PHILOSOPHISCHE DIGITALISIERUNGSFORSCHUNG

VERANTWORTUNG, VERSTÄNDIGUNG, VERNUNFT, MACHT

herausgegeben von
Rainer Adolphi, Suzana Alpsancar,
Susanne Hahn und Matthias Kettner

[transcript] Digitale Gesellschaft

Rainer Adolphi, Suzana Alpsancar, Susanne Hahn, Matthias Kettner (Hg.)
Philosophische Digitalisierungsforschung

Digitale Gesellschaft | Band 75

Rainer Adolphi (Prof. Dr.) war Professor für Philosophie an der Technischen Universität Berlin. Seine Arbeitsgebiete liegen in den Feldern Sozialphilosophie, Theorie der Kultur, politische Philosophie, Anthropologie sowie History of Ideas.

Suzana Alpsancar (Prof. Dr.) leitet die Fachgruppe für Angewandte Ethik am Heinz-Nixdorf Institut der Universität Paderborn. Aktuell forscht sie zur Ethik und Normativität erklärbarer KI sowie zum Verhältnis von Nachhaltigkeit und Digitalisierung.

Susanne Hahn (Prof. Dr.) lehrt Philosophie an der Heinrich-Heine-Universität Düsseldorf. Ihre Arbeitsschwerpunkte sind philosophische Fragen der Digitalisierung, Rationalität, Normativität und Wirtschaftsethik.

Matthias Kettner (Prof. Dr., Dipl.-Psych.) ist Professor für Philosophie an der Universität Witten/Herdecke. Seine Forschungsschwerpunkte sind Diskursethik, Psychoanalyse und Digitalisierung als Kulturprozess.

Rainer Adolphi, Suzana Alpsancar, Susanne Hahn, Matthias Kettner (Hg.)

Philosophische Digitalisierungsforschung

Verantwortung, Verständigung, Vernunft, Macht

[transcript]

Die Veröffentlichung wurde gefördert durch die Stiftung Mercator im Rahmen des Projekts »Digitale Ethik am CAIS«.

Bibliografische Information der Deutschen Nationalbibliothek
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über https://dnb.dnb.de/ abrufbar.

Erschienen 2024 im transcript Verlag, Bielefeld

Umschlaggestaltung: Maria Arndt, Bielefeld
Druck: Majuskel Medienproduktion GmbH, Wetzlar
https://doi.org/10.14361/9783839474976
Print-ISBN: 978-3-8376-7497-2
PDF-ISBN: 978-3-8394-7497-6
Buchreihen-ISSN: 2702-8852
Buchreihen-eISSN: 2702-8860

Gedruckt auf alterungsbeständigem Papier mit chlorfrei gebleichtem Zellstoff.

Inhalt

III
Vernunftverhältnisse

IV
Machtverhältnisse

Einleitung

Es ist mittlerweile ein Gemeinplatz, dass Digitalisierung ein alle Lebensbereiche erfassender Vorgang ist. Als Vorgang und Resultat ist Digitalisierung Gegenstand verschiedener Wissenschaften, z.B. empirischer Sozialwissenschaften, Kultur- und Medienwissenschaften, aber auch Gegenstand rechtlicher Regulierung und – als Desiderat – Objekt der Politik. Noch vor wenigen Jahren fanden Fragen der Digitalisierung kaum philosophische Aufmerksamkeit. Heute bilden sie den am schnellsten wachsenden Bereich neuer philosophischer Themen. Die neue wissenschaftliche Betriebsamkeit geht in der Regel mit einem Ruf nach Öffnung und Kooperation zwischen verschiedenen Disziplinen einher, um dem interdisziplinären Charakter vieler Problemlagen, die im gesellschaftlichen Großprozess der Digitalisierung entstehen, gerecht werden zu können. Der Kooperationsbedarf und die Notwendigkeit des interdisziplinären Austauschs sind unbestritten. Wir sehen jedoch auch einen spezifischen philosophischen Beitrag zu diesen Debatten: Philosophieren zeichnet sich durch die Analyse und (vorschlagende) Normierung von Begriffen, durch Identifikation und Reflexion normativer Fragen sowie durch Sortierungen und Einordnungen von Phänomenen unter einer bestimmten Perspektive aus. Die Beiträge dieses Bandes offerieren begriffliche Differenzierung, De- und Rekonstruktion der normativen und axiologischen Dimension sowie synthetische Reflexionen von Digitalisierung als kulturellem Prozess und Praktik. Die Analysen, Beschreibungen, Erklärungen, Bewertungen, Problematisierungen und Einordnungen, so unsere Hoffnung, mögen den interdisziplinären Austausch gleichermaßen beleben wie politische Regulierung informieren und eine aufgeklärte gesellschaftliche Debatte über wünschenswerte oder nicht wünschenswerte technokulturelle Zukünfte anregen.

Spätestens seit OpenAI ChatGPT am 30. November 2022 in die Öffentlichkeit entlassen hat, ist »KI« in aller Munde. Auch wenn KI die Handlungsfähigkeit menschlicher Akteure verändert und gegenwärtig in geschichtlich beispiellosem Ausmaß nichtmenschliche Agenzien ertüchtigt, würde es zu kurz greifen, die aktuellen digitaltechnisch induzierten Herausforderungen und Transformationen auf KI zu verengen. Nicht nur, dass »KI« als Buzzword unterbestimmt bleibt, auch in technischem Sinn haben wir es in der Regel mit mehr zu tun als KI, die

zumeist lediglich eine Komponente einer Software bzw. eines soziotechnischen Systems darstellt. Die mittlerweile veralltäglichten KI-Services (wie Sprachassistenzfunktionen, Textgenerierung, Klassifikation von Fotos, Routenauswahl von Wegen, Spamfilter etc.) basieren auch ökonomisch i.d.R. auf digitaltechnischen Ökosystemen, die in erheblicher Konzentration von einigen wenigen Tech-Unternehmen kontrolliert werden. Wir fassen den Blick deswegen weiter und gehen von Digitaltechnologien als Bezeichnung für ein umfassenderes Konglomerat solcher Technologien aus, die technisch auf Prozessen der *Digitisierung* aufruhen. Unter Digitisierung wird die technische Konversion von Signalen in ein solches Format verstanden, das sie von digitalen Maschinen verarbeitbar werden lässt, d.h. es geht um digital technische Infrastrukturen. Der uns hier interessierende Gegenstand ist die *Digitalisierung*, verstanden als Gesamtphänomen der Praktiken, Strukturen, und Prozesse, die sich auf den Gebrauch, den Umgang mit und die Integration von digitalen Infrastrukturen beziehen. Wir wollen hier Digitalisierung als Bündel *kultureller Prozesse und Praktiken* thematisieren, um von vornherein technozentrischen Blickverengungen zu entgehen.

Auch wenn digitalkulturelle Prozesse spezifische digitaltechnologische Voraussetzungen und Grundlagen haben, erlangen sie ihre volle Wirklichkeit erst in unseren Handlungen und Erfahrungen, also dort, wo jene technischen Voraussetzungen und Grundlagen allererst mit Sinn und Bedeutung erfüllt werden: in der ganzen Breite aller Praktiken, die die Lebenswelt ausmachen, die Handlungsmacht von Menschen und nichtpersonalen Agenzien verändern und die Spielräume möglicher Lebensweisen verschieben. Was für die Technisierung im Allgemeinen gilt, ist für die Digitalisierung ebenso in Anschlag zu bringen. So ist, *zunächst*, Digitalisierung kein rein technisch zu verstehender Vorgang; vielmehr ist er – als soziotechnischer Prozess – im Ensemble mit den individuellen Handlungen und sozialen Praxen sowie den institutionellen Rahmenbedingungen zu betrachten. *Zudem* ist eine für neue Techniken womöglich typische Dynamik zu beachten, die darin liegt, dass die Absichten, unter denen Individuen bei der Entwicklung und unter Nutzung der Technik agiert haben, zwar möglicherweise realisiert wurden, diese Technik aber im Einsatz Resultate hervorgebracht hat, die diese Absichten womöglich konterkarieren. *Schließlich* geht die Einordnung als kultureller Prozess, der letztlich auf zielorientiertes und regelerzeugendes individuelles und kollektives Handeln in institutionellen und ökonomischen Strukturen zurückzuführen ist, mit der Maxime der Gestaltbarkeit einher.

Die Einsicht, dass soziotechnische Vorgänge kein Schicksal, sondern gestaltbar sind, erlaubt eine grundsätzliche Reflexion über die neuen Handlungsmöglichkeiten und Zumutungen, dies sowohl in individueller wie in kollektiver Perspektive. Der Sammelband macht – ohne Vollständigkeitsanspruch – ein Panorama des philosophischen Nachdenkens über Digitalisierung auf. Wir ordnen die in den Beiträgen thematisierten Herausforderungen, die sich für unsere bewährten Überzeu-

gungen und sozialen Praktiken stellen, nach Fragen zu ›Verantwortung‹, ›Verständigung‹, ›Vernunft‹ und ›Macht‹.

Verantwortungsverhältnisse

Technische Fortschritte im Bereich des maschinellen Lernens alias Künstlicher Intelligenz und die Verbreitung von entsprechenden Produkten, z.B. KI-basierten Assistenzsystemen, algorithmischen Entscheidungssystemen oder personalisierten Chatbots, berühren inzwischen merklich eine für das menschliche Zusammenleben zentrale Praxis: die der Zuschreibung von Verantwortung. Dazu gehört, dass Akteure sich untereinander Verantwortung bzw. Pflichten zuschreiben und nach Maßgabe guter Gründe Rechenschaft verlangen und geben können. Vier Beiträge erörtern unter verschiedenen Rücksichten Herausforderungen dieser Praxis, die unter Digitalisierungsdruck hervortreten. Dazu gehören die Herausforderungen für die Handlungskontrolle durch die verflochtenen Handlungsketten bei Einsatz von KI, die mangelnde Nachvollziehbarkeit der algorithmischen Prognosen, aber auch der Akteursstatus von KI.

Susanne Hahn analysiert am Beispiel des Einsatzes von KI in der medizinischen Diagnostik und unter Rückgriff auf das klassische Modell der Verantwortungszuschreibung die Zurechnungslücken, die sich durch den Einsatz von KI bei möglicherweise entstehenden Schäden ergeben. Unter der Voraussetzung, dass eine Hauptfunktion der Verantwortungszuschreibung in der Handlungssteuerung besteht, fragt sie danach, ob und wenn ja, wie das Handeln so umorganisiert werden kann, dass sich erwünschte Resultate des KI-Einsatzes erhalten lassen und Schäden vermieden werden.

Suzana Alpsancar greift die Forderung nach einer ›erklärbaren‹ KI auf, die in der KI-Forschung immer lauter und intensiver erhoben wird. Die Debatte um erklärbare KI, auch die philosophische, krankt jedoch an einer pauschalen Aufwertung der Erklärbarkeit. Alpsancar kann zeigen, dass den divergierenden Motiven der Forderung nach Erklärbarkeit (wenigstens) drei unterschiedliche Forschungsparadigmen entsprechen: das Paradigma der epistemischen Qualität, das der effizienten Nutzung und das der Vereinbarkeit mit grundlegenden Werten und Normen. Die mangelnde Berücksichtigung der Verschiedenartigkeit dieser Paradigmen begünstigt vereinfachende ingenieurwissenschaftliche Lösungsansätze für Fragen, die aus politisch-philosophischer Sicht eher im Bereich der gesellschaftlichen Diskussion und politischen Entscheidungsfindung angesiedelt werden müssen.

Der Beitrag von *Natalia Juchniewicz* bringt Ressourcen der Sozialphilosophie der Anerkennung ins Spiel. Diese kann eine erhebliche Aufklärungskraft gewinnen, wenn sie auf die Beziehungen zwischen Menschen und maschinell intelligenten Systemen angewandt wird und die reflexhafte Rede von der ›Mensch-Maschine-Interaktion‹ beiseitelässt. Die Erweiterung des Anerkennungsparadigmas um

kollektive Intelligenz und verteiltes Handeln stellt die normative Ethik, die Sozialphilosophie, aber auch die empirische Sozialforschung vor die Herausforderung, heterogen verteilte Anerkennung und verteilte Verantwortung adäquat zu erfassen. Natalia Juchniewicz versucht dies mit Blick auf die interpersonalen Beziehungen, die Verantwortungszuschreibungen erst ermöglichen. Sie diskutiert, unter welchen Bedingungen wir nichtmenschliche Akteure als »partielle Personen« anerkennen und sogar emotional bedeutsame Beziehungen zu ihnen entwickeln können.

Dass Algorithmen ›hinter‹ digitaltechnischen Anwendungen stehen, ist inzwischen fast jedem bekannt. Algorithmen, so scheint es, sind verantwortlich – für das Verhalten der Anwendungen und, zumindest indirekt, auch für das der Nutzer. Die aufkommende pan-algorithmische Realität wird nicht nur begrüßt oder zumindest als unvermeidlich hingenommen, sondern hat auch ein neues Feld skeptischer Reflexionen hervorgebracht, die ›Algorithmenkritik‹. *Tobias Matzner* beschreibt drei gängige Muster dieser Kritik. Das Algorithmische stellt Technik in vielen Praxisfeldern unter den Verdacht der Undurchschaubarkeit, der politischen Unkontrollierbarkeit und der Subversion menschlicher Selbstbestimmung und Offenheit. Ausgehend von seiner technikphilosophischen Kritik am latenten Mensch-Technik-Dualismus der Kritikmuster stellt Matzner programmatische Überlegungen an, wie die Problematisierung von Phänomenen algorithmischer Überformung der Lebenswelt verbessert und für politische Regulierungsbemühungen nutzbar gemacht werden könnte: durch eine stärkere Kontextualisierung und größere Erfahrungsnähe der Kritik sowie durch ein vertieftes technikphilosophisches Verständnis für den unabgeschlossenen Möglichkeitsraum der Technik selbst.

Verständigungsverhältnisse

Es dürfte schwerfallen, Kommunikationspraktiken zu finden, die der kulturelle Prozess der Digitalisierung nicht bereits tiefgreifend verändert hat oder in Kürze zu verändern verspricht. Dies betrifft die öffentliche Kommunikation und Meinungsbildung ebenso wie die Vermittlung und Wahrnehmung wissenschaftlicher Expertise, aber auch die zwischenmenschliche Kommunikation bis in ihre privatesten und intimsten Formen hinein. Vier an sehr verschiedenen exemplarischen Bezugsproblemen arbeitende Beiträge zeigen, wie die philosophische begriffliche Arbeit, gepaart mit einer starken phänomenologischen Aufmerksamkeit für die Empirie, zu überraschenden Einsichten führt, welche Änderungen des gewohnten Denkens und Handelns angezeigt sind, um die identifizierten Herausforderungen anzugehen.

Am Bezugsproblem der Meinungsfreiheit zeigt *Micha Werner*, dass der digitale Wandel neue Fragen nach der Bedeutung, Institutionalisierung und Begründung kommunikativer Grundnormen aufwirft, zu deren Verständnis und vernünftiger Begründung die Philosophie beitragen kann. Werner thematisiert die digitalkul-

turelle Herausforderung für unser eingespieltes Verständnis von Meinungs- und Äußerungsfreiheit, die sich aus der drastischen Ausweitung der Äußerungsmöglichkeiten und der Verknappung der Rezeptionschancen ergibt. Das klassische politisch-liberale Verständnis von Meinungsfreiheit im Sinne eines Abwehrrechts gegenüber dem Staat erscheint angesichts dieser Verschiebung und der herausragenden Rolle sogenannter Intermediäre und ihres Geschäftsmodells der Kommunikationsplattformen als dringend ergänzungsbedürftig. Werner plädiert für eine Ergänzung durch Teilhaberechte an Kommunikationschancen und Mitwirkungsrechte an der Gestaltung von Kommunikationsstrukturen.

Die private und öffentliche Meinungsbildung zu schwierigen Fragen aller Art bedarf, wenn sie aufgeklärt sein soll, eines verlässlichen kommunikativen Zugangs zu glaubwürdigem Expertenwissen. *Nicola Mößner* untersucht am Beispiel der Corona-Pandemie, als sich eine beträchtliche Zahl von Menschen mit Verschwörungsnarrativen identifizierte und virologische und medizinische Verhaltensratschläge zur Reduzierung von Neuinfektionen bewusst ablehnte, die ethische und epistemische Einordnung eines weit verbreiteten Bedeutungsverlusts traditioneller Expertenkulturen. Die epistemische und moralische Verwundbarkeit vieler Menschen, die neue Medien als epistemische Alternativen wählen, anstatt sich auf Expertenmeinungen in etablierten Formen zu verlassen, ist beträchtlich. Mößner übt philosophische Metakritik: Gegenüber einigen tendenziell beschwichtigenden philosophischen Annahmen über soziale Medien als vermeintlich alternative Möglichkeiten der Informationsbeschaffung und Quelle moralischer Unterstützung in Gruppen von Gleichgesinnten zeigt sie die Haltlosigkeit dieser Annahmen auf.

Im Gegenzug zur Vervielfachung der Ausdrucksmöglichkeiten in digital-kulturellen Kommunikationsverhältnissen scheinen gerade diese Verhältnisse zur Einsamkeit als einem neuen, brisanten Massenphänomen beizutragen, das auch gesundheitspolitisches Handeln herausfordert. Mit Blick auf das technische Einsamkeitsmanagement, wie es derzeit vor allem in Japan erprobt wird, analysiert *Kerrin Artemis Jacobs* Versuche, Einsamkeit durch spezielle KI-Anwendungen zu überwinden. Zwar kann die Interaktion mit künstlich intelligenter Technik unter bestimmten Bedingungen eine starke Illusion zwischenmenschlicher Intersubjektivität erzeugen. Sie scheint daher auch für die Linderung leidvoller Einsamkeit durch ›soziale Künstliche Intelligenz‹ empfehlenswert. Warum wir auf diese Empfehlung nicht allzu große Hoffnungen setzen sollten, begründet Jacobs mit sozial- und medizinphilosophischen Argumenten.

Die digitalkulturell geprägten Kommunikationsverhältnisse begünstigen in unvergleichlichem Maße die Zirkulation von Gefühlsäußerungen, insbesondere von negativen. In der öffentlichen und empirisch-wissenschaftlichen Aufmerksamkeit für Feindseligkeit im Netz und politische Radikalisierungsprozesse mit Hilfe affektgeladener Internetkommunikation hat sich der Begriff ›Hate Speech‹ eingebürgert. *Kai Denkers* Analyse begründet Skepsis gegenüber dieser scheinbaren

Selbstverständlichkeit. Am Beispiel einer für politisch inkorrekte Kommunikation berüchtigten Imageboard-Website zeigt er, wie Affordanzen der Technik und bestimmte Nutzungsgewohnheiten wie Flüchtigkeit, Anonymität und offensiver ›Humor‹ die Entstehung und den Ausdruck aggressiver Gefühlszustände begünstigen und für politische Zwecke, vor allem zur Beförderung rechtsextremer Ideologien, ausgenutzt werden. Das vergleichsweise stärkste affektive Register in diesen Kommunikationen ist jedoch nicht Hass, sondern Wut.

