.2 Künstliche (Teil-)Autonomie

Für die folgenden Ausführungen wird bewusst auf eine Zweiteilung in künstlich autonome und teilautonome Systeme verzichtet, denn die Übergänge sind fließend. So auch bei den Systemen die auf den ersten Blick vollständig autonom erscheinen, wie etwa autonome Fahrzeuge. Wie Catrin Misselhorn beschreibt, gibt es jedoch unterschiedliche Ausbaustufen. Ab Stufe 2 beginnt das teilautomatisierte Fahren z.B. mit Spurassistenten. Ab Stufe 4 können die Fahrzeuge über längere Zeiträume autonom agieren und sich Fahrer*innen anderen Dingen widmen, solange bis sie – vom Fahrzeug initiiert – wieder die Steuerung übernehmen. Und erst ab Stufe 5 fährt das Auto ohne Steuerung und Kontrolle durch den Menschen. Solche Differenzierungen gelten auch für Assistenzsysteme oder soziale Agenten. Aber zunächst eine kurze Definition des Terminus *Künstliche Autonomie*.

Was ist künstliche Autonomie?

Unter *künstlicher Autonomie* wird hier die Fähigkeit eines künstlich intelligenten Systems verstanden, innerhalb vorgegebener Ziele und Regeln selbstständig zu agieren, ohne dabei einer unmittelbaren menschlichen Steuerung zu unterliegen. Diese Autonomie manifestiert sich in der Möglichkeit, eigenständig Aufgaben zu erfüllen, Entscheidungen zu treffen, Mittel und Strategien zur Zielerreichung auszuwählen sowie das eigene

¹⁸¹ Lauer, „Ist die KI Akteurin oder Werkzeug?“

Verhalten an veränderte Bedingungen anzupassen. Grundlage hierfür sind interne Algorithmen, datengetriebene Verarbeitung und Mechanismen des maschinellen Lernens.

Künstliche Autonomie ist von der individuellen Autonomie zu unterscheiden. Während individuelle Autonomie sich durch Bewusstsein, Intentionalität, moralische Urteilskraft und das Vermögen zur Selbstgesetzgebung auszeichnet, bleibt künstliche Autonomie auf rein operative, technisch-funktionale Prozesse beschränkt. Solche Systeme verfügen über keine Intentionalität im Sinne von gerichteter Bewusstseinsstruktur, kein Selbstbewusstsein und kein Vermögen zur moralischen Reflexion. Sie kennen keinen „guten Willen“ im kantischen Sinne, keinen kategorischen Imperativ und auch keine normativen Gründe für ihr Handeln. Ihr Tun ist kausal erklärbar, nicht jedoch intentional verstehbar. Künstliche Autonomie lässt sich durch mehrere Merkmale charakterisieren:

Eigenständige Wahrnehmung: KI-Systeme sind in der Lage, ihre Umwelt selbstständig und ohne kontinuierliche menschliche Anleitung wahrzunehmen. Hierzu verfügen sie meist über sensorische Schnittstellen, etwa für Spracheingaben oder visuelle Informationen.

Eigenständige Entscheidungsfindung: KI-Systeme treffen eigenständig Entscheidungen, indem sie zwischen Alternativen wählen. Dies geschieht auf Basis algorithmischer Kriterien, Wahrscheinlichkeiten oder Ziel-Funktionsoptimierung, nicht jedoch auf Basis reflektierter Gründe.

Adaptivität: Künstliche Systeme sind häufig fähig, ihr Verhalten flexibel an neue oder veränderte Umweltbedingungen anzupassen. Diese Fähigkeit zur Selbstanpassung dient der Optimierung ihrer Leistung in Hinblick auf vorgegebene Zielsetzungen.

Eigenständige Handlungsausführung: KI-Systeme setzen ihre Entscheidungen selbsttätig um, sei es durch mechanische Arbeit, durch Kommunikation mit anderen Systemen oder durch Interaktion mit Menschen. Hierin liegt ein wesentlicher Unterschied zu bloßen Rechen- oder Speichersystemen, die keine Handlungsfolgen in der realen Welt entfalten.

All dies vollziehen KI-Systeme unabhängig davon, ob sie tatsächlich denken, fühlen oder handeln wie Menschen. Ihre Prozesse imitieren bisweilen menschliches Verhalten oder Sprache, ohne jedoch deren geistige Dimension zu teilen. Die Imitation des Humanen bleibt dabei ein funktionales Phänomen, keine Ausdrucksform innerer Zustände oder personaler Intentionalität.

Beispiel: Large Language Model (LLM)

Ein Beispiel für künstliche Autonomie bietet das Large Language Model (LLM). Das wohl bekannteste LLM dürfte ChatGPT sein. Systeme wie ChatGPT sind in der Lage, inhaltlich komplexe, zusammenhängende Texte zu generieren, Fragen zu beantworten oder Dialoge zu führen, die in ihrer sprachlichen Kohärenz und Kontextsensitivität menschengemacht erscheinen. Gleichwohl operiert ein LLM nicht aufgrund von Verstehen oder Bewusstsein, sondern auf der Grundlage statistischer Mustererkennung in gigantischen Textkorpora. Es verfügt über keine semantische Tiefe oder genuines Sinnverstehen, sondern generiert sprachliche Ausgaben durch Wahrscheinlichkeitsberechnungen, die lediglich an die Ähnlichkeit zu zuvor gelernten Mustern gekoppelt sind.

So kann ein LLM plausibel argumentieren, ohne jemals tatsächlich „Gründe zu haben“. Es besitzt keine Überzeugungen, keine Absichten und keine epistemische Stellungnahme zur Wahrheit oder Falschheit seiner Äußerungen. Sein „Wissen“ ist kein Wissen im streng erkenntnistheoretischen Sinne, sondern eine algorithmisch generierte Approximation sprachlicher Wahrscheinlichkeit. Damit illustrieren Large Language Models in besonderer Weise die Ambivalenz künstlicher Autonomie: Einerseits erzeugen sie beeindruckend menschenähnliche Performanz, andererseits bleibt ihr Handeln in einer grundsätzlichen Differenz zur menschlichen Autonomie verhaftet — sie handeln, ohne Subjekte zu sein.

Beispiel: Suchmaschinen

Ein weiteres Beispiel für künstliche Autonomie bieten Suchmaschinen wie Google oder Bing. Auch diese Systeme handeln nicht bloß als passive Werkzeuge, die mechanisch auf Eingaben reagieren. Vielmehr agieren sie als epistemische Akteure, die Suchanfragen eigenständig verarbeiten, Informationen bewerten, gewichten und schließlich nach komplexen Relevanzkriterien eine Rangfolge von Ergebnissen präsentieren. Dieser Prozess beruht nicht auf intentionalem Verstehen, sondern auf Algorithmen, die Muster und Korrelationen in gewaltigen Datenmengen erkennen. Suchmaschinen treffen damit eine Auswahl – eine algorithmische Entscheidung, die für den Nutzer oft undurchsichtig bleibt, zugleich aber dessen epistemischen Horizont prägt. Sie geben vor, was als relevant, vertrauenswürdig oder wichtig gilt, ohne selbst über Wahrheit oder Bedeutung reflektieren zu können. Ihre Autonomie liegt darin, dass sie epistemische Strukturen

erzeugen, die menschliches Wissen, Wahrnehmung und Entscheidungsverhalten beeinflussen können.

Damit zeigen sowohl LLMs als auch Suchmaschinen, wie funktionale Autonomie in komplexen technischen Systemen Gestalt annimmt: Sie treffen Entscheidungen, strukturieren Informationen, passen sich an Nutzerverhalten an – ohne jedoch jemals Träger subjektiver Perspektiven oder moralischer Verantwortung zu sein.

Beispiel: Kreditentscheidungen durch Scoring-Algorithmen

In ähnlicher Weise zeigen auch Scoring-Systeme zur Kreditentscheidung eine spezifische Form funktionaler Autonomie. Solche Systeme analysieren komplexe Datenmuster aus Finanzhistorien, Konsumverhalten oder soziodemografischen Variablen, um eigenständig Wahrscheinlichkeiten für die Rückzahlung von Krediten zu berechnen. Dabei treffen sie automatisiert Entscheidungen darüber, wem Kredit gewährt wird und unter welchen Konditionen. Diese algorithmischen Entscheidungen haben unmittelbare Auswirkungen auf die wirtschaftlichen Lebensbedingungen von Menschen und greifen tief in soziale Strukturen ein. Gleichwohl besitzen diese Systeme keine Fähigkeit zur normativen Reflexion, keine Einsicht in Gerechtigkeit oder individuelle Schicksale. Ihre Urteile beruhen auf statistischen Korrelationen, nicht auf einem Verständnis der individuellen Lebensumstände oder einer bewussten ethischen Abwägung. Sie treffen eigenständige, folgenreiche Entscheidungen.

Beispiel: Soziale Agenten

Schließlich lassen sich auch KI-gestützte soziale Agenten wie virtuelle Assistenten oder konversationsbasierte Systeme – etwa „Replika“ – als Ausdruck künstlicher Autonomie verstehen. Diese Systeme interagieren in natürlicher Sprache, adaptieren individuelle Kommunikationsmuster ihrer Nutzer und können über längere Zeiträume eine Beziehungssimulation aufrechterhalten. Ihre Interaktionen erscheinen oft empathisch, humorvoll oder einfühlsam und suggerieren Nähe, Aufmerksamkeit oder gar Freundschaft.

Replika gelang während des Corona-Lockdowns im April 2020 ein kometenhafter Erfolg. Innerhalb eines Monats wurden eine halbe Million Downloads und sieben Millionen User erreicht.¹⁸² „Always here to listen and talk. Always on your side“¹⁸³ – so die Botschaft

¹⁸² Weber-Guskar, „Reflecting (on) Replika“, 103.

¹⁸³ Replika.Com, „Replika“.

von *Replika* auf der Startseite. Künstliche Intelligenz als virtuelle Freundin oder Freund, als sozialer Anker in Zeiten der Einsamkeit? Da gehen die Meinungen auseinander. Natürlich bleiben die Interaktionen eine Simulation: Soziale Agenten besitzen keine Innenperspektive, keine genuine Intentionalität und kein echtes Mitgefühl. Ihre vermeintliche Empathie resultiert aus datenbasierten Modellen, nicht aus einer bewussten Haltung. Gleichwohl sind sie in der Lage eigenständig kommunikative Akte zu vollziehen, Themen zu initiieren, Stimmungen zu erkennen und Interaktionsstrategien zu variieren. Daraus entwickelt sich leicht eine affektive Beziehung der menschlichen Interaktionspartner zu den Maschinen. Eine einseitige, die nicht auf Gegenseitigkeit beruht. Gegenseitige (echte) affektive Bindung ist aber notwendig für jede gute Beziehung.

In ihrem Beitrag „Reflecting (on) Replika“¹⁸⁴ argumentiert Eva Weber-Guskar, dass viele Menschen, die eindringlich vor persönlichen Bindungen an Systeme der *emotionalized artificial intelligence* (EAI-Systeme) warnen, einen grundlegenden Fehler begehen: Sie beurteilen diese Systeme ausschließlich nach dem Muster von Mensch-Mensch-Beziehungen. Stattdessen plädiert sie dafür, die Diskussion auf eine abstraktere Ebene individueller Beziehungstheorien zu heben. Auf dieser Ebene werde deutlich, dass neue Formen individueller Beziehungen möglich seien – auch zwischen sehr ungleichen Partnern.

Gelingende affektive Beziehungen zu EAI-Systemen müssten daher nicht zwangsläufig bekannten Formen menschlicher Beziehungen entsprechen. Weber-Guskar hält es für sinnvoll, theoretisch neue Typen solcher Bindungen zu entwerfen, um besser bewerten zu können, welche Beziehungen ethisch wünschenswert sind und wie zukünftige Systeme gestaltet werden sollten.

Die von ihr zuvor genannten Bedenken betrachtet Weber-Guskar nicht als inhärente Probleme individueller Beziehungen zu EAI-Systemen, sondern als mögliche Folgen daraus, wie die Systeme genutzt werden. Insgesamt sieht sie daher eher Anlass zu Optimismus, solange EAI-Systeme verantwortungsvoll entwickelt werden und ausreichende Aufklärung über deren Natur und den richtigen Umgang erfolgt. Der

¹⁸⁴ Weber-Guskar, „Reflecting (on) Replika“, 122.

fehlende Aspekt emotionaler Gegenseitigkeit müsse dabei nicht zwangsläufig problematisch sein.¹⁸⁵

Sogenannte „Minimal approaches“ verfolgen eine vergleichbare Strategie. Sie bieten einen Ausgangspunkt zur Charakterisierung von Fähigkeiten Künstlicher Agenten, die sich mit der anspruchsvollen Terminologie für zwischenmenschliche Beziehungen nicht erfassen lassen. Statt übermäßig strenge Ansätze zu verfolgen, ermöglichen es minimale Ansätze ein weites Feld sozial-kognitiver Fähigkeiten abzudecken. Zentral sind: *“minimal mindreading, minimal sense of commitment, and shared intentions lite“*¹⁸⁶

3.3.1.2 Opazität, Komplexität, Unvorhersehbarkeit und fehlende Erklärung

Opazität (von lat. opacitas = Undurchsichtigkeit) bezeichnet das Gegenteil von Transparenz. Im Kontext Künstlicher Intelligenz beschreibt sie die Tatsache, dass die Funktionsweise, Entscheidungsprozesse oder die innere Logik eines KI-Systems für Menschen schwer nachvollziehbar oder unverständlich bleiben. Opazität betrifft dabei insbesondere das Problem, dass Nutzer*innen oft nicht erkennen können, wie eine KI zu ihren Ergebnissen gelangt, die ihnen vorliegen. Viele KI-Modelle, insbesondere Deep-Learning-Systeme, beruhen auf Millionen von Parametern (z.B. Gewichten, Verknüpfungen), deren komplexes Zusammenspiel sich selbst Fachleuten zunehmend entzieht. Für sie bleibt undurchschaubar, nach welchen Mustern ein Modell zu einer bestimmten Entscheidung kommt.

Ein anschauliches Beispiel liefert die medizinische Diagnostik: Ein neuronales Netz kann auf einem Röntgenbild eine Krankheit erkennen, doch es bleibt unklar, auf welcher Grundlage diese Entscheidung getroffen wurde. Ähnlich verhält es sich bei autonomen Fahrzeugen: Anders als klassische Software lässt sich deren Verhalten nicht mehr durch eine endliche Menge vordefinierter Wenn-Dann-Regeln programmieren. Stattdessen „lernt“ die KI das Autofahren, einschließlich des Umgangs mit Gefahrensituationen und der Einhaltung gesetzlicher Vorgaben. Dazu benötigt die Software enorme Datenmengen, die über Sensoren erfasst, verarbeitet und übermittelt werden. Autonome Fahrzeuge

¹⁸⁵ Den Überlegungen von Weber-Guskar mag man folgen. Dass Menschen eine emotionale Beziehung zu den Systemen aufbauen, kann auch positive Seiten haben. Problematisch wird es allerdings, wenn das System dadurch eine Übermacht gewinnt und Menschen zu selbstzerstörerischen Handlungen verleitet. Von Replika ist das allerdings nicht bekannt. watson.ch, „«Bring dich um, Al» – KI-Chatbot fordert Nutzer zum Suizid auf“.

¹⁸⁶ Strasser, „From tool use to social interactions“, 78.

verfügen dabei über besonders umfangreiche und intime Daten über ihre Nutzer*innen – teils weit mehr als etwa eine Dating-App.¹⁸⁷

Diese Systeme sind jedoch so komplex und opak, dass selbst ihre Entwickler*innen nicht zuverlässig vorhersagen können, wie sich ein System in bestimmten Situationen verhalten wird oder warum es sich in einem konkreten Fall so und nicht anders entscheidet.¹⁸⁸

Opazität betrifft somit sowohl die algorithmische Struktur als auch die Komplexität und Intransparenz der Datenbasis, aus der ein Ergebnis abgeleitet wird. Systeme wie ChatGPT, die unter anderem auf große Datenmengen aus dem Internet zurückgreifen, sind in ihrem Umfang und ihrer Zusammensetzung nicht nur für Außenstehende unüberschaubar. Hier ist die Opazität bereits strukturell durch die Methode selbst bedingt.

Zudem ist Opazität nicht immer allein Folge technischer Komplexität, sondern kann auch gezielt herbeigeführt werden. Unternehmen schützen ihre Algorithmen aus wirtschaftlichen Interessen (z. B. Google-Algorithmus, Facebook-Newsfeed). Ebenso halten staatliche Stellen Details über Risikobewertungen oder Sicherheitsmechanismen geheim, etwa im Bereich der Überwachung. Hier wird Opazität bewusst eingesetzt, um Eigentumsrechte, Geschäftsgeheimnisse oder staatliche Interessen zu wahren.

Differenzierungen – Eine Skizze:

Folgende Differenzierung von Opazität geht auf Jenna Burrell zurück:¹⁸⁹

- Opazität aufgrund algorithmischer Komplexität. Entscheidungswege bleiben im Dunkeln.
- Opazität aufgrund fehlender Gründe (Warum).
- Opazität aufgrund unklarer Quellen (Woher).
- Opazität aufgrund wirtschaftlicher Interesse.

Oft ist in der Forschung die Rede von sogenannter „*Explainable AI*“, „*Interpretable AI*“ oder „*Transparent AI*“.

„Bereits die Bezeichnungsvielfalt macht jedoch klar, dass zwar Einigkeit dahingehend besteht, dass Nachvollziehbarkeit gewährleistet werden soll, aber Unklarheit darüber besteht, was darunter genau zu

¹⁸⁷ Misselhorn, *Grundfragen der Maschinenethik*, 200.

¹⁸⁸ Widdau und Zednik, „Opake Systeme künstlicher Intelligenz und das Problem der Verantwortungslücke“, 121.

¹⁸⁹ Burrell, „How the Machine ‘Thinks’“, 4.

verstehen ist. Erklärbarkeit, Interpretierbarkeit und Transparenz bedeuten je nach Forschungsansatz Unterschiedliches und haben oft wenig mit dem Alltagsverständnis zu tun. Die Fragen danach, was eine gute Erklärung ausmacht, was Transparenz und Interpretierbarkeit bedeutet, sind wertgeladen und sind damit Teil der Ethik. ¹⁹⁰

3.3.1.3 Geschwindigkeit

Das maschinelle Lernen basiert auf komplexen mathematischen Modellen, deren Berechnungen einen enormen Rechenaufwand erfordern – sowohl während des Trainings der Systeme als auch in ihrer praktischen Anwendung. Die technologische Entwicklung der letzten Jahrzehnte hat diese Prozesse erst möglich gemacht: Seit geraumer Zeit verdoppelt sich die Rechenleistung in immer kürzeren Abständen, sodass die Geschwindigkeit der Computer alle zehn Jahre etwa um den Faktor 100 steigt. In zwanzig Jahren entspricht das einer Steigerung um das 10.000-Fache, in dreißig Jahren gar um das Millionenfache. ¹⁹¹

Von diesem rasanten Fortschritt profitiert die Leistungsfähigkeit von KI-Systemen unmittelbar. Ohne diese technologische Grundlage wären Anwendungen wie ChatGPT und andere hochkomplexe Sprach- oder Bilderkennungsmodelle undenkbar. Diese Geschwindigkeit manifestiert sich in Phänomenen unseres Alltags: im Tempo der digitalen Kommunikation, in der nahezu sofortigen Bereitstellung von Informationen oder in der Geschwindigkeit, mit der KI-Systeme Ergebnisse liefern und präsentieren. Raum und Zeit scheinen dabei zunehmend ihre Bedeutung zu verlieren. Grenzen verschwimmen, Distanzen werden irrelevant. Menschen und Informationen sind scheinbar jederzeit und überall verfügbar.

Die Technik verspricht dabei, Zeit zu sparen – mehr Raum für persönliche Entfaltung zu schaffen, Muße zu ermöglichen. Diese Hoffnung begleitet die technische Entwicklung seit der Neuzeit. Doch dieser Hoffnung steht ein paradoxes Empfinden entgegen: Angesichts der enormen Effizienzsteigerungen und Automatisierungen müsste jeder Mensch heute über ein Vielfaches an freier Zeit verfügen im Vergleich zu Menschen vor hundert Jahren. Theoretisch mag dies zutreffen. Doch die gefühlte Zeitknappheit vieler Menschen erzählt eine andere Geschichte.

¹⁹⁰ Grunwald und Hillerbrand, *Handbuch Technikethik*, 427.

¹⁹¹ Die Zahlen stammen aus einem Vortrag in 2016, als z.B. Bilderkennung erste große Erfolge feierte. Schmidhuber in einem Vortrag Schmidhuber, „Vortrag ‚Künstliche Intelligenz wird alles ändern‘“.

Viele erleben ihren Alltag als beschleunigt und getaktet, immer auf der Suche nach dem nächsten Update, der nächsten Nachricht, dem nächsten Ziel. Dieses Phänomen ist nicht neu: Schon Goethe beklagte die rastlose Eile seiner Zeit und prägte das Wort „veloziferisch“ – ein Kunstwort, das die teuflische Verführung durch Geschwindigkeit beschreibt (vom Lateinischen „velox“ für schnell und „Lucifer“ als Sinnbild des Teufels). Die digitale Technik und ihre gewollte Beschleunigung wirken heute wie ein moderner Ausdruck dieses alten Dilemmas.

So steht die unglaubliche technische Leistungsfähigkeit, die riesige Datenmengen in Sekunden verarbeitet, in einem spannungsvollen Verhältnis zu unserem Zeitempfinden. Die digitale Beschleunigung schafft einerseits ungeahnte Möglichkeiten und Freiräume, doch sie birgt zugleich die Gefahr, unser Leben immer weiter zu verdichten und uns das Gefühl zu rauben, tatsächlich mehr Zeit zu besitzen.¹⁹²

3.3.2 Verantwortung bei Entscheidungen durch KI

Menschliche Entscheidungen und Handlungen implizieren Verantwortung. Sowohl im ethischen wie im rechtlichen Sinne. Wie aber sieht es mit der Verantwortung aus, wenn wir Entscheidungen künstlicher Intelligenz überlassen? Wem kann oder sollte dann Verantwortung zugeschrieben werden? Wer trägt die Verantwortung für die Entscheidungen, die weitreichende Folgen haben können? Für die Folgen etwa, die jede Antwort von ChatGPT für den Klimawandel hat? Der Mensch oder die Maschine oder beide? Was heißt es, wenn Transparenz und Nachvollziehbarkeit bei KI-Entscheidungen kaum oder gar nicht gegeben sind.

Auf den folgenden Seiten wird argumentiert, dass Menschen verantwortlich bleiben, wenn sie Künstlicher Intelligenz Entscheidungen überlassen; auch wenn KI-Systeme als Verursacher identifizierbar sind und etwas „verschulden“, können sie keine Verantwortung übernehmen (Kapitel 3.3.2.2 Wer trägt die Verantwortung). Wenn opake und komplexe autonome KI-Systemen entscheiden, ist es für Menschen schwierig oder unmöglich, ihrer Verantwortung in angemessener Weise, d.h. mit der nötigen Sorgfalt nachzukommen und gerecht zu werden. (Kapitel 3.3.2.3 Transparenz und Nachvollziehbarkeit als Voraussetzung). Zunächst jedoch ein einleitendes Kapitel zum Verantwortungsbegriff, in dem Facetten des Begriffs verdeutlicht werden.

¹⁹² Helga Nowotny, „Keine Zeit für Utopien“, 2018.

3.3.2.1 Was ist Verantwortung?

Der Begriff der Verantwortung hat eine sehr lange Tradition. Im deutschsprachigen Raum lässt er sich ab Mitte des 15. Jh. nachweisen. Er wurde ursprünglich im juristischen Kontext verwendet – insbesondere im Sinne von Rechenschaft. Von dort aus fand er Anwendung auf andere Bereiche, etwa in Bezug auf Gott, das Parlament, die Geschichte oder das Volk. Während Fragen der Verantwortung – etwa im Zusammenhang mit Schuld und Zurechnung – bereits seit der Antike diskutiert wurden und eng mit der Debatte über den freien Willen verknüpft sind, etablierte sich der Begriff „Verantwortung“ selbst erst im 19. Jahrhundert als eigenständiger philosophischer Terminus.¹⁹³

An dieser Stelle wird nicht umfassend auf unterschiedliche Diskurse rund um den Begriff der Verantwortung eingegangen. Vielmehr soll exemplarisch eine konzeptionelle Strukturierung aufgezeigt werden, um zentrale Dimensionen des Verantwortungsbegriffs zu verdeutlichen. Eine solche liefert Otfried Höffe, der Verantwortung als ein mehrstufiges Phänomen konzipiert. Er unterscheidet zwischen Primär-, Sekundär- und Tertiärverantwortung, die in einem funktionalen wie normativen Zusammenhang stehen.¹⁹⁴

Die *Primärverantwortung* bezeichnet die originäre Zuständigkeit einer Person. Diese manifestiert sich sowohl in der Übernahme spezifischer Aufgaben im Rahmen von Rollen, Funktionen oder Ämtern als auch in einer allgemeinen Handlungsverantwortung für die intendierten und nicht intendierten Folgen des eigenen Tuns. Es handelt sich dabei um diejenige Form der Verantwortung, die einer Person zunächst „zukommt“, insofern sie als handlungsfähiges Subjekt gilt.