Vernunftverhältnisse

Digitalisierung fordert die Praktiken, die als spezifisch für menschliche Vernunft betrachtet werden, in vielfältiger Weise heraus. Menschliches Denken und Handeln zielt auf Erkenntnis der Welt ab, um mit den Fährnissen in der Welt umgehen zu können. Die Hervorbringung von Erkenntnis ist durch Normen geleitet, die u.a. das Begründen, das empirische Feststellen, das Postulieren, das Prognostizieren bestimmen. Diese Erkenntnistätigkeiten werden z. B. durch generative Sprachmodelle oder auch allgemein durch KI herausgefordert. Statistische Zusammenhänge bringen Prognosen hervor, deren Zustandekommen nicht im Einzelnen nachzuverfolgen ist. Daneben werden die kommunikativen Praktiken, die sich bewährt haben, um Erkenntnis zu fördern, durch die unterliegenden Mechanismen sozialer Medien in den Hintergrund gedrängt.

Gabriele Gramelsberger erläutert ontologische, phänomenologische und epistemologische Aspekte des Digitalen. Während digitale Daten-Objekte durch die Verknüpfung von immer mehr Daten an kontextueller Dichte gewinnen, unterläuft das Digitale gleichzeitig immer mehr die menschliche Wahrnehmungsfähigkeit und wird zu einer unterschwelligen Parallelwelt der Maschinen, die nur durch ebenso unterschwellige Entscheidungsalgorithmen kontrollierbar und zugänglich bleibt, allerdings vorzugsweise im Backend auf Seiten der Digitaltechnikkonzerne. Gramelsberger skizziert ein Desiderat für die weitere philosophische Digitalisierungsforschung: Trotz zunehmender Abkoppelung der Digitaltechnik von unseren Sinnen sind wir Menschen immer enger mit dem zur Umwelt werdenden, gleichsam environmentalen Digitalen verwoben. Die Auswirkungen dieser Konstellation auf den Einzelnen sollten in ihrer ganzen Breite untersucht und prospektiv auf mögliche anthropologische Konsequenzen hin befragt werden.

Sybille Krämer thematisiert die Nutzung von generativer KI und zugleich das Verhältnis zwischen Vernunft und Verstehen. Sie zielt darauf ab, den Unterschied zwischen dem sinnhaften Sprachgebrauch des Menschen und der Tokenisierung von Text in den KI-Systemen aus der Familie der ›Large Language Models‹ (LLMs), die heute Furore machen, auf den Begriff zu bringen. Für die nahe Zukunft erwartet Krämer, dass durch die rasche Verbreitung von LLMs die dialogische Nutzung von KI zu einer im Alltag verankerten Kulturtechnik wird. Doch die Eleganz, Natür-

lichkeit und Plausibilität, mit der LLM-basierte Kommunikationsassistenten schon heute auf Eingaben reagieren, kann einer unangemessenen Anthropomorphisierung und einem problematischen Übervertrauen in unsere algorithmischen Maschinen Vorschub leisten.

Im philosophischen Rahmen der normativen Diskurstheorie problematisiert *Matthias Kettner* die Erwartung, dass der digitale Wandel einer ›kommunikativen‹ Rationalität entgegenkomme, die auf evaluierbare Geltungsansprüche und deren Begründungen zugeschnitten sei. Hatte man, wie in den utopischen Anfängen des Internets, glauben wollen, dass die entfesselte Konnektivität der neuen ›sozialen‹ Medien die Kommunikationsgemeinschaft, in der diese Rationalität zählt, enorm erweitern würde, so spricht die eingetretene Realität, bezogen auf die informelle Social-Media-Kommunikation, dieser Hoffnung Hohn. Kettner findet fünf Teilerklärungen für systematische Einschränkungen kommunikativ rationaler Kommunikation in solchen Medien, macht aber gegen einen naheliegenden pauschalen Irrationalitätsverdacht geltend, dass die berechtigten Ansprüche an die Rationalität von Kommunikationspraktiken mit den Zwecken dieser Praktiken variieren.

Vincent C. Müller gibt einen Überblick über die derzeit wichtigsten Themen, Argumente und Positionen des philosophischen Diskurses zur Künstlichen Intelligenz. (Die Ethik ist dabei bewusst ausgeklammert.) Neben den Grundbegriffen der Intelligenz und des Rechnens sind Wahrnehmung, Handlung, Bedeutung, rationale Wahl, Willensfreiheit, Bewusstsein und Normativität die aktuellen Themenschwerpunkte im Feld der künstlichen Vernunftleistungen. Der besondere Wert einer auf KI fokussierten philosophischen Digitalisierungsforschung erweist sich immer dann, wenn sich in ihrem Zuge unser bisheriges philosophisches Verständnis dieser zentralen Themen verbessert. Entsprechend schlägt Müller vor, ›KI-Philosophie‹ als eine neue Methode für die Philosophie insgesamt zu verstehen.

Machtverhältnisse

Wie bei anderen neuen Techniken gilt auch bei der Digitaltechnologie, dass mit signifikanten Veränderungen der verfügbaren technischen Mittel immer auch Veränderungen der verfügbaren Handlungsmacht und ihrer gesellschaftlichen Verteilung einhergehen. Drei Beiträge versuchen auf einer sehr grundsätzlichen Ebene zunächst zu explizieren, worüber wir sprechen müssen, wenn wir von der Macht der Digitalisierung, der Macht der KI oder der Macht der Algorithmen sprechen wollen und stellen sich damit der durchaus unbequemen Aufgabe, das Machtthema mit Bezug auf Digitalisierung zu behandeln.

Wo immer Verhältnisse des Ausgeliefertseins und Beherrschtwerdens nicht mehr dumpf bleiben, haben sich Weisen und Praktiken kritischer Verständigung und auch entsprechende semantische Formen herausgebildet – in Lebenswelt-

mentalitäten und dann wissenschaftlich reflektiert in Psychologie, Soziologie, Philosophie. Der Beitrag von *Rainer Adolphi* gilt dem, wie in der Lage zunehmend digitalisiert geprägter Wirklichkeiten bisherige Potenziale in massive Verunsicherung geraten und zugleich die Erfahrungsverwobenheiten Denkformen und Narrative induzieren, die Wesentliches unterbestimmt lassen. Neben einer Strukturierung und Analyse der Prozesse ist es die Perspektive des Beitrags, das, was alltagsweltlich in Sensorien für ›Macht‹ und was in Medienwissenschaften, Techniksoziologie und Technikphilosophie schon einmal erreicht war, nicht nun im Stadium des Digitaltechnischen der Lebenswirklichkeiten zu verlieren bzw. zu hinterschreiten. Als ein zentrales Phänomen werden die Verschmelzungen mit dem Digitalisierten sowie dessen Geräten und andrerseits Oberflächen-Schwellen diskutiert.

Digitalkultureller Wandel, wie er gegenwärtig betrieben und zugelassen wird, schafft sich allmählich verfestigende Abhängigkeiten und plötzliche Krisen. *Armin Grunwald* interpretiert mit technikphilosophischen Argumenten die alternativlos gewordenen Abhängigkeiten von digitaltechnischen Infrastrukturen als latente Disruptionen, die sich in epistemischen, kommunikativen, ethischen und pragmatischen Dimensionen beschreiben lassen. Besonders betroffen sind Praktiken, die – wie das demokratische Regieren – auf die Aufrechterhaltung langsamer Deliberation, korrigierbarer Lernprozesse und des Vertrauens in verantwortliche Gestaltung angewiesen sind. Die allseits geforderte Anpassung an disruptiv riskante Abhängigkeiten bedeutet einen Verlust von Zukunft im Sinne eines gestaltungsoffenen Raumes, damit aber auch eine Einbuße an politischer Gestaltungsmacht – ein Einspruch gegen das populäre technikoptimistische Narrativ, Digitalisierung ermächtige uns mehr als jede andere bisherige Technologie, unsere Zukunft vernünftig zu gestalten.

Ausgehend von einer Re-Analyse des klassischen Machtbegriffs von Max Weber entwickelt *Matthias Kettner* ein neues dynamisches Verständnis von Macht, das es erlaubt, Machtverhältnisse sowohl in Bezug auf Personen und quasipersonale korporative Akteure als auch in Bezug auf a-personale Software-Agenten und andere maschinelle Akteure zu analysieren. Die Macht, über die Akteure in einer Situation verfügen, wird modal und relational konzeptualisiert als die Fähigkeit jedes Akteurs, andere Kräfte mit Hilfe von Kräften, die er kontrafaktisch robust kontrolliert, so zu steuern, dass die Akteure ihren Zielen näherkommen. Es ist ein Versuch der grundbegrifflichen Klärung und Verbesserung. Was der dynamische Machtbegriff in der Anwendung auf Phänomene der Digitalisierung leistet, können erst konkrete Fallstudien des digitalen Wandels erweisen.

* * *

Die Beiträge des vorliegenden Bandes sind aus einer Reihe von Konferenzen entstanden, die wir im Rahmen unserer 2019 konstituierten Arbeitsgruppe zur philosophischen Digitalisierungsforschung am Bochumer *Center for Advanced Internet Studies* (CAIS) durchgeführt haben. Insgesamt wurden 48 Themenvorträge und Diskussionen durchgeführt. Die thematische Gliederung des vorliegenden Bandes bildet die vier Themenschwerpunkte der CAIS-Konferenzen ab. Im Fortgang der Konferenzen entstand ein beachtliches Netzwerk von Diskussionspartnern. Für vielfältigen Austausch und kritische Anregungen und Diskussionsbeiträge möchten wir uns an dieser Stelle ausdrücklich bedanken, sowohl bei den Autorinnen und Autoren unseres Bandes als auch bei Amrei Bahr, Jana Baum, Michael Baurmann, Thomas Bedorf, Christoph Bieber, Dieter Birnbacher, Eva Buddeberg, Christoph Hubig, Simone Dietz, Alexander Filipovic, Aline Franzke, Jonathan Geiger, Selin Gerlek, Thomas Grote, Gerald Hartung, Jessica Heesen, Ole Kliemann, Dirk Lanzerath, Christoph Lauer, Burkhard Liebsch, Rainer Mühlhoff, Boris Rähme, Alberto Romele, Magnus Schlette, Domenico Schneider, Lisa Schurrer, Jonathan Seim, Jan Siebold, Bernd Stahl, Tom Sterkenburg, Dieter Sturma, Tobias Vogel, Johanna Wagner, Eva Weber-Guskar, Carlos Zednik. Kira Boots danken wir für die Unterstützung bei der redaktionellen Arbeit. Dem *Center for Advanced Internet Studies* (CAIS) danken wir für die vielfältige und großzügige Förderung unserer Forschungsgruppe »Philosophische Digitalisierungforschung«, der Stiftung Mercator für die unkomplizierte Förderung der Drucklegung und dem transcript Verlag für die gute Zusammenarbeit bei der Veröffentlichung des vorliegenden Bandes.

Rainer Adolphi
Suzana Alpsancar
Susanne Hahn
Matthias Kettner

I

Verantwortungsverhältnisse

Verantwortung in Zeiten ›künstlicher Intelligenz‹

Eine Problemexposition am Beispiel medizinischer Diagnostik*

Susanne Hahn

Abstract: *The ascription of responsibility is a social practice with a high significance for human coexistence. The challenge posed to this practice by the use of artificial intelligence, for example in medical diagnostics, needs to be specified in more detail. An examination of the classic concept of responsibility makes it possible to identify ascription gaps. Assuming that an important function of the ascription of responsibility, namely the guidance of action to avoid harm, should be preserved, it is worth taking a look at the historical handling of ascription gaps using the sketch of the regulation of steam boilers in the 19th century. Some epistemological considerations on the predominantly prognostic function of algorithms in diagnostics as well as existing proposals for the certification of artificial intelligence serve to work out questions that are decisive for the assignment of responsibilities.*

Keywords: *responsibility; responsibility gap; role obligations; demand for transparency; social practice*

1. Das Ausgangsproblem: Herausforderung für die Verantwortungszuschreibung durch den Einsatz ›künstlicher Intelligenz‹

Die Zuschreibung von Verantwortung ist eine soziale Praxis mit einem hohen Stellenwert für das menschliche Zusammenleben. Die Zuordnung von Handlungsfolgen, insbesondere solchen unerwünschter Art, zu Akteuren als ihren Urhebern, hat

* Die Arbeit an diesem Beitrag wurde gefördert durch das Center for Advanced Internet Studies (CAIS) Bochum und die Stiftung MERCATOR. – Seit der Arbeit an diesem Thema im Rahmen eines Fellowships 2019 haben sich sowohl in der Sache selbst (Stichwort ChatGPT) als auch in der Reflexion und Regulierung von KI (Stichwort AI Act) viele weitere Neuerungen oder Beschleunigungen ergeben. An der Grundstruktur von Handeln, Technikeinsatz und Verantwortungszuschreibung hat sich dadurch nichts verändert. Der vorliegende Aufsatz versucht sich an einer Analyse dieser Grundstruktur.

sowohl präventive als auch wiedergutmachende Funktionen. Diese soziale Praxis, die auch rechtlich normiert ist, wird durch den Einsatz künstlicher Intelligenz herausgefordert, so beispielsweise in der medizinischen Diagnostik. Hier, wie in vielen anderen Bereichen, ist in den letzten Jahren eine stark anwachsende Entwicklung und Anwendung des maschinellen Lernens zu verzeichnen, insbesondere in der Bilderkennung, aber auch in der Verarbeitung anderer Merkmale zu Diagnosen bzw. Voraussagen.

So vielversprechend der Einsatz von Algorithmen auch sein mag, verhindert er jedoch nicht, dass Personen Schäden durch Fehldiagnosen erleiden. Dies ist der Ansatzpunkt für die Frage nach der Verantwortung: Wem ist Verantwortung zuzuschreiben, wenn ein Patient aufgrund einer Einschätzung durch einen Algorithmus beispielsweise nicht zu einer Präventivuntersuchung überwiesen worden war und er einen Herzinfarkt erlitten hat? Wer trägt Verantwortung, wenn eine Patientin nach einer Einschätzung durch eine maschinelle Bilderkennung als krebserkrankt gilt und zu weiteren Eingriffen wie Gewebeentnahmen geschickt wird, ohne dass sich ein Tumor bzw. eine Tumorvorstufe findet. Wer ist verantwortlich, wenn eine bösartige Veränderung nicht entdeckt wurde? Und – um neuere Entwicklungen zur Früherkennung von Alzheimer zu nennen – wer trägt Verantwortung, wenn jemand mit 55 Jahren unter Einsatz eines Algorithmus die Diagnose erhält, in der Zukunft mittelfristig wahrscheinlich an Alzheimer zu erkranken, er aber mit 89 an einer Lungenentzündung bei guter geistiger Gesundheit stirbt?

Durch die Einschaltung einer Technik – hier der Einsatz von Algorithmen zur Klassifikation bzw. Voraussage von medizinisch relevanten Zuständen – entstehen *Zurechnungslücken*: Es ist nicht mehr klar, welchem Akteur ein eintretender Schaden zuzurechnen ist, wodurch der Adressat für eine Kompensation und eine mögliche Sanktion ebenfalls nicht mehr identifizierbar ist. Durch diesen Umstand ist eine für das menschliche Zusammenleben wichtige soziale Praxis herausgefordert und die Frage, ob der Einsatz dieser Techniken entsprechend reguliert werden soll, auf der Tagesordnung.[1] Der vorliegende Aufsatz dient nicht dazu, eine affirmative oder negative Antwort auf diese Frage zu liefern. Es geht vielmehr ausschließlich darum, dieses Problem so zu analysieren, dass die aufgeworfene Frage spezifiziert werden kann und deutlich wird, dass zu ihrer adäquaten Behandlung zum einen weitere Fragen nach der technischen Handhabbarkeit von algorithmischen Voraussagemodellen zu beantworten sind und zum anderen Abwägungen hinsichtlich der Chancen und Risiken der Technik notwendig sind.

Diese Problemexposition soll vermitteln, dass zur Bearbeitung dieser Aufgaben die Expertise aus Mathematik, Statistik, Informatik, Wissenschaftsphilosophie, Moralphilosophie, Rechtswissenschaft und Medizin erforderlich ist. Das hierzu

1 Auf vielen Ebenen und in vielen Gremien wurden und werden bereits Richtlinien zum Umgang mit künstlicher Intelligenz entwickelt. Für einen Überblick vgl. Hagendorff 2020.

entwickelte Szenario umfasst fünf Schritte: Zunächst sind einige Klärungen zum Verständnis künstlicher Intelligenz zu leisten und zu erläutern, welche Formen in der medizinischen Diagnostik, die hier als Beispielfeld dienen soll, zum Einsatz gelangen könnten. Sodann ist das klassische Konzept von Verantwortung zu erläutern, um darzulegen, in welcher Weise diese Form der Verantwortungszuschreibung durch den Einsatz künstlicher Intelligenz herausgefordert wird. Unterstellend, dass eine wichtige Funktion der Verantwortungszuschreibung, nämlich die Handlungssteuerung zur Vermeidung von Schäden, erhalten bleiben soll, bietet sich ein Blick in den historischen Umgang mit Zurechnungslücken an. Eine Skizze der Regulierung von Dampfkesseln im 19. Jh. liefert strukturelle Anknüpfungspunkte. Einige erkenntnisphilosophische Überlegungen zur vorwiegend prognostischen Funktion von Algorithmen in der Diagnostik sowie vorliegende Vorschläge zur Zertifizierung von künstlicher Intelligenz dienen dazu, Fragen herauszuarbeiten, die für die Zuordnung von Verantwortlichkeiten entscheidend sind.

Angewandte Philosophie zeichnet sich dadurch aus, dass man verallgemeinernde Überlegungen auf spezifische Handlungsfelder bezieht oder auch umgekehrt eben solche Verallgemeinerungen aus diesen gewinnt. Diese Handlungsfelder sind durch einen eigenen Fundus an Wissen und Verfahren charakterisiert. Das gilt unabhängig davon, ob man sich mit Organtransplantation, Gentherapie, Umwelttechnik oder – im betrachteten Fall – mit Anwendungen maschinellen Lernens beschäftigt (vgl. Bayertz 1991: 28ff.). In allen Fällen stellt sich die Frage, wie tief und umfassend man sich mit dem jeweiligen Gegenstandsbereich auseinandersetzen muss, um begriffliche Sortierungen vornehmen zu können, die wiederum Voraussetzungen für normative Einschätzungen darstellen.

Die Antwort hängt davon ab, welche – beispielsweise – ethischen Prinzipien und Normen man potenziell durch eine Handlungsweise betroffen sieht. Wenn man z. B. vermutet, dass die Steuerbarkeit und Transparenz von Handlungen durch den Einsatz künstlicher Intelligenz beeinträchtigt werden und diese Eigenschaften für die moralphilosophische Betrachtung relevant sind, dann sollte man die zugrundeliegenden Verfahren so weit erfassen, dass diese Merkmale deutlich werden.

Damit ergibt sich der erste Schritt zur Realisierung des hier verfolgten Projekts, die Grundlage für die Formulierung handhabbarer Szenarien und Fragen zu liefern: Es gilt also zunächst, unterschiedliche Formen maschinellen Lernens so aufzubereiten, dass Fragen zu ihrem gerechtfertigten Einsatz bearbeitbar werden.[2] Die relative Explizitheit dieses Vorgehens dient zum einen dazu, Unterschiede in den

2 Die Nicht-Expertin auf diesem Gebiet ist auf Darstellungen angewiesen, die sich an interessierte Laien richten. Die hier vorgelegte Skizze orientiert sich zunächst – auch in den Abbildungen und Beispielen – an dem für Einsteiger sehr nützlichen Buch von Steven Finlay (Finlay 2017) und zieht daneben Bart Baesens (Baesens 2014) heran.

Spielarten künstlicher Intelligenz zu verdeutlichen und in ihrer Relevanz für die Verantwortungsthematik einzuordnen. Zum anderen sollen die Darstellungen Anhaltspunkte für das Verständnis von Schlagworten wie Transparenz, Erklärbarkeit, Nachvollziehbarkeit etc. liefern.

2. Formen maschinellen Lernens als derzeit realisierte Form ›künstlicher Intelligenz‹

Mit dem Begriff künstliche Intelligenz, abgekürzt KI (»artificial intelligence«, AI), ist für einige Menschen die Erwartung verbunden, dass Menschen, als Träger »natürlicher« Intelligenz (jetzt/in Kürze/in absehbarer Zeit/langfristig) durch Maschinen ersetzbar werden. Diesem *starken* Verständnis künstlicher Intelligenz im Sinne von Menschenähnlichkeit (hier allerdings oft nur in Bezug auf die erwünschten Eigenschaften) steht ein *schwaches* Verständnis gegenüber, das auf bestimmte, vor allem auf Kalkulationen bezogene, Fertigkeiten abstellt.[3] Im Zentrum dieses Verständnisses stehen Algorithmen maschinellen Lernens.[4] Wenn im Folgenden von künstlicher Intelligenz die Rede ist, dann ist damit dieser problembezogene Einsatz maschinellen Lernens gemeint. Diese auf statistischen Verfahren basierende Technologie ist durch enorm gestiegene Rechnerkapazitäten auf der einen Seite und zugleich stark wachsender Verfügbarkeit von Daten in den letzten Jahren zu einem dominanten Zweig in der Informatik und Datenwissenschaft geworden (vgl. Engemann 2018: 254; Lepri/Oliver/Letouzé/Pentland/Vinck 2018: 612). – Die Folgefrage lautet: Was ist maschinelles Lernen?

> »Machine learning is the use of mathematical procedures (algorithms) to analyze data. The aim is to discover useful patterns (relationships or correlations) between different items of data. Once the relationships have been identified, these can be used to make inferences about the behavior of new cases when they present themselves.« (Finlay 2017: 5)

Auch wenn die Erfolge in der Anwendung maschinellen Lernens häufig eine mystische Aura autonom agierender, aber nicht identifizierbarer Akteure hervorrufen, ist – wie in dem obigen Zitat ablesbar – festzuhalten, dass diese Verfahren auf der *An-*

3 Unterschiedliche Bedeutungen von künstlicher Intelligenz werden z.B. dargelegt bzw. erörtert in Bringsjord/Govindarajulu 2020; Mainzer 2016.

4 Zur Charakterisierung von Algorithmen sowie zur Unterscheidung regelbasierter, »klassisch« programmierter Algorithmen und Algorithmen maschinellen Lernens vgl. Fry 2018.

wendung mathematischer und statistischer Verfahren beruhen.[5] Die Zielsetzung besteht darin, in den vorliegenden Daten, d.h. in den bisherigen Verläufen, Merkmalskonstellationen zu finden, die von den zur Mustererkennung betrachteten Fällen auf neue Fälle übertragen werden können. In übertragener Redeweise handelt es sich um Lernen aus Erfahrung: »In den-und-den Fällen hat die-und-die Merkmalskonstellation zum Verlauf A geführt. Dieser Fall weist eine große Ähnlichkeit zu dieser Merkmalskonstellation auf. Also ist Verlauf A wahrscheinlich.« Eine typische Anwendung ist die Erstellung von Prognosen, ob eine Person einen Kredit zurückzahlen wird. Die ermittelten Beziehungen zwischen Merkmalen sind allerdings, wie in statistischen Verfahren üblich, Korrelationen; über Kausalverhältnisse ist damit noch nichts gesagt. Ein klassisches illustratives Beispiel liefert die Merkmalskonstellation zwischen der Anzeige eines Barometers, dem atmosphärischen Druck und den daraus abgeleiteten Prognosen für die Wetterlage. Die Korrelationen, die in »beiden Richtungen« zwischen den Barometeranzeigen und dem atmosphärischen Druck bestehen, sind nur in einer Richtung auch Kausalbeziehungen: Nicht die Veränderung der Barometeranzeige verursacht die Druckveränderung, sondern die Druckveränderung verursacht umgekehrt die Änderung der Anzeige des Barometers. Werden reine Korrelationen betrachtet, ist weder etwas über den tatsächlichen Einfluss einer Größe auf künftige Entwicklungen gesagt, noch über das Zustandekommen von Mustern.[6]

Der warnende Hinweis auf die lediglich ermittelten Korrelationen deutet bereits an, dass im Hintergrund des maschinellen Lernens verschiedene erkenntnis- und wissenschaftsphilosophische Probleme stehen. Neben dem genannten Problem von Korrelation vs. Kausalität ist dies allgemein das Problem der Induktion: Unter welchen Bedingungen ist man berechtigt, von einer endlichen Zahl beobachteter Fälle mit bestimmten Merkmalskonstellationen auf eine allgemeine Aussage über Typen dieser Fälle – und damit auch auf bislang unbeobachtete Fälle – zu schließen (vgl. Abschnitt 6.1)?

Eine weitere für normative Betrachtungen wichtige Eigenschaft des Einsatzes maschinellen Lernens sind die enthaltenen Verfahrensschritte. Üblicherweise sind dies die folgenden (vgl. Finlay 2017: 48ff.):

5 »[…] it [sc. AI] is only ›intelligent‹ in the narrowest sense of the word. It would probably be more useful to think of what we've been through as a revolution in computational statistics than a revolution in intelligence.« (Fry 2018: 14)

6 Die Nicht-Berücksichtigung von Kausalitätsbeziehungen hat auch damit zu tun, dass die mathematische Darstellung von Merkmalsbeziehungen eine Unterscheidung zwischen bloßer Korrelation und Kausalität erschwert. Diesem Defizit will Judea Pearl abhelfen, indem er versucht zu zeigen, wie man Kausalbeziehungen formal-mathematisch darstellen kann (Pearl 2018).

- Zusammenstellung eines Dateninputs
- Aufbereitung der Daten
- Entwicklung der Voraussagemodelle
- Formulierung von Entscheidungsregeln bzw. von Wenn-Dann-Verknüpfungen[7]
- resultierende Maßnahme.

Für die spätere Diskussion ist es wichtig zu bemerken, dass ein großer Teil der analytischen Arbeit in der inhaltlichen Durchdringung einer Fragestellung, welche Merkmalsbeziehungen es zwischen unterschiedlichen Zuständen oder Ereignissen gibt, bereits in Prozessen liegen, die *vor* der Bereitstellung des Dateninputs und der Aufbereitung der Daten liegen. Daneben wird mit dieser Identifizierung von Verfahrensschritten deutlich, dass notwendig Entscheidungsregeln formuliert werden müssen. Dieser Verfahrensschritt liegt wiederum nicht in unbeeinflussbaren technischen Prozessen, sondern ist Gegenstand einer menschlichen Entscheidung (vgl. weiter unten zur Bildung von Risikoklassen und Bestimmung eines Schwellwertes). Der hier zunächst interessierende Einsatz des eigentlichen maschinellen Lernverfahrens besteht lediglich im Schritt zur Entwicklung des Voraussagemodells.

> »A predictive model (or just model going forward) is the output generated by the machine learning process. The model captures the relationships (patterns) between which have been uncovered by the analytics process. Once a model has been created it can be used to generate new predictions. Organizations then use the model's predictions to decide what to do or how to treat people. *So machine learning is a process, and a predictive model is the end product of that process*.« (Finlay 2017: 38f.; Hervorhebung – SH)

Für die Fragen der Verantwortungszuschreibung ist der Umstand bedeutsam, dass es – anders als durch die Bezugnahme auf neuronale Netze in öffentlichen Debatten nahegelegt – *verschiedene Techniken maschinellen Lernens* gibt, um ein Voraussagemodell zu gewinnen. Gemeinsam ist allen Modellen, dass die generierte Voraussage durch eine Zahl repräsentiert wird, die die Wahrscheinlichkeit angibt, mit der das in Frage stehende Verhalten oder Ereignis eintritt. Die Verfahren maschinellen Lernens unterscheiden sich jedoch hinsichtlich ihrer Nachvollziehbarkeit.[8]

---

7 Zur Frage, ob die Rede von »Entscheidungen« durch Algorithmen berechtigt und adäquat ist, vgl. Hahn 2024.

8 Die »Opakheit« maschineller Mustererkennung ist nicht nur Gegenstand der Suche nach technischen Lösungen, sondern auch der kritischen Reflexion. – Bezogen auf Arten der Opakheit, die unterschieden werden, ist hier im Folgenden die Art gemeint, die sich aus dem maschinellen Lernverfahren selbst ergibt, nicht etwa aus eigentumsrechtlichen Gegebenheiten (vgl. Burrell 2016; Durán/Jongsma 2021).

Dieser Unterschied lässt sich an einem – fiktiven – Beispiel aus der medizinischen Diagnostik illustrieren, das der erwähnten Arbeit zur Darstellung maschinellen Lernens entnommen ist (vgl. Finlay 2017). Die ausführliche Darstellung dieses Beispiels dient dem oben bereits erwähnten Ziel, nachvollziehbar darzustellen, an welchen Stellen die etablierte Praxis der Verantwortungszuschreibung herausgefordert wird und Anschauungsmaterial für die Verwendung der »Transparenzbegrifflichkeit« zu liefern. Der Dateninput besteht im Beispiel aus zufällig ausgewählten Berichten von 500.000 Patienten, die noch keine Anzeichen einer Herzkrankheit aufweisen. Die Daten umfassen Angaben zu Alter, Geschlecht, Vorerkrankungen, Blutdruck, Body-Mass-Index, Alkoholkonsum, Rauchen, Gewicht sowie zum Einkommen.

Diese Daten werden mit dem weiteren Verlauf in den darauffolgenden fünf Jahren konfrontiert: Wer von diesen Personen entwickelt eine Herzkrankheit und wer nicht? In der Sprache der Voraussagemodelle hat man an dieser Stelle zwei Sorten von Daten: *Beobachtungsdaten* – das sind in diesem Fall die Daten aus den medizinischen Berichten und *Ergebnisdaten* – in diesem Fall die Dokumentation des Gesundheitsverlaufs über fünf Jahre, konzentriert darauf, ob diese Personen eine Herzkrankheit entwickeln oder nicht.

Zusammen ergeben diese Daten das sogenannte *Entwicklungssample*. Die Zusammenstellung des Entwicklungssamples enthält bereits die ersten beiden genannten Schritte bei der Anwendung von Verfahren maschinellen Lernens, nämlich die Bereitstellung des Dateninputs und die Aufbereitung der Daten. Auf dieser sollen aussagekräftige Parameter ermittelt werden, die es erlauben, bei neuen Fällen auf die Entwicklung oder Nicht-Entwicklung einer Herzkrankheit zu schließen. In diesem Beispiel soll diese Einschätzung als Entscheidungsbasis für die Einladung zu einer umfassenderen Herz-Kreislauf-Untersuchung dienen (im Fünfer-Schritt der Anwendung von Verfahren maschinellen Lernens ist dies der letzte Schritt, d.h. die resultierende Maßnahme).

Im Beispiel stellt man fest, dass von den 500.000 Personen innerhalb von fünf Jahren 30.000 eine Herzkrankheit entwickeln. An dieser Stelle kommen die Verfahren maschinellen Lernens zum Einsatz. Gesucht werden Algorithmen für ein Voraussagemodell, das die Beobachtungsdaten mit den Ergebnisdaten korreliert. Welche Parameter sind in welchem Maße ausschlaggebend dafür, eine Herzkrankheit zu entwickeln oder eben nicht, welche *Muster* ergeben sich? Die Gegenüberstellung von zwei Voraussagemodellen soll dazu dienen, die Problematik der Nachvollziehbarkeit und Debattierbarkeit und letztlich der Verantwortungszuschreibung zu erläutern.

Das erste Voraussagemodell ist eine *Scorecard*, ein lineares Modell, das auf dem statistischen Verfahren der logistischen Regression beruht. Das andere Verfahren ist ein sogenanntes *künstliches neuronales Netz*.

Abbildung 1

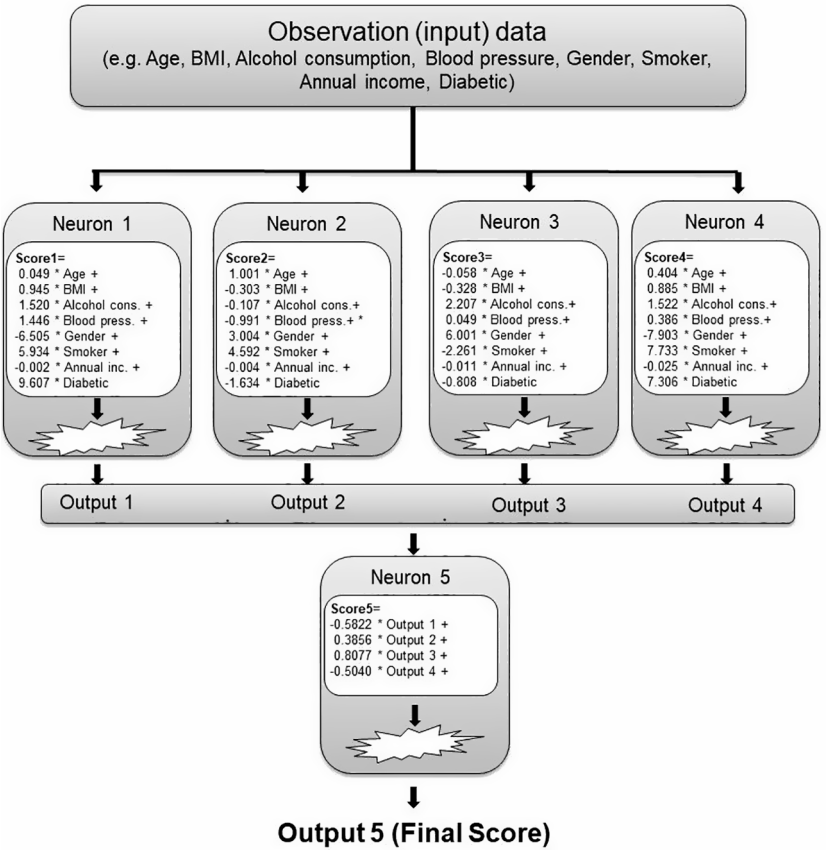
Starting score (constant)	350		
Age (years)		**Gross annual income ($)**	
<23	-57	< $22,000	11
23 - 32	-26	$22,001 - $38,000	6
33 - 41	0	$38,001 - $60,000	0
42 - 48	7	$60,001 - $94,000	-3
49 - 57	15	$94,001 - $144,000	-5
58 - 64	24	>$144,000	-6
65 - 71	31		
>71	65	**Smoker ?**	
		Yes	37
BMI (weight in kg / {height in metres}^{2})		No	0
<19	2		
19 - 26	0	**Diabetic ?**	
27 - 29	8	Yes	21
30 - 32	14	No	0
>32	29		
		Cholesterol level (mg per decilitre of blood)	
Gender		Low (< 160 mg)	-2
Male	2	Normal (160 - 200 mg)	0
Female	-4	High (201 - 240 mg)	19
		Very high (>240 mg)	32
Alcohol consumption (units/week)			
0	4	**Blood pressure**	
1 - 12	0	Low (below 90/60)	3
13 - 24	5	Average (between 90/60 and 140/90)	0
25 - 48	10	High (above 140/90)	36

Quelle: Finlay 2017: 33

Beide Verfahren nutzen dieselben Kriterien, von Alter bis Einkommen. Die Angaben, welche Ausprägung eines Attributs ein Patient hat, z.B. ein Alter von 53 Jahren und Blutdruckwerte von 150 zu 90 liefern den Input für die Verfahren. Bei einem neuronalen Netz, wie es oben abgebildet ist, gehen diese Daten mehrfach in die Berechnung ein. Die Werte der neun Attribute beliefern vier Neuronen bzw. Knoten, in denen die Werte mit einer Gewichtung versehen und zu einem Output-Wert addiert werden. Die Output-Werte von diesen liefern den Input für – in diesem Fall – ein weiteres Neuron, in dem diese vier Werte wiederum gewichtet und addiert werden. Der letzte Schritt besteht in einer Transformation dieser Werte auf eine Zahl aus einem bestimmten Intervall, z.B. auf Werte zwischen 0 und 1. Dieser transformierte Wert gibt die Wahrscheinlichkeit dafür an, in den nächsten fünf Jahren an einem Herzleiden zu erkranken. Im sehr einfachen Beispielnetzwerk kommen dabei 36 (4 x 8 + 4 = 36) Gewichtungen zum Einsatz. In künstlichen neuronalen Netzen, wie sie insbesondere bei der Bilderkennung zum Einsatz kommen, geht die Zahl der Gewichtungen in die Tausende.[9]

9 Für grundlegende Informationen zu neuronalen Netzen, Arten des Lernens sowie der Analogie zu biologischen neuronalen Vorgängen vgl. Mainzer 2016: Kap. 7.2.

Abbildung 2



Quelle: Finlay 2017: 52

Bei der Scorecard handelt es sich um das Resultat eines statistischen Standardverfahrens, der logistischen Regression. Im Vergleich mit den künstlichen neuronalen Netzen kann man dieses als ein einzelnes Neuron auffassen.[10] Mit der logistischen Regression wird der Einfluss der hier genannten Merkmale auf das Eintreten eines Herz-Kreislauf-Ereignisses bestimmt. Im Unterschied zu den künstlichen neuronalen Netzen werden die Werte für die Attribute nur einmal zur Bestimmung

10 Baesens setzt der verbreiteten Auffassung, künstliche neuronale Netzwerke seien eine Nachbildung physiologischer Neuronen, eine »realistischere« Perspektive entgegen: »A first perspective on the origin of neural networks states that they are mathematical representations inspired by the functioning of the human brain. Another more realistic perspective sees neural networks as generalizations of existing statistical models.« (Baesens 2014: 48)

dieser Wahrscheinlichkeit herangezogen. Die Gewichtungen sind als solche erkennbar. Auch bei der Übertragung in das Scorecard-Modell sind die Gewichtungen weiterhin nachvollziehbar: So trägt beispielsweise das Geschlecht weniger zur Erstellung der Prognose bei als das Alter der Patienten.[11]

Beide Verfahren sind so zu interpretieren, dass gefragt wird, welche Gewichte für die einzelnen Attribute angenommen werden müssen, um sich den Ergebnisdaten möglichst weit anzunähern. Nachdem in einem *Trainingsdatenset* ein Voraussagemodell entwickelt wurde, wird dieses an einem *Testdatenset*, das verschieden ist vom Trainingsdatenset, erprobt und eventuell adjustiert. Von den erwähnten Schritten in der Anwendung maschinellen Lernens ist der dritte Schritt mit der Formulierung von Voraussagemodellen abgeschlossen. Die Anwendung der Voraussagemodelle auf neue Fälle ergibt Wahrscheinlichkeitsaussagen für das betrachtete Ereignis.