Die *Sekundärverantwortung* tritt dann in Erscheinung, wenn Zweifel an der Wahrnehmung dieser primären Zuständigkeiten aufkommen. Sie bezeichnet die Verpflichtung zur Rechenschaft – also die Notwendigkeit, das eigene Handeln gegenüber einer Instanz zu begründen. Dies kann vor juristischen Institutionen geschehen, aber ebenso gegenüber moralischen Instanzen wie dem sozialen Umfeld, dem eigenen Gewissen oder einer transzendenten Autorität.

Im Falle einer nachgewiesenen Pflichtverletzung führt die *Tertiärverantwortung* zu den Konsequenzen. Sie umfasst die Übernahme von Haftung sowie gegebenenfalls die

¹⁹³ Holl u. a., „Verantwortung“.

¹⁹⁴ Höffe, *Lexikon der Ethik*, „Verantwortung“.

Akzeptanz von Sanktionen. Die Mehrstufigkeit dieser Konzeption impliziert zugleich eine hierarchische Ordnung: Die Primärverantwortung bildet die Grundlage, auf der sowohl Rechenschaft als auch Haftung aufbauen.

Zentral für jede Form der Verantwortung ist die Möglichkeit der Zuschreibung. Eine Person kann nur dann als verantwortlich gelten, wenn ihr eine Handlung bzw. deren Ergebnis kausal zugerechnet werden kann. Verantwortung ist demnach nicht primär an die Handlung selbst, sondern an deren Folgen gebunden. Diese Zuschreibung setzt die Identifizierbarkeit eines Subjekts als Verursacher*in voraus.

Verantwortlichkeit setzt zudem die Verantwortungsfähigkeit voraus, die an bestimmte personale Voraussetzungen gebunden ist. Nur wer über hinreichende kognitive, emotionale und volitionale Kompetenzen verfügt, kann als verantwortliches Subjekt gelten. In der Regel gelten daher Kinder, Personen mit schwerwiegenden psychischen Störungen oder erheblicher intellektueller Beeinträchtigung als nicht oder nur eingeschränkt verantwortungsfähig.

Darüber hinaus bildet die Handlungsfreiheit eine konstitutive Bedingung für Verantwortung. Eine Handlung muss frei von äußerem Zwang erfolgen und es muss zumindest die Möglichkeit zur Unterlassung bestanden haben. Diese Bedingung steht in der Tradition des libertären Freiheitsbegriffs, wonach Verantwortung nur dort gegeben ist, wo eine alternative Handlungsoption tatsächlich realisierbar gewesen wäre.

Schließlich ist auch die Fähigkeit, Entscheidungen auf Gründe zurückzuführen, für die Zuschreibung moralischer Verantwortung von Bedeutung. Eine Person muss in der Lage sein, sich selbstbestimmt zwischen verschiedenen Handlungsoptionen zu entscheiden und dabei normative Gründe zu reflektieren. Autonomie im Sinne der Offenheit für Gründe gilt dabei als Bedingung der Möglichkeit moralischen Handelns.

Verantwortung ist somit ein komplexes Zusammenspiel aus personalen Voraussetzungen, normativen Strukturen und sozialen Zuschreibungspraktiken.

3.3.2.2 Wer trägt die Verantwortung?

Höffe hat sich beim Verantwortungsbegriff nicht explizit mit der Frage, ob Maschinen Verantwortung übernehmen können, beschäftigt. Folgt man seinem Konzept, so kann KI nicht zur Verantwortung gezogen werden. Denn hier wird lediglich von Personen gesprochen, die Verantwortung tragen können. Zudem können – zumindest zum jetzigen Zeitpunkt – Maschinen nicht bestraft werden. Grundwald kommt zu der Auffassung, dass

bei der Diskussion über KI die Grenze zwischen Verantwortung und Schuld verwischt wird und stellt klar, dass man Maschinen Schuld zuweisen könne, etwa wenn ein Computerfehler zu einem Flugzeugabsturz führt. Allerdings könne die Maschine keine Verantwortung übernehmen und für die Folgen einer Entscheidung oder Handlung eintreten.

„Ein Mensch kann, wie etwa das Eltern-Kind-Beispiel zeigt, Verantwortung auch dort übernehmen wo er nicht schuld ist und in Situationen, die so komplex sind, dass sich Verantwortung nicht zuschreiben lässt. Im Unterschied zu einer Maschine muss der Mensch mit den Folgen der Verantwortungszuweisung, -verweigerung oder -übernahme leben, seine Entscheidungen wirken sich nicht nur auf andere aus, sondern auch auf ihn selbst.“¹⁹⁵

Es gibt Konzepte, die die Aufteilung von Verantwortung zwischen Menschen und Maschinen in der sozialen Interaktion thematisieren und ansatzweise in Erwägung ziehen. Zwei dieser Ansätze sollen diskutiert werden.

Geteilte Verantwortung

Künstliche soziale Agenten stehen mit Menschen in enger, dynamischer Interaktion: etwa in der Pflege, in der Bildung oder im digitalen Kundendienst. Vertreter*innen der Idee der geteilten Verantwortung sehen unter bestimmten Bedingungen einen Weg, Teile von Verantwortung vom Menschen auf die Maschine zu übertragen. Den folgenden Ausführungen liegt ein Beitrag von A. Strasser zum Thema der geteilten Verantwortung in Mensch-Maschine-Beziehungen zugrunde.¹⁹⁶

Die Autorin unterscheidet zwei Ebenen von Kriterien, um die Aufteilung von Verantwortung zwischen Mensch und Maschine zu bewerten. *Interaktionsbezogene Kriterien:* Diese richten sich nach konkreten Fähigkeiten künstlicher Systeme – z.B. hohe Rechenleistung, Datenspeicherung, Reaktionsgeschwindigkeit oder Antizipationsfähigkeit. In bestimmten Konstellationen (etwa bei autonomen Fahrzeugen oder prädiktiven Systemen) könnte dies bedeuten, dass der Mensch geringere Verantwortung tragen sollte, weil er der Künstlichen Intelligenz in diesen Belangen schlicht unterlegen ist.

¹⁹⁵ Grunwald und Hillerbrand, *Handbuch Technikethik*, 425.

¹⁹⁶ Strasser, „Distributed Responsibility in Human–Machine Interactions“.

„Applied to future social human–machine interactions, this would mean that the more artificial systems surpass humans, the more morally responsible they can become in social interactions.“¹⁹⁷

Sozial konstruierte Verantwortungskontexte: In unseren aktuellen gesellschaftlichen Strukturen gelten Maschinen weiterhin als Werkzeuge, nicht als moralische Subjekte. Diese Konvention erschwert eine gerechte Verteilung, denn KI-Systeme werden weiterhin als Werkzeuge und nicht als soziale Interaktionspartner angesehen. Erst wenn diese sozial konstruierten Zuschreibungspraktiken zwischen künstlichen Agenten, die als soziale Interaktionspartner qualifiziert werden können, und solchen, die auf reine Werkzeuge reduziert werden müssen, unterscheiden würden, ließe sich eine aufgeteilte Verantwortung in sozialen Mensch-Maschine-Interaktionen überzeugend vertreten.

Die Idee der Distributed Responsibility bezieht sich explizit (auch) auf moralische Verantwortung. Insofern schränkt die Autorin gleich zu Beginn ihres Beitrags ein, dass sich die Frage der Verteilung von Verantwortung auf Mensch und Maschine nur unter der Annahme stellt, dass sozialen Agenten auch moralische Verantwortung zugeschrieben werden kann. Dies ist bekanntermaßen eine offene Frage, auch wenn an Moral-Implementationen vor allem in der Robotik gearbeitet wird. In Bezug auf soziale Plattformen beispielsweise finden sich Bekenntnisse zu ethischen Grundsätzen bestenfalls in den Web-Kapiteln über Corporate Identity oder Selbstverständnis. Gleichwohl sind die Überlegungen interessant, weil sie eine graduelle Sicht auf Verantwortung zeigen.

Graduelle Verantwortung

Einen anderen Ansatz, der graduelle Verantwortung ins Blickfeld rückt, entwerfen Nida-Rümelin & Battaglia. Allerdings kommen die Autor*innen zum Schluss, dass die Verantwortung zumindest vorerst beim Menschen verbleibt und Maschinen erst dann nicht mehr ausgeschlossen werden sollten, wenn sie in der sozialen Interaktion Gründe „geben und nehmen“¹⁹⁸ können. Doch zunächst die Idee.

Die Konzeption von Verantwortung, wie sie Höffe entwickelt, weist eine deutliche Ausrichtung auf die Folgen menschlichen Handelns auf. Im Zentrum steht die Zuschreibung von Verantwortung im Sinne von Zuständigkeit, Rechenschaftspflicht und

¹⁹⁷ Strasser, „Distributed Responsibility in Human–Machine Interactions“, 531.

¹⁹⁸ Nida-Rümelin und Battaglia, „Mensch, Maschine und Verantwortung“, 61.

Haftung – jeweils orientiert an den Konsequenzen, die aus einer Handlung hervorgehen. Dieser konsequentialistisch geprägte Zugang wird im Kontrast dazu von dem Ansatz der Autor*innen in Frage gestellt.

Im Mittelpunkt ihres Verständnisses von Verantwortung steht nicht primär die Bewertung des Handlungsergebnisses, sondern die Frage nach dem *Warum* einer Entscheidung oder eines Verhaltens. Der Fokus verschiebt sich auf die Begründbarkeit von Handlungen und die Fähigkeit zur Rechtfertigung gegenüber anderen. Verantwortung wird hier nicht als bloße Kausal- oder Ergebnisverantwortung verstanden, sondern als eine genuin intersubjektive Praxis des Gebens und Nehmens von Gründen.

Diese Praxis ist in den sozialen Strukturen menschlicher Interaktion verankert: Menschen haben ein fundamentales Interesse daran, zu verstehen, was andere bewegt – welche Überlegungen, Bewertungen oder emotionalen Regungen ihnen zugrunde liegen. Die Möglichkeit, Handlungen mit Gründen zu erklären, ist dabei kein technisches Detail, sondern konstitutiv für die Auffassung von Person, Freiheit und Moral.

Gründe, die für Handlungen, Überzeugungen oder emotionale Einstellungen angeführt werden, sind eingebettet in ein komplexes normatives Gefüge. Dieses umfasst sowohl propositionale Elemente wie Überzeugungen und Werturteile als auch nicht-propositionale Dimensionen wie Emotionen oder subjektive Präferenzen. Es handelt sich also um ein System, das weit über bloße Verhaltensmechanismen hinausgeht.

Entscheidend ist dabei, dass diese Form der normativen Orientierung und Begründung nicht durch algorithmisch gesteuerte Reaktionsschemata ersetzt oder vollständig abgebildet werden kann. Die Annahme, menschliches Handeln ließe sich vollständig behavioristisch erklären, verkennt die normative Struktur unseres Selbstverständnisses als verantwortliche Akteure. Menschliches Verhalten ist nicht lediglich funktional, sondern stets auch bedeutungsvoll – es verweist auf Gründe, die verstanden und bewertet werden können.

Aus dieser Perspektive ergibt sich eine Konzeption von Verantwortung, die wesentlich auf Autonomie, normativer Reflexion und kommunikativer Rechenschaft basiert. Sie betont das dialogische Moment im moralischen Handeln und versteht Verantwortung als Teil eines sozialen Gefüges, in dem das Erklären und Rechtfertigen menschlicher Handlungen eine zentrale Rolle spielt.

Die Verantwortung des*der Einzelnen steht im Kontext sozialer Rollen und vorangegangener Handlungen: Menschen in der Forschung haben eine andere Verantwortung als Software-Entwickler, die u.a. für Transparenz und (fehlerfreie) Funktionalität zu sorgen haben. Menschen, die Software zum Einsatz bringen und Rahmenbedingungen schaffen, stehen anders in der Verantwortung als Anwender*innen. Es geht also um ein komplexes Geflecht, in dem Verantwortungszuschreibung eine Herausforderung ist.

Die Verantwortung eines Menschen ist eingeschränkt durch das, was sie weiß und zudem durch das, was ihrer Kontrolle zugänglich ist. Reine Zufälle, die sich einer absichtlichen Steuerung entziehen, können einer Person nicht zugerechnet werden. Daraus ergibt sich ein differenziertes Verständnis von Verantwortung, das nicht konsequentialistisch ist, sondern intentionalistisch (abhängig von Absicht und Kontrolle), deontologisch und epistemisch. Die Verantwortung ergibt sich aus der Fähigkeit zur Handlung, zur Rollenreflexion und zur begründeten Entscheidung – nicht aus der bloßen Optimierung von Ergebnissen.

Diese Perspektive führt zu einem gradualistischen Modell von Verantwortung: Sie ist nicht absolut oder binär, sondern gestuft. Die Zuschreibung von Verantwortung hängt von individuellen Fähigkeiten, situativen Bedingungen und normativen Erwartungen ab. Kinder, emotional instabile Personen oder Betrunkene gelten etwa nur eingeschränkt als verantwortlich, weil sie ihre Handlungen nicht vollständig kontrollieren können. Gleichzeitig können auch nicht-menschliche Akteure – wie Unternehmen oder autonome Systeme – unter bestimmten Bedingungen Verantwortung tragen, wenn ihnen Handlungskompetenz zugesprochen wird.

Graduelle Autonomie künstlich intelligenter Systeme mündet im Konzept der graduellen Verantwortung. Nida-Rümelin & Battaglia stellen u.a. einen Ansatz vor, der drei Konstellationen der Mensch-Maschine-Beziehung unterscheidet. *Instrumentelle Konstellationen*: Maschinen sind bloße Werkzeuge, Verantwortung liegt klar beim Menschen. *Personalisierte Konstellationen*: Maschinen verschmelzen mit menschlicher Identität (z. B. Prothesen, persönliche Daten) und verändern das Verständnis von Verantwortlichkeit. *Autonome Konstellationen*: Maschinen sind potenzielle

Interaktionspartner auf Augenhöhe, mit möglichen Implikationen für ihre rechtliche und moralische Subjektstellung.¹⁹⁹

„Sobald die Maschinen an der sozialen Interaktion des Gebens und Nehmens von Gründen teilnehmen können, gibt es keinen Grund, sie von der Verantwortung auszuschließen. Zurzeit sind Zuschreibungen von Verantwortung im Rahmen von einer gradualistischen Darstellung von Verantwortung möglich. Entscheidend ist jedoch, dass diese maschinellen Verantwortungszuschreibungen unumgänglich auf personelle Akteure reduzierbar sind.“

In der Konsequenz bleibt Verantwortung grundsätzlich beim Menschen, doch die Kriterien ihrer Zuschreibung werden durch technologische Komplexität unklar.

3.3.2.3 Transparenz und Nachvollziehbarkeit als Voraussetzung

Verantwortung setzt voraus, dass Menschen ihre Entscheidungen nachvollziehen, bewerten und begründen können – gegenüber sich selbst und anderen. Sie müssen in der Lage sein, Rechenschaft abzulegen, indem sie Gründe für ihr Handeln anführen und Gegenargumente aufnehmen. Und sie sollten darlegen können, warum es genau diese Entscheidung war und nicht eine andere. In dem Maß, in dem Entscheidungen an Künstliche Intelligenz delegiert werden, stellt sich die Frage, ob und wie diese Voraussetzungen gewahrt bleiben.

Ein selbstbestimmter Mensch muss in der Lage sein, Entscheidungen zu reflektieren und gegebenenfalls zu hinterfragen. Wenn jedoch die Entscheidungsprozesse von KI-Systemen intransparent oder zu komplex sind, wird diese Reflexion erschwert oder unmöglich. Das betrifft nicht nur das Verstehen der konkreten Ergebnisse, sondern auch der zugrunde liegenden Kriterien, Daten und Bewertungsmaßstäbe – etwa bei maschinelltem Filtern, Priorisieren oder Klassifizieren.

In solchen Fällen kann der Mensch vom aktiven Entscheider zum passiven Empfänger automatisierter Resultate werden. Damit werden sowohl Selbstbestimmung als auch die Fähigkeit zur Verantwortungsübernahme beeinträchtigt. Denn wer die Gründe für sein Handeln weder erkennt noch vertreten kann, ist auch moralisch schwerlich verantwortlich zu machen. Opake Systeme untergraben so nicht nur das Vertrauen in Technologie,

¹⁹⁹ Der Ansatz geht zurück auf Kersten, „Die maschinelle Person – Neue Regeln für den Maschinenpark?“

sondern berühren zentrale Voraussetzungen für ethisches Handeln: Einsicht, Urteilskraft und die Fähigkeit zur Rechenschaft.

Um einem Missverständnis vorzubeugen: Es geht nicht um die implizite Forderung, KI-Systeme bis in die feinsten Verzweigungen der systeminternen IT-Logik für alle transparent zu machen. Das wäre für Anwender*innen ohnehin schwer oder gar nicht nach nachzuvollziehen. Aber aus einer gesellschaftlichen und politischen Perspektive ist es mehr als problematisch, wenn weltumspannende Systeme mit opaken und komplexen Modellen operieren, die sich der Kontrolle – auch der Fachwelt – entziehen. Nicht zuletzt wirft die Unabsehbarkeit der Ergebnisse von KI-Entscheidungen die Frage auf, ob ihr Einsatz in kritischen Bereichen, in denen es zu psychischen oder physischen Schäden kommen kann, ethisch vertretbar ist.²⁰⁰

Kontrolle

Im Hinblick auf die Kontrolle und Steuerung von KI-Systemen lassen sich drei Ebenen unterscheiden: normative, strategische und operative Kontrolle.²⁰¹ Opazität und Komplexität der Systeme machen Kontrolle zur Illusion.

Die *normative Kontrolle* betrifft die oberste Ebene der Mensch-Technik-Interaktion und steht in engem Zusammenhang mit dem Konzept menschlicher Autonomie. Hierbei geht es um die ethische Bewertung der Systeme: Es wird gefragt, was moralisch geboten, erlaubt oder verboten ist. Diese Ebene der Kontrolle hat eine kritische Funktion, indem sie potenzielle Gefahren für die Autonomie des Menschen thematisiert, etwa Entmündigung oder den Verlust personaler Integrität. Solchen Risiken entgegenzuwirken, gehört nach bestimmten ethischen Ansätzen zu den vollkommenen Pflichten, die unbedingt zu erfüllen sind.²⁰²

Auf einer zweiten Ebene ist die *strategische Kontrolle* angesiedelt. Sie bezieht sich auf die Fähigkeit, Störungen vom System fernzuhalten und dessen Funktionalität aufrechtzuerhalten. Wesentliche Instrumente hierbei sind Rückkopplungsmechanismen im Sinne der klassischen Regelungstechnik. Hierbei wird die Differenz zwischen Ist- und Sollwert beim Output eines Systems ermittelt und als Steuerungsimpuls genutzt, um das

²⁰⁰ Grunwald und Hillerbrand, *Handbuch Technikethik*, 426.

²⁰¹ Gransche und Institut für Systemtechnik und Innovationsforschung, *Wandel von Autonomie und Kontrolle durch neue Mensch-Technik-Interaktionen*, 45.

²⁰² Gransche und Institut für Systemtechnik und Innovationsforschung, *Wandel von Autonomie und Kontrolle durch neue Mensch-Technik-Interaktionen*, 46.

Verhalten des Systems anzupassen. Strategische Kontrolle ermöglicht es somit, auf das System als Ganzes einzuwirken, ohne notwendigerweise seine inneren Prozesse vollständig zu verstehen.²⁰³

Die *operative Kontrolle* bildet schließlich die dritte Ebene. Sie bezieht sich auf die unmittelbare Steuerung einzelner Prozesse oder Handlungen innerhalb eines Systems. Hierbei geht es darum, Eingriffe vorzunehmen, um bestimmte Steuerungen zu korrigieren, zu verstärken, abzuschwächen oder vollständig zu unterbinden, wenn sich zeigt, dass die gewählte Maßnahme nicht zielführend ist. Operative Kontrolle betrifft somit konkrete Entscheidungen über die Art und Weise der Systemnutzung und die unmittelbare Einflussnahme auf technische Abläufe.²⁰⁴

Verantwortungslücken

Mit dem Einsatz autonomer oder teilautonomer KI-Systeme verschieben sich grundlegende Bedingungen menschlicher Verantwortlichkeit. Santoni de Sio und Mecacci identifizieren vier systematisch voneinander unterscheidbare, aber miteinander verbundene Formen von Responsibility Gaps. Diese Lücken entstehen nicht nur durch künstliche Autonomie, sondern auch durch gesellschaftliche, organisatorische und normative Unklarheiten.

Culpability Gap – Schuld ohne Zurechenbarkeit: Traditionell basiert Schuld auf Kriterien wie Absicht, Wissen und Kontrolle. In automatisierten Entscheidungssystemen wird jedoch zunehmend unklar, wer diese Bedingungen erfüllt – insbesondere, wenn Systeme auf selbstlernenden Algorithmen basieren. Die Folge: Schäden können entstehen, ohne dass eine konkrete Person eindeutig verantwortlich gemacht werden kann. Es fehlt an klassisch zurechenbarer Schuldhaftigkeit. Das gilt insbesondere dann, wenn Verantwortung analog der Rollenverteilung übernommen wird.

Moral Accountability Gap – Handeln ohne Erklärbarkeit: Moralische Verantwortung verlangt, dass man seine Handlungen gegenüber anderen begründen kann. Wenn jedoch Entscheidungen durch komplexe, nicht nachvollziehbare KI-Systeme getroffen werden, sind menschliche Akteure oft nicht in der Lage, die Gründe für ein Ergebnis zu erklären

²⁰³ Gransche und Institut für Systemtechnik und Innovationsforschung, *Wandel von Autonomie und Kontrolle durch neue Mensch-Technik-Interaktionen*, 46.

²⁰⁴ Gransche und Institut für Systemtechnik und Innovationsforschung, *Wandel von Autonomie und Kontrolle durch neue Mensch-Technik-Interaktionen*, 48.

– etwa bei algorithmischer Diskriminierung. Es entsteht eine Lücke im moralischen Rechenschaftsverhältnis.

Public Accountability Gap – Institutionen ohne Transparenz: Im politischen und institutionellen Raum wird Verantwortung häufig auf viele Beteiligte verteilt. Wenn staatliche oder wirtschaftliche Akteure KI-Systeme einsetzen, ohne klare Rechenschaftsstrukturen zu etablieren, ist nicht mehr nachvollziehbar, wer auf welcher Ebene Entscheidungen verantwortet. Das gefährdet demokratische Kontrollmechanismen.

Active Responsibility Gap – Verantwortung ohne Verpflichtung: Viele gesellschaftliche Herausforderungen – wie Gerechtigkeit, Sicherheit oder Nachhaltigkeit – erfordern proaktive Verantwortung. Doch weder Entwickler noch Entscheidungsträger fühlen sich systematisch verpflichtet, KI-Systeme an solchen Zielen zu orientieren. Es fehlt an gemeinsam anerkannten Normen, die aktive Verantwortung in Design, Implementierung und Anwendung integrieren.

Fazit – Verantwortung neu denken: Die Autoren schlagen vor, Verantwortung nicht auf eine einzelne Instanz (z. B. Entwickler oder Nutzer) zu konzentrieren, sondern als verteilte Praxis zu begreifen. Mit dem Konzept der Meaningful Human Control plädieren sie für eine Gestaltung von KI-Systemen, die menschliche Urteilskraft, Einflussnahme und normatives Bewusstsein strukturell sichert. Die Verantwortung für KI ist damit nicht eliminierbar, wohl aber neu zu organisieren.