Wenn man diese Voraussagen mit Maßnahmen verknüpfen möchte, sind weitere Schritte erforderlich, die jenseits des maschinellen Lernverfahrens liegen. Im Beispiel ist unterstellt, dass die Zeit, die Ärzten zu großflächigen Präventionsuntersuchungen zur Verfügung steht, begrenzt ist und somit eine Auswahl unter den Patienten zu treffen ist. Hierzu dient zunächst die Bildung von Risikoklassen, um diejenigen zu identifizieren, die am stärksten von einer Präventivuntersuchung profitieren.

Abbildung 3

Group	Score range From	To	Number of people	% of population	Number with heart disease after	% with heart disease after 5 yrs.
1	0	300	55,950	11.19%	40	0.07%
2	301	320	56,606	11.32%	68	0.12%
3	321	340	59,700	11.94%	129	0.22%
4	341	360	58,706	11.74%	216	0.37%
5	361	380	64,429	12.89%	403	0.63%
6	381	400	52,749	10.55%	575	1.09%
7	401	420	34,089	6.82%	600	1.76%
8	421	440	21,107	4.22%	632	2.99%
9	441	460	17,269	3.45%	878	5.09%
10	461	480	23,364	4.67%	2,020	8.65%
11	481	500	17,477	3.50%	2,553	14.61%
12	501	520	13,554	2.71%	3,366	24.84%
13	521	540	7,103	1.42%	3,463	48.76%
14	541	560	8,260	1.65%	6,587	79.74%
15	561	999	9,637	1.93%	8,469	87.88%
Total	**Total**		**500,000**		**30,000**	**6.0%**

Quelle: Finlay 2017: 41

11 Aus der Formel der logistischen Regression lässt sich – auch wenn eine detailliertere Einsicht statistische Kenntnisse erfordert – ablesen, dass die Gewichte für die Merkmale nur einmal in die Berechnung des Wahrscheinlichkeitswertes eingehen (vgl. Baesens 2014: 48).

Hier sind dies beispielhaft 15 Klassen aus den Scorecard-Werten (vgl. Abbildung 3). Von den Personen aus der Risikoklasse 15, dies sind 9637 Personen, entsprechend einem Populationsanteil von 1,93%, sind 87,88% in den nächsten fünf Jahren erkrankt. Wenn man nun dieses Erfahrungswissen auf neue Fälle anwenden möchte, und die Beschränkung annimmt, dass aufgrund endlicher Zeit der Ärzte nur 5% der Population eingeladen werden können, ist zu fragen, welche Risikoklassen zu einer Vorsorgeuntersuchung eingeladen werden sollen und welche nicht.

Die Angehörigen der Risikoklassen 13–15, also alle Personen mit einer Punktzahl größer gleich 521, entsprechen 5% der Population. Schaut man auf die Krankheitsentwicklung dieser Personen in der zweitletzten Spalte, sieht man, dass 18519 davon erkranken – das sind 62% aller 30000 Erkrankten. Indem man 5% der Population einlädt, kann man somit 62% der potentiell Erkrankten identifizieren.

Geht man davon aus, dass diese großen Datenmengen von Computern in kurzer Zeit verarbeitet werden können, dann sieht man zunächst, dass der Einsatz maschinellen Lernens hier sehr effizient ist, d.h. dass viele Personen von präventiven Maßnahmen profitieren, bei zugleich vertretbarem ärztlichem Aufwand.

Dieses Geschehen als *Entscheidungen* eines Algorithmus zu beschreiben, ist jedoch irreführend.[12] *Ex ante* werden Regeln formuliert, wie mit den algorithmisch ermittelten Prognosen zu verfahren ist, z.B. »Alle Patientinnen, die zu den Risikoklassen 13–15 gehören, sollen zu einer umfassenden Vorsorgeuntersuchung eingeladen werden.« Diese Festlegung auf bestimmte Risikoklassen stellt eine *vorgezogene Entscheidung* dar. Der Algorithmus hat hier keine diskretionären Spielräume wie das bei Entscheidungen von Akteuren im engeren Sinne der Fall ist. So könnte ein Arzt von einer Behandlungsleitlinie abweichen, wenn er bei einem Patienten im zeitlichen Verlauf eine ungewöhnliche Steigerung bei einem Merkmal sieht. Er könnte den Patienten zu einer Präventivuntersuchung überweisen, obwohl er noch unter dem dafür vorgesehenen Wert liegt. Bei einer automatisierten Maßnahme hingegen lautet die Vorgabe, dass ab einer Wahrscheinlichkeit von 49 Prozent, in den nächsten fünf Jahren ein Herzkreislaufereignis zu erleiden, eine Patientin zu einer umfassenden fachärztlichen Untersuchung einzuladen ist. Die Verknüpfung zwischen dem Resultat maschinellen Lernens und einer Maßnahme wird aufgrund *von ex ante* angestellten Überlegungen unter bestimmten Zielen *gesetzt*.

Das Beispiel demonstriert die Leistungsfähigkeit der Verfahren maschineller Mustererkennung. Wenn man solche Verfahren jedoch umfassend einschätzen will, sind nicht nur die dadurch geschaffenen neuen Handlungsmöglichkeiten und ihre

12 Das Stichwort lautet »Automated Decision Making« oder »Automatisiertes Entscheiden«. Eines von vielen Beispielen ist die Kapitelüberschrift »Die nächste Stufe der Automatisierung: Maschinen treffen Entscheidungen« in dem informativen Bändchen von Ramge (Ramge 2018: 13). – Zu einer Kritik an dieser Übertragung von Ausdrücken aus den Kontexten menschlichen Handelns auf Maschinen vgl. Hahn 2024.

Effizienz zu betrachten. Vielmehr ist zu fragen, ob und inwiefern diese Verfahren dort, wo sie eingesetzt werden, negativ beurteilte Konsequenzen für das Handeln haben können.

Vergleicht man die beiden Verfahren unter diesem Gesichtspunkt, dann ist festzuhalten, dass Scorecards einfach handhabbar und transparent sind. Damit ist gemeint, dass jeder – auch der Laie – aus seinen oder den Merkmalsausprägungen einer anderen Person leicht einen Wert berechnen und sehen kann, wie viel ein Parameter zum Gesamtergebnis beiträgt. Diese einfache Kontrollierbarkeit und Nachvollziehbarkeit geht bei künstlichen neuronalen Netzen verloren.[13] Angesichts der 36 Gewichtungen im Beispiel lässt sich nicht im Einzelnen nachvollziehen, warum welcher Wahrscheinlichkeitswert für eine Patientin berechnet wurde. Viele Anwendungen kommen auf mehr Variablen und mehr Schichten, so dass sich leicht eine Zahl von 1000 Gewichtungen ergibt. Wie sich die berechneten Werte im Einzelnen zusammensetzen, ist – auch für die Programmierer des Algorithmus – *nicht mehr nachvollziehbar*. Hier findet die Rede von der *black box* ihre Anwendung.

Neuronale Netzwerke genießen jedoch besondere Wertschätzung, weil sie – anders als Verfahren wie die logistische Regression – feinere Muster in den Daten auffinden können. Man setzt sie bereits sehr erfolgreich bei vielerlei Arten von Mustererkennung, sei es von gesprochener Sprache, Schrift und Bildern, ein.

3. Wer ist verantwortlich? – Das klassische Verantwortungskonzept

Die Zuschreibung von Verantwortung ist eine soziale und auch rechtlich normierte Praxis. Üblicherweise und in einem ersten Zugriff wird Verantwortung zugeschrieben, wenn ein *Schaden* entstanden ist, der wesentlich auf menschliches *Handeln* zurückgeht. Mit dieser ersten Charakterisierung soll nicht ausgeschlossen werden, dass man auch für handelnd herbeigeführte *positive* Zustände Verantwortung zuschreibt; diese Variante ist aber an dieser Stelle weniger interessant. Mit dem Zusatz »normiert« ist gemeint, dass nicht jede faktisch geäußerte Verantwortungszuschreibung auch eine solche ist. Vielmehr haben sich in Gemeinschaften Kriterien für die *korrekte* Zuschreibung von Verantwortung herausgebildet. In der Strafrechtsprechung beispielsweise sind solche Bedingungen für die Zurechnung von Schäden systematisch etabliert. Die folgende Charakterisierung orientiert sich an der

13 Vgl. zu dieser speziellen Opakheit verfahrensinhärenter Art: Burrell 2016. Zu den Vor- und Nachteilen neuronaler Netzwerke in Bezug auf Effizienz und Interpretierbarkeit: Baesens 2014: 48ff.

deutschen Strafrechtsprechung, doch gehören die genannten Elemente anerkanntermaßen zum Kern des neuzeitlichen Verantwortungsbegriffs.[14]

Zunächst ist noch einmal hervorzuheben, dass die Verantwortungszuschreibung dort ansetzt, wo Schäden durch das *Handeln* von Akteuren hervorgerufen wurden. Vorgänge, die nicht durch Handeln beeinflussbar sind, sind kein Gegenstand der Verantwortungszuschreibung. So wird beispielsweise niemand für den Ausbruch eines Vulkans verantwortlich gemacht. Verantwortungszuschreibung ist jedoch immer eine soziale Praxis auf dem Hintergrund von Handlungsmöglichkeiten und Wissen: Wenn es irgendwann möglich wäre, durch technische Maßnahmen einen Vulkanausbruch zu verhindern, könnte eine entsprechende Unterlassung zum Gegenstand der Verantwortungszuschreibung werden.

Das klassische Konzept der Zuschreibung von Verantwortung in Bezug auf einen Schaden enthält die Bedingungen Kausalität, Norm und Normverletzung, Willentlichkeit und Wissentlichkeit.[15] Jemand wird genau dann korrekt für einen Schaden verantwortlich gemacht, wenn er durch sein Handeln diesen Schaden hervorgebracht hat, die Kausalitätsbedingung also erfüllt ist, und wenn er dabei eine bestehende Norm verletzt hat und diese Hervorbringung wissentlich und willentlich erfolgt ist. Es reicht somit nicht, lediglich ein Element in einer Kausalkette zu identifizieren, die zum Schaden führt. Sonst wäre beispielsweise auch der Inhaber eines Messergeschäftes für einen Mord verantwortlich zu machen, der mit einem in diesem Geschäft erworbenen Filetiermesser verübt wurde. Vielmehr ist für die korrekte Verantwortungszuschreibung entscheidend, dass zum Zeitpunkt der Tat eine Norm in Kraft war, die die Herbeiführung eines Tatbestands verbietet und der Akteur diese Norm verletzt hat (vgl. Heinrich 2005: 73ff.).[16] Der Verkauf von Küchenmessern zählt jedoch zu den erlaubten Tätigkeiten.

Dieses klassische Konzept – obwohl bereits in wesentlichen Anteilen in Aristoteles' Nikomachischer Ethik angedeutet (Aristoteles 2006: Drittes Buch) – ist selbst

14 Für die Bedingungen der objektiven und subjektiven Zurechnung vgl. Heinrich 2005: 73ff. Für die Herausbildung des Verantwortungskonzepts vgl. Bayertz 1995.

15 Vgl. für eine Einbettung und die besondere Rolle von Normen sowie weitere Literaturhinweise Hahn 2014.

16 In der – angelsächsischen – philosophischen Diskussion findet sich auch eine bloße »Kausalverantwortung«, bei der schwer vorstellbar ist, wie sich damit die Steuerungswirkung von Verantwortungszuschreibungen realisieren ließe: vgl. z.B. Braham/van Hees 2012. – Hier unterscheidet sich die deutschsprachige Tradition, wie sie bei Bayertz (Bayertz 1995) in ihrer historischen Entwicklung und sozialen Konstruiertheit nachgezeichnet wird und auch in Handbuchartikeln (vgl. Werner 2011) weitergegeben wird. Die Rede von bloßer »Kausalverantwortung« wird als abgeleitete, eher metaphorische Redeweise aufgefasst, der gegenüber die korrekte Zuschreibung retrospektiver Verantwortungskonzept über die Kausalität hinaus weitere Bedingungen erfordert.

das Resultat, oder besser, das Zwischenresultat einer langfristigen kulturellen Entwicklung. So hat es beispielsweise die Zurechnung eines Schadens zu ganzen Gruppen gegeben, denen der eigentliche Täter angehört hat, oder aber, noch bis ins Mittelalter hinein, auch die Verurteilung von Tieren oder Gegenständen.[17] Insgesamt ist festzuhalten, dass die Praxis der Verantwortungszuschreibung eingebettet ist in ein Verständnis vorsätzlichen, kausal wirksamen Handelns durch Akteure und in Vorstellungen, welche Handlungsweisen (nicht) erlaubt sein sollen.

4. Verantwortungszuschreibung beim Einsatz künstlicher Intelligenz in der medizinischen Diagnostik

Inwiefern fordert der Einsatz künstlicher Intelligenz in der medizinischen Diagnostik die Praxis der Verantwortungszuschreibung heraus? Ärztliches Tun, d.h. Handeln und Unterlassen, kann – wie jedes Handeln – nicht nur erwünschte Folgen herbeiführen, sondern eben auch Schädigungen. Typischerweise handelt es sich bei Schädigungen durch ärztliches Tun nicht um vorsätzliche Handlungen. Schädigungen in medizinischen Kontexten sind vielmehr in der Regel darauf zurückzuführen, dass Individuen ihre rollenbezogenen Sorgfaltspflichten nicht oder nur mangelhaft wahrgenommen haben. Im Strafrecht werden sie den sogenannten Fahrlässigkeitsdelikten zugeordnet. Die Verletzung einer ärztlichen Sorgfaltspflicht erhält eine entscheidende Rolle bei der Zuschreibung von Verantwortung (vgl. Heinrich 2005: 66). Wenn Patienten in Folge ärztlichen Handelns geschädigt wurden, ist für die korrekte Verantwortungszuschreibung erforderlich, dass mit dem Handeln oder Unterlassen Sorgfaltspflichten verletzt wurden. Beispiele stellen Behandlungsfehler dar, die auf fehlerhafte oder nicht dem Stand des Wissens entsprechende Diagnosen zurückgehen oder auf Therapien, die nicht dem Stand der Kunst entsprechen.[18] Wenn ein Patient geschädigt wird und sich diese Schädigung auf die Verletzung ärztlicher Sorgfaltspflichten zurückführen lässt, dann wird der Schaden korrekt der ärztlichen Person zugeschrieben.

Die medizinische Diagnostik ist ein wachsendes Einsatzfeld für maschinelles Lernen. Mittlerweile gibt es Studien, die vermuten lassen, dass insbesondere Varianten des sogenannten *Deep Learning*, zu denen auch neuronale Netze gehören, den menschlichen Diagnosefertigkeiten ebenbürtig sind.[19]

---

17 Vgl. Bayertz 1995: 6f. Das Auspeitschen des Meeres wegen der durch Sturm zerstörten Brücken durch den Perserkönig Xerxes, von dem Herodot berichtet, ist ein Beispiel für die Verurteilung eines Objekts. Prozesse gegen Tiere hat es im Mittelalter gegeben.

18 Rechtlich wird das Erfordernis, die aktuellen Standards zu beachten z.B. in § 630a BGB geregelt.

19 Vgl. zu dieser Aussage die Metastudie Liu et al. 2019. Darin finden sich Anhaltspunkte für die Ebenbürtigkeit der maschinellen Lernverfahren. Zugleich weisen die Autoren auf erheb-

Im präsentierten Beispiel geht es um die Prävention von Herzerkrankungen. Die Verfahren der Scorecard und die neuronalen Netze werden diagnostisch eingesetzt, und zwar, um zu bestimmen, welche Personen zu einer Untersuchung eingeladen werden sollen.

Angenommen, Person A, die einen Herzinfarkt erleidet, hatte zwei Jahre zuvor bei ihrem Hausarzt alle Daten angegeben, die für den Präventionscheck erforderlich waren, ohne eine Einladung zu einer Untersuchung zu bekommen. Es ist davon auszugehen, dass der Herzinfarkt mit entsprechenden Maßnahmen vermeidbar gewesen wäre. Person A will ihren Arzt für diesen Schaden verantwortlich machen. Die Korrektheit dieser Zuschreibung hängt davon ab, ob die erwähnten Bedingungen für die Verantwortungszuschreibung erfüllt sind. Wesentlich ist hier die Prüfung, ob der Hausarzt bei seiner Diagnose *eine Sorgfaltspflicht verletzt* hat. Die Zurückweisung der Verantwortungszuschreibung kann nicht lediglich mit dem Verweis erfolgen, dass eine Prognose »Patient A wird in den nächsten Jahren wahrscheinlich kein Herzkreislaufereignis erleiden« als Prognose mit den üblichen Unsicherheiten behaftet ist. Zwar ist festzuhalten, dass, wer etwas prognostiziert, dieses Ereignis nicht als mit Sicherheit eintreffend angibt. Prognosen sind aber nicht beliebig, sondern an *Korrektheitsstandards* zu prüfen. Die ärztlichen Prognosen erfolgen im Rahmen der entsprechenden Diagnostik und den aus der Vergangenheit bekannten Verläufen: Wenn die-und-die Merkmale in der-und-der Ausprägung vorliegen, dann darf man aufgrund aus der Vergangenheit bekannter Verläufe mit dem-und-dem-Verlauf rechnen. Verstöße gegen die Sorgfaltspflicht liegen vor, wenn der Arzt die relevanten Merkmale oder die bekannten Verläufe nicht adäquat berücksichtigt hat, anders gesagt: wenn Ärzte Fehler gemacht haben. Kann der Arzt nachweisen, dass er diesen Sorgfaltspflichten nachgekommen ist, dann ist er für die Schädigung nicht verantwortlich. Er hat eine gemäß den Standards korrekte Prognose abgegeben, die sich – wie bei Prognosen möglich – nicht bewahrheitet hat. Hat eine Ärztin jedoch z.B. nicht adäquat berücksichtigt, dass eine Patientin einen ungünstigen BMI und hohe Cholesterinwerte hat, dann ist die Sorgfaltspflicht verletzt.

Wie ist jedoch zu verfahren, wenn bei der Diagnose Verfahren künstlicher Intelligenz zum Einsatz gekommen sind? Wie ist zu prüfen, ob der an die Ärztin gerichtete Vorwurf der Schädigung korrekt ist? *Welche Sorgfaltspflichten* sind überhaupt zu

---

liche methodische Mängel der dieser Metastudie zugrunde gelegten Studien hin, die es auszuräumen gilt, um belastbare und Glaubwürdigkeit schaffende Beurteilungen zu ermöglichen. Beispielsweise wurden häufig keine menschlichen Kontrollgruppen vorgesehen; wenn es Kontrollgruppen gab, wurde nicht sichergestellt, dass Algorithmen und Mediziner dasselbe Datenmaterial beurteilten; es wurde häufig nicht angegeben, wie man mit fehlenden Daten umgeht etc. (vgl. Liu et al. 2019: 291ff.). – Eric Topol (Topol 2019) gibt einen umfassenden Einblick in die Möglichkeiten des Einsatzes maschinellen Lernens in der Medizin.

unterstellen, wenn sie die Prognose nicht selbst erstellt hat, sondern diese nach Dateneingabe in ein Voraussagemodell entstanden ist? Welche Rolle spielt dabei die Art des eingesetzten maschinellen Lernens? Wenn man unterstellt, dass die Sorgfaltspflicht nunmehr lediglich in der korrekten Dateneingabe besteht, dann muss die Ärztin dokumentieren, dass sie diese Pflicht erfüllt hat. Kann sie die korrekte Dateneingabe nachweisen und dann darauf verweisen, dass der berechnete Wert für den Patienten unterhalb der für die Einladung gesetzten Grenze zu einer Präventivuntersuchung gelegen hat, ist sie für den Schaden nicht verantwortlich, da keine Pflichtverletzung vorliegt. Dies trifft für den Einsatz der Scorecard genauso zu wie für die künstlichen neuronalen Netze.