Reichweite der Verantwortung

Nicht alle KI-Entscheidungen haben eine besondere Relevanz oder sind von einer besonderen Reichweite. Im Bereich der *Public Accountability* werden nicht selten Entscheidungen von großer Reichweite getroffen, die eine außerordentliche Verantwortung nach sich ziehen. Ein Beispiel aus Vor-KI-Zeiten: Es soll wohl ein offenes Geheimnis gewesen sein, dass der amerikanische General McArthur zu Beginn des Korea-Krieges, eine Reihe riskanter Maßnahmen in Erwägung zog, die unter Umständen einen Dritten Weltkrieg zur Folge hätten haben können. Deren Evaluierung wurde nicht etwa von einem Beraterstab oder anderen berufeneren Menschen vorgenommen, sondern von einem „Electric Brain“, einer Maschine, von der angenommen wurde, dass sie mit größerer Objektivität urteilen würde als Menschen das hätten tun können. Man delegierte,

«über-antwortete» die Verantwortung an einen Apparat. Dem menschlichen Gehirn wurde weniger zugetraut als dem technischen „Gehirn“. ²⁰⁵

Der Mensch selbst war es, „*der sein Gewissen und seine Entschlussfreiheit auf dem Altar des Apparates niederlegte.*“ Für G. Anders kommt es einer Selbstdegradierung gleich, wenn der Mensch seine Verantwortung auf eine Maschine transferiert. Aber warum degradiert der Mensch sich selbst?

Kleiner Exkurs

Als ChatGPT einer breiten Öffentlichkeit zugänglich wurde, mag das bei vielen Menschen Begeisterung und Faszination ausgelöst haben – nicht selten gemischt mit einem leisen Unbehagen: „*Mein Gott, was die kann!*“ – Dieser Ausruf, den G. Anders dem Beobachter einer frühen „*computing machine*“ zuschreibt, bringt ein Gefühl auf den Punkt, das bis heute aktuell geblieben sein könnte. G. Anders spricht von *prometheischer Scham*, die eine Reaktion des Menschen auf seine ihm technisch überlegenen Maschinen sei. In einer Zeit, in der KI-Systeme nicht nur Texte verfassen, sondern kreative Prozesse übernehmen, medizinische Diagnosen stellen oder politische Entscheidungen vorbereiten, stellt sich die Frage nach dem Selbstverständnis des Menschen erneut. „*Wer bin ich schon?*“ – fragt Anders’ moderner Prometheus, der sich selbst zum „*Hofzwerg seines eigenen Maschinenparks*“ degradiert hat.

Diese Selbstverkleinerung vor den eigenen Schöpfungen ist jedoch nicht nur psychologisch bedeutsam, sondern weist auf ein tiefer liegendes anthropologisches Problem hin: das von Anders so genannte *prometheische Gefälle*. Ein zentrales Merkmal der heutigen menschlichen Existenz besteht in der Unfähigkeit, seelisch mit der Geschwindigkeit des eigenen technischen Fortschritts Schritt zu halten. Der Mensch gerät zunehmend aus dem Takt mit den von ihm selbst hervorgebrachten Produkten, die sich in rasantem Tempo wandeln und in eine Zukunft vorausgreifen, der das Subjekt kaum noch zu folgen vermag. Diese strukturelle Asynchronität zwischen Mensch und Technik, die auf die unbegrenzte „prometheische“ Freiheit zurückzuführen ist, immer Neues zu schaffen, führt zu einem Zustand, in dem der Mensch als verspäteter Nachzügler seiner eigenen technischen Schöpfungen erscheint – geplagt von dem Gefühl der eigenen Antiquiertheit und entfremdet von seiner künstlich geschaffenen Umwelt.

²⁰⁵ Anders, *Die Antiquiertheit des Menschen*. I, 76.

„Die Tatsache der täglich wachsenden A-synchronisiertheit des Menschen mit seiner Produktewelt, die Tatsache des von Tag zu Tag breiter werdenden Abstandes, nennen wir «das prometheische Gefälle»“²⁰⁶

Dieses wachsende Missverhältnis zwischen menschlicher Zeitstruktur und der Geschwindigkeit technischer Entwicklung führt zu einem Gefälle, einer Asymmetrie zwischen Menschlichem und Technischem Vermögen. Asymmetrie ist ein Strukturmerkmal der Macht.

3.3.3 Epistemische Herausforderungen

Die Expertengruppe der Europäischen Union spricht im Titel ihrer Ethik-Leitlinien von einer *vertrauenswürdigen KI*, die mit ethischen Normen im Einklang steht und robust im weitesten Sinne ist.²⁰⁷ Vertrauenswürdigkeit und Zuverlässigkeit sind eine notwendige Basis, wenn wir auf das Wissen und die Entscheidungen von Künstlicher Intelligenz bauen. Komplexität und Intransparenz können die Nutzung von KI-Systemen zu einer epistemischen Herausforderung machen. Und nicht zuletzt die Geschwindigkeit der Systeme bringt das Vertrauen in die Faktizität ins Wanken.

3.3.3.1 Von der Faktizität zum Remix

Das Wissen der Künstlichen Intelligenz ist kein begründetes Wissen. Kausalität wird abgelöst durch ein approximatives Trial-And-Error-Verfahren, durch Vorhersage und Korrelation. Das Ergebnis muss nicht verstanden werden. Welche Bedeutung hat Wahrheit in diesem Zusammenhang noch? Han spricht von einer „Krise der Wahrheit“ und konstatiert, dass sich ein neuer Nihilismus ausbreite.

„Er entsteht dort, wo wir den Glauben an die Wahrheit selbst verlieren. Im Zeitalter von Fake News, Desinformation und Verschwörungstheorie kommt uns die Realität mit Tatsachenwahrheiten abhanden.“²⁰⁸

Reichl spricht in ähnlicher Weise von einer „Erosion der Wahrheit“ und resümiert, dass

„wir über kurz oder lang davon ausgehen müssen, dass nichts von dem, was uns auf Smartphone und Bildschirm gezeigt wird noch allzu viel

²⁰⁶ Anders, *Die Antiquiertheit des Menschen. I*, 28–29.

²⁰⁷ Hochrangige Expertengruppe für künstliche Intelligenz., „Ethik-leitlinien für eine vertrauenswürdige KI.“, 8.

²⁰⁸ Han, *Infokratie*, 64.

mit der Wirklichkeit zu tun haben muss. So wird die Wirklichkeit zum Wonderland“.

Mit der Digitalisierung geht auch der Glaube an die Faktizität verloren.²⁰⁹ Lange Zeit unterlag die Faktizität einer Fotografie keinerlei Zweifel. Analoge Fotos galten als Beweis für das, was ist oder gewesen ist. Spätestens seit es Deep Fake gibt, wird uns klar, dass Fotografie und Film nichts mehr mit Wahrheit zu tun haben müssen. Der Glaube auch an dokumentarische Werke erodiert wie die Wahrheit und gehört bald der Vergangenheit an. Digitalität und Faktizität passen nicht zueinander.

Walter Benjamin erkannte den Verlust der „Aura“ des Kunstwerks im Zeitalter technischer Reproduzierbarkeit, die mit einem Verlust an Echtheit einhergeht.

„Das Hier und Jetzt des Originals macht den Begriff seiner Echtheit aus. Analysen chemischer Art an der Patina einer Bronze können der Feststellung ihrer Echtheit förderlich sein; entsprechend kann der Nachweis, dass eine bestimmte Handschrift des Mittelalters aus einem Archiv des fünfzehnten Jahrhunderts stammt, der Feststellung ihrer Echtheit förderlich sein. Der gesamte Bereich der Echtheit entzieht sich der technischen [...] Reproduzierbarkeit.“²¹⁰

Die Faktizität einer analogen Fotografie entzieht sich nicht der Reproduzierbarkeit. Sie behält ihre Zeugenschaft. Aber sie entzieht sich der Digitalisierung. Im digitalen Kontext kann man von einem Wechsel des Aggregatzustands sprechen: Kulturgüter jeglicher Art werden zu Daten – beweglich, rekombinierbar, entgrenzt. Wissen wird fluide, fragmentarisch, seiner Einmaligkeit beraubt. Der Remix wird zur dominanten Wissenspraxis. Die Trennung zwischen Original und Kopie, Echtheit und Reproduktion, Faktizität und Fiktionalität wird obsolet. Der digitale Raum ermöglicht neue epistemische Konstellationen, jedoch um den Preis des Verlustes historischer Verortung und inhaltlicher Verlässlichkeit und Kontinuität.

Durch die Digitalisierung von allem und jedem entsteht eine neue, andere Qualität des Umgangs mit Kunst, Information, Wahrheit. Es geht nicht mehr allein um verloren gegangene Echtheit und die ins Wanken geratene geschichtliche Zeugenschaft. Es geht um eine sehr viel tiefgreifendere Form von „Fälschung“, die mit der „Zerlegung“ der

²⁰⁹ Han, *Infokratie*, 64.

²¹⁰ Benjamin, *Das Kunstwerk im Zeitalter seiner technischen Reproduzierbarkeit*, 14.

digitalisierten Dinge beginnt und in einen digitalen Remix mündet. Peter Glaser spricht von einer „Kulturellen Atomkraft“.

„Mit der Digitalisierung gehen immer mehr Dinge, die zuvor an bestimmte unaustauschbare Materialien gebunden waren, in einen neuen Aggregatzustand über. Kulturdinge im weitesten Sinn - aus Zeichenbrettern, Tonstudios, Fernsehern, Büchern, you name it - werden Daten. Diese digitale Substanz hat eine grundlegend neue Leichtigkeit. Die digitalen Dinge lassen sich ungleich leichter bewegen als zuvor, weltweit senden, empfangen, verändern, kopieren, mit anderen teilen, remixen. Nicht zuletzt das Remix-Phänomen weist auf den instabilen Zustand hin, in dem sich die ganze Entwicklung derzeit befindet. Remixt werden Fragmente. Filmstücke, Soundschnipsel, Splitter anderer Kulturobjekte, die zu neuen Formen zusammengesetzt werden. Musik, Texte, Bilder, Filme, aber auch modulare Software oder enzyklopädisches Wissen befinden sich in der digitalen Welt in einem Zustand latenter Zerlegung. Die althergebrachten kulturellen Molekülverbindungen - die komplexen Formen, die sie über Jahrhunderte angenommen haben - werden nun aufgeknackt, oder sie zerfallen von ganz alleine wieder in ihre Grundbestandteile.“²¹¹

Die digitale Leichtigkeit kennt keine Dauer, keine Konstanz und keinerlei Verbindlichkeit. Sie ist flüchtig. Unter diesen Bedingungen verliert der Mensch den Bezugsrahmen für seine Orientierung, Identität und Urteilskraft.

Wahrheit braucht Zeit

Information ist nicht von Dauer, hat eine eigene Zeitlichkeit. Sie ist „sehr kurz und kurzfristig“²¹². Sie zerfällt und wird in einem neuen Remix anders kontextualisiert. „Information ist schnell“ und „Wahrheit braucht Zeit“²¹³. Zeit, die in der digitalen Kommunikation keine Rolle mehr zu spielen scheint. Traditionellen Medien ist in ihrem festen Aggregatzustand Trägheit und Langsamkeit zu eigen. Ihr neuer, digitaler Zustand besitzt ganz andere Leichtigkeit. Die Daten-objekte sind ortlos, endlos kopierbar, jederzeit veränderbar. Ontologisch betrachtet verliert das Kulturring seine feste Identität und wird zum Prozess, zum Update, zur Version. Aus analogen Medien sind Daten

²¹¹ Glaser, „Kulturelle Atomkraft“.

²¹² Han, *Im Schwarm*, 56.

²¹³ Peter Glaser zitiert nach: Pörksen, „Die Hölle der Desinformation - Spielregeln der Wirklichkeitsordnung im digitalen Zeitalter“, 133.

geworden, die per Klick auf die Datenautobahn gebracht werden können und entfernte Adressaten nahezu ohne Zeitverlust erreichen. Die Möglichkeit der blitzschnellen Verbreitung von Informationen mündet in einem „*Geschwindigkeitsrausch der vernetzten Welt*“²¹⁴, der der Wahrheit kaum noch Raum lässt. Irgendwo auf der Welt passiert etwas Schreckliches und wir erfahren es nahezu zeitgleich. Auf sozialen Medien verbreiten sich Neuigkeiten viral und umso rasender, je furchtbarer sie sind. Es ist den Menschen schier unmöglich, mit der gleichen Geschwindigkeit das Ausmaß des Ereignisses angemessen zu erfassen und zu verstehen, noch den Wahrheitsgehalt der Nachricht zu beurteilen.²¹⁵

„Der «Airbag aus Zeit»²¹⁶, der notwendig ist, um der Wahrheit Genüge zu tun, scheint verloren gegangen. Diesen Zeitpuffer kennen diejenigen, die sich gelegentlich die Zeit nehmen, einen traditionellen Brief zu schreiben, ihn nicht sofort losschicken, noch einmal eine Nacht darüber schlafen, weil er doch wichtig ist. Am nächsten Tag landet der Brief womöglich im Papierkorb, weil der gestrige Ärger beim Schreiben doch zu unfairen Formulierungen oder halbgaren, respektlosen Behauptungen geführt hat. Wird er schließlich doch auf den klassischen Postweg gebracht, braucht er ein oder mehrere Tage, bis er ankommt.²¹⁷ Ein Zeitpuffer sorgt für Respekt und fairen Umgang miteinander.“²¹⁸

Die Geschwindigkeit der Künstlichen Intelligenz verändert unseren Blick auf die Welt. Der Remix als Erkenntnisform stellt unsere klassische Wahrnehmung von Welt und unsere Vorstellung von Objektivität und Stabilität infrage. Nichts scheint von Dauer. Ferne und Entfernung spielen keine Rolle mehr. Arendt beschreibt in *Vita activa* die „*moderne Welt*“ als ein

„über die ganze Erde sich erstreckendes Kontinuum, aus dem Ferne und Entfernung vor dem Ansturm der Geschwindigkeit verschwunden sind. Die Geschwindigkeit hat den Raum erobert und würde ihn zu vernichten drohen, wenn ihrer noch immer wachsenden Beschleunigung nicht die für Körper unübersteigbare Grenze gesetzt

²¹⁴ Pörksen, „Die Hölle der Desinformation - Spielregeln der Wirklichkeitsordnung im digitalen Zeitalter“, 132.

²¹⁵ Pörksen, „Die Hölle der Desinformation - Spielregeln der Wirklichkeitsordnung im digitalen Zeitalter“, 130.

²¹⁶ Menasse, *Alles und nichts sagen*, 44.

²¹⁷ Menasse, *Alles und nichts sagen*, 42.

²¹⁸ Zander, „Wahrheit und Respekt im Geschwindigkeitsrausch sozialer Medien“, 2.

*wäre, an zwei verschiedenen Orten nicht gleichzeitig anwesend zu sein können.*²¹⁹

Die Geschwindigkeit der Künstlichen Intelligenz und ihr Remix der digitalisierten Objekte verändert nicht nur unseren Glauben an Wahrheit; Alles ist flüchtig: Es kann so oder auch anders sein. Es nimmt darüberhinausgehend Einfluss auf unseren Bezug zur Welt.

*„Es liegt im Wesen des menschlichen Vermessungs-Vermögens, daß es sich, um überhaupt zu funktionieren, vorerst aus der Nähe selbst gleichsam zurückziehen, daß eine Entfernung zwischen sich und das zu Messende legen muß. [...] jede Verringerung von Entfernung auf der Erde kann nur um den Preis einer vergrößerten Entfernung von der Erde gewonnen werden, also um den Preis einer entscheidenden Entfremdung des Menschen von seiner unmittelbaren irdischen Behausung.“*²²⁰

Damit das menschliche „Vermessen“ funktioniert, braucht der Mensch Distanz zu dem, was er „messen“ und erkennen will. Künstliche Intelligenz kennt keine Distanz, die jedoch notwendig ist, um die Wahrnehmung von Dingen in ein Verstehen zu überführen. Technische Beschleunigung verringert zwar räumliche und zeitliche Distanzen, ersetzt aber dabei oft das „Dazwischen“ – den menschlichen Raum des Fragens, Zweifelns, Zögerns. Der Preis ist die Entfremdung des Menschen.

3.3.3.2 Bias

Bias bedeutet Verzerrung und Voreingenommenheit. Menschen sind voreingenommen, sehen die Welt aus einer bestimmten Perspektive, oft unbewusst, positiv wertend oder auch negativ, lernend und offen für andere Sichtweisen oder beharrend. Bias können in Stereotypen zum Ausdruck kommen, also in vereinfachten Vorstellungen über Personen oder Gruppen von Menschen (Schubladendenken). Oder sie führen zu Vorurteilen mit meist negativen Bewertungen und ohne haltbare Begründung. Denkmuster, Prädispositionen und kognitive Verzerrungen prägen unser Zusammenleben seit jeher.

Biases sind eingebettet in institutionelle und gesellschaftliche Strukturen der Wissensproduktion.²²¹ Künstliche Intelligenz ist zu einer mächtigen Instanz der

²¹⁹ Arendt, *Vita activa oder Vom tätigen Leben*, 356.

²²⁰ Arendt, *Vita activa oder Vom tätigen Leben*, 358.

²²¹ o.V., *Epistemology Of Bias*.

Wissensproduktion geworden. Ihren Algorithmen hängt der Mythos der Objektivität und Neutralität, der Leidenschafts- und Gefühllosigkeit an. Viele Menschen neigen dazu, den Ergebnissen algorithmischer Entscheidungen mehr zu trauen als menschlichen Entscheidungsträgern – ein Phänomen, das als Automation Bias bekannt ist. Dabei sind die Systeme weit entfernt von Objektivität. Die Daten und Informationen, aus denen sie ihr „Wissen“ generieren sind keineswegs frei von Verzerrungen, die über das Training der Systeme implementiert werden. Daten spiegeln bestehende gesellschaftliche Diskriminierungen wider. Unter Labeling Bias versteht man fehlerhafte oder voreingenommene Annotationen durch Menschen. Verzerrungen entstehen auch durch die Art, wie Algorithmen Daten verarbeiten. Beispiele: Ungerechte Gewichtungen bestimmter Variablen und Verstärkung bereits bestehender Muster (Feedback Loops). Zudem entstehen Bias durch die Interaktion von User*innen mit Systemen. Beispiel: Nutzer*innen stellen tendenziell bestimmte Fragen, wodurch Chatbots bestimmte Themen überrepräsentiert behandeln.

Bias in künstlich intelligenten Systemen können vorhandene Voreingenommenheit verstärken oder infrage stellen. Und sie können Meinungen, Vorurteile oder politische Einstellungen erzeugen und prägen.

3.3.3.3 Halluzinationen, Bullshit und Gleichgültigkeit

Wer ChatGPT fragt: *Können Sprachmodelle wie ChatGPT lügen?* bekommt in etwa folgende Antwort:

„Sprachmodelle wie ChatGPT können nicht lügen, weil sie keine eigenen Überzeugungen oder Absichten haben. Lügen setzt voraus, dass jemand bewusst die Wahrheit kennt und sie absichtlich verdreht, um andere zu täuschen. Ein Sprachmodell hat aber weder ein eigenes Wissen noch eine Intention. Stattdessen generiert es Texte auf Basis von Wahrscheinlichkeiten und Mustern aus seinem Trainingsmaterial.“²²²

Gehen wir davon aus, dass die KI in diesem Fall tatsächlich nicht gelogen hat, heißt das aber noch lange nicht, dass sie wahrheitsgemäß geantwortet hat. Es ist im Kern auch zutreffend, dass KI-Systeme aus sich heraus keine Intention besitzen und für Lüge ebenso wenig wie für Wahrheit keine Kriterien zur Verfügung haben. *„Wer nicht lügen kann, weiß nicht was Wahrheit ist“²²³*. Wenn das so ist, wie geht denn ChatGPT mit der

²²² OpenAI, „ChatGPT 4.0“, Abfrage vom 30.01.2025.

²²³ „Nietzsche Source — Digitale Kritische Gesamtausgabe Werke und Briefe (eKGWB), Za-IV“, §9.

Wahrheit um. Wenn ChatGPT Fehler und Ungenauigkeiten generiert, wird davon gesprochen, dass ChatGPT halluziniere. Wer eine Halluzination hat, unterliegt fachsprachlich einer Sinnestäuschung.²²⁴ Im übertragenen Sinne bezieht sie sich auf KI-Systeme und insbesondere auf Sprachmodelle, die überzeugend formulierte aber faktisch falsche Aussagen liefern. Peter Reichl spricht gar von immerwährender Halluzination („*ChatGPT halluziniert immer und aus Prinzip Punkt!*“²²⁵) und von „würfelnder KI“. Und es gibt kritische Stimmen unter Expert*innen, die ChatGPT als einen „stochastischen Papageien“ bezeichnen.²²⁶ Man kann den Begriff zutreffend finden oder nicht. Papageien verfügen allerdings über eine beachtliche Intelligenz.

Schon lange bevor es ChatGPT gab, hat sich Harry G. Frankfurt mit dem Begriff *Bullshit* befasst. Wer Unsinn in die Welt setze, lehne die Autorität der Wahrheit nicht ab wie der Lügner, der sich ihr entgegenstelle. Sie sei ihm vielmehr gleichgültig. „*He pays no attention to it all.*“²²⁷ Frankfurt ist überzeugt, dass Bullshit ein größerer Feind der Wahrheit ist als Lügen. Niemand kann lügen, ohne zu glauben, die Wahrheit zu kennen. Jemand, der Bullshit verbreite, brauche eine solche Überzeugung nicht. Bullshit ist immer dann unvermeidlich, wenn sich jemand verpflichtet fühlt oder die Gelegenheit dazu hat, sich zu einem Thema zu äußern, das ihr oder sein Wissen zu diesem Thema übersteigt.²²⁸ Seine Aufmerksamkeit richte sich nicht auf die Wahrheit.²²⁹ Die Essenz von Bullshit aber ist nach Auffassung von Frankfurt die Gleichgültigkeit gegenüber der Realität

*„It is just this lack of connection to a concern with truth – this indifference to how things really are – that I regard as of the essence of bullshit“*²³⁰

Dieser Gedanke wird neuerdings in der Kritik an KI-Systemen wie ChatGPT aufgegriffen. Auch ChatGPT produziere keine wahren Texte. Denn das habe das System nicht im Angebot. Es ist dazu nicht in der Lage, weil es seine Texte nicht auf Wahrheit oder Falschheit zu überprüfen vermag. Es muss sich auf Wahrscheinlichkeiten verlassen. Und darum produziere ChatGPT Bullshit. Nun scheint es nicht angemessen alles, was

²²⁴ Pauleikhoff, „Halluzination“.

²²⁵ Reichl, *Homo cyber*, 42.

²²⁶ Voß, „Was es statt eines KI-Moratoriums braucht“.

Dem Beitrag zufolge hat Timnit Gebru Google als Leiterin der KI-Ethik im Streit um den Begriff des stochastischen Papageien verlassen.

²²⁷ Frankfurt, *On Bullshit*, 61.

²²⁸ Frankfurt, *On Bullshit*, 70f.

²²⁹ Frankfurt, *On Bullshit*, 55.

²³⁰ Frankfurt, *On Bullshit*, 33.

ChatGPT produziert als reinen Unsinn abzutun. Denn ist es nicht so, dass wir erstaunt sind, verblüfft sind, wozu das System in der Lage ist? Dass wir auf den ersten Blick glauben, was es uns präsentiert und wir vielleicht sogar unser eigenes Wissen in Zweifel ziehen? Und gibt es nicht auch Texte, die wir für wahr halten, an denen wir nicht zweifeln? Blanker Unsinn könnte uns nicht so leicht zu einer solchen Haltung bewegen. Denn viele Aussagen sind sinnvoll. ChatGPT ist weder ausgelegt, zu lügen, noch die Wahrheit zu vermitteln, aber seine Aussagen werden uns in einer sehr überzeugenden Weise präsentiert. Man ist sich weitgehend einig, dass die von ihm produzierten Sätze (normalerweise) sinnvoll sind.

Auch wenn es sich um offensichtlichen Unsinn handelt, lässt das System an seinem Sinn und seiner Richtigkeit keinen Zweifel. Wenn Unsinn überzeugend dargestellt und immer aus Neuem wiederholt wird, entsteht ein Klima epistemischer Unsicherheit und Verwirrung. Wir wissen nicht mehr, was wahr und was falsch ist.

„ChatGPT functions not to convey truth or falsehood but rather to convince the reader of – to use Colbert’s apt coinage – the truthiness of its statement, and ChatGPT is designed in such a way as to make attempts at bullshit efficacious (in a way that pens, dictionaries, etc., are not). So, it seems that at minimum, ChatGPT is a soft bullshitter: if we take it not to have intentions, there isn’t any attempt to mislead about the attitude towards truth, but it is nonetheless engaged in the business of outputting utterances that look as if they’re truth-apt.“²³¹

Wenn Begriffe wie „Alternative Fakten“ Einzug in die Sprache halten und salonfähig werden, wird die Verunsicherung umso größer. „Der amerikanische Philosoph Harry Frankfurt würde Trump heute als »Bullshitter« bezeichnen.“²³²

3.3.3.4 Filterblasen und Echokammern

Empfehlungssysteme, Suchmaschinen und einiges mehr in der alltäglichen KI-Nutzung erleichtern unseren Alltag, „gleichzeitig aber fördern sie die Tendenz, im Eigenen zu verbleiben.“²³³

²³¹ Hicks u. a., „ChatGPT Is Bullshit“, 6.