Wenn man jedoch eine weitergehende ärztliche Pflicht unterstellt, die neben der korrekten Dateneingabe auch die Einschätzung der Plausibilität der maschinell erstellten Prognose einschließlich einer Abweichung von dieser Einschätzung umfasst, dann unterscheiden sich die beiden Verfahren in Bezug darauf, inwiefern die Ärztin dieser Pflicht nachkommen kann. Hier findet die unterschiedliche Nachvollziehbarkeit ihren Niederschlag: Bei der Scorecard kann die Ärztin die Gewichtungen der Merkmale im Voraussagemodell mit den üblichen Standards, wie sie über die Aus- und Fortbildung in der Medizin sowie in den Leitlinien der Fachgesellschaften vertreten werden, vergleichen. Relativ auf diese Standards kann sie einen möglichen »Fehler« der maschinellen Prognose feststellen und entsprechend darauf reagieren. Wenn die Ärztin im Beispielfall dieser unterstellten Pflicht nicht nachgekommen ist, und tatsächlich eine Abweichung z.B. in der Gewichtung des Body-Mass-Index durch den Algorithmus vorliegt, den sie durch eine abweichende Einschätzung nicht korrigiert hat, dann ist sie für den Schaden verantwortlich. Ist sie hingegen dieser Pflicht nachgekommen und hat keine Abweichung festgestellt, ist sie für den Schaden nicht verantwortlich.

Bezogen auf dieses Verfahren maschinellen Lernens, das auf die logistische Regression zurückzuführen ist, ist festzuhalten, dass das Handeln der Ärztin sowie das Voraussagemodell des Algorithmus – auch für die Ärztin – nachvollziehbar ist. Basiert das Voraussagemodell hingegen auf einem künstlichen neuronalen Netz, ist die Nachverfolgung eines Fehlers für die Ärztin nicht möglich. In der Abbildung des einfachen neuronalen Netzes (vgl. Abbildung 2) sieht man, dass aufgrund der vielen Gewichtungen die Nachvollziehbarkeit der Prognose und damit die Handhabbarkeit dieses Instruments im Detail auf der Strecke bleibt.[20] Diese mangelnde

20 Mit dem Stichwort der Nachvollziehbarkeit ist die Debatte um die Konzepte Opakheit, Transparenz und Interpretierbarkeit aufgeworfen. Dazu gehören sowohl Fragen zu den Begrifflichkeiten als auch zu den damit verknüpften Forderungen (vgl. weiter unten, 6.; sowie Burell 2016; Durán/Jongsma 2021; Rudin 2019; Grote/Berens 2020). – Im Anschluss an begriffliche Klärungen ist auch der Hinweis zu diskutieren, dass die Forderung nach Erklärbarkeit oder Nachvollziehbarkeit ärztlichen Handelns eine überzogene Forderung sei, da es sich beim ärztlichen Handeln ohnehin um eine eher auf impliziten Kenntnissen beruhende Pra-

Nachvollziehbarkeit lässt eine ärztliche Pflicht zur Plausibilitätsüberprüfung des Voraussagemodells im Einzelnen als nicht erfüllbar erscheinen. Ärzte könnten lediglich eine Abweichung von ihrer auf Standards beruhenden Einschätzung feststellen, nicht aber, aufgrund welcher Gewichtung eines Merkmals diese Abweichung beruht und damit auch nicht, wo ein potenzieller Fehler vorliegen könnte. Da im diskutierten fiktiven Beispiel beide Verfahren maschinellen Lernens zur Verfügung stehen, könnte man unter der Voraussetzung, dass bei ihrem Einsatz auch die Nachvollziehbarkeit der Prognose oder Diagnose gesichert sein soll, verlangen, auf künstliche neuronale Netze zu verzichten und stattdessen auf andere Verfahren wie logistische Regression oder Entscheidungsbäume zu setzen. In Anwendungsfeldern, in denen solche Alternativen zur Verfügung stehen, könnte dies eine sinnvolle Forderung sein.[21] Allerdings lassen sich diese nachvollziehbaren Verfahren nicht für die Bilderkennung, die gerade von besonderer Bedeutung für die medizinische Diagnostik ist, nutzen. In diesen Zusammenhängen sind Verfahren wie die künstlichen neuronalen Netze und andere erforderlich.

Welche *Herausforderungen für die Zuschreibung von Verantwortung* für Schäden im Rahmen medizinischen Handelns lassen sich hier identifizieren? Die Bedingungen für die korrekte Zuschreibung von Verantwortung enthalten wesentlich die *Pflichten* des Akteurs sowie die Bedingungen der *Wissentlichkeit* und *Willentlichkeit*. Der Einsatz von Verfahren maschinellen Lernens fordert die sonst üblichen ärztlichen Sorgfaltspflichten heraus. Im Fall neuronaler Netze ist die Forderung nach einer vergleichenden Einschätzung der Resultate des Algorithmus durch den behandelnden Arzt nicht nur im Ergebnis, sondern bezüglich einzelner Merkmale, obsolet. Wegen der mangelnden Nachvollziehbarkeit des Zustandekommens dieser Resultate lässt

xis handle (vgl. London 2019). *Prima facie* kann dieser Einwurf nicht vollständig überzeugen: Die Formulierung von Standards, wie sie sich insbesondere in medizinischen Leitlinien niederschlägt und auch der Verweis auf »objektive Sorgfaltsmaßstäbe« als Grundlage für Sorgfaltspflichten zeigen den Bedarf, auch ärztlichem Handeln die Pflicht zum Korrektheitsnachweis aufzuerlegen. Allerdings muss sich dieser Nachweis korrekten Handelns nicht zwingend darauf beziehen, die Funktionsweise der eingesetzten Technik im Einzelnen nachvollziehen zu können, sondern kann auch darin bestehen, zu zeigen, dass ein bestimmter geforderter Umgang mit dem technischen Medium eingehalten wurde.

21 Rudin weist daraufhin, dass es viele proprietäre Anwendungen gibt, deren Opakheit sich dem Wunsch der Hersteller bzw. Vertreiber verdankt, nicht ohne weiteres nachahmbar zu sein. Am Beispiel des Systems COMPAS, mit dem in vielen US-Staaten die Rückfallprognosen für straffällig gewordene Personen berechnet werden, konnten Informatiker zeigen, dass sich die Prognosen mit einer einfachen Regelprogrammierung, die nur zwei bis drei Merkmale verwendet, nachgebildet werden können. Die Wahl eines nicht-nachvollziehbaren maschinellen Lernverfahrens ist somit vermutlich nicht immer von dem Ziel der besten Leistungsfähigkeit getragen (vgl. Rudin 2019). – Rudin bestreitet, dass es zwingend einen Trade-Off gibt zwischen der Interpretierbarkeit eines Modells und der Leistungsfähigkeit.

sich zudem die Frage aufwerfen, inwiefern weitere Kriterien für korrekte Verantwortungszuschreibungen erfüllbar sind. Kann man davon sprechen, dass jemand wissentlich und willentlich handelt, wenn das Instrument, das er dabei einsetzt, nicht durchschaubar ist? Bei der einzelnen behandelnden Ärztin geht die oben erwähnte Forderung nach der Fehlereinschätzung des Algorithmus zu weit und damit auch die Forderung, im Detail zu wissen, was man tut, wenn man ein bestimmtes Voraussagemodell einsetzt.[22]

Insgesamt ergeben sich damit bei Schäden, die im ärztlichen Handeln mit Unterstützung künstlicher Intelligenz entstehen, *Zurechnungslücken*: Es ist nicht klar, *welchen* Akteuren die Verantwortung für diese Schäden zuzuschreiben ist. Wie sollte man mit dieser Herausforderung für die Verantwortungszuschreibung umgehen? *Verschiedene Optionen* sind als Reaktion möglich: Man könnte aufgrund der unbefriedigenden Zurechnungslücken den Einsatz dieser Technologie in essenziellen Lebenssituationen verbieten. Man könnte das Verantwortungskonzept in der Anwendung künstlicher Intelligenz suspendieren und mit den Zurechnungslücken leben. Schließlich könnte man versuchen, durch eine Umorganisation des Handelns auf diese Herausforderung zu reagieren.[23]

Die folgende Erwägung der dritten Option verdankt sich einer Reflexion der *Zwecke*, die mit der Zuschreibung von Verantwortung verfolgt werden. Zwar spielen dabei auch Aspekte wie die Vergeltung von Taten bzw. die Entschädigung von Opfern eine Rolle; letztlich dient das Verantwortungskonzept vor allem aber einer *Steuerung des Handelns* (vgl. Hahn 2014; Seebaß 2001). Normen und Pflichten werden in Kraft gesetzt, um bestimmtes Handeln zu fördern und anderes zu unterbinden. Die Verantwortungszuschreibung und eine darauf gründende Entschädigungsforderung bzw. eine Strafe dient der Stützung des sozialen Drucks, sich an die Normen zu halten und die Pflichten zu erfüllen. Viele Normen wie z.B. die, andere nicht zu verletzen oder sie nicht zu bestehlen oder zu belügen, dienen dazu, Schäden zu vermeiden. Die Herausforderung der Verantwortungszuschreibung durch künstliche Intelligenz lässt sich als eine – in der Technikgeschichte übliche – Situation auffassen, in der es darum geht, *erlaubtes* Handeln, das zu Schädigungen führt bzw. führen kann, so umzuorganisieren, dass die Vorteile der technischen Handlungsmöglichkeiten möglichst weitgehend erhalten bleiben, und die Nachteile nicht mehr entstehen. Ein Beispiel aus der Technikgeschichte soll dazu dienen, eine solche Umor-

22 Allerdings gilt vermutlich für viele Technikanwendungen, dass die Anwendenden die Funktionsweise nicht durchschauen und in diesem Sinn »nicht wissen, was sie tun«. Dennoch ist das Nicht-Betätigen der Bremse eines Autos eine Handlung bzw. Unterlassung, die den Autolenkern in ihren Wirkungen bekannt ist. Eine Frage lautet, ob dies bei Algorithmen ebenso der Fall ist.

23 Vgl. zum sogenannten »responsibility gap« im Zusammenhang mit künstlicher Intelligenz Matthias 2004; und möglichen Reaktionsweisen Coeckelbergh 2020; Santoni de Sio/Mecacci 2021.

ganisation in den Blick zu nehmen und auf mögliche strukturelle Ähnlichkeiten im Umgang mit künstlicher Intelligenz zu untersuchen.

5. Ein historisches Beispiel für die Herausforderung der Verantwortungszuschreibung

Die erwähnten Bedingungen bei der Zuschreibung von Verantwortung – Kausalität, Norm, Normverletzung und Vorsätzlichkeit – sind auch schon in der Vergangenheit herausgefordert worden. So wird – prägnant von Kurt Bayertz formuliert – gegen Ende des 18. Jahrhunderts ein tiefgreifender Wandel der Struktur des Handelns verortet, der sich auch im Verantwortungskonzept niedergeschlagen hat. Vor allem zwei Elemente kennzeichnen diesen Wandel: Die zunehmende Arbeitsteilung und der Einsatz von Technik.

> »Die Struktur der gesellschaftlichen Arbeit – als einer paradigmatischen Form menschlichen Handelns – wird im Zuge der Industrialisierung vor allem durch zwei Prozesse in immer kürzeren Zeitabständen revolutioniert: zum einen durch die intensivierte Arbeitsteilung, zum zweiten durch den Fortschritt der Technik. *Damit schieben sich zwischen das handelnde Individuum und die durch dieses Handeln bewirkten Effekte vermittelnde Instanzen, die eine Zurechnung der Handlungsfolge auf bestimmte Individuen erschweren oder gar unmöglich machen.* Zwar hat es solche vermittelnden Instanzen immer schon gegeben, sie gewinnen im Zuge der Industrialisierung seit dem 18. Jahrhundert aber ein solches Gewicht und verstärken die Effekte des Handelns in einem solchen Maße, daß man von einer neuen Qualität sprechen muß: Es besteht keine direkte und lineare Beziehung mehr zwischen dem Akteur und der von ihm hervorgerufenen Folge. Damit wird die Reichweite und Effektivität des Handelns in einem bis dahin unbekannten Maße gesteigert.« (Bayertz 1995: 25)

Die angedeutete Änderung der Handlungsstruktur hat jedoch nicht zu einer Verwerfung des Verantwortungskonzepts geführt, sondern zu einer Veränderung bzw. Ergänzung, die im Folgenden skizziert wird.

5.1 Der Dampfkesselbetrieb

Die Rolle, die der Einsatz von Technik bei veränderten Handlungsstrukturen spielt, lässt sich am Beispiel des Dampfkessels illustrieren: In der ersten Hälfte des 19. Jahrhunderts nahm die Zahl der Kesselexplosionen auf Dampfschiffen enorm zu. Das klassische Modell der Zuschreibung von Verantwortung versagt an dieser Stelle, es ergeben sich *Zurechnungslücken*. Die Dampfkesselunfälle sind keine Resultate absichtlicher Schädigungshandlungen von Individuen an Individuen. Es handelt sich

auch nicht um Pflichtverletzungen, da das Betreiben und Benutzen der Technik ein erlaubtes Handeln darstellt. Vielmehr wird der *Betrieb* dieser von Menschen produzierten technischen Mittel von solchen Unfällen »begleitet«. Dennoch will man sich nicht einfach mit diesen schweren Schäden als bloßen Kollateralschäden erlaubten technischen Handelns abfinden, sondern will die Zurechnungslücken schließen, um Opfer entschädigen zu können und zukünftig Schäden zu vermeiden.

In diesem Rahmen entstanden sogenannte Kesselgesetze, mit denen tiefgreifend in den Betrieb solcher Anlagen eingegriffen wurde (vgl. Lueger 1904). Diese Bestimmungen enthalten verschiedene Arten von Vorgaben, die zum einen *Konstruktionsvorgaben* für Kessel und zum anderen *Normen für die Inbetriebnahme* und *für den laufenden Betrieb* formulieren. So werden für die Zulassung einer Anlage Prüfungen durch zuständige Kesselprüfer gefordert. Die weiter spezifizierten berechtigten Personen müssen für die Inbetriebnahme eine Konstruktionsprüfung, eine Wasserdruckprobe und eine Abnahmeprüfung vornehmen, wobei diese Elemente ebenfalls weiter spezifiziert werden. Außerdem werden in der Folge für die Kesselwärter Verhaltensregeln formuliert, die in den Kesselhäusern aufgehängt werden. Die Dokumentationspflicht schlägt sich in dem Gebot der Führung eines Revisionsbuches nieder: »Jeder einem Dampfkesselüberwachungsverein angehörige oder unter staatlicher Aufsicht stehende besitzt ein Revisionsbuch, in dem der Zeitpunkt und die Ergebnisse aller Prüfungen und Untersuchungen einzutragen sind.«[24]

Bezogen auf den Eisenbahnbetrieb wird schließlich der *Gefährdungstatbestand* eingeführt: Unabhängig vom Verschulden haftet der Betreiber für Schäden, die sich aus dem Eisenbahnbetrieb ergeben.

> »Die Gesellschaft ist zum Ersatz verpflichtet für allen Schaden, welcher bei der Beförderung auf der Bahn, an den auf derselben beförderten Personen und Gütern, oder auch an anderen Personen und deren Sachen, entsteht und sie kann sich von dieser Verpflichtung nur durch den Beweis befreien, dass der Schaden entweder durch die eigene Schuld des Beschädigten oder durch einen unabwendbaren äußeren Zufall bewirkt worden ist. Die gefährliche Natur der Unternehmung selbst ist als ein solcher, von dem Schadenersatz befreiender Zufall nicht zu betrachten.« (Preußisches Eisenbahngesetz, 3.11. 1938, § 25; zit. nach von Gadow 2002: 68)

Mit der Einführung von Gefährdungstatbeständen wurde die bisherige Praxis der Verantwortungszuschreibung wesentlich erweitert. Der Verzicht auf den Nachweis, dass der Schaden vorsätzlich (oder fahrlässig) herbeigeführt wurde, kann vermutlich als eine kleine Revolution der Verantwortungszuschreibung betrachtet werden.

---

24 [http://www.zeno.org/Lueger-1904/A/Dampfkesselbetrieb+%5B1%5D] (Zugriff: 12.06.2024).

5.2 Strategien für den Umgang mit Zurechnungslücken: Konstruktionsbedingungen, Rollenpflichten, Gefährdungshaftung

Die Maßnahmen stellen eine Reaktion auf die durch Technik und Arbeitsteilung entstandenen Zurechnungslücken dar. Die erweiterten Handlungsmöglichkeiten bringen nicht nur als positiv bewertete Resultate hervor, nämlich die industrielle Fertigung sowie die neuen, effizienteren Transportmöglichkeiten, sondern auch negative, nämlich Schäden an Leib und Leben.

Ausgehend davon, dass die Zuschreibung von Verantwortung eine soziale Praxis ist, kann man anhand dieses Beispiels die Funktion dieser sozialen Praxis und ihre Realisierung erläutern. Die Funktion liegt *letztlich* in der Handlungssteuerung. Akteure sollen beispielsweise Handlungen unterlassen, die zu bestimmten unerwünschten Folgen führen. In der rechtlichen Systematisierung dieser Praxis wird im Fall von Körperverletzung, Diebstahl, Betrug etc. diese Zielsetzung verfolgt, indem man Tatbestände als rechtswidrig identifiziert und ihre Herbeiführung verbietet. Es wird also eine entsprechende Verbotsnorm in Kraft gesetzt. Führt jemand diesen rechtswidrigen Zustand herbei und erfüllt die Bedingungen für die Verantwortungszuschreibung, ist die Grundlage für eine Schadenersatzforderung und eine Bestrafung gelegt. *Ex ante* wird dieses Handeln somit als unerwünschtes verboten. Kommt es trotzdem zu solchen Handlungen, wird auf dieser Basis *ex post* Verantwortung zugeschrieben und der Täter bestraft. Die Sanktion versieht das handlungssteuernde Verbot mit einem zusätzlichen sozialen Druck. Soweit zum klassischen Modell.