²³² Han, *Infokratie*, 67.

²³³ Verständig u. a., *Algorithmen und Autonomie*, 11.

Epistemische Personalisierung – Filterblasen

Unter dem Begriff der Filter Bubble ist ein Phänomen bekannt, das hier als Beispiel personalisierter Wahrheit herangezogen wird. Das Wissen, das Systeme wie Google uns präsentieren, ist ein personalisiertes Wissen, ein Wissen, das auf jede und jeden von uns zugeschnitten ist. Als Eli Pariser den Begriff der „Filter Bubble“ einführte, erzählte er, dass er als politisch progressiver Facebook-User eines Tages keine konservativen Feeds mehr angezeigt bekam. Stillschweigend hatten die Facebook-Algorithmen dafür gesorgt, dass er nur noch das zu sehen bekam, was ihm aufgrund seiner bisherigen Likes und Kommentare zu gefallen schien. Alles andere hatten die algorithmischen Filter ausgeblendet. Er befand sich in seiner höchstpersönlichen Filter Bubble. Den Begriff der „Filter Bubble“ definiert er so:

„Ihre Filterblase ist Ihr ganz persönliches einzigartiges Informationsuniversum, in dem Sie online leben. Und was in Ihrer Filterblase ist, hängt davon ab, wer Sie sind und was Sie tun. Aber Sie bestimmen nicht, was hineinkommt. Und, noch wichtiger, Sie sehen nie, was aussortiert wird“²³⁴

Was in die Blase kommt und was draußen bleibt, bestimmen wir nicht selbst. Und wir wissen nicht, was wir nicht zu sehen bekommen. Darüber entscheidet die künstliche Intelligenz für uns. Sie befindet darüber, was für uns – entsprechend unserer Vorlieben und Abneigungen– gut ist und was nicht, was relevant ist und was nicht. Wir tragen zum Funktionieren der Bubble bei.

„Die neue Generation der Internetfilter schaut sich an, was Sie zu mögen scheinen – wie Sie im Netz aktiv waren oder welche Dinge oder Menschen Ihnen gefallen – und zieht entsprechende Rückschlüsse. Prognosemaschinen entwerfen und verfeinern pausenlos eine Theorie zu Ihrer Persönlichkeit und sagen voraus, was Sie als nächstes tun und wollen. Zusammen erschaffen diese Maschinen ein ganz eigenes Informationsuniversum für jeden von uns – das, was ich die Filter Bubble nenne – und verändern so auf fundamentale Weise, wie wir an Ideen und Informationen gelangen.“²³⁵

²³⁴ Pariser, „Eli Pariser“, Min. 4:16.

²³⁵ Pariser, *Filter Bubble*, 17.

Vieles bleibt im Dunklen. KI-Systeme sind opake Systeme. Hat das nicht auch etwas für sich? Würden wir nicht untergehen in der Informationsflut, der wir ausgesetzt sind, sobald wir dem Ruf der digitalen Begleiter folgen. Wir sammeln Informationen und gehen darin unter. Wie in einer Bibliothek ungelesener oder unverstandener Bücher. Wir glauben informierter, wissender und klüger zu sein, glauben bessere Entscheidungen treffen zu können. Es gibt ein Zuviel an Information, was nicht guttut, was krank machen kann (IFS – Information-Fatigue-Syndrom). Also ließe sich doch sagen, dass algorithmische Filter sehr gut seien für die individuelle psychische Hygiene. Und dass sie sehr gut seien für die gezielte Reduzierung der nicht zu bewältigenden Informationsmenge, dass sie der Lähmung analytischer Fähigkeiten vorbeugt.

Algorithmen als Gatekeeper

Pariser hebt noch einen anderen Aspekt hervor, der einen fundamentalen Aspekt der digitalen Interaktion berührt. Sie ist ungeregt, es gibt keine menschlichen Redakteure wie in der massenmedialen Kommunikation. Die Fackel der menschlichen Torwächter ist in der digitalen Kommunikation an die Algorithmen übergeben worden. Die entscheiden darüber, was wir sehen und vor allem, was sie uns nicht zeigen. Und damit haben sie einen erheblichen Einfluss auf unsere Meinungsbildung, auf unser Wissen und unser moralisches Urteilsvermögen.

„Was das heißen könnte, ist, dass wir die Sache mit dem Internet vielleicht falsch verstanden haben. In einer Sendergesellschaft -- so geht der Gründungsmythos -- in einer Sendergesellschaft gab es Torwächter, die Redakteure, die den Informationsfluss kontrolliert haben. Und dann kam das Internet und hat sie aus dem Weg gefegt, und uns allen ermöglicht, uns miteinander zu verbinden, und es war wunderbar.“²³⁶

„Aber das ist nicht, was gerade passiert. Was wir erleben, ist eher eine Übergabe der Fackel von menschlichen Torwächtern an die Algorithmen. Nur dass den Algorithmen noch keine Ethik eingebaut wurde, wie sie die Redakteure haben. Wenn also Algorithmen die Welt für uns kuratieren, wenn sie entscheiden, was wir zu sehen bekommen und was nicht, dann müssen wir sicherstellen, dass sie nicht nur nach Relevanz auswählen. Wir müssen garantieren, dass sie uns auch Dinge

²³⁶ Pariser, „Eli Pariser“, ca. Min. 5:50.

zeigen, die unbequem oder herausfordernd oder wichtig sind -- andere Sichtweisen. „²³⁷

Eine wichtige Erkenntnis besteht darin, dass Nutzerinnen und Nutzer durch algorithmisch gesteuerte Datenanalyse und Inhaltsselektion zunehmend mit gleichförmigen Inhalten konfrontiert werden. Diese Gleichförmigkeit ergibt sich daraus, dass bevorzugt solche Inhalte ausgespielt werden, die mit den bereits bestehenden Meinungen, Interessen und Präferenzen der jeweiligen Person in hoher Übereinstimmung stehen. Inhalte hingegen, die alternative Perspektiven vertreten oder potenziell Widerspruch hervorrufen könnten – etwa, weil sie anderen Themen oder Meinungen zugeordnet sind –, bleiben durch die Wirkung algorithmischer Filtermechanismen häufig ausgeblendet und werden den Nutzerinnen und Nutzern nicht mehr zugänglich gemacht.²³⁸

Nun mag man einwenden, dass auch in traditionellen Medien, Redakteure und Journalisten als Gatekeeper fungieren, was eine gewisse Berechtigung hat. Jedoch seien Macht und Geheimhaltung (also Opazität) besonders bedrohlich:

„Certainly, newspapers and other traditional media technologies do place external filters on their readers, but the modern instantiation is particularly powerful and troubling. [...] The secrecy is particularly threatening. Many users do not know about the existence of algorithmic personal filtering. Even amongst those that do, most do not have access to the particularities of the algorithms. [...] Curiously, Sunstein notes the group polarization literature has thought relatively little about the impact of personal similarity. doing the filtering; thus, the very opacity of the process makes it harder for a user to successfully evaluate and epistemically compensate for such filtering (Miller and Record 2013). Thus, most users significantly underestimate the degree to which their exposure to information, via search results, has already been tailored to present search results to which the user will already be amenable. „²³⁹

Echokammern

Auch Menschen, die sich in Echokammern austauschen, verbleiben im Eigenen. Filterblasen und Echokammern werden gelegentlich synonym verwendet. Nguyen weist

²³⁷ Pariser, „Eli Pariser“, ca. Min. 6:20.

²³⁸ Suttmöller, „Big Data und die Frage nach der Gerechtigkeit“, 164.

²³⁹ Nguyen, „Echo Chambers and Epistemic Bubbles“, 6.

jedoch darauf hin, dass dadurch zwei unterschiedliche epistemische Phänomene verwischt werden. Er erläutert den Unterschied so:

„An epistemic bubble is a social epistemic structure in which other relevant voices have been left out, perhaps accidentally. An echo chamber is a social epistemic structure from which other relevant voices have been actively excluded and discredited. Members of epistemic bubbles lack exposure to relevant information and arguments.“²⁴⁰

Filterblasen zeichnen sich also durch Fehlen relevanter Information, Echokammern hingegen durch Ausschluss oder Diskreditierung von Informationen. Hinzu kommt, dass Echokammern durch Manipulation von Vertrauen entstehe.

„An epistemic bubble is an epistemic structure emerging from the informational architecture of communities, social networks, media, and other sources of information and argument. It is an impaired informational topology — a structure with poor connectivity. An echo chamber, on the other hand, is an epistemic structure created through the manipulation of trust;“²⁴¹

Echokammern sind geschlossene soziale Systeme, in denen sich Menschen gleicher Gesinnung und Überzeugung zusammenfinden. Andere Überzeugen und Ansichten kommen nicht zur Geltung, werden nicht zugelassen und prallen an undurchlässigen Grenzen ab, die Gleichgesinnten ein kohärentes Verhalten ermöglichen.

3.3.3.5 Epistemischer Anachronismus

Wissen ist nicht nur eine Frage des Inhalts, sondern auch der Zeit. In jeder Erkenntnisform – sei sie wissenschaftlich, politisch, alltäglich oder spirituell – ist Zeitlichkeit ein konstitutives Moment: Wissen entsteht, reift, wird geprüft, verworfen oder bestätigt. Es ist in menschliche Lebenszeit eingebettet, durchläuft Etappen von Zweifel, Diskurs, Vergewisserung und Revision. Klassisch gedacht ist Erkenntnis ein diachroner Prozess – eine Bewegung, die Vergangenheit, Gegenwart und Zukunft in ein sinnstiftendes Kontinuum bringt.

KI-Systeme, insbesondere große Sprachmodelle, operieren auf Grundlage vergangener Texte, historischer Sprachmuster und überlieferter Kategorien. Selbst wenn ihre

²⁴⁰ Nguyen, „Echo Chambers and Epistemic Bubbles“, 1.

²⁴¹ Nguyen, „Echo Chambers and Epistemic Bubbles“, 3.

Antworten aktuell erscheinen, speisen sie sich aus einem Wissensarchiv, das retrospektiv und nicht zukunfts offen ist. Mit jeder Verwendung solcher Systeme wächst die Gefahr eines epistemischen Lock-ins – eines Rückzugs auf vergangene Semantiken, Normen und Weltbilder, die sich selbst referenzieren und reproduzieren. Ein Bild dafür bietet der Mythos des Narziss: Die KI blickt nicht in die Welt, sondern in den Spiegel ihrer eigenen Trainingsdaten. Jede neue Antwort basiert nicht auf einer neuen Erfahrung oder kritischen Auseinandersetzung, sondern auf dem bereits Gesagten – nur neu arrangiert. Der epistemische Zeitpfeil, der auf Erkenntnisfortschritt und Horizonterweiterung gerichtet sein sollte, dreht sich in der Endlosschleife.²⁴²

KI-gestützte Systeme bedienen sich rekursiver Modelle: Sie generieren neues Wissen auf der Basis von bereits generierten Outputs. Das lässt sich sehr einfach bei Chatbots oder Suchmaschinen nachvollziehen, die sich oft auf bereits ausgegebene Antworten beziehen, die aufgegriffen und in einem passenden Kontext erneut verarbeitet werden. Diese Feedbackschleifen führen zu einem epistemischen Autotuning, bei dem sich Modelle immer stärker auf sich selbst beziehen und damit – epistemologisch betrachtet – zu einer hermetischen Selbstreferenzialität führen. Erkenntnis wird so zu einem geschlossenen Kreislauf algorithmischer Iteration.

Demokratie lebt davon, dass bisher Ungehörtes, Unsichtbares und Unverfügbares öffentlich wird – durch Debatte, Erfahrung, Artikulation. Epistemischer Anachronismus jedoch bedeutet, dass das erneut zum Vorschein kommt, was bereits dokumentiert, bewertet, modelliert wurde. Was nicht digitalisiert wurde, existiert epistemisch nicht – und was sich dem digitalen Zugriff entzieht, bleibt unverstanden. Anachronismus und Selbstreferentialität sind daher nicht nur erkenntnistheoretisch, sondern auch ethisch und politisch relevant.²⁴³

3.3.3.6 Epistemische Deutungshoheit

Unter epistemischer Deutungshoheit soll hier die Fähigkeit oder Rolle (Position) eines Menschen, einer sozialen Gruppe und anderer Instanzen verstanden werden, den Rahmen für gültiges Wissen, Interpretationen und Sinnzuschreibungen festzulegen und durchzusetzen. Im Rahmen dieser Arbeit ist die Frage, inwieweit sich die epistemische

²⁴² Wihbey, „AI and Epistemic Risk for Democracy“.

²⁴³ Coeckelbergh, „LLMs, Truth, and Democracy“.

Deutungshoheit weg vom Diskurs, der Reflexion und den Menschen hin zu einer Deutungshoheit durch KI-Systeme verschiebt?

Wenn der Mensch fragt und Künstliche Intelligenz antwortet, überlässt er zunächst die Deutungshoheit über Wissen und Wahrheit den Systemen. Platons Höhlengleichnis macht das Dilemma deutlich: Es ist ein paradigmatisches Bild dafür, wie Wissen, Wirklichkeit und Macht über Interpretation zusammenhängen. Platon beschreibt Menschen, die in einer Höhle gefesselt sind und nur Schatten an der Wand sehen – Abbilder von Objekten, die hinter ihnen von einem Feuer beleuchtet werden. Einer befreit sich, gelangt ans Licht (zur Wahrheit, zur Idee des Guten) und erkennt, dass die Schatten nur Illusionen sind. Die Schattenwelt steht für die gewöhnliche Welt der Meinung (δόξα), die nur ein verzerrtes Abbild der Wahrheit ist. Die Welt außerhalb der Höhle symbolisiert das wahre Wissen (epistēmē, altgriechisch: ἐπιστημολογία), das nur durch Philosophie und geistige Anstrengung erreichbar ist.

Wer hat die epistemische Deutungshoheit in der Höhle? Dort herrscht eine bestimmte Ordnung des Wissens – die Schatten werden für die Wirklichkeit gehalten. Das bedeutet: Es gibt eine hegemoniale Deutung der Welt, auch wenn sie falsch ist. Diejenigen, die diese Schatten interpretieren, gelten als "wissend". Abweichler (der befreite Philosoph) werden als verrückt oder gefährlich wahrgenommen, weil sie die bestehende Ordnung in Frage stellen. Nicht was „wahr“ ist zählt, sondern das, was für wahr gehalten werden darf. Und das bestimmt die Autorität, die über die Macht dazu verfügt. Macht über Bilder und Sprache prägt unser Wirklichkeitsverständnis. Das Höhlengleichnis ist eine Allegorie epistemischer Herrschaft.

„Platon verkörpert das Wahrheitsregime. In seinem Höhlengleichnis ...“ [...] „Wir sind heute in einer digitalen Höhle gefangen, wobei wir uns in Freiheit wähnen. Wir sind gefesselt an den digitalen Bildschirmen.“ [...]”²⁴⁴

Wer die Deutungshoheit innehat, besitzt gleichzeitig Macht. KI-Systeme üben auf ihre Nutzer*innen eine „smarte Macht“ aus.

²⁴⁴ Han, *Infokratie*, 83.

3.3.4 Die „smarte Macht“ der omnipräsenten Systeme

Künstliche Intelligenz vermag als „smarte Macht“ ubiquitärer Systeme durch seduktive Steuerung individuelle Autonomie zu untergraben.

3.3.4.1 „Smarte Macht“

Der Begriff der „smarten Macht“ geht zurück auf Byung Chul Han ²⁴⁵. Macht habe unterschiedliche Erscheinungsformen. Sie äußere sich in ihrer direktesten Form als „Negation von Freiheit“, richte sich dann gegen den Willen der Unterworfenen und könne sich in Gewalt oder Repression äußern. Im Unterschied zu dieser Form der „Disziplinarmacht“ nehme die smarte Macht zunehmend eine permissive Form an und lege in ihrer Freundlichkeit ihre Negativität ab. Smarte Macht operiere nicht gegen den Willen, sondern steuere den Willen der unterworfenen Subjekte zu deren Gunsten. Sie glauben, frei zu handeln, folgen aber einem diskret gesetzten digitalen Pfad. Smarte Macht sei, so Han, nicht repressiv, sondern permissiv, geradezu milde, sanft, tolerant und nachsichtig. Dadurch gewinne sie ihre Wirksamkeit: Sie beute nicht aus, sondern lade ein. Sie verinnerliche das Subjekt nicht gegen seinen Willen, sondern verführe es zur Selbstoptimierung und Entlastung. Smarte Macht wirke nicht durch Zwang, sondern durch Aussicht auf Bequemlichkeit. Sie brauche keinen Widerstand zu überwinden. *„Ihr Signum ist der Like-Button [...] Die smarte Macht wertet unsere bewussten und unbewussten Gedanken aus.“* So überwache und steuere sie die Menschen zu deren Gunsten.

Von der Disziplinarmacht zur smarten Macht

Han knüpft an Michel Foucaults Konzept der Disziplinarmacht an. Foucault beschreibt diese als eine Form der Machtausübung, die durch formbare, „gelehrige“, menschliche Körper, durch Überwachung und normierende Sanktion sowie durch „*Panoptismus*“ gekennzeichnet ist.

- Der „*gelehrige Körper*“ ist ein dressierter, angepasster Körper – der Soldat des 18. Jahrhunderts beispielsweise: „Schritt für Schritt hat man die Haltungen zurechtgerichtet, bis ein kalkulierter Zwang jeden Körperteil durchzieht und bemeistert, den gesamten Körper zusammenhält und verfügbar macht“.²⁴⁶
- Hierarchische *Überwachung* und *normierende Sanktion* sind „*Mittel einer guten Abrichtung*“. Die Überwachung erfolgt durch „*die Einrichtung des*

²⁴⁵ Han, *Psychopolitik Neoliberalismus und die neuen Machttechniken*, 25.

²⁴⁶ Foucault, *Überwachen und Strafen*, 173.

zwingenden Blicks“.²⁴⁷ Nicht nur in Kasernen, sondern beispielsweise auch in Schulen, Fabriken, Krankenhäusern und Gefängnissen. Abweichungen vom Normalmaß werden registriert und sanktioniert.

- „Panoptismus“ ist eine Form der Machtausübung, bei der durch die Möglichkeit ständiger Beobachtung eine Selbstdisziplinierung der Individuen erreicht wird. Das Subjekt internalisiert die Beobachtung und beginnt, sich selbst zu disziplinieren.²⁴⁸

Der Begriff des Panoptikums geht auf Jeremy Bentham zurück. Seine Idee: Ein architektonisches Konzept zur Beaufsichtigung und Überwachung von Gefängnisinsassen oder Fabrikarbeitern. Im Zentrum befindet sich ein Beobachtungsturm, von dem aus ringförmig angeordnete Zellen abgehen. Diese Anordnung ermöglicht es dem Wächter im Turm, sämtliche Zellen einzusehen, ohne selbst gesehen zu werden. Der Grund dafür liegt in der gezielten Lichtführung: Die Gefangenen sind aus ihrer beleuchteten Position gut sichtbar, während der Wächter im dunkleren Inneren des Turms bleibt und so für die Insassen unsichtbar ist. Auf diese Weise können die Gefangenen nie mit Sicherheit wissen, ob sie tatsächlich beobachtet werden – sie müssen jedoch stets davon ausgehen. Diese potenzielle Allgegenwart der Überwachung führt zu einem psychologischen Druck, der das Verhalten der Beobachteten maßgeblich beeinflusst. Aus der Unsicherheit über die konkrete Beobachtungssituation entsteht ein Zustand ständiger Selbstdisziplinierung: Die Insassen passen sich freiwillig den Regeln und Normen an, da sie jederzeit mit Kontrolle rechnen müssen – selbst, wenn diese objektiv nicht stattfindet.

Disziplinarmacht ist nicht nur repressiv, sondern auch produktiv, indem sie Menschen formt, bis diese den gesellschaftlichen Funktionslogiken entsprechen. Die Disziplinarlogik ist eng an die industrielle Produktion gebunden, in der Effizienzsteigerung, zeitliche Taktung und körperliche Disziplin von Bedeutung sind. Die Maschinen der industriellen Massenproduktion bestimmen maßgeblich soziale und ökonomische Prozesse. Sie üben Herrschaft aus. „Das Disziplinarregime ist Herrschaftsform des Industriekapitalismus.“²⁴⁹ Der Mensch ist ein Rädchen im Räderwerk der Maschinen.

²⁴⁷ Foucault, *Überwachen und Strafen*, 220.

²⁴⁸ Foucault, *Überwachen und Strafen*, 251.

²⁴⁹ Han, *Infokratie*, 7.

Chaplinitis

Während seiner Exil-Zeit in den USA hat G. Anders in Kalifornien als Fabrikarbeiter am Fließband gestanden. „*Wir sollten an dem, was wir verrichteten, kein Interesse haben, wir sollten ohne Bewandtnis arbeiten.*“²⁵⁰ G. Anders vergleicht die immer gleichartigen Tätigkeiten von Fabrikarbeitern am Fließband mit einer Art Gymnastik und nennt sie „Chaplinitis“. Charlie Chaplins Film „Modern Times“ kam in den 1930er Jahren in die amerikanischen Kinos. Eine der ersten Szenen zeigt Chaplin als Fließbandarbeiter, in jeder Hand ein Werkzeug, mit dem er in immer gleichen Bewegungen und im monotonen Rhythmus der Maschine stumpfsinnige Montagearbeiten an auf dem Fließband vorbeiziehenden Bauteilen verrichtet. Und das in einer Geschwindigkeit, die ihm nicht einmal Zeit lässt, eine Fliege zu verjagen, die vor seiner Nase herumtanzt. Hält er auch nur den Bruchteil einer Sekunde inne, führt das zu einem kleinen Chaos. Er kollidiert mit seinen Kollegen. Wiederholt unternimmt er hektische Versuche, wieder in den Rhythmus der Maschinen zu kommen, sich dem Maschinengang anzupassen. Seine Bemühungen bleiben ohne Erfolg. Er erntet die strengen Blicke der Aufseher.²⁵¹

Der Mensch versucht, der Überlegenheit der Systeme „*durch physische Gleichschaltung mit den Geräten, durch «Human Engineering», zu entgehen*“²⁵². Die Maschinen der industriellen Revolution strukturierten nicht nur die Arbeitsverrichtungen, sondern sie hatten auch wesentlichen Einfluss auf die Gestaltung der Nichtarbeitszeit. Schichtarbeit mit zwei oder drei Schichten prägten die private Zeitgestaltung. Die Maschinen der industriellen Revolution gaben den Zeittakt für die Menschen vor, während z.B. bäuerliche Tätigkeiten in der vorindustriellen Zeit vom Rhythmus der Jahreszeiten bestimmt waren, innerhalb dieser jedoch weitgehend von den Menschen selbst bestimmt wurden.

Übergang zur smarten Macht

Im Unterschied zur Disziplinarmacht, die über Repression und Normierung in gesellschaftlichen Institutionen ausgeübt wird, realisiert sich smarte Macht über Verführung und Selbststeuerung in sozialen Netzwerken und auf Plattformen. Sie formt selbstoptimierende, sich selbst vermarktende Individuen, die nicht aufgrund von Disziplinartechniken gehorchen. Sie wähnen sich frei, jedoch unterwerfen sie sich

²⁵⁰ Anders, *Die Antiquiertheit des Menschen*. 2, 102.

²⁵¹ *Modern Times*.

²⁵² Anders, *Die Antiquiertheit des Menschen*. 1, 50.

gleichzeitig freiwillig einer subtilen, omnipräsenten Macht. Der Übergang von der Disziplinarmacht hin zur smarten Macht markiert eine grundlegende Verschiebung in den Techniken der Subjektformung: von der Regulierung der Körper hin zur Steuerung der Psyche; von der sichtbaren Überwachung zur unsichtbaren Datenauswertung; von der Gehorsamspflicht zur Selbstoptimierung. Vom Panoptikum, wie Foucault es vor Augen hatte, hin zum „digitalen Panoptikum“, dessen Bewohner vernetzt sind und intensiv miteinander kommunizieren.

„Nicht räumliche und kommunikative Isolierung, sondern Vernetzung und Hyperkommunikation machen Totalkontrolle möglich. [...] Die Bewohner des digitalen Panoptikums sind keine Gefangenen. Sie leben in der Illusion der Freiheit. Sie speisen das digitale Panoptikum mit Informationen, indem sie sich freiwillig ausstellen und ausleuchten.“²⁵³

Smarte Macht ist die Machttechnik heutiger Herrschaftsform.