Bei den geschilderten Dampfkesselunfällen liegt die Lage anders: Der Betrieb von Dampfkesseln ist ein *erlaubtes technisches Handeln mit erwünschten Folgen* – allerdings können auch unerwünschte Nebenfolgen eintreten. Wenn die Funktion der Verantwortungszuschreibung, nämlich die Handlungssteuerung, erfüllt werden soll, ist zu fragen: *Welche Handlungsbeschränkungen bzw. Handlungslenkungen des eigentlich erlaubten und prinzipiell erwünschten Handelns sollen eingeführt werden, um die unerwünschten Folgen, die Schäden, zu vermeiden? Welche Akteure können und sollen in ihrem Handeln wie angeleitet werden?* – Hier handelt es sich um eine *ethische* (wenn es um die Umsetzung in Regulierungen geht, auch rechtliche) Fragestellung: Es geht darum, welche Handlungsbeschränkungen im Sinne von zu unterlassenden, aber auch auszuführenden Handlungen man Akteuren auferlegen darf/sollte/könnte, um die Beeinträchtigungen anderer (oder auch gleicher), die sich ohne diese Regulierung einstellen würden, zu vermeiden. Diese Handlungsbeschränkungen durch entsprechende Normen – im Grenzfall auch das vollständige Verbot dieses Handelns, im Beispiel also das Betreiben von Dampfkesseln –, werden den Akteuren gegenüber als Forderung und mit Zwang aufrechterhalten, auch wenn diese weder der Zielsetzung der Regulierung noch ihrer Rechtfertigung zustimmen. Wie diese Fragen entschieden werden, d.h. wie die Abwägung von erzielbaren erwünschten

Folgen gegenüber den potenziellen Schädigungen ausfällt, ist nicht vorgegeben. Es handelt sich um eine verhandelbare Abwägung, bei deren Durchführung wenigstens zwei Elemente entscheidend sind: *Erstens* müssen die jeweiligen Optionen mit ihren abgeschätzten Konsequenzen für die Entscheider klar dargelegt sein: Wer profitiert in welcher Weise, wer hat Einschränkungen und potenzielle Schäden zu gewärtigen? Mit welchem Grad an Gewissheit lassen sich diese Aussagen treffen? – Hierzu ist in den betrachteten Fällen die entsprechende wissenschaftlich-technische Expertise erforderlich. *Zweitens* muss eine normative Analyse die Regulierungsoptionen auf Verträglichkeit mit bereits vorhandenen Rechten, Pflichten und Verfahrensvorschriften prüfen.[25]

Im Fall der erwähnten technischen Entwicklung, d.h. beim Betrieb von Dampfkesseln und Eisenbahnen, hat die Regulierung auf verschiedenen Ebenen angesetzt. Erstens wurde wissenschaftlich-technische Expertise herangezogen, um Vorschriften für die *Konstruktion* der Kessel zu formulieren. Zweitens wurden Personen mit bestimmter Expertise identifiziert, nämlich Ingenieure, die berechtigt sind, Kessel für die Inbetriebnahme und auch den laufenden Betrieb zu prüfen. Drittens wurde eine Personengruppe identifiziert, nämlich die Kesselwärter, und Vorschriften für ihr Handeln formuliert. Die letzten beiden Maßnahmen lassen sich als Etablierung von *Rollenpflichten* lesen. Rollenpflichten können darin bestehen, bestimmte Zustände, z.B. die Funktionsfähigkeit einer Maschine, aufrechtzuerhalten und geben dabei die auszuführenden Handlungen nicht im Einzelnen vor. Vielmehr ist unterstellt, dass diejenigen, an die sich die Pflicht richtet, über die notwendige Expertise verfügen, um die spezifizierte Aufgabe zu erfüllen. In der Debatte um das Verantwortungskonzept wird der Umstand, dass z.B. der Einsatz von Technik die Anwendung des klassischen individualistischen Verantwortungskonzepts verhindert, mit der Forderung nach einem neuen Verantwortungskonzept verbunden. Für den Umgang mit Zurechnungslücken ist jedoch nicht ein neues Verantwortungskonzept erforderlich, sondern es werden vielmehr *neue Pflichten* benötigt. Diese Pflichten bestehen vermehrt in Rollenpflichten, auch um dynamische Entwicklungen des technischen Handlungsfeldes zu erfassen und zu vermeiden, immer wieder dem jeweiligen Entwicklungsstand angepasste Handlungsnormen zu verabschieden. Die Rollenpflichten bauen auf einer Expertise der angesprochenen Akteure auf; diese haben beispielsweise den »Stand der Technik« zu beachten.[26]

Viertens, und das ist die deutlichste Veränderung gegenüber der bisherigen sozialen Praxis, wurden *Gefährdungstatbestände* formuliert und damit zugleich

25 Zur Berücksichtigung empirischer Sachverhalte in normativen Argumentationen vgl. Bayertz 1991: 27ff. – Der Abgleich neuer normativer Forderungen mit bestehenden Rechten und Pflichten lässt sich methodisch durch die Herstellung von Überlegungsgleichgewichten anleiten (vgl. Hahn 2016).

26 Vgl. zur Debatte um das Verantwortungskonzept Bayertz 1995; Hahn 2014.

von den Bedingungen der Verantwortungszuschreibung abgewichen. In diesen Fällen ist die Kausalitätsbedingung nicht mehr relevant bzw. abgeschwächt, d.h. es ist nicht nachzuweisen, dass der Betrieb der Eisenbahn den Schaden konkret hervorgerufen hat. Es reicht, dass der Schaden im Umfeld dieses Betriebes eingetreten ist. Ähnliche Bedingungen finden sich auch bei der später eingeführten Produkthaftung.[27]

Die Formulierung von Konstruktionsbedingungen, von Rollenpflichten sowie vor allem des Gefährdungstatbestands lassen sich im Sinne der Handlungssteuerung lesen: Die externen Bedingungen des Handelns werden durch diese Forderungen und die Androhung von Schadenersatzforderungen bzw. sogar Strafen so verändert, dass gerade die Individuen, die Einfluss auf den Technikbetrieb nehmen können, entsprechende Anreize bekommen, mögliche Schäden zu verhindern. Anders gesagt, handelt es sich um eine Umstrukturierung des Handelns, die auch die Motivationen der Akteure miteinbezieht – eine Vorgehensweise, die Moritz Schlick als zentral für die Verantwortungszuschreibung hervorhebt:

> »Die Frage nach der Verantwortung ist nun die: Wer ist denn im gegebenen Fall eigentlich zu bestrafen? Wer ist als wahrer Täter der Handlung anzusehen? Die Frage ist nicht einfach identisch mit der nach dem Urheber der Handlung überhaupt, denn als solche könnten sich schließlich ebensogut die Urgroßeltern des Täters gelten, denen er durch Vererbung seinen Charakter verdankt, ferner die Staatsmänner, die sein soziales Milieu geschaffen haben usf. – Sondern ›Täter‹ heißt derjenige, *an dem die Motive hätten einsetzen müssen*, um die Tat sicher zu verhindern (bzw. hervorzurufen). […] Die Frage nach dem Verantwortlichen ist die Frage nach dem *richtigen Angriffspunkt der Motive*.« (Schlick 1984[1930]: 161f.)

6. Übertragbarkeit der Strategie auf den Umgang mit Algorithmen?

Lässt sich dieses Vorgehen im Umgang mit Zurechnungslücken auf die Zurechnungslücken übertragen, die durch den Einsatz von Algorithmen in der medizinischen Diagnose entstanden sind? Selbst wenn das nicht der Fall sein sollte, kann die Erörterung möglicherweise dazu dienen, neue Fragen oder neue mögliche Lösungsstrategien zu formulieren.

27 Vgl. dazu unter dem Stichwort »Zurechnungsexpansion« Lübbe 1998. Für eine detailliertere Darstellung des Technikrechts und der durchgreifenden Änderungen im Zeitalter der Industrialisierung vgl. Vec 2011.

6.1 Algorithmen zur Entwicklung von Vorhersagemodellen – Wissenschaftsphilosophische Bemerkungen

Die genannten Fragen erfordern zunächst die Überlegung, welche Funktionen mit dem Einsatz der Algorithmen erfüllt werden sollen. Eine Hauptaufgabe besteht darin, auf der Basis verfügbarer Merkmale die Manifestation einer Krankheit *vorauszusagen*. Grundlage für die Voraussage ist die vorhandene Erfahrung. Die bisherigen Vorkommnisse der Erkrankung werden in einen allgemeinen Zusammenhang zu Merkmalen gesetzt, die die Erkrankten aufweisen. Anders gesagt: Man will durch die Bildung allgemeiner Hypothesen die Mittel bereitstellen, um neue Fälle voraussagen zu können, letztlich um entsprechend intervenieren zu können. Damit ist ein *wissenschaftsphilosophisches Kernthema* berührt, nämlich das von *Erklärung und Voraussage*, und damit auch der Gewinnung von Gesetzeshypothesen durch *Induktion* (vgl. Chalmers 2007: 35ff.; Schurz 2006: 47ff.). Auf diesen Wegen zur Gewinnung von Erkenntnis liegen zentrale wissenschaftsphilosophische Problemstücke, die Gegenstand vieler Überlegungen waren und sind. Dazu gehören neben dem bereits erwähnten Problem des Unterschieds von Korrelation und Kausalität der Umstand der Gehaltserweiterung durch induktive Schlüsse, der Umgang mit Wahrscheinlichkeiten in Anwendung auf Einzelfälle sowie das Problem der Gesetzesartigkeit.[28]

Wissenschaftsphilosophische Reflexionen begleiten die Bemühungen der Wissenschaften, die Erkenntniswege zu sichern. Die Bedingungen für kontrollierte Experimente oder für die Gewinnung statistischer Hypothesen lassen sich als Korrektheitsstandards für die Qualität der allgemeinen Aussagen und der daraus abgeleiteten Voraussagen deuten. Wissenschaftliche Äußerungen mit dem Anspruch auf Erkenntnis sind mit dem Verweis auf das Einhalten der Standards als korrekte Äußerungen zu rechtfertigen. Wer eine Voraussage macht, muss zu ihrem Korrektheitsnachweis auf entsprechende allgemeine Aussagen verweisen und die Gewin-

28 »Daß ein gegebenes Stück Kupfer den elektrischen Strom leitet, erhöht die Glaubwürdigkeit von Aussagen, daß andere Kupferstücke den Strom leiten und damit wird die Hypothese bestätigt, daß alles Kupfer den Strom leitet. Doch die Tatsache, daß ein bestimmter Mann, der sich jetzt in diesem Zimmer befindet, ein dritter Sohn ist, erhöht nicht die Glaubwürdigkeit von Aussagen, daß andere Männer, die sich jetzt in dem Zimmer befinden, auch dritte Söhne sind, und bestätigt also nicht die Hypothese, daß alle Menschen, die sich jetzt in diesem Zimmer befinden, dritte Söhne sind. Doch in beiden Fällen ist unsere Hypothese eine Verallgemeinerung der Datenaussage. Der Unterschied liegt darin, daß im ersten Fall die Hypothese eine gesetzesartige Aussage ist, im zweiten dagegen bloß eine zufällige allgemeine Aussage. Nur eine gesetzesartige Aussage – unabhängig von ihrer Wahrheit oder Falschheit oder ihrer wissenschaftlichen Bedeutung – kann durch einen ihrer Anwendungsfälle bestätigt werden, zufällige Aussagen können es nicht. Offenbar müssen wir uns also nach einer Möglichkeit umsehen, gesetzesartige von zufälligen Aussagen zu unterscheiden.« (Goodman 1975: 97)

nung dieser allgemeinen Aussagen mit dem Nachweis rechtfertigen, beispielsweise kontrollierte Experimente durchgeführt zu haben.[29]

Die unter Einsatz von Algorithmen entwickelten Voraussagemodelle sollen ihren Befürwortern zufolge den menschlichen Prognosen äquivalente Leistungen erbringen. Sie werden im hier betrachteten Beispielbereich zur Erstellung von Diagnosen verwendet und werden somit zum Bestandteil von Handlungsketten, an deren Ende unerwünschte Folgen stehen können. Die Rechtfertigungsbedürftigkeit der Prognosen sowie in einem weiteren Schritt der Bildung von gesetzesartigen Aussagen oder »Muster« führt auf die Frage, auf welche Weise und durch das Handeln welcher Akteure sich die Qualität der Voraussagemodelle sicherstellen lässt.[30] Kann die Strategie des Umgangs mit Zurechnungslücken, die in Bezug auf die Dampfkesselunfälle identifiziert wurde, hier übertragen werden?

6.2 Konstruktionsbedingungen für Algorithmen?

Kern der Herausforderung für die Verantwortungszuschreibung ist die mangelnde Nachvollziehbarkeit der Resultate von Algorithmen maschinellen Lernens. Inzwischen gibt es vielerlei Bemühungen, gerade an diesem Defekt Reparaturmaßnahmen vorzunehmen. Die Erklärbarkeit, Interpretierbarkeit, Nachvollziehbarkeit, Verständlichkeit oder auch Transparenz von Algorithmen werden als von Betroffenen einforderbare Bedingungen in Katalogen zum Umgang mit KI festgeschrieben.[31] Diese Bedingungen ließen sich als Analogon zu den Konstruktionsbedingungen für Dampfkessel diskutieren.[32]

Eine Einschätzung bezüglich der Realisierbarkeit dieser Verständlichkeitseigenschaften scheint aber unter den beteiligten Wissenschaftlern nicht einhellig zu sein. So kommt beispielsweise in der einschränkenden Formulierung »Wenn keine anderen Modelle zur Verfügung stehen, muss die Nachvollziehbarkeit durch

29 Für die Rechtfertigung im statistischen Fall sowie zu Überlegungen zu Korrelation und Kausalität vgl. Schurz 2006: 133ff.

30 Dazu gehört eine Bewertung, wie zuverlässig die Prognosen sind im Vergleich zu menschlichen Voraussagen. Bislang gibt es lediglich Ansätze für einen solchen Vergleich, die die Leistungen der Algorithmen als vielversprechend beurteilen. Allerdings fehlen hierzu noch umfassende und methodisch sorgfältige Vergleichsstudien. (Vgl. Liu et al. 2019) – Eine Zusatzschwierigkeit entsteht für die allgemeine Betrachtung der Leistungsfähigkeit von Algorithmen dadurch, dass die Vorhersagemodelle häufig in privatwirtschaftlichen Arrangements mit entsprechenden Eigentumsrechten entwickelt werden.

31 Vgl. die Übersicht bei Hagendorff 2020.

32 Allerdings wäre hier zunächst eine begriffliche Abgrenzung dieser Eigenschaften und in einem weiteren Schritt eine kritische Betrachtung ihrer Verwendung in normativen Zusammenhängen angezeigt (vgl. Alloa 2019).

den Gebrauch von post-hoc Erklärungen gesteigert werden«[33] zum Ausdruck, dass Verständlichkeit im engeren Sinn einer Vorsehbarkeit und Erklärbarkeit von Input und Output nicht immer realisierbar ist. Dem stehen Aussagen gegenüber, die den häufig behaupteten trade-off zwischen Akkuratheit der Vorhersagen und Transparenz – also der zwingenden Verminderung der Transparenz als Preis der besseren Vorhersagen – als Mythos bezeichnen. Es gebe auch Vorgehensweisen, gute Prognosen oder Klassifikationen zu bekommen, ohne diesen ›Preis‹ zu entrichten.[34] Beim jetzigen Stand lässt sich angesichts dieser Erörterungen lediglich festhalten, dass in der Forschungscommunity keine Einigkeit darüber herrscht, welche Arten der Verständlichkeitsforderungen bezüglich der Konstruktion von KI-Modulen in welcher Form realisierbar sind.

Neben den genannten Bedingungen werden inzwischen aber auch weitere Kriterien für die Qualität von KI-Algorithmen genannt, die das Vertrauen in diese Technologie stärken sollen. Im Anschluss an die wissenschaftsphilosophische Frage, wie sich die Qualität von algorithmischen Voraussagemodellen sichern lässt, und damit letztlich auch das übergeordnete Problem, das Handeln so zu steuern, dass möglichst viel von den erwünschten Konsequenzen des Technikeinsatzes erhalten bleiben, während unerwünschte Folgen vermieden werden, wären diese weiteren Kriterien zu erörtern.

So werden beispielsweise in der erwähnten Richtlinie des Deutschen Instituts für Normung (neben der Nachvollziehbarkeit) Funktionalität und Leistungsfähigkeit einerseits sowie Robustheit andererseits als weitere Bedingungen für einen Qualitätsnachweis von Algorithmen genannt.[35] Dabei werden diese Qualitätsmerkmale erläutert mit Forderungen, die sich auf die adäquate »Problemformalisierung,

33 DIN SPEC 92001–1: Artificial Intelligence – Life Cycle Processes and Quality Requirements – Part 1: Quality Meta-Model. In dieser DIN-Richtlinie wird ein Modell vorgestellt, das den allgemeinen Rahmen für spezielle Qualitätsanforderungen für AI-Entwicklungen bieten soll. Ein Teil dieser speziellen Anforderungen wird in einer weiteren Richtlinie ausgearbeitet: DIN SPEC 92001–2: Artificial Intelligence – Life Cycle Processes and Quality Requirements – Part 2: Robustness. – Die DIN-Richtlinien sind nur ein Beispiel von vielen Initiativen zur Sicherstellung der Qualität von Algorithmen.

34 Vgl. das Plädoyer von Rudin, das sich bereits im Aufsatztitel niederschlägt (Rudin 2019) – Für eine skeptische Position bezüglich der Einlösbarkeit der Verständlichkeitsforderungen vgl. Durán/Jongsma 2021. – Die Beispielbetrachtung der Herz-Kreislauf-Diagnostik sollte illustrieren, dass künstliche neuronale Netze, erst recht, wenn sie mehr Ebenen enthalten als im Beispiel, durch die Vielzahl der beteiligten Gewichtungen nicht mehr hinsichtlich des Weges, den eine Dateneingabe (Merkmalsset) bis zu einem Ergebnis (Prognose) nimmt, nachvollziehbar sind.

35 DIN SPEC 92001–1:2019-04: 21. – Ähnliche Anforderungen, vor allem im Hinblick auf die Einbeziehung des Expertenwissens werden im Übrigen auch von Rudin genannt, wenn es darum geht, die Interpretierbarkeit der Modelle sicherzustellen (vgl. Rudin 2019).

Aufgabenanalyse, Datensammlung, Datenanalyse und Datenverarbeitung«[36] beziehen. Hinzu kommt die Bewertung der Leistungsfähigkeit und die Auswahl des Modells, die wiederum adäquate Kriterien, einschließlich passender Metrisierung voraussetzen. Das Qualitätsmerkmal der Robustheit von Modulen künstlicher Intelligenz soll sicherstellen, dass diese mit irritierenden Daten umgehen können, d.h. z.B. mit solchen Daten, die über die Trainings- und Testdatensets hinausgehen.[37]

Eine ähnliche Richtung schlägt ein Vorschlag ein, der unter dem Titel »Computational Reliabilism« geführt wird.[38] Unter den vier Indikatoren für die Zuverlässigkeit finden sich Verifizierung und Validierung, Robustheitsanalyse, Betrachtung der erfolgreichen bzw. erfolglosen Implementationsgeschichte und die Einbeziehung von Expertenwissen (vgl Durán/Jongsma 2021: 332).

Die Erläuterungen zur sorgfältigen Formulierung der zu lösenden Aufgabe sowie der Datensammlung, -analyse und -verarbeitung im Rahmen des Qualitätsmerkmals »Funktionalität und Leistungsfähigkeit« lassen sich als Pendant zu den Standards bei der »manuellen Entwicklung« von Voraussagemodellen betrachten. Dabei wird deutlich, dass neben dem eigentlichen Einsatz von Verfahren wie logistischer Regression oder künstlichen neuronalen Netzen die *fachliche Expertise* im Umgang mit dem jeweiligen Problem, aus bestehenden Daten und Erfahrungen Voraussagen für neue Fälle abzuleiten, ausschlaggebend sind. Im Beispiel sind es nicht beliebige Daten, die für die Entwicklung des Voraussagemodells herangezogen werden, sondern solche, bei denen Mediziner bereits die Vermutung haben, dass sie mit Krankheitsverläufen korreliert sind. Selbst wenn Verfahren maschinellen Lernens verwendet werden, um erste Hinweise auf Korrelationen zu bekommen, sind diese Vermutungen zum Gegenstand weiterer Analysen und Beurteilungen zu machen.