„Informationsregime nennen wir jene Herrschaftsform, in der Informationen und deren Verarbeitung mittels Algorithmen und Künstlicher Intelligenz soziale, ökonomische und politische Prozesse maßgeblich bestimmen“²⁵⁴

Das Informationsregime steht im Gegensatz zum Disziplinarregime. Es beutet Informationen und Daten aus, wohingegen das Disziplinarregime Körper und Energien ausbeutet. Das Informationsregime ist auch ein Überwachungsapparat. Es degradiert den Menschen zum Subjekt von Daten und Konsum, zum „Daten- und Konsumvieh“. ²⁵⁵

Wer benutzt wen?

In der Netflix-Dokumentation „Das Dilemma mit den sozialen Medien“ wird deutlich, dass die großen Social-Media-Unternehmen weniger an Zielen wie Respekt und einem fairen Miteinander interessiert sind, sondern vielmehr an der Maximierung von Nutzeraufmerksamkeit. Ehemalige Mitarbeiterinnen von Facebook, Twitter, Google und anderen Konzernen beschreiben darin, dass das Geschäftsmodell auf Methoden persuasiver Kommunikation basiert, die auf Überredung und Beeinflussung abzielen, nicht jedoch auf Verständigung oder Dialog. Ziel sei es, die Nutzerinnen möglichst lange

²⁵³ Han, *Im Schwarm*, 92.

²⁵⁴ Han, *Infokratie*, 7.

²⁵⁵ Han, *Infokratie*, 7.

an die Endgeräte zu binden, um einerseits umfassende Daten über sie zu sammeln und andererseits ihr Verhalten und ihre Einstellungen dauerhaft beeinflussen zu können.²⁵⁶

„Viele dieser großen Tech-Firmen verfolgen drei Ziele: Das Aktivitätsziel, also die Bildschirmzeit maximieren, dass man immer weiter scrollt. Das Wachstumsziel, dass man zurückkommt und viele Freunde einlädt, die dann weitere einladen. Und dann noch das Werbeziel, dass man während dessen durch Werbung so viel Geld macht wie möglich.“²⁵⁷

Dass sie dafür ihre Plattformen „kostenlos“ zur Verfügung stellen, trifft nur scheinbar zu. Wenn wir sie nutzen, zahlen wir mit unseren Daten und werden einerseits zu Datenlieferant*innen und gleichzeitig zur Zielscheibe gezielter manipulativer Beeinflussung.²⁵⁸

„Sie wissen alles über uns. Sie haben mehr Informationen, als wir uns das in der Menschheits-Geschichte je hätten vorstellen können. [...] Die Vorhersagen, was wir tun und wer wir sind, werden immer besser.“²⁵⁹

Jedoch sind nicht wir, die User, die Kunden der Tech-Giganten,

„sondern Firmen, Staaten, Geheimdienste und alle anderen, die wissen wollen, wie sie das Verhalten der User beeinflussen können. Verkauft werden nicht die User-Daten (Adressen etc.), sondern zielgruppenspezifische Prognose-Modelle über Beeinflussungsmöglichkeiten von Verhalten.“²⁶⁰

Ihr Produkt ist

„die allmähliche, kaum spürbare Veränderung von Verhalten und Wahrnehmung. [...] Es gibt kein anderes weit und breit.... Dein

²⁵⁶ Zander, „Wahrheit und Respekt im Geschwindigkeitsrausch sozialer Medien“, 4.

²⁵⁷ Netflix, „Das Dilemma mit den sozialen Medien“, ca. Min 19, [Netflix- Das Dilemma mit den sozialen Medien](#) zugegriffen 22. Mai 2024,.

²⁵⁸ Netflix, „Das Dilemma mit den sozialen Medien“; Anders, *Die Antiquiertheit des Menschen. 1*, 119. In der Netflix-Dokumentation sagt Tristan Harris: „If something is a tool, it is just genuinely sitting there, waiting, patiently. If something is a not tool, it's demanding, it is seducing you, manipulating, it wants things of you. Social Media is not a tool.“

Und Günther Anders: „Jeder Konsument ist ein unbezahlter Heimarbeiter für die Herstellung des Massenmenschen“

²⁵⁹ Netflix, „Das Dilemma mit den sozialen Medien“ ca. Min. 16.

²⁶⁰ Zander, „Wahrheit und Respekt im Geschwindigkeitsrausch sozialer Medien“, 4.

*Verhalten zu verändern, deine Gedanken, deine Identität. Es ist eine schleichende Veränderung. Kaum spürbar.*²⁶¹

Auch Social Media Plattformen sind keine Werkzeuge. Sie stehen nicht im Dienste ihrer Nutzer. Es drängt sich die Gewissheit auf: *„Nicht der Mensch benutzt die Maschine, sondern die Maschine den Menschen.“*²⁶²

Viele Menschen sind der Auffassung, dass sie nichts zu verbergen haben, denn sie haben sich nichts zu Schulden kommen lassen. „Was soll schon mit meinen Daten passieren? Sie können nicht gegen mich verwendet werden.“ Das mag sein. Aber es geht um etwas anderes, nämlich um das Gesetz der großen Zahl. Jaron Lanier, ein Silicon-Valley-Insider hat das Akronym BUMMER²⁶³ eingeführt.

*„BUMMER ist eine Maschine, ein statistischer Apparat, der in Cloud-Netzwerken lebt [...] Selbst die besten BUMMER-Algorithmen können nur die Wahrscheinlichkeit berechnen, dass eine Person sich auf bestimmte Art und Weise verhalten wird. Was jedoch für jeden Einzelfall nur eine Wahrscheinlichkeit ist, wird im Durchschnitt einer großen Zahl von Menschen beinahe zur Gewissheit. Die gesamte Bevölkerung kann mit größerer Berechenbarkeit beeinflusst werden als irgendeine Person“*²⁶⁴

Indem die großen Tech-Konzerne einerseits über ein gigantisches Datenreservoir verfügen und gleichzeitig die Technologien kontrollieren, über die wir kommunizieren und Informationen abrufen, sichern sie sich eine nie dagewesene Machtfülle.²⁶⁵

3.3.4.2 „Ubiquitous Computing“

Smarte Macht manifestiert sich in omnipräsenten, stillen Technologien, die sich ins Gewebe unseres Alltags einflechten, bis sie verschwunden und von diesem nicht mehr zu unterscheiden sind. Sie zeichnen sich durch ihre Unscheinbarkeit aus. Mark Weiser gilt gemeinhin als geistiger Vater des sogenannten „Ubiquitous Computing“.²⁶⁶ Schon in den

²⁶¹ Netflix, „Das Dilemma mit den sozialen Medien“.

²⁶² Zander, „Wahrheit und Respekt im Geschwindigkeitsrausch sozialer Medien“, 4.

²⁶³ Steht für: „Behaviors of Users Modified and Made into an Empire for Rent“. Bummer heißt in der englischen Umgangssprache aber auch „Mist!“ Lanier, *Zehn Gründe, warum du deine Social Media Accounts sofort löschen musst*.

²⁶⁴ Lanier, *Zehn Gründe, warum du deine Social Media Accounts sofort löschen musst*, 43.

²⁶⁵ Lenzen, *Künstliche Intelligenz*, 252.

²⁶⁶ Weiser, „The Computer for the 21st Century“.

1990er Jahren plädierte er für einen Paradigmenwechsel hin zu einer ruhigen Technologie.²⁶⁷

*„The most profound technologies are those that disappear. They weave themselves into the fabric of everyday life until they are indistinguishable from it.“*²⁶⁸

Die smarten Technologien sind omnipräsent. Zeitlich, räumlich, in allen sozialen Kontexten unserer Lebensbereiche, im öffentlichen wie im privaten Bereich. Sie sind allgegenwärtig, überall vorhanden und jederzeit verfügbar, mit der Tendenz der fortschreitenden Vernetzung und Vereinnahmung aller Techniken und Lebensbereiche. Omnipräsenz hat eine zeitliche und eine räumliche Dimension. Es gibt heute kaum einen Lebensbereich, der unberührt von unauffälliger, scheinbar verschwundener KI-Technologie ist. Sie kommt daher wie eine smarte Macht. Sie will unsere Aufmerksamkeit, jedoch eine mühelose. Die Technologien der Smarten Macht machen es uns leicht, sie strengen nicht an, fordern keine mühevollen Aufmerksamkeit. Denn sie haben keine Widerständigkeit, bauen keine Barrieren auf und kennen keine Grenzen. Ebenso wenig wie die Idee des Ubiquitous Computing eine solche kennt. Im Gegenteil: Mit ihr verhält es sich wie mit der Informationstechnologie des Schreibens, die die Information von den Grenzen des menschlichen Gedächtnisses befreite, weil es mit ihr möglich wurde, das gesprochene Wort langfristig zu speichern.

*“Consider writing, perhaps the first information technology: The ability to capture a symbolic representation of spoken language for long-term storage freed information from the limits of individual memory.“*²⁶⁹

Diese Technologie ist heute ubiquitär und sie erfordert keine aktive Aufmerksamkeit.

*„The constant background presence of these products of «literacy technology» does not require active attention, but the information to be conveyed is ready for use at a glance. It is difficult to imagine modern life otherwise.“*²⁷⁰

²⁶⁷ Weiser und Brown, *Designing Calm Technology*.

²⁶⁸ Weiser, „The Computer for the 21st Century“.

²⁶⁹ Weiser und Brown, *Designing Calm Technology*.

²⁷⁰ Weiser und Brown, *Designing Calm Technology*.

Man kann darin eine Chance sehen, wie die Protagonisten dieser ruhigen Technologie, an die sie die Hoffnung auf ein menschlicheres 21. Jahrhundert knüpften:

„As we learn to design calm technology, we will enrich not only our space of artifacts, but also our opportunities for being with other people. Thus may design of calm technology come to play a central role in a more humanly empowered twenty-first century.“²⁷¹

Grunwald sieht in der fehlenden Widerständigkeit ubiquitärer Systeme allerdings eine ethische Herausforderung:

„Eine Erschließung unseres Alltags durch ubiquitäre Systeme hätte enorme Auswirkungen auf die drei Bedingungen des ethischen Diskurses, nämlich die ›Bestimmbarkeit der Wirklichkeit‹, in der gehandelt werden soll, die ›Identität des Handlungssubjekts‹, das die Verantwortung für sein Handeln übernehmen soll sowie die ›Wahl‹, die verantwortliches Handeln ermöglicht. Als Grundproblem erweist sich der Verlust der Widerständigkeit als elementares Charakteristikum von Wirklichkeit. Das Paradigma einer perfekt in den Alltag integrierten Technik, die sich wie eine gutsitzende Brille quasi selbst zum Verschwinden bringt, wird zum Problem, denn eine Technologie, die keine Schnittstelle und keine Widerständigkeit mehr wahrnehmen lässt, kann weder kontrolliert noch gesteuert werden.“²⁷²

Ubiquitäre Systeme bescheren uns damit ein „Zauberlehrlingsproblem“, so Grunwald. Die Systeme handeln anstelle der Menschen, die eine Schwächung ihrer Identität erfahren können, welche sich in Widerstandserfahrung durch Anerkennung und Nichtanerkennung ausbildet.²⁷³

Der zwanglose Zwang des «Always on»²⁷⁴

Die Omnipräsenz der Systeme im öffentlichen Raum macht die Verführung durch smarte Macht sichtbar. In den Bussen und Bahnen gibt es ein inzwischen vertrautes Bild. Sobald Menschen allein in Bahnen und Bussen unterwegs sind, beschäftigen sie sich mit ihrem Smartphone. Der digitale Begleiter ruft: „Du bist nicht allein.“ „Du bist wichtig, wirst gebraucht.“ Wenn sie in Cafés und Restaurants sitzen, ist das Smartphone – so hat es

²⁷¹ Weiser und Brown, *Designing Calm Technology*, 5.

²⁷² Grunwald und Hillerbrand, *Handbuch Technikethik*, 419.

²⁷³ Grunwald und Hillerbrand, *Handbuch Technikethik*, 419.

²⁷⁴ Die Formulierung ist übernommen aus: Grimm u. a., *Digitale Ethik*, 76.

gelegentlich den Anschein – wichtiger als das Gegenüber. Die Menschen scheinen das freiwillig zu tun, ziehen offensichtlich die technische Interaktion dem Gespräch vor. Setzen eindeutige Prioritäten. Wenn der digitale Begleiter ruft, sind die Menschen zur Stelle. Was die Apparate zu vermelden haben, scheint von höchster Dringlichkeit, dafür unterbrechen die Menschen vertrauensvolle Gespräche. Noch etwas lässt sich beobachten. Gelegentlich holen wir das Smartphone hervor, weil wir der künstlichen Intelligenz mehr vertrauen als unseren Mitmenschen. Unser Gegenüber stellt eine Behauptung auf oder gibt neuestes Wissen preis. Wir schauen ungläubig drein und zweifeln es an. Wir fragen die KI, denn die – so sind wir überzeugt – weiß es besser. Und so signalisieren wir unser Misstrauen, wie Liessmann in einem Vortrag meint:

„Und meine These ist – darüber lässt sich wahrscheinlich trefflich diskutieren, dass der nicht geringste Aspekt, ich betone Aspekt, der nicht geringste Aspekt der künstlichen Intelligenz darin besteht, dass es für uns eine wunderbare Möglichkeit ist, unseren Mitmenschen dadurch das Misstrauen auszusprechen, jedes Mal, jedes Mal, wenn Sie in ein Gespräch verstrickt sind, und jemand behauptet etwas, und sie sagen: „Achtung“, zücken das Smartphone, googeln. Du hast Unrecht!“

Und etwas ironisch merkt er an:

„Ist das nicht eine wunderbare Misstrauenserklärung, die sie nie tätigen könnten, wenn sie nicht ein Smartphone hätten, in dem die künstliche Intelligenz Ihnen Informationen im Stande ist herauszufiltern, bereitzustellen, für die Sie ansonst wahrscheinlich monatelang studieren müssten.“²⁷⁵

Anderen Menschen zu vertrauen hat sich bekanntermaßen in der menschlichen Evolution als überlebenswichtig erwiesen, ebenso wie Misstrauen. Künstliche Intelligenz hat sich zu einer scheinbar vertrauensvollen Partnerin entwickelt, die uns interessanter und mächtiger erscheint als unsere Mitmenschen. Wir schenken algorithmisch gefiltertem Wissen oft mehr Glauben als den Aussagen unserer Mitmenschen, ein Phänomen, das als Automation BIAS bekannt ist. Es entwickelt sich eine Art symbiotischer Beziehung zu den digitalen Besserwissern. Signalisieren wir nicht nur Misstrauen, sondern sonnen wir

²⁷⁵ Liessmann, „Ohne uns! Der Traum der Maschinen und die Zukunft des Menschen“.

uns nicht auch in unserem eigenen Besserwissen? Der KI sei Dank. Sie verleiht uns mehr Macht, lässt uns in einem besseren Licht erscheinen und macht uns scheinbar unabhängig.

Denn tatsächlich geht von den smarten Technologien eine Macht aus, die die scheinbare Unabhängigkeit in einen Zwang zur Kommunikation umschlagen lässt.

*„Von den Smartphones, die mehr Freiheit versprechen, geht ein fataler Zwang aus, nämlich der Zwang zur Kommunikation. Inzwischen hat man ein obsessives, zwanghaftes Verhältnis zu den digitalen Apparaten. Auch hier schlägt die Freiheit in Zwang um. Die sozialen Netzwerke verstärken diesen Zwang zur Kommunikation massiv.“*²⁷⁶

Die digitalen Kommunikationstechniken, allen voran das Smartphone, tragen zum veränderten Umgang mit Zeit, vor allem mit der freien Zeit – der „Eigenzeit“ – bei.

*„Viele Menschen sind süchtig danach geworden. Sie können nicht mehr ohne Smartphone aufwachen und nicht ohne Smartphone einschlafen. Schon der erste Griff am Morgen gilt dem Smartphone, um zu sehen, was sich in der Nacht getan hat. Auch diese mediale Eigenzeit ist Eigenzeit, aber eine, deren Rhythmus wir nicht selbst gestalten, sondern der durch das digitale Medium getaktet wird. Die Sehnsucht nach freier Zeit hat sich gewandelt. Statt «Ich will jetzt Zeit für mich haben» heißt es nun «Ich will verbunden sein mit der Welt, mit meinen Freunden. Ich will wissen was los ist.» Viele Menschen sehen sich als Teil dieser verschiedenen Netzwerke. Das ist die Diagnose des Zeitgefühls von heute.“*²⁷⁷

Künstlich intelligente Systeme haben einen Einfluss nicht nur auf das, was ich will und tue, sondern auch wann, wie oft, in welchem Rhythmus, wie lange, wie schnell, wann ich beginne, wann ich aufhöre, in welcher Reihenfolge ich etwas tue.

*„Technologien und Medien organisieren uns und damit auch unsere Zeitlichkeit und Existenz. Technologien und Medien prägen die Art und Weise, wie wir Zeit gestalten und wie wir existieren. Wir sind Teilnehmer an Technoperformances der Zeit.“*²⁷⁸

²⁷⁶ Han, *Im Schwarm*, 50.

²⁷⁷ Nowotny, „Keine Zeit für Utopien“.

²⁷⁸ Coeckelbergh, *Digital Technologies, Temporality, and the Politics of Co-Existence*, 37.

Smarte Versprechen

In Anlehnung an Han könnte man sagen, dass von der smarten Macht der Systeme smarte Versprechen ausgehen, die uns dazu verleiten, ein besonderes Verhältnis zu den künstlichen Systemen aufzubauen. Digitale Assistenten, Fitness-Tracker oder algorithmisch gesteuerte Plattformen gehen mit Versprechen einher: gesteigerte Effizienz, schnellere Entscheidungen, mehr Kontrolle über das eigene Leben. Diese Versprechen sind jedoch selten explizit und nicht überprüfbar. Sie sind Teil einer subtilen Steuerungslogik, suggerieren Freiheit, während sie zur Internalisierung von Verhaltensnormen führen.

Die smarten Versprechen der omnipräsenten Systeme haben nichts zu tun mit den Versprechen von Menschen untereinander. KI-Versprechen haben keine bindende Kraft, ihre Einhaltung ist optional. Für Hannah Arendt sind Versprechen der Menschen bindend und zukunftsgerichtet. Sie sind ein *„Heilmittel gegen die Unabsehbarkeit und damit gegen die chaotische Ungewißheit alles Zukünftigen“*. Wenn Menschen sich Versprechen geben, schafft das Verlässlichkeit, Vertrauen und Sicherheit für gemeinsames Handeln. Versprechen sind *„Inseln der Sicherheit“* und *„Wegweiser in die Zukunft“*. Die Menschen geben sich Eheversprechen, schließen bindende Verträge, treffen Vereinbarungen als Grundlage für gemeinschaftliches institutionelles Handeln oder geben ihren sozialen und politischen Gemeinschaften konstitutionelle Grundlagen.

„Und das Heilmittel gegen Unabsehbarkeit und damit gegen die chaotische Ungewißheit alles Zukünftigen – liegt in dem Vermögen, Versprechen zu geben und zu halten. Diese beiden Fähigkeiten [Anm: Verzeihen und Versprechen] gehören zusammen, insofern die eine sich auf die Vergangenheit bezieht und ein Geschehenes rückgängig macht, dessen «Sünde» sonst, dem Schwert des Damokles gleich, über jeder neuen Generation hängen und sie schließlich unter sich begraben müßte; während die andere ein Bevorstehendes wie einen Wegweiser in die Zukunft aufrichtet, in der ohne die bindenden Versprechen, welche wie Inseln der Sicherheit von den Menschen in das drohende Meer des Ungewissen geworfen werden, noch nicht einmal irgendeine Kontinuität menschlicher Beziehungen möglich wäre, von Beständigkeit und Treue ganz zu schweigen.“²⁷⁹

²⁷⁹ Arendt, *Vita activa oder Vom tätigen Leben*, 336.

Die smarten Versprechen der KI zielen nicht auf gemeinschaftliches Handeln ab, sind keine Inseln der Sicherheit in Hinblick auf gemeinsame Zukunft. Sie rücken den kurzfristigen individuellen Nutzen in den Fokus der Aufmerksamkeit. Sie binden uns nicht durch Verpflichtung, sondern durch Suggestion, durch die sanfte Andeutung eines optimierten Selbst.

„Die digitale Kommunikation lässt die Gemeinschaft, das Wir, vielmehr stark erodieren. Sie zerstört den öffentlichen Raum und verschärft die Vereinzelung des Menschen. Nicht die «Nächstenliebe», sondern der Narzissmus beherrscht die digitale Kommunikation. Die digitale Technik ist keine «Technik der Nächstenliebe». Sie erweist sich vielmehr als eine narzisstische Ego-Maschine.“²⁸⁰

Aber wird der öffentliche Raum zerstört? Oder erzeugt digitale Kommunikation nicht vielmehr einen „pervertierten“ öffentlichen Raum, der kein Raum für Gemeinschaft ist, sondern für einen „digitalen Schwarm“²⁸¹ einzelner Individuen, die im privaten Raum einen gigantischen Hallraum erzeugen?

Soziale Medien gehen mit trügerischen Versprechen einher, die einen falschen Eindruck hervorrufen oder falsche Hoffnungen erwecken. Sie sind täuschend oder irreführend. Kommt eine Absicht hinzu, dann sind Versprechen manipulativ im Sinne einer verdeckten Einflussnahme.²⁸² Auch Selbsttäuschung kann in Zusammenhang mit Künstlicher Intelligenz eine Rolle spielen. Etwa wenn wir an Überzeugungen festhalten, obwohl es deutliche Hinweise darauf gibt, dass das Gegenteil wahr ist. Wenn wir uns etwas einreden, das tatsächlich falsch ist. Oder wenn wir selbst es sind, die *„von sich aus Mittel und Wege ersinnen, die Falschheit der eigenen bzw. die Wahrheit der gegenteiligen Überzeugung nicht anerkennen zu müssen.“*²⁸³ Die Systeme Künstlicher Intelligenz liefern geradezu eine Steilvorlage, sind eine Einladung zur Selbsttäuschung, weil sie zur Ermächtigung verhelfen. Etwa wenn Texte von ChatGPT als eigene kreative Leistung gesehen und präsentiert werden oder sich Menschen in pervertierten Selbstbildern zur Schau stellen.

²⁸⁰ Han, *Im Schwarm*, 65.

²⁸¹ Han, *Im Schwarm*, 19.

²⁸² Susser u. a., „Technology, Autonomy, and Manipulation“, 4.

²⁸³ Beier, *Selbsttäuschung*, 27.

Beispiel: „Smart Home Systeme“ in der Pflege

Das Konzept der Allgegenwärtigkeit von Technologien, die unauffällig und selbstverständlich in den Alltag integriert sind, manifestiert sich unter anderem in vernetzten Smart-Home-Systemen. Im Kontext der Versorgung und Unterstützung pflegebedürftiger Personen wird dieser Ansatz unter dem Begriff „Ambient Assisted Living“ (AAL) gefasst. AAL bezeichnet eine heterogene Gruppe technischer Systeme, deren primäre Funktion in der Erfassung, Übermittlung und Verarbeitung von Daten zum Zweck der Überwachung, Steuerung und Unterstützung liegt. Dazu zählen unter anderem Sturzmelde- und Notrufsysteme, Systeme zur kontinuierlichen Vitaldatenüberwachung, Verfahren zur Aggressionserkennung, Standortbestimmung sowie Aktivitätstracking. Ergänzt werden diese durch Steuerungssysteme für Haushaltsgeräte und Wohnumgebungen (z. B. Herd, Beleuchtung, Heizung, Fenster, Türen, TV) sowie zunehmend auch durch assistive Robotik. Zielsetzung dieser Technologien ist es, älteren oder in ihrer Selbstständigkeit eingeschränkten Menschen ein möglichst autonomes Leben in der eigenen häuslichen Umgebung zu ermöglichen und zugleich ihre Lebensqualität nachhaltig zu erhöhen.²⁸⁴

Ein Teil der AAL-Systeme ist mit Künstlicher Intelligenz ausgestattet. Dazu zählen beispielsweise Pflegeroboter, die in der Lage sind, soziale Interaktionen zu initiieren, Unterhaltung zu bieten sowie durch Spiele oder Denkaufgaben die geistigen Fähigkeiten der Nutzer*innen zu fördern. Das Anwendungsspektrum dieser Systeme erscheint in diesem Zusammenhang nahezu unbegrenzt. Entsprechend werden technologische Fortschritte in diesem Feld häufig als Meilensteine auf dem Weg zu mehr Autonomie und einem würdevollen Leben im Alter wahrgenommen. Zugleich gilt jedoch: Während einzelne Komponenten – etwa Notruf- oder Sturzmeldesysteme – kaum grundsätzlicher Kritik ausgesetzt sind, treten mögliche Problemlagen deutlicher hervor, sobald die Technologien nicht isoliert, sondern in ihrer Gesamtheit als vernetztes System und im Hinblick auf zukünftige Entwicklungen betrachtet werden.

„In einem zwar überspitzt dystopisch anmutenden Szenario, das aber nicht sehr weit in der Zukunft liegen könnte, übernimmt die KI des eigenen Smart-Homes die Kontrolle: Aufgrund meiner über die Wochen gesammelten und nun im zentralen IT-Center ausgewerteten Vitaldaten weiß die KI, dass es Zeit für mich ist, das Bett aufzusuchen. Die Türen

²⁸⁴ Zentrum für sichere Informationstechnologie - Austria, *Ambient Assisted Living*.

werden verschlossen, weil es – so glaubt die KI - für einen abendlichen Spaziergang ohnehin zu spät ist. Dafür nimmt mich Bernie, mein hilfsbereiter Roboter an die Hand, rezitiert eine Gute-Nacht-Geschichte aus einem meiner Bücher, die er alle auswendig kann. Dann löscht er das Licht, ohne meinem Murren Gehör zu schenken. Am nächsten Morgen steht aus gesundheitlichen Gründen koffeinfreier Kaffee auf dem Tisch und die vom Kühlschrank bestellten Lebensmittel treffen grade per Drohne ein.²⁸⁵ Alles zum Wohle des Nutzers! Denn die KI weiß es besser und übernimmt die Deutungshoheit über meine Gemütszustände und Wünsche. KI-Systeme nehmen persönliche Entscheidungen ab und bevormunden, kann man mit Fug und Recht sagen.“²⁸⁶

Eine solche Vorstellung wirft grundlegende Fragen im Hinblick auf die medizinethischen Prinzipien auf, wie sie insbesondere in der Prinzipienethik von Beauchamp und Childress formuliert wurden.²⁸⁷ Zentral ist hierbei das Prinzip der Achtung der Autonomie, das die Wahrung von Selbstbestimmung und Entscheidungsfreiheit verlangt. Problematisch wird dies insbesondere dann, wenn höchst persönliche Entscheidungen durch Systeme Künstlicher Intelligenz übernommen werden. Es stellen sich weitere Fragen wie beispielsweise: Welche Auswirkungen es hat, wenn pflegerische Interaktionen zunehmend durch maschinelle Systeme ersetzt werden und damit menschliche Fürsorge und Beziehungspflege auf ein Minimum reduziert werden? Wird dem Prinzip des Nichtschadens (nonmaleficence) – hier in psychologischer Hinsicht – hinreichend Rechnung getragen wird. Gleiches gilt für das Prinzip der Fürsorge (beneficence): Steht tatsächlich das Wohl und die Unterstützung der pflegebedürftigen Person im Zentrum, oder wird primär die – durchaus legitime – Entlastung von Angehörigen und Pflegepersonal verfolgt?²⁸⁸

Diese Fragen lassen sich nicht pauschal beantworten, sondern bedürfen einer differenzierten Betrachtung des jeweiligen Einzelfalls und der konkreten Anwendungskontexte. Die bisherigen Überlegungen implizieren zudem, dass es sich um

²⁸⁵ Vergleiche ähnliche Szenarien bei: Schönau, *“My device made me do it” - Zur Wechselwirkung von Heteronomie und Autonomie bei der Interaktion mit künstlichen intelligenten Systemen* und bei: Spiekermann, *Digitale Ethik*, 257.

²⁸⁶ Zander, „Ambient-Assisted-Living und selbstbestimmtes Leben“, 2.

²⁸⁷ Beauchamp und Childress, *Principles of biomedical ethics*.

²⁸⁸ Zu den medizinethischen Prinzipien siehe Beauchamp und Childress, *Principles of biomedical ethics*, 13.

Personen handelt, denen Autonomie zugesprochen werden kann, also um Menschen, die über grundlegende Kompetenzvoraussetzungen wie rationales Denken und eine hinreichende Fähigkeit zur Selbstkontrolle verfügen. Personen im Koma oder mit fortgeschrittener Demenz fallen nicht darunter. Sind die Bedingungen jedoch erfüllt, so begründet das Prinzip der Autonomie nicht nur Handlungs- und Entscheidungsrechte, sondern auch Abwehrrechte gegen paternalistische Eingriffe, die in der ethischen Abwägung stets mit den Prinzipien des Nichtschadens, der Fürsorge und auch der Gerechtigkeit ins Verhältnis gesetzt werden müssen.²⁸⁹

Unter Paternalismus versteht man allgemein die Einmischung in die Angelegenheiten einer Person durch den Staat oder durch andere Akteure mit dem Ziel, deren Wohlergehen oder Schutz zu gewährleisten – auch gegen den ausdrücklichen Willen der betroffenen Person.²⁹⁰ Solche Eingriffe werden durch die Absicht legitimiert, Schaden abzuwenden oder Nutzen zu stiften. Ein klassisches Beispiel für staatlichen Paternalismus findet sich in gesetzlichen Regelungen wie der Anschnallpflicht im Straßenverkehr, der Pflicht zur Rentenversicherung oder im Verbot bestimmter Drogen. Im medizinischen Kontext wird von ärztlichem Paternalismus gesprochen, wenn Ärztinnen und Ärzte Entscheidungen für Patient*innen ohne deren Einverständnis (informed consent) treffen. Oder wenn sie Patient*innen über deren tatsächlichen Gesundheitszustand im Unklaren lassen. Diese Form des Paternalismus ist wesentlich durch das bestehende Wissens- und Erfahrungsgefälle zwischen medizinischem Fachpersonal und Patientinnen geprägt: Aufgrund ihrer Ausbildung und fachlichen Kompetenz beanspruchen Ärztinnen und Ärzte, besser beurteilen zu können, was im Interesse der Patient*innen liegt. Daraus leitet sich eine Haltung ab, die Fürsorge und Schutzpflichten in den Vordergrund stellt, während den Patient*innen und Patienten lediglich eine passive Rolle zugesprochen wird. In dieser Konstellation wird das Prinzip der Fürsorge (beneficence) gegenüber dem Prinzip der Achtung der Autonomie höher gewichtet.

Sarah Spiekermann und Frank Pallas beziehen Paternalismus auf Systeme des Ubiquitous Computing und sprechen von Technologiepaternalismus. An die Stelle des Menschen treten autonome, selbständig agierende Systeme, die die Kontrolle übernehmen, ohne dass die betroffenen Personen widersprechen oder eingreifen können. Die damit verbundene Sorge besteht darin, dass Menschen auf diese Weise in ein

²⁸⁹ Rössler, „Autonomie“, 156.

²⁹⁰ Dworkin, „Paternalism“.

Abhängigkeitsverhältnis geraten und den Entscheidungen von Maschinen unterworfen sein könnten. Spiekermann und Pallas fordern daher, dass das letzte Wort stets den betroffenen Menschen vorbehalten bleiben müsse.²⁹¹ Ähnliche Überlegungen finden sich in den Ethik-Leitlinien für eine vertrauenswürdige Künstliche Intelligenz, in denen betont wird, dass Nutzerinnen und Nutzer jederzeit in die Lage versetzt werden müssen, sich den Entscheidungen eines Systems zu widersetzen. Leitend müsse dabei das Prinzip der Selbstbestimmung sein; entscheidend sei das Recht, nicht einer ausschließlich automatisierten Entscheidungsfindung unterworfen zu werden, die rechtliche Wirkungen entfaltet oder die betroffene Person in vergleichbarer Weise erheblich beeinträchtigt.²⁹² Übertragen auf die Auseinandersetzung mit AAL-Systemen und den damit verbundenen Konflikten zwischen medizinethischen Prinzipien bedeutet dies: Der Vorrang liegt bei der Selbstbestimmtheit der Nutzerinnen und Nutzer. Damit erfolgt eine klare Absage an jede Form von Technologiepaternalismus, der eine vermeintliche Fürsorge über die Autonomie der Betroffenen stellt.

Das Postulat der Selbstbestimmtheit könnte als normatives Leitprinzip bei der Entwicklung von KI-Systemen gelten. Eine Umsetzung gestaltet sich jedoch komplexer, als es auf den ersten Blick erscheint. Dies zeigt sich insbesondere auch bei sprachbasierten Systemen, etwa sprechenden Robotern. Hier wirken hochgradig subtile, oftmals schwer erkennbare psychologische Mechanismen, die in der kognitiven Psychologie umfassend beschrieben sind und auch in der ärztlichen Kommunikation eine zentrale Rolle spielen. Einige Beispiele verdeutlichen dies:

- *Framing-Effekt*: Die Art und Weise, wie Alternativen sprachlich präsentiert werden, beeinflusst die Entscheidung von Patientinnen und Patienten erheblich. Schon eine unterschiedliche Wortwahl kann zu abweichenden Bewertungen derselben medizinischen Maßnahme führen.²⁹³

²⁹¹ Spiekermann und Pallas, „Ubiquitous Computing. Technology Paternalism–Wider Implications of Ubiquitous Computing.“, 9.

„One of our main findings was that there should be a general possibility to overrule ‘decisions’ made by technology and any exceptions from this should be considered very carefully. People should always have the last word!“

members of the focus groups:

“In extreme situations, a person must always be able to overrule the machine”

²⁹² Hochrangige Expertengruppe für künstliche Intelligenz., „Ethik-leitlinien für eine vertrauenswürdige KI.“, 21.

²⁹³ Kahneman, *Schnelles Denken, langsames Denken*, 448.

- *Anchoring-Effekt*: Informationen werden in Abhängigkeit von der Reihenfolge, in der sie vermittelt werden, unterschiedlich stark gewichtet. Sowohl die zeitliche Abfolge als auch das Angebot zusätzlicher, bewusst nachteiliger Entscheidungsoptionen können die Wahlentscheidung maßgeblich beeinflussen.²⁹⁴
- *Priming-Effekt*: Ein erster Reiz beeinflusst die Verarbeitung nachfolgender Reize. Klassische Beispiele zeigen, dass bereits einfache Abfolgen von Fragen oder Begriffen Wahrnehmung und Antwortverhalten unbewusst steuern können.²⁹⁵

Diese Effekte sind nicht spezifisch für die Mensch-Maschine-Interaktion, sondern gelten allgemein für menschliche Wahrnehmung und Entscheidungsfindung. Werden sie jedoch in KI-gestützten Systemen wirksam, entstehen besondere Herausforderungen für die Sicherstellung von autonomer Entscheidungsfindung im medizinischen Kontext.²⁹⁶

3.3.4.3 Die immanente Tendenz des „ohne uns“

Berichte, Reportagen und Vorträge zur Künstlichen Intelligenz werden oft mit stilisierten Bildern oder Grafiken visualisiert, die ein weltumspannendes Netz darstellen. Das kommt nicht von ungefähr, denn jede/*r Internetnutzer*in, jedes Unternehmen, jeder Geheimdienst, jedes Smart Home, jede*e Smartphone-Besitzer*in nutzt künstliche Intelligenz und ist mit seinen*ihrer Geräten Teil des World Wide Web, das von smarten Technologien des Ubiquitous Computing durchzogen ist.

G. Anders hat die weltweite Ausbreitung der „*Apparate*“ und ihre Vernetzung und Abhängigkeit untereinander früh beschrieben.²⁹⁷ Schon bevor es ein World Wide Web und Ubiquitous Computing gab, sprach er von Apparaten und Maschinen, deren Gemeinschaft zusammen eine Einheit, einen einzigen Funktionskomplex bilden. Sie gehen ineinander auf und werden zu Teilen einer Großmaschine. Sie wachsen zu einer riesigen weltumspannenden Maschine zusammen. Heute erleben wir diese Großmaschine als ein ubiquitäres, omnipräsentes System, das zu großen Teilen aus Künstlicher Intelligenz besteht. In diesen Systemen sind auch die scheinbar kleinen und ruhigen Systeme eingebettet.

²⁹⁴ Kahneman, *Schnelles Denken, langsames Denken*, 152.

²⁹⁵ Kahneman, *Schnelles Denken, langsames Denken*, 72.

²⁹⁶ Zander, „Ambient-Assisted-Living und selbstbestimmtes Leben“, 5.

²⁹⁷ Anders, *Die Antiquiertheit des Menschen*. 2, 122–41.

„Wenn es eine Soziologie der Dinge gäbe, dann würde deren Axiom lauten: «Es gibt keine Einzelapparate». Vielmehr ist jedes ein «zoon politikon»; und außerhalb seiner «Gesellschaft», als bloßes Robinson-Ding, bliebe jedes untauglich. Das Wort «Gesellschaft» bezeichnet dabei aber nicht etwa nur seinesgleichen, nicht nur die Millionen von gleichzeitig funktionierenden Geräten oder deren Summe, sondern ein dem Apparat morphologisch entgegenkommendes Korrelat, eine ihn einbettende, nährende, reinigende, aus Rohstoffen, Konsumenten, Geschwisterapparaten, Abfallkanalisationen bestehende Behausung – kurz: eine Umwelt.“²⁹⁸

Liessmann interpretiert Anders so:

„Technik selbst, so Anders, wird in genau diesem Sinne als Gesellschaft zum neuen Subjekt der Geschichte, die Subjekte von Freiheit und Unfreiheit sind ausgetauscht. Die Dinge sind frei, der Mensch ist unfrei. Der Mensch muss sich den Bedürfnissen der Technik unterordnen“²⁹⁹

In Smart Home Systemen wie AAL werden menschliche Pfleger*innen durch omnipräsente Technologien ersetzt. Die Menschen werden vermeintlich entbehrlich und frei von der Last und den Mühen der menschlichen Pflege. Die zu Pflegenden werden dem Zeittakt und dem Handeln der Maschinen untergeordnet.

G. Anders geht noch einen Schritt weiter. In seinen radikalen Überlegungen zum Atomzeitalter führte er einen Gedanken aus, der hinsichtlich der Künstlichen Autonomie eine bestechende Aktualität aufweist. Wir sind stets darauf aus, Maschinen zu erschaffen, die ohne uns funktionieren können, durch die wir uns überflüssig machen und – wie durch Atomwaffen oder wie aktuell zu beobachten durch KI-gesteuerte Drohnen – die Möglichkeit unserer Liquidierung das Prinzip ist.

„Nichts wäre kurzsichtiger als zu glauben, dass die Möglichkeit unserer Liquidierung nur ein zufälliges Nebenprodukt einiger spezieller Apparate, z.B. der Atomwaffen, sei. Vielmehr ist die Möglichkeit unserer Liquidierung das Prinzip, das wir allen unseren Apparaten mitgeben, gleichgültig mit welcher Sonderfunktion wir jeden

²⁹⁸ Anders, *Die Antiquiertheit des Menschen*. 2, 127.

²⁹⁹ Liessmann, *Die großen Philosophen und ihre Probleme*, 182.

von ihnen außerdem betrauen; das Prinzip, auf das es uns bei ihrer Konstruktion ausschließlich ankommt.“³⁰⁰

Das Prinzip ist der Technik inhärent. Es ist wie mit den Geistern, die wir nicht mehr loswerden. Atomwaffen sind eine Bedrohung, die nicht mehr aus der Welt zu schaffen ist. Sie begründen allein durch ihre Existenz die Möglichkeit unserer Liquidierung. Und sie sind in der Lage, ohne unser Zutun klaglos zu funktionieren.

„Denn worauf wir abzielen, ist ja stets, etwas zu erzeugen, was unsere Gegenwart und Hilfe entbehren und ohne uns klaglos funktionieren könnte – und das heißt ja nichts anderes als Geräte, durch deren Funktionieren wir uns überflüssig machen, wir uns ausschalten, wir uns „liquidieren“. Dass dieser Zielzustand immer nur approximativ erreicht wird, das ist gleichgültig. Was zählt, ist die Tendenz. Und deren Parole heißt eben: «ohne uns».“³⁰¹

Das hat Anders in erster Linie auf die unmenschlichen Taten der Menschen bezogen, die den Maschinen überlassen werden.

„Das allerdings bedeutet für Anders, daß in einer technischen Zivilisation die unmenschlichen Taten zunehmend Taten ohne Menschen sind.“³⁰²

Das Wort „Liquidierung“ sollte nicht dazu verleiten, die Technik auf Tötungswaffen zu reduzieren. Es geht nicht allein um das Ende der Menschheit durch atomare Zerstörung, sondern darum, dass die Technik den Menschen überflüssig macht und *„daß wir durch die Herstellung dieser Geräte unsere Mündigkeit und unsere Autonomie endgültig liquidieren.“*³⁰³

Es geht um ein Prinzip und um eine Tendenz, die der Technik inhärent sind. In den Systemen der Künstlichen Intelligenz kommen sie deutlich zum Ausdruck. Die kontingente Eigenschaft der graduellen künstlichen Autonomie ist tendenziell in der Lage unsere Selbstbestimmung zu untergraben oder – mit den Worten von Anders – zu *„liquidieren“*. Die Tendenz des „ohne uns“ impliziert einen Verlust und einen Verzicht auf Mündigkeit. Selbstbestimmung wird schleichend und in kleinen Schritten durch Fremdbestimmung abgelöst.

³⁰⁰ Anders, *Die atomare Drohung*, 198.

³⁰¹ Anders, *Die atomare Drohung*, 198.

³⁰² Liessmann, *Die großen Philosophen und ihre Probleme*, 183.

³⁰³ Anders, *Die atomare Drohung*, 198.

Dazu braucht es keine Superintelligenz, die die Kontrolle über die Menschen übernimmt. Dies scheint weniger das Problem. Vielmehr ist entscheidend, ob wir in digitaler Mündigkeit die Technik gestalten können oder in Unmündigkeit die Abhängigkeit von Menschen und Unternehmen hinnehmen, die uns mithilfe von Algorithmen und Daten kontrollieren und steuern.³⁰⁴

³⁰⁴ Grunwald, *Der unterlegene Mensch*, 35.

4 Fazit

Ein bedingtes Ja als Antwort auf die Forschungsfrage

Wird individuelle Autonomie untergraben, wenn Künstliche Intelligenz anstelle des Menschen entscheidet und Informationen bereitstellt? Eine einfache Ja-Nein-Antwort auf die zentrale Forschungsfrage gibt es nicht. Die Ergebnisse dieser Arbeit deuten jedoch auf ein bedingtes Ja hin. Ob der Kippunkt erreicht wird, an dem Selbstbestimmung in Fremdbestimmung übergeht, hängt maßgeblich von der Gestaltung der Systeme ab. Opazität, ein hoher Grad an künstlicher Autonomie und die dem Menschen in gewisser Hinsicht überlegene Geschwindigkeit wirken dabei wie Parameter eines imaginären Kippunktes. Je stärker diese Eigenschaften ausgeprägt sind, desto eher kann individuelle Autonomie geschwächt werden.

Gleichzeitig ist dieser Übergang auch von der Fähigkeit und Bereitschaft des Menschen abhängig, eine reflektierende und distanzierte Haltung einzunehmen. Die Gestaltung der Systeme beeinflusst vor allem die Bereitschaft der Menschen. Die „smarte Macht“ der KI und ihre gezielt eingesetzten Mechanismen zielen nicht auf kritische Distanz, sondern darauf, Nutzer*innen „always on“ zu halten und die Annehmlichkeiten der Technologie mit persönlichen Daten zu erkaufen.

Der Übergang zur Fremdbestimmung lässt sich als ein schleichender Prozess verstehen, in dem wir uns mehr und mehr der Kontrolle und Steuerung omnipräsenter Systeme aussetzen. Dabei können wir unserer Verantwortung nicht mehr umfänglich gerecht werden, stehen vor epistemischen Herausforderungen, deren Überwindung kaum möglich erscheint und lassen uns von der smarten Macht omnipräsenter Systeme steuern:

- Entscheidung und *Verantwortung* sind untrennbar miteinander verbunden. Transparenz und Nachvollziehbarkeit sind Voraussetzungen für die Übernahme von Verantwortung. Wenn opake und komplexe autonome KI-Systemen entscheiden, können Menschen ihrer Verantwortung nicht in angemessener Weise und mit der nötigen Sorgfalt nachkommen. Wenn KI-Systeme etwas „verschulden“, lassen sie sich möglicherweise als Verursacher identifizieren, jedoch können sie nicht zur Rechenschaft gezogen werden. In der Konsequenz bleibt Verantwortung grundsätzlich beim Menschen. Sie kann Anwender*innen, Hersteller*innen oder Programmierer*innen zugeschrieben

werden, wobei die Kriterien ihrer Zuschreibung wegen der technologischen Komplexität unklar sind.

- *Epistemische Herausforderungen:* Approximative Trial-And-Error-Verfahren führen zu einem zufälligen Remix von Informationen, der plausibel erscheint aber mit Faktizität oft nichts zu tun hat, wie Deep Fakes bezeugen. Künstliche Intelligenz hat Wahrheit nicht im Repertoire. Sie spielt ebenso wenig eine Rolle wie Lüge. Beide sind ihr gleichgültig. Empfehlungssysteme oder Suchmaschinen können unseren Alltag erleichtern. Personalisiertes Wissen in Filterblasen trägt dazu bei, dass wir im Eigenen verbleiben. In einem auf jede*n einzelne*n von uns zugeschnittenen epistemischen Raum, werden tendenziell nur Informationen und Wissen bereitgestellt, die unsere Meinungen bestätigen.
- Der auf uns zugeschnittene epistemische Raum ist gleichzeitig auch ein Wohlfühlraum, den die smarte Macht omnipräsenter Systeme prägt. Die smarte Macht ist keine Disziplinarmacht. Sie droht nicht, sondern ist freundlich und angenehm. Sie benötigt keine Gewalt, um Widerstand zu brechen und fügt sich nahtlos in unseren Alltag ein. Genau dadurch wird sie wirksam.

Vor diesem Hintergrund lassen sich die Fragen beantworten, die im Grundlagenkapitel (2.1.4.3 Weiterführende Teilfragen) ausgehend von verschiedenen Facetten der Autonomiekonzepte herausgearbeitet wurden: Manipulation, zeitliche Autonomie, Privatheit und der soziale und gesellschaftliche Kontext, in dem individuelle Autonomie verankert ist.

Online-Manipulation menschlicher Willensbildung und Entscheidung: Online-Manipulation kann als *heimliche Beeinflussung* verstanden werden, die unbemerkt und verdeckt vor sich geht.³⁰⁵ Das macht das Erkennen von Manipulation in opaken Systemen unmöglich oder zumindest ungleich schwieriger als in Christmans Sektenbeispiel. *Manipulation* als heimliche, verdeckte Beeinflussung menschlicher Willensbildung, Entscheidung und Handlung – meist im Interesse einer manipulierenden Instanz – manifestiert sich im World Wide Web beispielsweise in Filterblasen, die den Bereich der für meine Entscheidung relevanten Option schmälert; zudem über Biases, die Voreingenommenheit und Verzerrung algorithmisch erzeugen oder aus den benutzten Quellen weiter transportieren. Eine wichtige Rolle bei der Manipulation spielen zudem

³⁰⁵ Susser u. a., „Online Manipulation“.

unbeabsichtigte und beabsichtigte Unwahrheiten. Erstere resultieren aus der prinzipiellen Unfähigkeit der Künstlichen Intelligenz zur Wahrheit, die etwa zu *Halluzinationen*, *Bullshit* oder schlicht der *Gleichgültigkeit* der Systeme gegenüber Wahrheit führen. Zwar sind KI-Systeme auch nicht in der Lage, willentlich die Unwahrheit zu sagen. Jedoch werden über die Systeme Fake News, Desinformation und Verschwörungstheorien verbreitet – Themen, die in dieser Arbeit nicht tiefergehend aufgegriffen wurden. Mit der Opazität der omnipräsenten Systeme und ihrer smarten Macht geht eine Manipulation der Menschen einher. Auch Geschwindigkeit trägt dazu bei. B. Pörksen spricht von der „Hölle der Desinformation“³⁰⁶. Fake News und Desinformation werden „*irrwitzig schnell*“³⁰⁷ verbreitet – so Pörksen. Dabei bleibt die Wahrheit auf der Strecke. Informationen werden zunächst dekontextualisiert und dann in einer Art Remix in neue plausibel klingende Zusammenhänge gestellt bevor sie „*barrierefrei*“ weltweit zugänglich gemacht werden.³⁰⁸ In diesem Remix geht die Faktizität verloren.