Die genannten Qualitäts- oder Zuverlässigkeitskriterien lassen sich – jedenfalls u.a. – als Qualitätskriterien für die Gewinnung von Erkenntnis auffassen. So werden datenbezogene Erfordernisse für die Verbesserung der Verallgemeinerung erwähnt. Man kann vermuten, dass z.B. Modelle für die Wirksamkeit von Medikamenten, die in ihrer Datengrundlage nicht repräsentativ sind für die angezielte Patientengruppe, diesen Erkenntnisnormen nicht genügen. Anders als häufig in der öffentlichen Diskussion anzutreffen, ist anzuraten, diese Forderung nicht unter die Rubrik »Ethik« einzusortieren. Nicht jede Norm ist eine moralische Norm. Für die Sicherstellung der Qualität wissenschaftlicher Resultate sind nicht Moralphiloso-

36 DIN SPEC 92001–1:2019-04: 21; Übersetzung – SH.

37 DIN SPEC 92001–1:2019-04: 21.

38 Vgl. Durán/Jongsma 2021. – In diesem Artikel wird auf diesen Ansatz im Kontext medizinischer KI-Anwendungen verwiesen.

phinnen zuständig, sondern die Wissenschaftlerinnen, die sich an den Standards zur Erkenntnisgewinnung orientieren müssen

Mit den aufgeführten Bedingungen, die die Qualität der Vorhersagemodelle sicherstellen sollen, wird zum Teil an Selbstverständlichkeiten wissenschaftlicher Sorgfalt erinnert. An wen richten sich die so formulierten Pflichten? Es scheint – und das ist vermutlich jedenfalls in dieser Umfassendheit eine Neuheit –, dass hier Dateningenieure und Wissenschaftler des jeweiligen Gebiets, für das ein KI-Modul entwickelt wird, in Kooperation verpflichtet werden (vgl. Rudin 2019).

Liegt mit diesen geforderten Merkmalen ein Analogon zu den Konstruktionsbedingungen für Kessel im 19. Jahrhundert vor? Sind Richtlinien solcher Art *geeignet, das Handeln in gewünschter Weise zu steuern*, d.h. so zu steuern, dass man vom Einsatz der KI-Module profitieren kann, ohne ihren Schäden ohne weitere Vorkehrungen ausgesetzt zu sein? Von der Antwort auf diese Frage ist auch die Erörterung des Problems betroffen, inwiefern man den Einsatz der KI-Module in risikoreichen Situationen zumuten kann.

Lassen sich Funktionalität, Leistungsfähigkeit und Robustheit sicherstellen, können diese Anforderungen somit handlungssteuernd wirken? Lässt sich auf diese Weise Vertrauen in die Technologie stärken? Könnten Initiativen wie die TÜV-Zertifizierung von KI-Anwendungen diese Bedingungen nutzen?

Diese Fragen sind zu klären, um einschätzen zu können, ob die genannten Bedingungen ein leistungsäquivalentes Gegenstück zu den Konstruktionsbedingungen für Kessel darstellen. Wenn dies der Fall ist, sie also eine handlungssteuernde Wirkung entfalten können, können auch Verstöße gegen diese Qualitätsstandards festgestellt und zur Grundlage von Verantwortungszuweisungen gemacht werden.

6.3 Rollenpflichten und Gefährdungshaftung

Neben den Konstruktionsbedingungen waren Rollenpflichten für Ingenieure und Kesselwärter sowie das Institut der Gefährdungshaftung Reaktionen auf die entstandenen Zurechnungslücken. Letzteres wird beispielweise inzwischen auch in der Produkthaftung realisiert. Eine Überlegung wäre also, die KI-Module, die in Hochrisikosituationen eingesetzt werden, mit einer solchen Gefährdungshaftung zu verknüpfen. Bezogen auf den Eisenbahnbetrieb hat diese Regulierung offenbar nicht dazu geführt, dass die Gesellschaften ihren Betrieb aus Furcht vor nicht bewältigbaren Entschädigungsleistungen bzw. auch strafrechtlichen Folgen eingestellt haben. Will man die Übertragbarkeit auf KI-Module weiter ausloten, könnten weitere Untersuchungen die Voraussetzungen dieser historischen Verantwortungszuschreibung für Schäden, unabhängig von nachgewiesenem Verschulden eruieren. Es ist allerdings zu vermuten, dass man bei der Regulierung davon ausgegangen ist, dass die Betreiber über das notwendige Know-how verfügen, um Schäden sowohl durch die Konstruktion der Anlagen als auch durch die Normierung der Handha-

bung zu verhindern. Eine entscheidende Frage lautet: Gilt dies auch für Algorithmen? Die Frage verweist erneut auf die Bedingungen für die Handhabbarkeit und damit auf die Funktionsweise.[39]

Wenn man sich gegen eine Gefährdungshaftung entscheidet und stattdessen auf Rollenpflichten setzt, ist zu fragen, an wen sich Rollenpflichten im Fall von KI-Modulen in der medizinischen Diagnostik richten würden? Es liegt vermutlich nahe, hier Pflichten von Dateningenieurinnen und Medizinern bei der Entwicklung der KI-Module zu formulieren, die im Sinne der beiden genannten Qualitätsstandards liegen. Was darüberhinausgehend die Pflicht der einzelnen behandelnden Ärztin angeht, die – möglicherweise gezwungenermaßen – mit den Empfehlungen des Algorithmus umgehen muss, wäre es denkbar zu fordern, dass sie die Empfehlung des KI-Moduls missachtet, wenn sie dies für richtig hält. Ob diese Forderung adäquat ist, ist fraglich – immer auf dem Hintergrund der Zielsetzung, dass man vom Einsatz der Algorithmen profitieren möchte. Zunächst hieße das, dass die behandelnde Ärztin einerseits einem Algorithmus vertrauen, andererseits aber misstrauen soll. Wenn nicht klar ist, aufgrund welcher Merkmale ein Algorithmus zu einem Ergebnis kommt, scheint diese Forderung schwierig einzulösen.[40] Zudem ist auf das allgegenwärtige Problem übermäßigen Vertrauens gegenüber übermäßigem Misstrauen zu verweisen: Häufig stellen sich Routinen ein, in denen das Funktionieren des Algorithmus nicht hinterfragt wird (übermäßiges Vertrauen). Demgegenüber ist auch festzuhalten, dass die Forderung, das Ergebnis des Algorithmus zu missachten, das Ziel, von seinem Einsatz zu profitieren, konterkarieren

39 Ob die Entwicklung, der Betrieb und die Nutzung von KI-Algorithmen und von mit ihnen betriebenen Robotern »nicht mehr mit den Mitteln individualistischer Handlungstheorien zu erfassen sind, sondern auf Theorieangebote zurückgreifen müssen, die imstande sind, neue soziotechnische Ensembles von Menschen als zentrale Phänomene des Gesellschaftlichen zu begreifen« (Gruber 2013: 366) – das soll hier bezweifelt werden: Wie im oben aufgeführten Schlick-Zitat dargelegt, geht es bei der Handlungssteuerung darum, diejenige Person zu identifizieren, die die Tat hätte verhindern können. Solange Maschinen keine Motive aufweisen, an denen Anreize oder Sanktionen ansetzen können, bleibt nur der Ansatz bei den agierenden Personen (oder Personenkollektiven wie Unternehmen). Diesen Weg schlägt Gruber (der interessanterweise in seinem Aufsatz ebenfalls mit dem Vergleich zur Gefährdungshaftung des Eisenbahnbetriebs operiert) mit seiner Überlegung zur Haftungsverteilung selbst ein: Er unterstellt, dass die Personen, die wissen, dass sie haften werden, die entsprechenden Sicherungsmaßnahmen ergreifen bzw. Versicherungen abschließen (vgl. Gruber 2013: 370).

40 Auch die Forderung von Hannah Fry, Mathematikerin, das Beste aus beiden Welten in einer »AI-alliance« zu vereinigen, nämlich die Fähigkeit des Algorithmus, Veränderungen von Gewebe zu entdecken und die Fähigkeit der Ärztinnen, falsch-positive Resultate zu identifizieren, ist auf diesem Hintergrund zu hinterfragen (vgl. Fry 2018: bes. 102ff.). – Für eine Diskussion des Umgangs abweichender Einschätzungen von Algorithmus und Ärztinnen vgl. Grote/Berens 2020.

kann. Es ist gerade nicht so, dass der menschliche Eingriff immer zu besseren Resultaten führt.

7. Zwei Fragen als Resümee

Der Einsatz von Algorithmen in der medizinischen Diagnostik fordert wegen resultierender Zurechnungslücken überkommene Praxen der Verantwortungszuschreibung hinaus. Das dadurch aufgeworfene Problem, ob sie wegen dieser Herausforderung überhaupt zum Einsatz kommen sollen oder dürfen, und wenn ja, ob diese Anwendungen unter der Maßgabe regulatorischer Vorgaben stehen sollen, lässt sich aufgrund der vorangegangenen Analyse in zwei Fragen überführen:

A. *Sollen Handlungsbeschränkungen des eigentlich erlaubten und prinzipiell erwünschten Handelns eingeführt werden, um die unerwünschten Folgen, die Schäden, zu vermeiden?*

Will man diese Abwägungsentscheidung von ethischer Relevanz behandeln, dann gehen die verfolgten Ziele, die auf dem Spiel stehenden Interessen von Individuen aus unterschiedlichen Gruppen, das Verhältnis zu bestehenden Rechten und Pflichten sowie auch Überlegungen zur Umsetzbarkeit entsprechender Regulierungen ein. Der letzte Gesichtspunkt stellt die Verbindung zur zweiten Frage her:

B. *Welche Akteure können und sollen in ihrem Handeln wie angeleitet werden?*

Hierzu ist Auskunft von Informatikern, Datenwissenschaftlern und Fachwissenschaftlern erforderlich:
Lässt sich Transparenz herstellen?
Lassen sich andere Gütekriterien für Algorithmen zur Entscheidungsfindung formulieren?

Antworten auf diese Fragen sind wesentlich, um darüber entscheiden zu können, welche Regulierungen sich umsetzen lassen und zu welchen Zielen sie beitragen können, d.h. inwiefern sich das Handeln in gewünschter Weise lenken lässt. Normative Überlegungen, so eine Lehre daraus, umfassen immer auch Überlegungen nicht-normativer Art.

All das zusammen liefert die Grundlage für aufgeklärte Urteile.

Literatur

Alloa, E. (2019): Das Unbehagen in der Transparenz, in: *Internationales Jahrbuch für Medienphilosophie*, 5(1), 155–182.

Aristoteles (2006): Nikomachische Ethik. übersetzt und herausgegeben von Ursula Wolf, Reinbek bei Hamburg: Rowohlt Verlag.

Baesens, B. (2014): Analytics in a big data world. The essential guide to data science and its applications, Minneapolis: John Wiley & Sons, Inc.

Bayertz, K. (1991): Praktische Philosophie als angewandte Ethik, in: Ders. (Hg.), Praktische Philosophie. Grundorientierungen angewandter Ethik, Reinbek bei Hamburg: Rowohlt Verlag, 7–47.

Bayertz, K. (1995): Eine kurze Geschichte der Herkunft der Verantwortung, in: Ders. (Hg.), Verantwortung: Prinzip oder Problem?, Darmstadt: Wissenschaftliche Buchgesellschaft, 3–71.

Braham, M.; van Hees, M. (2012): An anatomy of moral responsibility, in: *Mind*, 121(483), 601–634.

Bringsjord, S.; Govindarajulu, N.S. (2020): Artificial intelligence, in: Zalta, E.N.; Nodelman, U. (Hg.), The Stanford Encyclopedia of Philosophy, Stanford (CA): The Metaphysics Research Lab. [https://plato.stanford.edu/archives/win2019/entries/artificial-intelligence/] (Zugriff: 22.02.2022).

Burrell, J. (2016): How the machine ›thinks‹. Understanding opacity in machine learning algorithms, in: *Big Data & Society*, 3(1), 1–12.

Chalmers, A.F. (5. Auflage 2007): Wege der Wissenschaft. Einführung in die Wissenschaftstheorie, Berlin: Springer-Verlag.

Coeckelbergh, M. (2020): Artificial Intelligence, Responsibility Attribution, and a Relational Justification of Explainability, in: *Science and Engineering Ethics*, 26(4), 2051–2068.

Durán, J.M.; Jongsma, K.R. (2021): Who is afraid of black box algorithms? On the epistemological und ethical basis of trust in medical AI, in: *Journal of Medical Ethics*, 47(5), 329–335.

Engemann, C. (2018): Rekursionen über Körper. Machine Learning-Trainingsdatensätze als Arbeit am Index, in: Engemann, C.; Sudmann, A. (Hg.), Machine Learning. Medien, Infrastrukturen und Technologien der Künstlichen Intelligenz, Bielefeld: transcript Verlag, 247–268.

Finlay, S. (2017): Artificial intelligence and machine learning for business. A no-nonsense guide to data driven technologies, Lancashire: Relativistic.

Fry, H. (2018): Hello world. How to be human in the age of the machine, New York: W. W. Norton & Company.

von Gadow, O. (2002): Die Zähmung des Automobils durch die Gefährdungshaftung, Berlin: Duncker & Humblot.

Goodman, N. (1975): Tatsache, Fiktion, Voraussage, Frankfurt a.M.: Suhrkamp.

Grote, T.; Berens, P. (2020): On the ethics of algorithmic decision-making in healthcare, in: *Journal of medical ethics*, 46(3), 205–211.

Gruber, M.-C. (2013): Gefährdungshaftung für informationstechnologische Risiken: Verantwortungszurechnung im ›Tanz der Agenzien‹, in: *Kritische Justiz*, 46(4), 356–371.

Hagendorff, T. (2020): The ethics of AI ethics. An evaluation of guidelines, in: *Minds and Machines*, 30(1), 99–120.

Hahn, S. (2014): Norm und Verantwortung, in: *Archiv für Rechts- und Sozialphilosophie*, 100(4), 429–449.

Hahn, S. (2016): From Worked-Out Practice to the Justification of Norms by Producing a Reflective Equilibrium, in: *Analyse & Kritik*, 38(2), 339–369.

Hahn, S. (2024): Algorithmische ›Entscheidungen‹ in der Medizin? Eine Reflexion zu einem handlungsbezogenen Ausdruck, in: Ruschemeier, H.; Steinrötter, B. (Hg.), Der Einsatz von KI & Robotik in der Medizin. Interdisziplinäre Fragen, Nomos: Baden-Baden, 13–26.

Heinrich, B. (2005): Strafrecht – Allgemeiner Teil I, Stuttgart: Kohlhammer Verlag.

Lepri, B.; Oliver, N.; Letouzé, E.; Pentland, A.; Vinck, P. (2018): Fair, transparent and accountable algorithmic decision-making processes. The Premise, the proposed solutions, and the open challenges, in: *Philosophy & Technology*, 31(4), 611–627.

Liu, X.; Faes, L.; Kale, A.U.; Wagner, S.K.; Fu, D.J.; Bruynseels, A.; Mahendiran, T.; Moraes, G.; Shamdas, M.; Kern, C.; Ledsam, J.R.; Schmid, M.K.; Balaskas, K.; Topol, E.J.; Bachmann, L.M.; Keane, P.A.; Denniston, A.K. (2019): A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging. A systematic review and meta-analysis, in: *The Lancet Digital Health*, 1(6), 271–297.

London, A.J. (2019): Artificial intelligence and black-box medical decisions. Accuracy versus explainability, in: *Hastings Center Report*, 49(1), 15–21.

Lübbe, W. (1998): Verantwortung in komplexen kulturellen Prozessen, Freiburg i.Br.: Verlag Karl Alber.

Mainzer, K. (2016): Künstliche Intelligenz – Wann übernehmen die Maschinen?, Berlin: Springer Verlag.

Matthias, A. (2004): The responsibility gap. Ascribing responsibility fort he actions of learning automata, in: *Ethics and Information Technology*, 6(3), 175–183.

Pearl, J. (2018): The book of why: The new science of cause and effect, New York: Basic Books.

Ramge, T. (2018): Mensch und Maschine. Wie Künstliche Intelligenz und Roboter unser Leben verändern, Stuttgart: Reclam Verlag.

Rudin, C. (2019): Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, in: *Nature Machine Intelligence*, 1(5), 206–215.

Santoni de Sio, F.; Mecacci, G. (2021): Four Responsibility Gaps with Artificial Intelligence. Why they Matter and How to Address them, in: *Philosophy & Technology*, 34(4), 1057–1084.

Schlick, M. (1984[1930]): Fragen der Ethik, Frankfurt a.M.: Suhrkamp.

Schurz, G. (2006): Einführung in die Wissenschaftstheorie, Darmstadt: Wissenschaftliche Buchgesellschaft.

Seebaß, B. (2001): Kollektive Verantwortung und individuelle Verhaltenskontrolle, in: Wieland, J. (Hg.), Die moralische Verantwortung kollektiver Akteure, Berlin: Physica Verlag, 79–99.

Topol, E. (2019): Deep medicine. how artificial intelligence can make healthcare human again, New York: Basic Books.

Vec, M. (2011): Kurze Geschichte des Technikrechts, in: Schulte, M.; Schröder, R. (Hg.), Handbuch des Technikrecht, Berlin: Springer Verlag, 3–92.

Werner, M.H. (2011): Verantwortung, in: Düwell, M.; Hübenthal, C.; Werner, M.H. (Hg.), Handbuch Ethik, Stuttgart/Weimar: J.B. Metzler, 541–548.

Warum und wozu erklärbare KI?

Über die Verschiedenheit dreier paradigmatischer Zwecksetzungen

Suzana Alpsancar

Abstract: *Currently, explainable AI (XAI) is often touted as a remedy for the so-called black box problem of machine learning without any distinction. However, both the problem of seeing the issue in the blackness, i.e. opacity, of certain AI systems, and the dominant strategy to solve this problem by decreasing the opacity with XAI are misleading in their general applicability. This article calls for a nuanced and reflected investigation into the usefulness of XAI. To do this, it is not only important to name purposes in the first place, but also to make them concrete and take their paradigmatic differences seriously. To demonstrate this dissimilarity, the article adopts a perspective inspired by Th. Kuhn and highlights three paradigms of current XAI research, all of which, despite their fundamental differences, are currently being dealt with as if they represent a sort of ›normal scientific puzzle solving‹ in machine learning.*

Keywords: *black box; AI Ethics; explainable AI (XAI); human centered AI; value alignment*

1. Einleitung: Schwarze Kisten als Problemstellung

›Explainable AI‹ (XAI) ist ein in den letzten Jahren schnell wachsendes Forschungsfeld, welches sich darum bemüht, schwarze KI-Kisten zu lichten und eine Palette von Methoden vorgelegt hat, die verschiedene formale Aspekte der Funktionsweise von KI-Systemen beschreiben, die sonst im Verborgenen blieben (Hu 2020; Vilone/Longo 2021).[1] Das Forschungsfeld und die Debatte um erklärbare KI versteht sich selbst primär als Reaktion auf einen Bedarf, der aus der Forschung und Entwicklung und der gesellschaftlichen Anwendung von KI-Systemen hervorgegangen ist

1 Die Arbeit an diesem Beitrag wurde gefördert von der Deutschen Forschungsgemeinschaft (DFG): TRR 318/1 2021 – 438445824. Viele Gedanken sind im Austausch mit meinen wundervollen Kolleg:innen im TRR 318 entstanden, denen ich an dieser Stelle herzlich danke. – Für die Redaktion des Textes bedanke ich mich bei Amber Sophie Kieffer und Sebastian Mantsch für ihre Unterstützung.