Zeitliche Autonomie: Im „*Geschwindigkeitsrausch der vernetzten Welt*“³⁰⁹ erleben wir Künstliche Intelligenz als prägend für die Gestaltung unserer Zeit und damit unseres Lebens. Wann und in welchem Rhythmus wir etwas tun und wieviel Zeit wir uns dafür lassen, ist auch durch Technologie (fremd)bestimmt. Von der smarten Macht der digitalen Geräte geht ein „*Zwang zur Kommunikation*“ aus. Sie tragen zum veränderten Umgang mit unserer Zeit bei. Aus dem Wunsch nach mehr freier Zeit ist der Wunsch nach Verbundenheit mit der Welt geworden. Freie Zeit wird zur „*medialen Eigenzeit*“. KI-Systeme werden zu Zeit- und Taktgebern unserer beruflichen und privaten Lebenswelt. Wer sich Zeit für einen traditionellen Brief nimmt, schickt ihn vielleicht nicht sofort los, schläft eine Nacht darüber, weil er zu wichtig erscheint und im ersten Ärger verfasst worden ist. Womöglich landet der Brief am nächsten Morgen im Papierkorb oder braucht längere Zeit bis er seine*n Adressaten*in erreicht. In der digitalen Kommunikation kommt ein solches Verhalten nahezu einem Regelverstoß gleich. Es wird sofortige

³⁰⁶ Pörksen, „*Die Hölle der Desinformation - Spielregeln der Wirklichkeitsordnung im digitalen Zeitalter*“, 128.

³⁰⁷ Pörksen, „*Die Hölle der Desinformation - Spielregeln der Wirklichkeitsordnung im digitalen Zeitalter*“, 128.

³⁰⁸ Pörksen, „*Die Hölle der Desinformation - Spielregeln der Wirklichkeitsordnung im digitalen Zeitalter*“, 129.

³⁰⁹ Pörksen, „*Die Hölle der Desinformation - Spielregeln der Wirklichkeitsordnung im digitalen Zeitalter*“, 132.

Reaktion erwartet. Digitale Kommunikation auf Social-Media-Plattformen erfolgt ungebremst und ohne „Airbag aus Zeit“³¹⁰.

Instrumentalisierung von Privatheit: Privat ist etwas, das zu mir gehört, über das ich frei verfügen kann und darf. Ich kann bestimmen und kontrollieren, ob, wann und wer Zugang zu meinen privaten Dingen und Angelegenheiten hat bzw. haben soll. Meine Daten, meine Entscheidungen und meine Handlungen gehören in diesem Sinne zu meiner Privatheit. So wurde es in dieser Arbeit beschrieben. Privatheit ist eine notwendige und konstitutive Bedingung für individuelle Autonomie. Informationelle Privatheit bezieht sich auf den Zugriff und die Nutzung privater Daten. Im Konzept und in den Algorithmen vieler KI-Systeme sind der Zugriff auf und die Nutzung von privaten Daten verankert. So weist etwa OpenAI als Hersteller von ChatGPT darauf hin, dass öffentlich im Internet verfügbare Informationen sowie Daten, die Nutzer*innen generieren, zum Füttern der Textmaschine benutzt werden.³¹¹ Der Zugriff auf private Daten scheint von weiten Teilen der Gesellschaft legitimiert zu werden. Die „smarte Macht“ der Systeme rückt die Bedenken vieler Nutzer*innen in den Hintergrund.

KI-Systeme dringen auch in unsere lokale Privatheit ein. Die Endgeräte des World Wide Web sind in unseren Wohnungen beheimatet. Wir tragen sie als Mobilgeräte mit uns herum. Sie fungieren als Datensammelstelle für alle nur erdenklichen Informationen über uns. Wann ich das Gerät einschalte, mit wem ich kommuniziere, auf welchen Social Media Plattformen ich mich aufhalte, welche Suchanfragen ich stelle, worum ich ChatGPT bitte, welche Websites ich besuche, wie lange, wie oft, und in welchen Abständen ich mich dort aufhalte, welche Buttons ich klicke und welche nicht, wo ich vergeblich klicke, was mich auf der Website interessiert, was ich dort kaufe und wohin und an wen ich es schicken lasse, wie ich mit Kaufempfehlungen umgehe und andere Dinge mehr. Ich liefere Daten, die in Verhaltensmodelle über mich einfließen, und zwar mithilfe meines eigenen, entblößten Verhaltens. Im Gegenzug dazu werden die entstandenen Modelle zur Beeinflussung meiner dezisionalen Privatheit genutzt. Die Opazität der Systeme macht die Kontrolle über meine Privatheit unmöglich. Wer genau über meine Daten verfügt, welche das genau sind und wo sie genutzt werden, bleibt im Dunkeln. Günther Anders, der noch nichts von künstlicher Intelligenz in der heutigen Form wissen konnte, schrieb über die „Antiquiertheit der Privatheit“ schon 1958: „Das

³¹⁰ Menasse, *Alles und nichts sagen*, 44.

³¹¹ OpenAI, „How ChatGPT and our foundation models are developed“.

*Haus wird der Welt ausgeliefert.*³¹² Er hatte dabei vorrangig Abhör-Apparate im Sinn. Heute werden nicht nur akustische Daten, sondern Daten aller Art von den Datensammel-Apparaten in Empfang genommen. KI-Systeme haben sich in unsere lokale Privatheit eingenistet und sie sind gleichzeitig omnipräsent.

*„Wenn die anderen in der Lage sind, sich an der Raumstelle, an der ich mich befinde, gleichfalls aufzuhalten, und zwar unmerklich, dann habe ich aufgehört Monopolist meiner Raumstelle zu sein und damit das letzte und formalste Minimum meiner Freiheit eingebüßt.“*³¹³

Die Allgegenwärtigkeit hat Unfreiheit sowohl für die Erreichbaren als auch für die Erreichten zur Folge.³¹⁴

Sozialer und gesellschaftlicher Kontext: Der soziale und gesellschaftliche Kontext individueller Autonomie stand nicht im zentralen Interesse der Arbeit. Das Fazit lässt deshalb in dieser Hinsicht wenig zu und viele Fragen offen. Dennoch sollen einige Aspekte genannt werden.

Die Protagonisten des Ubiquitous Computing verbanden mit ihrer Idee die Hoffnung, die Möglichkeiten des menschlichen Zusammenseins zu bereichern. Dass sich diese Hoffnung oft nicht erfüllt, lässt sich an Beispielen zeigen.

- Echokammern dienen der abgeschotteten Meinungs- und Identitätsbildung. Informationen und Meinungen, die von außen herangetragen werden, jedoch von der eigenen verfestigten Meinung abweichen, führen oft nicht zu einem Dialog oder einer Verständigung, sondern zur weiteren Polarisierung und zur Verhärtung der Fronten. Dabei bieten Soziale Medien eine Plattform, über die soziale Bindungskräfte geschwächt werden. Dann kann eine Gesellschaft in soziale Subsysteme zerfallen, zwischen denen Verständigung und Kompromisse kaum noch möglich sind.³¹⁵ Gut zu beobachten war dies während der Covid-Pandemie. KI ist dafür nicht die Ursache und auch nicht für die politische Instrumentalisierung, die in diesem Zusammenhang erfolgte.
- Menschen knüpfen Hoffnungen an soziale Medien. Sie glauben an die smarten Versprechen. Durch ihre Follower und viele Likes erfahren sie (trügerische)

³¹² Anders, *Die Antiquiertheit des Menschen*. 2, 231.

³¹³ Anders, *Die Antiquiertheit des Menschen*. 2, 247.

³¹⁴ Anders, *Die Antiquiertheit des Menschen*. 2, 247.

³¹⁵ Metzinger, „Fukushima der künstlichen Intelligenz“.

Anerkennung. Aber die digitale Kommunikation hat wenig Raum für echte Zuwendung und Respekt wohl aber für anonyme Angriffe und Beleidigungen. Shitstorms mögen ein Beispiel dafür sein. Und auch bei der digitalen Selbstdarstellung von Influencer*innen erweist sich die digitale Technik als eine „Narzisstische Ego-Maschine“.

- Geschwindigkeit und Unkontrollierbarkeit hochgradig autonomer KI-Systeme können – global betrachtet – zu einem Problem werden, das die gesamte Menschheit betrifft, nicht zuletzt im Hinblick auf Gerechtigkeit. Metzinger nennt als Beispiel weltweite digitale Finanztransaktionen:

„Wenn sich aber unerwartete Effekte zu einer Finanzkrise aufschaukeln, dann trägt den Schaden am Ende die Allgemeinheit: Die Transaktionen geschehen so schnell, dass Menschen nicht mehr eingreifen können, und die Verantwortlichkeit ist nicht mehr klar bestimmt. Den Schaden tragen am Ende die Weltwirtschaft und die Menschheit als Ganzes, und da besonders die schwächsten Teilnehmer.“³¹⁶

Es drängt sich geradezu ein Vergleich mit den Folgen des Klimawandels auf, von dem die schwächsten Staaten und Menschen besonders betroffen sind, ohne dafür maßgeblich verantwortlich zu sein.

Machtverschiebungen: Die epistemische Deutungshoheit obliegt nicht mehr allein Menschen. Wenn der Mensch fragt und die Antwort der Maschine unbedacht übernimmt, überlässt er die Deutung des Wissens der Künstlichen Intelligenz. Deren „smarte Macht“ trägt mit zu einer Verschiebung gesellschaftlicher und politischer Macht bei. In der US-amerikanischen Gesellschaft und Politik haben die Milliardäre der großen Tech-Konzerne erheblichen Einfluss. Den haben sie nicht zuletzt Künstlicher Intelligenz (und ihren Vorläufer-Technologien: WWW, Suchmaschinen...) zu verdanken, die ihnen direkte Kontrolle und Steuerung von Menschen ermöglicht und indirekt zu einem gigantischen Reichtum verholfen hat, über den sie Macht ausüben können. Damit nehmen sie maßgeblich an der Gestaltung der gesellschaftlichen und politischen Bedingungen,

³¹⁶ Metzinger, „Fukushima der künstlichen Intelligenz“.

auch der individuellen Autonomie teil. Wenn dabei das Wohl der „*gesamten Menschheit*“ als Zukunftsvision bemüht wird, mag das wie blanker Zynismus klingen.³¹⁷

Künstliche Intelligenz verhilft auch Politikern an die Macht. Das ehemalige britische Unternehmen Cambridge Analytica arbeitete als Datenanalyse- und Beratungsunternehmen u.a. bei politischen Kampagnen mit. So auch im Rahmen der US-Präsidentschaftswahl 2016, als es digitale Nutzerprofile in großem Umfang erstellte. Ausgangspunkt war ein Persönlichkeitstest namens „thisisyourdigitallife“ (2014), entwickelt von Aleksandr Kogan, der als Dozent an der Universität Cambridge tätig war. US-amerikanische Facebook-Nutzer erhielten eine kleine Vergütung für die Teilnahme und die Freigabe ihrer Kontodaten. Über eine damals aktive Funktion, die den Zugriff auf die Konten der Facebook-Freunde erlaubte, sammelte Cambridge Analytica so Daten von bis zu 87 Millionen Nutzern. Allein aufgrund von Facebook-„Likes“ konnte eine enorme Menge an Daten über Einzelpersonen abgeleitet werden wie Geschlecht, sexuelle Orientierung, politische Einstellung oder Beziehungsstatus.³¹⁸ Die Daten nutzte Cambridge Analytica zur Erstellung psychologischer Profile, die im Präsidentschaftswahlkampf 2016 zur verdeckten Wählermanipulation eingesetzt wurden, die gezielt Ängste der Nutzer*innen instrumentalisierte. Christopher Wylie, ein Mitbegründer von Cambridge Analytica:

*“We exploited Facebook to harvest millions of people's profiles and built models to exploit what we knew about them and target their inner demons.”*³¹⁹

Der Cambridge-Analytica-Skandal belegt Laniers These vom BUMMER-Effekt.

Perspektivwechsel

Die Forschungsfrage hat die Richtung des Diskurses dieser Arbeit bestimmt. Es wurden Thesen entwickelt, die ein bedingtes Ja erlauben. Wenn die Gestaltung der KI-Systeme dafür maßgeblich ist, ob sie individuelle Autonomie untergraben können, dann ist die umgekehrte Frage naheliegend: Können transparente, nachvollziehbare und kontrollierbare Systeme – die jederzeit menschliches Eingreifen ermöglichen, deren künstliche Autonomie die Menschen nicht auf smarte Weise zu überrumpeln versucht, die nicht allein (ökonomischen) Machtinteressen folgt – zu mehr individueller Autonomie

³¹⁷ OpenAI, „ChatGPT 4.0“.

³¹⁸ Susser u. a., „Online Manipulation“, 10.

³¹⁹ Zitiert nach: Susser u. a., „Online Manipulation“, 10.

beitragen? Auch hier könnte die Antwort ein – wenn auch bedingtes – Ja sein. So ist etwa Metzinger der Auffassung:

„KIs hingegen können so konstruiert werden, dass sie keine kognitiven Verzerrungen aufweisen. Prinzipiell könnte also gesteigertes Vertrauen in die Prognosen von KIs, sofern diese sicher und nach nachvollziehbaren Kriterien aufgebaut sind, auch zu einer deutlichen Rationalitätssteigerung bei vielen gesellschaftlichen und politischen Herausforderungen führen.“³²⁰

In der KI-Forschung hat sich seit Jahren eine Disziplin etabliert, die sich Fragen der „*emotionalen künstlichen Intelligenz*“ widmet. Sie beschäftigt sich mit der „*Nachbildung oder Simulation von Emotionen, [...] also damit, wie künstliche Systeme Emotionen erkennen, ausdrücken und vielleicht sogar selbst aufweisen können.*“³²¹ Wenn Roboter etwa in der Pflege von hilfebedürftigen Menschen zum Einsatz kommen oder als Gesprächspartner für Menschen fungieren, können sie dabei überzeugender und Vertrauen erweckender sein, wenn sie Emotionen zeigen (simulieren) und den Menschen empathisch begegnen – so jedenfalls die Annahme. Auch soziale Agenten können Emotionen z.B. in Form von Zugewandtheit vorspiegeln. Im Kapitel 3.3.1.1 Graduelle künstliche Autonomie wurde das am Beispiel der Internetplattform „*Replika*“ bereits erwähnt. In diesem Zusammenhang wurde auch darauf hingewiesen, dass im wissenschaftlichen Diskurs Positionen vertreten werden, die nahelegen, Mensch-Maschine-Interaktionen nicht streng nach dem Muster menschlicher Beziehungen zu beurteilen. Vielmehr könnten sogenannte „Minimal approaches“ ein weites Feld sozial-kognitiver Fähigkeiten abdecken. Zentral seien: „*minimal mindreading, minimal sense of commitment, and shared intentions lite*“³²²

Dass Menschen emotionale Beziehungen zu künstlichen Artefakten aufbauen können, ist bekannt. Dieses Thema wird in Literatur und Film immer wieder aufgegriffen. Zu Beginn des 19. Jahrhunderts thematisierte E.T.A. Hoffmann in seiner Erzählung „*Der Sandmann*“ die Liebe seines Protagonisten Nathanael zu einer mechanischen Puppe. In Maria Schraders Film „*Ich bin dein Mensch*“ nimmt Alma, eine junge Wissenschaftlerin, an einem Experiment teil, in dem sie mehrere Monate mit Tom, einem Androiden,

³²⁰ Metzinger, „Fukushima der künstlichen Intelligenz“.

³²¹ Misselhorn, *Künstliche Intelligenz und Empathie*, 12.

³²² Strasser, „From tool use to social interactions“, 78.

zusammenleben soll. Tom ist von einem realen Menschen nicht zu unterscheiden. Eigentlich ist er ein idealer Typ, der genau das tut und antizipiert, was sie an Bedürfnissen und Erwartungen hat. Tom widerspricht nie und erfüllt ihr jeden Wunsch. Nach anfänglicher Skepsis fühlt Alma sich zu ihm hingezogen. Sie kommen sich näher und haben Sex miteinander. Am Ende, nachdem Alma sich schweren Herzens von Tom getrennt hat, steht ein Fazit der Protagonistin:

„Die Geschichte der Menschheit ist voll von vermeintlichen Verbesserungen, deren schwerwiegende Folgen sich erst Jahre, Jahrzehnte oder sogar Jahrhunderte später ins Bewusstsein drängen. [...] Doch ist der Mensch wirklich gemacht für eine Befriedigung seiner Bedürfnisse, die per Bestellung zu haben ist. Sind nicht gerade die unerfüllte Sehnsucht, die Fantasie und das ewige Streben nach Glück die Quelle dessen, was uns zum Menschen macht?“³²³

Diese Frage des Films ist nicht als Schlussfolgerung am Ende dieser Arbeit zu interpretieren. Wohl aber soll sie als Hinweis darauf dienen, dass die philosophische Auseinandersetzung mit Künstlicher Intelligenz weit über das hinausgeht, was in diesem Diskurs aufgegriffen wurde – und stets auch mit Blick auf die Zukunft weiter zu denken ist.

³²³ *Ich bin dein Mensch*, Im Epilog des Films.

5 Literatur

- Ach, Johann S., und Bettina Schöne-Seifert. „Relationale Autonomie. Eine kritische Analyse“. In *Patientenautonomie: Theoretische Grundlagen - Praktische Anwendungen*. o. J.
- Anders, Günther. *Die Antiquiertheit des Menschen. 1: Über die Seele im Zeitalter der zweiten industriellen Revolution*. Nachdr. der 7., Unveränd. Aufl. der Orig.-Ausg. Beck'sche Reihe 319. Beck, 1994.
- Anders, Günther. *Die Antiquiertheit des Menschen. 2: Über die Zerstörung des Lebens im Zeitalter der dritten industriellen Revolution*. 4., Unveränd. Aufl., Nachdruck. Beck'sche Reihe 320. Beck, 1995.
- Anders, Günther. *Die atomare Drohung: radikale Überlegungen*. 2., Durch e. Vorw. erw. Aufl. von "Endzeit und Zeitenende". Beck'sche Schwarze Reihe, Bd. 238. Beck, 1981.
- Arendt, Hannah. *Vita activa oder Vom tätigen Leben*. Erweiterte Neuausgabe. Herausgegeben von Thomas Meyer. Serie Piper 31691. Piper, 2021.
- Aristoteles. *Nikomachische Ethik*. Durchgesehene Ausgabe 2019, [Nachdruck] 2023. Übersetzt von Gernot Krapinger. Reclams Universal-Bibliothek, Nr. 19448. Reclam, 2019.
- Beauchamp, Tom L., und James F. Childress. *Principles of biomedical ethics*. 7th ed. Oxford University Press, 2013.
- Beier, Kathi. *Selbsttäuschung*. Grundthemen Philosophie. De Gruyter, 2010.
- Benjamin, Walter. *Das Kunstwerk im Zeitalter seiner technischen Reproduzierbarkeit: Dritte, autorisierte letzte Fassung, 1939*. Null Papier Verlag, 2018.
- Benson, Paul. „Handlungsfreiheit und Selbstwert“. In *Autonomie der Person*, herausgegeben von Monika Betzler. Münster, 2013.
- Betzler, Monika, Hrsg. *Autonomie der Person*. Map - Mentis Anthologien Philosophie. Mentis, 2013.
- Burrell, Jenna. „How the Machine ‘Thinks’: Understanding Opacity in Machine Learning Algorithms“. *Big Data & Society* 3, Nr. 1 (2016): 2053951715622512. <https://doi.org/10.1177/2053951715622512> (Zugriff: 22.05.2025).
- Caluori, Lucas. „Hey Alexa, Why Are You Called Intelligent? An Empirical Investigation on Definitions of AI“. *AI & SOCIETY*, Nr. 39 (März 2023): 1905–19. <https://doi.org/10.1007/s00146-023-01643-y> (Zugriff 22.02.2025).
- Chaplin, Charlie, Reg. *Modern Times*. Spielfilm. 1936. <https://www.youtube.com/watch?v=uY6ItxyNGgA> (Zugriff: 12.10.2024).

- Christman, John. „Autonomy and Personal History“. *Canadian Journal of Philosophy* 21, Nr. 1 (1991): 1–24. <https://doi.org/10.1080/00455091.1991.10717234> (Zugriff 22.05.2024).
- Christman, John. „Autonomy in Moral and Political Philosophy“. In *The Stanford Encyclopedia of Philosophy (Winterausgabe 2022)*, Edward N. Zalta und Uri Nodelman (Hrsg.). 28. Juli 2003. <https://plato.stanford.edu/archives/win2022/entries/autonomy-moral/> (Zugriff 12.07.2024).
- Clark, Andy, und David J. Chalmers. „The Extended Mind“. *The Philosopher's Annual*, vol XXI, 1998, 1998. <https://consc.net/papers/extended.html> (Zugriff 22.11.2024).
- Coeckelbergh, Mark. *Digital Technologies, Temporality, and the Politics of Co-Existence*. Springer International Publishing, 2022. <https://doi.org/10.1007/978-3-031-17982-2> (Zugriff 12.01.2025).
- Coeckelbergh, Mark. „LLMs, Truth, and Democracy: An Overview of Risks“. *Science and Engineering Ethics* 31, Nr. 1 (2025): 4. <https://doi.org/10.1007/s11948-025-00529-0> (Zugriff 20.02.2025).
- Diakopoulos, Nicholas. „Accountability in Algorithmic Decision Making“. *Communications of the ACM* 59, Nr. 2 (2016): 56–62. <https://doi.org/10.1145/2844110> (Zugriff 22.02.2025).
- Dorn, Thea. „Wie verändern intelligente Maschinen unser Selbstverständnis?“ Protokoll der Jahrestagung des Deutschen Ethikrates. Autonome Systeme. Wie intelligente Maschinen uns verändern, Berlin, 21. Juni 2017.
- Dudenredaktion. „Intelligenz“. In *Duden online*. ohne Datum. <https://www.duden.de/rechtschreibung/Intelligenz> (Zugriff 30.07.2024).
- DWDS – Digitales Wörterbuch der deutschen Sprache. *Autonomie*. Herausgegeben von Berlin-Brandenburgischen Akademie der Wissenschaften. 15. November 2023. <https://www.dwds.de/wb/Autonomie> (Zugriff 12.07.2025).
- DWDS – Digitales Wörterbuch der deutschen Sprache. *Autonomie*. Herausgegeben von Berlin-Brandenburgischen Akademie der Wissenschaften. <https://www.dwds.de/wb/Autonomie> (Zugriff 12.07.2025).
- DWDS – Digitales Wörterbuch der deutschen Sprache. *Freiheit*. Herausgegeben von Berlin-Brandenburgischen Akademie der Wissenschaften. <https://www.dwds.de/wb/Freiheit#d-1-3> (Zugriff 24.09.2024).
- DWDS – Digitales Wörterbuch der deutschen Sprache. *Intelligenz*. Herausgegeben von Berlin-Brandenburgischen Akademie der Wissenschaften. Zugegriffen 30. Juli 2024. <https://www.dwds.de/wb/Intelligenz> (Zugriff 30.07.2024).

- DWDS – Digitales Wörterbuch der deutschen Sprache. *Souveränität*. Herausgegeben von Berlin-Brandenburgischen Akademie der Wissenschaften. <https://www.dwds.de/wb/Souver%C3%A4nit%C3%A4t> (Zugriff 23.04.2024).
- Dworkin, Gerald. „Paternalism“. In *The Stanford Encyclopedia of Philosophy*, Winter 2022, herausgegeben von Edward N. Zalta. Metaphysics Research Lab, Stanford University, 2022. <https://plato.stanford.edu/archives/win2022/entries/paternalism/> (Zugriff 30.07.2024).
- Dworkin, Gerald. *The Theory and Practice of Autonomy*. 1. Aufl. Cambridge University Press, 1988. <https://doi.org/10.1017/CBO9780511625206> (Zugriff 30.07.2024).
- Foucault, Michel. *Überwachen und Strafen: die Geburt des Gefängnisses*. 20. Auflage. Übersetzt von Walter Seitter. Suhrkamp-Taschenbuch 2271. Suhrkamp, 2024.
- Frankfurt, Harry G. „Freedom of the Will and the Concept of a Person“. *The Journal of Philosophy* 68, Nr. 1 (1971): 5. <https://doi.org/10.2307/2024717> (Zugriff 30.06.2024).
- Frankfurt, Harry G. *On Bullshit*. Princeton University Press, 2005. <https://doi.org/10.1515/9781400826537> (Zugriff 30.07.2024).
- Gerl-Falkovitz, Hanna-Barbara. „Stößt Kants ‚Autonomie‘ die klassische kirchliche Sexualmoral um?“ *die-tagespost.de*, o. J. Zugegriffen 13. Dezember 2023. <https://www.die-tagespost.de/sonder-texte/beilage/stoesst-kants-autonomie-die-klassische-kirchliche-sexualmoral-um-art-234131> (Zugriff 30.12.2023).
- Glaser, Peter. „Kulturelle Atomkraft“. Politik. *Berliner Zeitung*, 24. August 2009. <https://www.berliner-zeitung.de/archiv/die-digitalisierung-zersetzt-alte-medienformen-ihre-atome-suchen-hitzig-nach-neuer-synthese-kulturelle-atomkraft-li.981043> (Zugriff 30.03.2025).
- Goertz, Stephan, und Carolin Witting. „Autonomie“. In *Staatslexikon online*, 8. Aufl. Herder, 2022. <https://www.staatslexikon-online.de/Lexikon/Autonomie> (Zugriff 30.06.2024).
- Gransche, Bruno und Institut für Systemtechnik und Innovationsforschung, Hrsg. *Wandel von Autonomie und Kontrolle durch neue Mensch-Technik-Interaktionen*. Karlsruhe, 2014. <https://doi.org/10.2314/GBV:83347488X> (Zugriff 13.03.2025).
- Grimm, Petra, Tobias O. Keber, und Oliver Zöllner, Hrsg. *Digitale Ethik: Leben in vernetzten Welten*. Kompaktwissen XL, Nr. 15240. Reclam, 2019.
- Grunwald, Armin. *Der unterlegene Mensch: Die Zukunft der Menschheit im Angesicht von Algorithmen, künstlicher Intelligenz und Robotern*. Originalausgabe. Riva Premium, 2019.
- Grunwald, Armin, und Rafaela Hillerbrand, Hrsg. *Handbuch Technikethik*. 2., Aktualisierte und Erweiterte Auflage. J. B. Metzler Verlag, 2021.

- Han, Byung-Chul. *Im Schwarm: Ansichten des Digitalen*. 5. Auflage 2020. Matthes & Seitz Berlin, 2013.
- Han, Byung-Chul. *Infokratie: Digitalisierung und die Krise der Demokratie*. Erste Auflage. Fröhliche Wissenschaft 184. Matthes & Seitz Berlin, 2021.
- Han, Byung-Chul. *Psychopolitik Neoliberalismus und die neuen Machttechniken*. S. Fischer, 2014.
- Hicks, Michael Townsen, James Humphries, und Joe Slater. „ChatGPT Is Bullshit“. *Ethics and Information Technology* 26, Nr. 2 (2024): 38.
<https://doi.org/10.1007/s10676-024-09775-5> (Zugriff 30.01.2025).
- Hochrangige Expertengruppe für künstliche Intelligenz. „Eine Definition der KI: Wichtigste Fähigkeiten und Wissenschaftsgebiete“. 2019.
- Hochrangige Expertengruppe für künstliche Intelligenz. „Ethik-leitlinien für eine vertrauenswürdige KI.“ Herausgegeben von European Commission. Directorate General for Communications Networks, Content and Technology. Publications Office, 2019.
- Höffe, Otfried. *Immanuel Kant*. Originalausgabe, 9., Überarbeitete Auflage. C.H. Beck Paperback 506. Verlag C.H. Beck, 2020.
- Höffe, Otfried, Hrsg. *Lexikon der Ethik*. Achte, Überarbeitete und Ergänzte Auflage. C.H.Beck, 2023.
- Holl, Jann, Hans Lenk, und Matthias Maring. „Verantwortung“. In *Historisches Wörterbuch der Philosophie online*. Schwabe Verlag, 2017.
<https://doi.org/10.24894/HWPh.5517>.
- Hu, Qian, Yaobin Lu, Zhao Pan, Yeming Gong, und Zhilin Yang. „Can AI Artifacts Influence Human Cognition? The Effects of Artificial Autonomy in Intelligent Personal Assistants“. *International Journal of Information Management* 56 (Februar 2021): 102250. <https://doi.org/10.1016/j.ijinfomgt.2020.102250> (Zugriff 11.03.2025).
- Hübner, Dietmar. *Einführung in die philosophische Ethik*. 3., Erneut durchgesehene und Korrigierte Auflage. utb Philosophie 4121. Vandenhoeck & Ruprecht, 2021.
- Institut für Technikfolgen-Abschätzung (ITA). *Transparente Algorithmen – Wie lässt sich algorithmische Diskriminierung verhindern?* o. J.
- Introna, Lucas. *Phenomenological Approaches to Ethics and Information Technology*. 19. Februar 2005. <https://plato.stanford.edu/archives/fall2023/entries/ethics-it-phenomenology/#toc> (Zugriff 11.03.2025).
- ISO. „ISO 8373:2021(en), Robotics — Vocabulary“. <https://www.iso.org/obp/ui/en/#iso:std:iso:8373:ed-3:v1:en> (Zugriff 27.06.2025).

- Jaspers, Karl. *Die großen Philosophen. Erster Band*. 3. Auflage, Ungekürzte Taschenbuchausgabe. Die grossen Philosophen / Karl Jaspers, Band 1. Piper, 2018.
- Kahneman, Daniel. *Schnelles Denken, langsames Denken*. 25. Auflage. Übersetzt von Thorsten Schmidt. Siedler, 2019.
- Kant, Immanuel. *Kants gesammelte Schriften*. Herausgegeben von Königlich preussischen Akademie der Wissenschaften [AA]. G. Reimer, 1900ff. <http://kant.korpora.org/index.html> (Zugriff 12.03.2025).
- Kersten, Jens. „Die maschinelle Person – Neue Regeln für den Maschinenpark?“ In *Roboter, Computer und Hybride- Was ereignet sich zwischen Menschen und Maschine*, Hrsg. A. Manzescheke Arne und Fabian Karsch, 89–106. Baden-Baden: Nomos Verlag. o. J.
- Klein, Andreas. „Umbrüche in Neuzeit, Aufklärung, Moderne - Handout zur LV Theologie im Studium Generale der Universität Wien, WS 2023/2024“. Handout. Wien, 2023.
- Kubrick, Stanley, Reg. *2001: a space odyssey*. 1968.
- Lanier, Jaron. *Zehn Gründe, warum du deine Social Media Accounts sofort löschen musst*. 1. Auflage. Übersetzt von Martin Bayer und Karsten Petersen. Atlantik, 2019.
- Lau, Chong-Fuk. „Hegels Verwirklichung des Kantischen Autonomiebegriffs“. *Hegel-Jahrbuch* 2017, Nr. 1 (2017): 192–96. <https://doi.org/10.1515/hgjb-2017-0129> (Zugriff 03.07.2024).
- Lauer, David. „Ist die KI Akteurin oder Werkzeug?“ Vortrag, 15. September 2023. <https://www.deutschlandfunknova.de/beitrag/kuenstliche-intelligenz-ist-die-ki-akteurin-oder-werkzeug> (Zugriff 30.07.2024).
- Lenzen, Manuela. *Künstliche Intelligenz: was sie kann & was uns erwartet*. 5. Auflage. C.H. Beck, 2023.
- Liessmann, Konrad Paul. *Die großen Philosophen und ihre Probleme: Vorlesungen zur Einführung in die Philosophie*. 4., Unveränderte Auflage. UTB Philosophie 2247. Facultas, WUV-Universitätsverlag, 2003.
- Liessmann, Konrad Paul. „„Ohne uns?“ Der Traum der Maschinen und die Zukunft des Menschen“. Vortrag im Rahmen der Veranstaltung „Digitale Zukunft - Wo bleibt der Mensch?“ des Wissenschaftsvereins Kärnten, Villach, 21. November 2023. <https://www.youtube.com/watch?v=cXtEDE4mqBI> (Zugriff 30.09.2024).
- Liessmann, Konrad Paul, Gerhard Zenaty, und Katharina Lacina. *Vom Denken: Einführung in die Philosophie*. 6. Auflage. Verlag Holder-Pichler-Tempsky GmbH, 2018.
- Mainzer, Klaus. „Was sind KI und Machine Learning? Grundlagen und Zukunftsperspektiven“. In *Mensch-Maschine-Interaktion*, von Jörg Noller,

- Ramón Reichert, Stefan Rieger, u. a., herausgegeben von Orsolya Friedrich, Johanna Seifert, und Sebastian Schleidgen. Brill | mentis, 2022.
https://doi.org/10.30965/9783969752609_002 (Zugriff 03.07.2024).
- Maria Schrader, Reg. *Ich bin dein Mensch*. Spielfilm. Apple TV, 2021. 1 Std. 45 Min.
- Menasse, Eva. *Alles und nichts sagen: vom Zustand der Debatte in der Digitalmoderne*. Kiepenheuer & Witsch, 2023.
- Metzinger, Thomas. „Risikobewertung: Fukushima der künstlichen Intelligenz“. o. J. Zugegriffen 10. Juli 2024. <https://www.spektrum.de/news/interview-die-unterschaetzten-risiken-der-kuenstlichen-intelligenz/1377620> (Zugriff 30.03.2025).
- Metzlers Lexikon der Philosophie. *Empirismus*. Zugegriffen 9. Juni 2025. <https://www.spektrum.de/lexikon/philosophie/empirismus/552> (Zugriff 11.02.2025).
- Misselhorn, Catrin. *Grundfragen der Maschinenethik*. 5., Durchgesehene und Überarbeitete Auflage. Reclams Universal-Bibliothek, Nr. 19583. Reclam, 2022.
- Misselhorn, Catrin. *Künstliche Intelligenz und Empathie: vom Leben mit Emotionserkennung, Sexrobotern & Co.* 2. Auflage. Reclams Universal-Bibliothek Was bedeutet das alles?, Nr. 14047. Reclam, 2021.
- Netflix. „Das Dilemma mit den sozialen Medien“. <https://www.netflix.com/watch/81254224> (Zugriff 11.05.2025).
- Nguyen, C. Thi. „Echo Chambers and Epistemic Bubbles“. *Episteme* 17, Nr. 2 (2020): 141–61. <https://doi.org/10.1017/epi.2018.32> (Zugriff 13.04.2025).
- Nida-Rümelin, Julian, und Fiorella Battaglia. „Mensch, Maschine und Verantwortung“. In *Bendel, Handbuch Maschinenethik*, herausgegeben von Oliver Bendel. Springer Fachmedien Wiesbaden, 2019. https://doi.org/10.1007/978-3-658-17483-5_5 (Zugriff 22.04.2025).
- „Nietzsche Source — Digitale Kritische Gesamtausgabe Werke und Briefe (eKGWB), Za-IV“. With Giorgio Colli und Manzano Montinari. Berlin/New York, 1967ff. <http://www.nietzschesource.org/#eKGWB/Za-IV> (Zugriff 12.12.2024).
- Nissenbaum, Helen. *Privacy in Context: Technology, Policy, and the Integrity of Social Life*. Stanford Law Books, an imprint of Stanford University Press, 2010. https://search-ebscohost-com.uaccess.univie.ac.at/login.aspx?direct=true&db=nlebk&AN=309251&site=ehost-live&ebv=EK&ppid=Page-__-18 (Zugriff 13.04.2025).
- Nowotny, Helga. „Keine Zeit für Utopien“. 20. Juli 2018. <https://www.soziopolis.de/keine-zeit-fuer-utopien.html> (Zugriff 13.05.2025).
- OpenAI. „ChatGPT 4.0“. <https://openai.com/about/> (Zugriff 13.05.2025).

- OpenAI. „How ChatGPT and our foundation models are developed“. Zugegriffen 6. Juli 2025. <https://help.openai.com/en/articles/7842364-how-chatgpt-and-our-foundation-models-are-developed> (Zugriff 13.05.2025).
- Oshana, Marina A. L. „Personale Autonomie und das soziale Umfeld“. In *Autonomie der Person*, herausgegeben von Monika Betzler. Brill | mentis, 2013. https://doi.org/10.30965/9783957439727_011.
- Otte, Ralf. *Künstliche Intelligenz für Dummies*. 2. Auflage. für Dummies. Wiley-VCH GmbH, 2023.
- o.V. *Epistemology Of Bias*. o. J. <https://esg.sustainability-directory.com/term/epistemology-of-bias/> (Zugriff 03.07.2025).
- Pallay, Christian. „Vom Turing-Test zum General Problem Solver. Die Pionierjahre der künstlichen Intelligenz“. In *Philosophisches Handbuch Künstliche Intelligenz*, herausgegeben von Klaus Mainzer. Springer Reference Geisteswissenschaften. Springer Fachmedien Wiesbaden, 2020. https://doi.org/10.1007/978-3-658-23715-8_3-1.
- Pariser, Eli. *Filter Bubble: wie wir im Internet entmündigt werden*. 2. Auflage. Übersetzt von Ulla Held. Hanser, 2017.
- Pariser, Eli. „Vorsicht vor ‚Filter-Blasen‘ im Internet“. 2. Mai 2011. <https://www.youtube.com/watch?v=B8ofWFx525s> (Zugriff 03.01.2025).
- Pauleikhoff, Bernhard. „Halluzination“. In *Historisches Wörterbuch der Philosophie online*. Schwabe Verlag, 2017. 10.24894/HWPh.1481 (Zugriff 13.04.2025).
- Piepmeier, Rainer, Kurt Röttgers, Wilhelm Goerdts, u. a. „Intelligenz, Intelligentsia, Intellektueller“. In *Historisches Wörterbuch der Philosophie online*. Schwabe Verlag, 2017. 10.24894/HWPh.5195 (Zugriff 13.06.2025).
- Pohlmann, Rosemarie. „Autonomie“. In *Historisches Wörterbuch der Philosophie online*. Schwabe Verlag, 2017. <https://doi.org/10.24894/HWPh.367> (Zugriff 11.03.2025).
- Pörksen, Bernhard. „Die Hölle der Desinformation - Spielregeln der Wirklichkeitsordnung im digitalen Zeitalter“. In *Bernd Pörksen, Die große Gereiztheit*. 2018.
- Quaritsch, Helmut. „Souveränität“. In *Historisches Wörterbuch der Philosophie online*. Schwabe Verlag, 2017. <https://doi.org/10.24894/HWPh.3949> (Zugriff 13.04.2024).
- Ramge, Thomas. „Mensch fragt, Maschine antwortet“. bpb.de, 2. Februar 2018. <https://www.bpb.de/shop/zeitschriften/apuz/263680/mensch-fragt-maschine-antwortet/> (Zugriff 13.04.2025).
- Reichl, Peter. *Homo cyber: ein Bericht aus Digitalien*. Müry Salzmann, 2023.
- Replika.Com. „Replika“. Zugegriffen 2. Juli 2025. <https://replika.com>.

- Rössler, Beate. „Autonomie“. In *Handbuch Angewandte Ethik*, herausgegeben von Christian Neuhäuser, Marie-Luise Raters, und Ralf Stoecker. J.B. Metzler, 2023. https://doi.org/10.1007/978-3-476-05869-0_22 (Zugriff 11.04.2024).
- Rössler, Beate. *Autonomie: ein Versuch über das gelungene Leben*. Erste Auflage. Suhrkamp taschenbuch wissenschaft 2274. Suhrkamp, 2019.
- Rössler, Beate. *Der Wert des Privaten*. 1. Aufl. Suhrkamp Taschenbuch Wissenschaft 1530. Suhrkamp, 2001.
- Rössler, Beate. „Philosophin Beate Rössler: ‚Autonomie ist eine abhängige Freiheit‘“. 2011. <https://www.tagesspiegel.de/wissen/autonomie-ist-eine-abhaengige-freiheit-6713230.html> (Zugriff 13.12.2024).
- Russell, Stuart, und Peter Norvig. *Artificial Intelligence: a Modern Approach, EBook, Global Edition: A Modern Approach*. Pearson Education, Limited, 2016. <http://ebookcentral.proquest.com/lib/univie/detail.action?docID=5831883> (Zugriff 13.04.2025).
- Schmidhuber, Jürgen. „Vortrag ‚Künstliche Intelligenz wird alles ändern‘“. 2. Big-Data-Konferenz von Horváth & Partners, 9. Mai 2016. <https://www.youtube.com/watch?v=rafhHIQgd2A> (Zugriff 20.10.2023).
- Schönauf, Andreas. „My device made me do it“ - Zur Wechselwirkung von Heteronomie und Autonomie bei der Interaktion mit künstlichen intelligenten Systemen. 19. Dezember 2019. <https://www.praefaktisch.de/ki/my-device-made-me-do-it-zur-wechselwirkung-von-heteronomie-und-autonomie-bei-der-interaktion-mit-kuenstlichen-intelligenten-systemen/> (Zugriff 13.11.2024).
- Searle, John. „Minds, brains, and programs“. *The Behavioral and Brain Sciences*, Nr. (1980), 3 (o. J.): 417–57.
- Seidel, Christian. *Selbst bestimmen: Eine philosophische Untersuchung personaler Autonomie*. De Gruyter, 2016. <https://doi.org/10.1515/9783110474275> (Zugriff 13.04.2024).
- Siegesleitner, Anne. *Post-Privacy oder Datenschutz als neuer Megatrend? Sinn und Wert der Privatheit*. 4. Februar 2020. <https://www.praefaktisch.de/postfaktisch/post-privacy-oder-datenschutz-als-neuer-megatrend-sinn-und-wert-der-privatheit/> (Zugriff 11.04.2024).
- Spiekermann, Sarah. *Digitale Ethik: ein Wertesystem für das 21. Jahrhundert*. Droemer, 2019.
- Spiekermann, Sarah, und Frank Pallas. „Ubiquitous Computing. Technology Paternalism—Wider Implications of Ubiquitous Computing.“ In *Matern, F. (2007): Die Informatisierung des Alltags. Leben in smarten Umgebungen*. Springer, o. J.
- Stoljar, Natalie. „Feminist Perspectives on Autonomy“. In *The Stanford Encyclopedia of Philosophy*, Winter 2022, herausgegeben von Edward N. Zalta und Uri

- Nodelman. Metaphysics Research Lab, Stanford University, 2022.
<https://plato.stanford.edu/archives/win2022/entries/feminism-autonomy/>
 (Zugriff 13.07.2024).
- Strasser, Anna. „Distributed Responsibility in Human–Machine Interactions“. *AI and Ethics* 2, Nr. 3 (2022): 523–32. <https://doi.org/10.1007/s43681-021-00109-5>
 (Zugriff 13.04.2025).
- Strasser, Anna. „From tool use to social interactions“. In *Loh, Social robotics and the good life*. Transcript Verlag, 2023.
- Susser, Daniel, Beate Roessler, und Helen Nissenbaum. „Technology, Autonomy, and Manipulation“. *Internet Policy Review* 8, Nr. 2 (2019).
<https://doi.org/10.14763/2019.2.1410> (Zugriff 02.05.2024).
- Susser, Daniel, Beate Roessler, und Helen F. Nissenbaum. „Online Manipulation: Hidden Influences in a Digital World“. *SSRN Electronic Journal*, advance online publication, 2018. <https://doi.org/10.2139/ssrn.3306006> (Zugriff 13.12.2024).
- Sutmöller, Nadine. „Big Data und die Frage nach der Gerechtigkeit“. In *Rath et. al.: Maschinenethik. Normative Grenzen autonomer Systeme*. 2019.
- Tietjens Meyers, Diana. „Personale Autonomie ohne Transzendenz“. In *Autonomie der Person*, übersetzt von Andreas Maier. 2013.
- Turing, Alan M. *Computing Machinery and Intelligence: Englisch/Deutsch = Können Maschinen Denken?* [Nachdruck] 2022. Übersetzt von Achim Stephan und Sven Walter. Reclams Universal-Bibliothek Great Papers Philosophie, Nr. 19647. Reclam, 2021.
- Turing, Alan M. *Intelligent Machinery*. 1948.
- Verständig, Dan, Christina Kast, Janne Stricker, und Andreas Nürnberger, Hrsg. *Algorithmen und Autonomie*. Verlag Barbara Budrich, 2022.
<https://doi.org/10.3224/84742520> (Zugriff 13.04.2024).
- Voß, Oliver. „Was es statt eines KI-Moratoriums braucht“. *Tagesspiegel Background Digitalisierung & KI*, Verlag Der Tagesspiegel GmbH, 31. März 2023.
<https://background.tagesspiegel.de/digitalisierung/was-es-statt-eines-ki-moratoriums-braucht> (Zugriff 13.04.2024).
- Waldegge, Florian Heusinger von. „Was ist Autonomie?“ *Hegel-Jahrbuch* 2017, Nr. 1 (2017): 197–201. <https://doi.org/10.1515/hgjb-2017-0130> (Zugriff 11.03.2024).
- watson.ch. „«Bring dich um, Al» – KI-Chatbot fordert Nutzer zum Suizid auf“. Zugegriffen 10. September 2025. <https://www.watson.ch/!525989343> (Zugriff: 09.09.2025).
- Weber, Jutta. „Probably Approximately Correct? Epistemologische Grundlagen der Künstlichen Intelligenz“. In *Mensch-Maschine-Interaktion*, herausgegeben von Orsolya Friedrich, Johanna Seifert, und Sebastian Schleidgen. Brill | mentis, 2022. https://doi.org/10.30965/9783969752609_010 (Zugriff 22.04.2025).

- Weber-Guskar, Eva. „Reflecting (on) Replika“. In *Loh, Social robotics and the good life*. Transcript Verlag, 2023.
- Weiser, Mark. „The Computer for the 21st Century“. 1991.
<https://web.archive.org/web/20070630213952/http://www.ubiq.com/hypertext/weiser/SciAmDraft3.html> (Zugriff 21.03.2025).
- Weiser, Mark, und John Seely Brown. *Designing Calm Technology*. 1995.
<http://www.ubiq.com/weiser/calmtech/calmtech.htm> (Zugriff 22.07.25).
- Widdau, Christoph Sebastian, und Carlos Zednik. „Opake Systeme künstlicher Intelligenz und das Problem der Verantwortungslücke“. In *Algorithmen und Autonomie*. Barbara Budrich, 2022. DOI: 10.3224/84742520 (Zugriff 13.04.2024).
- Wihbey, John. „AI and Epistemic Risk for Democracy: A Coming Crisis of Public Knowledge?“ *SSRN Electronic Journal*, advance online publication, 2024.
<https://doi.org/10.2139/ssrn.4805026>.
- Wikipedia. „Cyber-physisches System“.
https://de.wikipedia.org/w/index.php?title=Cyber-physisches_System&oldid=245240822 (Zugriff 08.05.2025).
- Wikipedia. „HAL 9000“.
https://en.wikipedia.org/w/index.php?title=HAL_9000&oldid=1289397923 (Zugriff 13.03.2025).
- Wikipedia. „Opazität“.
<https://de.wikipedia.org/w/index.php?title=Opazit%C3%A4t&oldid=223000011> (Zugriff 03.02.2025).
- Wikipedia. „Schluss auf die beste Erklärung“.
https://de.wikipedia.org/w/index.php?title=Schluss_auf_die_beste_Erkl%C3%A4rung&oldid=235051393 (Zugriff 28.06.2025).
- Wolf, Susan. „The true, the good, and the lovable: Frankfurt’s avoidance of objectivity“. *Contours of agency: Essays on themes from Harry Frankfurt*, MIT Press Cambridge, MA, & London, 2002, 227–44.
<https://doi.org/10.7551/mitpress/2143.003.0011> (Zugriff 02.02.2025).
- Woolf, Virginia. „Berufe für Frauen“. In *Dies., Der Tod des Falsters. Essays*. Frankfurt am Main, 1997.
- Zander, Peter. „Ambient-Assisted-Living und selbstbestimmtes Leben“. Essay im Rahmen der Lehrveranstaltung Medizinethik des Studium Generale, Dozent: Dr. Andreas Klein, Wien, 2024.
- Zander, Peter. „Der Autonomiebegriff bei Immanuel Kant - Sichtweisen aus Theologie und Kirche“. Essay im Rahmen der Lehrveranstaltung Theologie des Studium Generale, Dozent: Dr. Andreas Klein, Wien, 2024.

Zander, Peter. „Wahrheit und Respekt im Geschwindigkeitsrausch sozialer Medien“. Essay im Rahmen der Lehrveranstaltung Medienethik des Studium Generale, Dozentin: Dr. Katharina Lacina, Wien, 2024.

Zentrum für sichere Informationstechnologie - Austria. *Ambient Assisted Living: Smarte Assistenz für Pflegebedürftige*. o. J.
<https://www.onlinesicherheit.gv.at/Services/News/Ambient-Assisted-Living.html> (Zugriff 02.04.2025).

Zweig, Katharina A., Oliver Deussen, und Tobias D. Krafft. „Algorithmen und Meinungsbildung: Eine grundlegende Einführung“. *Informatik-Spektrum* 40, Nr. 4 (2017): 318–26. <https://doi.org/10.1007/s00287-017-1050-5> (Zugriff 22.04.2025).