(Capel/Brereton 2023). Dieser Bedarf wird als das Problem verstanden, für welches XAI die, bzw. eine wichtige Lösung ist. Damit fußen Forschung und Debatte auf einer bestimmten Problemkonzeption. Die gängige Fassung dieses Problems ist die Konzeption als ›Black-Box-Problem von Machine Learning‹. Diese Konzeption findet sich zum Beispiel bei Kamath und Liu, die in ihrem Buch *Explainable AI. An Introduction* (Kamath/Liu 2021) die Notwendigkeit des Forschungsfelds XAI mit dem Nachteil von Machine Learning (ML), opak zu sein, begründen:

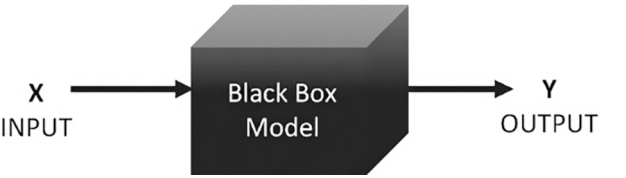
> »The field of explainable AI addresses one of the most significant shortcomings of machine learning and deep learning algorithms today: the interpretability of models.« (Kamath/Liu 2021: ix)

In dieser Problemkonzeption werden der jüngste Erfolg von KI sowie die Opazität ML-Verfahren allein zugewiesen. Die Opazität wird als Nachteil bewertet und als Preis deklariert, den man für den Gewinn von höherer Akkuratheit bezahlen müsse – ML-Systeme bringen »greater accuracy at the expense of complexity and explainability« (Kamath/Liu 2021: 2). Opazität wird so als Kehrseite des Erfolgs angenommen und beides als genuin und rein technische Angelegenheit gerahmt. XAI wird als eine zentrale bzw. die Lösungsstrategie eingeführt mit dem Problem der Undurchsichtigkeit umzugehen, abermals als rein technische Angelegenheit. Die verbreitete Metaphorik von ML-Systemen als Black-Box verstärkt den Eindruck, dass das Grundproblem, für das XAI die Lösung sein soll, allein auf Eigenschaften des technischen Systems zurückzuführen sei:

> »Many machine learning and deep learning models are essentially ›black-boxes‹ that do not reveal the internal mechanisms and nuances to their predictions.« (Kamath/Liu 2021: 2)

Dieser Blackbox-Charakter wird typischerweise als schwarze, geschlossene Kiste visualisiert (siehe Abbildung 1). Diese Problemkonzeption ist in zwei Hinsichten unzulänglich. Zunächst ist die Problemstellung ungenau und pauschal. Die Ursachen der Undurchsichtigkeit liegen nicht ausschließlich in technischen Eigenschaften (Burrell 2016).

Abbildung 1: Darstellung des Black-Box Problems



Quelle: Kamath/Liu 2021: 2

Burrell hat den Blick auf die Kontexte und die soziale, ökonomische und politische Einbettung der Technologien gelenkt und vorgeschlagen, drei verschiedene Typen von Opazität zu unterscheiden (Burrell 2016): Zunächst können Systeme für Dritte oder Außenstehende opak erscheinen, weil sie unter das Geschäftsgeheimnis fallen und damit wichtige Informationen über ihre Funktionsweise, ihren Aufbau, die verwendeten Datensätze u.ä. nicht einsichtig sind. Sodann ist eine Opazität relativ zum Kenntnisstand verschiedener Akteur:innen zu unterscheiden. Systeme können aufgrund einer fehlenden »digital literacy« für bestimmte Gruppen opaker als für andere sein. Von diesen beiden Hinsichten, die die soziotechnische Anwendung der Systeme in den Blick nimmt, unterscheidet Burrell einen dritten Typ, den sie epistemische Opazität nennt und der auf die Eigenschaften des Systems abhebt: sind die Eigenschaften des Systems generell nachzuvollziehen oder treten hier Grenzen und besondere Herausforderungen auf? Liptons Überlegungen (Lipton 2018) setzen bei diesem dritten Typ von Opazität an. In seinem kritischen Kommentar zum Mythos der Model-Interpretierbarkeit stellt er zunächst fest, dass es in der XAI-Debatte zwei grundverschiedene Bedeutungen von Erklärbarkeit/Opazität gibt: nämlich zum einen die *ex ante transparency* und zum anderen die *post hoc interpretability*. Der Unterschied ist ein zeitlicher: geht es um eine Einschätzung des Systems vor seinem Gebrauch, ex ante, und damit generell oder geht es darum rückblickend, von einem gegebenen Ergebnis des Systems ausgehend besser nachvollziehen zu können, wie es zu diesem Ergebnis kam? Lipton zufolge sind die Bemühungen im XAI-Bereich auf den zweiten Typ von Erklärbarkeit gerichtet. Die vielen verschiedenen Tools stellen Mittel dar, Systemergebnisse post hoc verstehbarer zu machen. Darüber hinaus führt er eine wichtige Differenzierung zur Rede von *ex ante transparency* ein, die verdeutlicht, dass es zu einfach ist von einer pauschalen Opazität von bestimmten Systemtypen, z.B. dem Machine Learning zu sprechen und spiegelbildlich von einer pauschalen Transparenz bestimmter Systemtypen, etwa der symbolischen KI. Lipton vertritt ein vergleichbares Anliegen wie Burrell, doch anders als sie hebt er nicht auf die Kontextualisierung der Systeme ab, sondern auf das Zusammenspiel der technischen Komponenten:

> »[…] what constitutes transparency? You might look to the algorithm itself: Will it converge? Does it produce a unique solution? Or you might look to its parameters: Do you understand what each represents? Alternatively, you could consider the model's complexity: Is it simple enough to be examined all at once by a human?« (Lipton 2018: 6)

Lipton unterscheidet drei Level der epistemischen Opazität, die je nach technischem Zusammenspiel mehr oder weniger opak sein können: (a) »algorithmic transparency«, »decomposability«, »simultability«. Die algorithmische Transparenz bezieht sich auf die Struktur der verwendeten Algorithmen. Diese werden häufig in transparente und opake Methoden eingeteilt. Zum Beispiel gelten Entscheidungsbäume, lineare Regressionen oder regelbasierte Systeme als ›white box models‹, während neuronale Netze per se als opak gelten. Diese Einteilung ist aber zu hinterfragen. Den Grund, den Lipton anführt, ist der der Komplexität, denn es macht praktisch einen Unterschied, wie komplex Modelle bzw. Systeme werden. Mit der Komplexität steigt die Undurchsichtigkeit:

> »A lack of transparency and interpretability is arguably less problematic for other ML methodology with a stronger ›white-box‹ character, most notably symbol-oriented approaches such as rules and decision trees. Yet, even for such methods, interpretability is far from being guaranteed, especially because accurate models often require a certain size and complexity. For example, even if a decision tree might be interpretable in principle, a tree with hundreds of nodes will hardly be understandable by anyone.« (Hüllermeier 2020: 206)

Liptons »algorithmic transparency« lässt sich sinnvoll auf Muster-Algorithmen (z.B. in Lehrbüchern) anwenden, die dann in transparente und opake sortiert werden können. Diese Zuschreibung ist für konkretisierte Algorithmen dann zu relativieren. Bei der Nutzung von Algorithmen stellt sich die Frage, wie nachvollziehbar das Zusammenspiel der wichtigen Komponenten des algorithmischen Systems sind. Lassen sich »input, parameter and calculation« (Lipton 2018: 14) für sich und in ihrem Zusammenwirken verstehen? Dann hätten wir eine Transparenz im Sinne der »decomposibility«. Davon unterscheidet Lipton weiter den Aspekt der ›simultability‹. Hiermit bezieht sich Lipton auf die Frage, ob das Systemverhalten im Ganzen und simultan zu verstehen ist:

> »[…], for a model to be fully understood, a human should be able to take the input data together with the parameters of the model and in reasonable time step through every calculation required to produce a prediction.« (Lipton 2018: 13)

Unter dem Gesichtspunkt der »simultability« hat es keinen Sinn, davon zu sprechen, Entscheidungsbäume seien per se transparent, neuronale Netze aber nicht, weil hier Größe, Komplexität, Zusammenspiel und Zeit in Betracht kommen.

Zur ungenauen und pauschalen Problemstellung kommt hinzu, dass die anvisierte Problemlösung (Erklärbarkeit) ebenfalls zu undifferenziert bleibt (Krishnan 2020; Freiesleben/König 2023; Alpsancar et al. 2024). Transparenz ist kein Selbstzweck. Erklärungen sind nicht per se hilfreich, nötig oder nützlich. Die Debatte um erklärbare KI kränkelt an einer pauschalen Inwertsetzung von Erklärbarkeit. Krishnan (Krishnan 2020) hat die Existenz und Relevanz des Black-Box-Problems in Frage gestellt und angeregt, die XAI-Debatte stärker an Fragen der Zwecksetzung von Verstehbarkeit auszurichten. Diesen Vorschlag aufgreifend lässt sich zunächst nach den gängigen Zwecken fragen, denen XAI dienen soll. Da es bisher wenig Erfahrungsberichte aus der Praxis gibt, d.h. konkrete Fallstudien zur Tauglichkeit von XAI, sondern das Feld weitestgehend im Zustand hoher Erwartungen, florierender Forschung und spekulativen Begründungen vibriert, blicke ich dazu in die Forschungsliteratur. Viele Forschungsartikel im Feld der XAI präsentieren hier, wie sie bestimmte Techniken weiterentwickelt haben, stellen neue Ansätze des technischen Erklärens vor oder geben einen Überblick über das Forschungsfeld (Adadi/Berrada 2018; Ras et al. 2018; Meske et al. 2022; Gilpin et al. 2018). Typischerweise findet sich in diesen Artikeln im Einleitungsteil (bzw. den einleitenden Sätzen) die Motivation für diese Forschung, d.h. hier wird aus dem Feld heraus artikuliert, warum und wozu es erklärbare KI gibt, und diese weiterentwickelt werden muss. Die Motivation beginnt i.d.R. mit der generalisierten Problemstellung, d.h. der Undurchsichtigkeit von ML-Systemen. Teilweise wird diese Undurchsichtigkeit als solche als Anlass genug gesehen, gegen sie vorzugehen, ohne das weiter erläutert würde, warum diese problematisch ist:

> »Explainable Artificial Intelligence (XAI) has experienced a significant growth over the last few years. This is due to the widespread application of machine learning, particularly deep learning, that has led to the development of highly accurate models that lack explainability and interpretability.« (Vilone/Longo 2021: 89)

Darüber hinaus werden häufig instrumentelle Gründe angegeben, d.h. die Dienlichkeit von XAI benannt, i.d.R. allerdings in abstrakter und generalisierter Form:

> »This opacity has created the need for XAI architectures that is motivated mainly by three reasons, […] (i) the demand to produce more transparent models; (ii) the need of techniques that enable humans to interact with them; (iii) the requirement of trustworthiness of their inferences. Additionally, […], models induced

> from data must be liable as liability will likely soon become a legal requirement.« (Vilone/Longo 2021: 89)

Ähnlich formulieren es Kamath und Liu, die vier Gründe auflisten: Erstens geht es um den Nutzen, den ML-Entwickler:innen aus XAI ziehen können: »We need interpretability to explain the model's working from both the diagnosis and debugging perspective«. Zweitens geht es darum, die Ergebnisse bzw. das Verhalten von KI-Systemen für Endnutzer:innen verständlich zu machen, damit diese die Systeme überhaupt bzw. besser nutzen können: »We need explanations for the end-user to explain the decisions made by the model and the rationale behind the decisions« (Kamath/Liu 2021: ix). Als drittes Motiv für XAI heben sie das Problem der ›biases/Verzerrungen‹ hervor, die es aufzudecken gelte:

> »Most datasets or models have been shown to have biases, and investigating these biases is imperative for model deployment. Explainability is one way of uncovering these biases in the model.« (Kamath/Liu 2021: ix)

Als vierten Grund stellen sie rechtliche Vorgaben heraus, die aus Sicht von Betreiber:innen und Hersteller:innen ein Compliance-Problem darstellen:

> »Many industries such as finance and healthcare have legal requirements on transparency, trust, explainability, and faithfulness of models, thus making interpretability of models a prerequisite.« (Kamath/Liu 2021: ix)

Rechtliche Regulierungen zu KI-Systemen finden sich zahlreiche (Nannini et al. 2023; Cath et al. 2018; Chakrabarti/Sanyal 2020; Roberts et al. 2021). In der EU wird insbesondere mit Bezug auf den *EU AI Act*, der Ende 2023 verabschiedet wurde und dessen Umsetzung noch ausgearbeitet werden muss (Laux et al. 2023), sowie mit Bezug auf die im Jahr 2018 in Kraft getretene *General Data Protection Right* darüber diskutiert, ob diese Regulationen ein sogenanntes ›right to explanation‹ fordern oder doch nur schwächere Auflagen wie Transparenzpflichten für bestimmte Anwendungen, Risiken oder ein ›right to information‹ (Goodman/Flaxman 2017; M.E. Kaminski 2019; Sovrano et al. 2022; Gyevnar et al. 2023; Panigutti et al. 2023). Es ist noch nicht ausgemacht, ob und für welche Fälle Erklärbarkeit von KI eine notwendige Bedingung für ihre Anwendung darstellt, geschweige denn, was genau unter Erklärbarkeit technisch, sozial und politisch zu verstehen sei (Doshi-Velez/Kim 2017; Gilpin et al. 2018; Ribera/Lapedriza García 2019; Rudin 2019).

Beim Problem der ›biases‹ (Verzerrungen) geht es darum, Diskriminierungsrisiken zu minimieren (vgl. Kolleck/Orwat 2020, für einen Überblick). Man spricht hier von diskriminierender KI (›racist algorithm‹) oder, wenn keine diskriminierenden Verzerrungen vorliegen, von ›ethical aglorithms‹ bzw. ›fair ML‹. Insofern Dis-

kriminierungen aufgrund von geschützten Merkmalen wie Alter, Rasse, Ethnie, Geschlecht, sexuelle Orientierung und Religion rechtlich verboten sind, sind der dritte und der vierte Grund Kamaths und Lius miteinander verbunden.[2] Ich schlage vor, beide unter dem allgemeineren Gesichtspunkt des Wunsches nach einer Vereinbarkeit von KI mit anerkannten Grundwerten und Normen zu fassen. Erklärbare KI wird hier als ein Mittel angesehen, welches (entscheidend) dazu beitragen kann, *KI-Systeme gesellschaftsfähig zu machen*. Hierbei geht es sowohl um rechtliche Normen und ethische Prinzipien als auch Fragen der sozialen Angemessenheit (Lipton 2018; Bellon et al. 2022).

Mit diesen Überlegungen lassen sich die von Kamath und Liu (Kamath/Liu 2021) sowie von Vilone und Longo (Vilone/Longo 2021) genannten Gründe, die sich in der einen oder anderen Formulierung immer wieder in der XAI-Debatte finden (Samek et al. 2017; Ribera/Lapedriza García 2019; Gilpin et al. 2018), zu drei *Hauptmotiven* zusammenfassen, die besagen, warum und wozu erklärbare KI entwickelt wird (Alpsancar et al. 2024):[3]

a. um KI-Systeme zu optimieren,
b. um eine effiziente Nutzung von KI-Systemen zu gewährleisten,
c. um KI-Systeme gesellschaftsfähig zu machen.

In der Zwecksetzung a. artikuliert sich die Perspektive der Entwickler:innen und Forschenden, denen es darum geht, Systeme besser einschätzen, verändern und optimieren zu können, z. B. in der Fehlersuche (›debugging‹). Genealogisch scheint mir dieser Zweck am Ursprung der derzeitigen XAI-Debatte in den Computer Sciences zu liegen, auch wenn historisch gesehen Erklärungen für Informationssysteme schon älteren Datums und aus Anwendungsbezügen erwachsen sind, insbesondere für Expertensysteme im medizinischen Bereich (de Bruijn et al. 2022; Meske et al. 2022). Dem zweiten Hauptmotiv liegt eine anwendungsbezogene Perspektive zugrunde. Es geht es um den gelingenden Gebrauch, der Voraussetzung für die erhofften Effizienzsteigerungen ist. Auf diese Weise kommen Endnutzer:innen und

2 Die Rechtslage divergiert freilich von Staat zu Staat. In der EU hat z. B. der Europäische Rat, zwischen 2000 und 2004, vier Gleichbehandlungsrichtlinien beschlossen, die in Deutschland durch das Allgemeine Gesetz zur Gleichbehandlung (AGG) umgesetzt werden. Eine gute Übersicht zu dieser Debatte bieten Kraus und Ganschow 2022.

3 Freilich lassen sich weitere wichtige Kategorien an Zwecksetzungen finden, etwa die Dienlichkeit von XAI für die Forschung in anderen Bereichen wie der Medizin oder Biologie, wo es darum geht, neue Einsichten und Erkenntnisse zu gewinnen (Markus et al. 2021; Guidotti et al. 2018; Miller 2019). Darüber hinaus könnte man diese Zweck-Kategorien binnendifferenzieren, etwa in ›XAI for contestability‹ oder ›subjects understanding‹ (Mittelstadt et al. 2019). Beides soll hier aber außen vor bleiben.

teilweise auch Anwendungskontexte in den Blick. Hierin artikuliert sich ein ökonomischer oder auch militärisch motivierter Wille. Es geht um Effizienz, Akzeptanz und ›Usability‹. Das dritte Hauptmotiv evoziert eine kollektive Perspektive. Es geht um die Frage nach den Regeln für den angemessenen Gebrauch und Einsatz von KI in diversen gesellschaftlichen Bereichen. Konkret betrifft dies die Sprechposition derjenigen, die KI regulieren wollen oder sollen bzw. diese Vorgaben oder Überlegungen umsetzen müssen oder wollen. Auch schließt hier eine gesellschaftliche Perspektive an, und zwar als Frage, welcher Einsatz und Gebrauch der Technologien eigentlich wünschenswert wäre. Die Frage, ob man diese Technologien überhaupt braucht oder haben möchte, kommt nicht vor: Die Technik ist da, nun muss sie eingehegt werden.

Mit den folgenden Überlegungen möchte ich zu einer kritischen Auseinandersetzung mit diesem motivationalen Horizont von XAI anregen. Hierzu werde ich an exemplarischen Fällen drei Problemlagen herausstellen, aus denen sich die drei genannten Hauptmotive nähren. Sie bilden meiner Ansicht nach den Grund von drei Paradigmen, die den Sinnhorizont der Forschung und Entwicklung von XAI stiften und damit formen, was im XAI-Bereich passiert: das Paradigma der epistemischen Güte von ML, das Paradigma der effizienten Handhabbarkeit von KI-Anwendungen sowie das Paradigma der Vereinbarkeit von KI mit anerkannten Grundwerten und Normen. Meine These ist, erstens haben wir es mit *drei verschiedenen Paradigmen* zu tun und zweitens wurde diese Verschiedenheit von der XAI-Forschung bislang zu wenig beachtet.

Lose angelehnt an Kuhn (Kuhn 1976), prägen Paradigmen in einer Forschungsgemeinschaft das, was als ein typisches und damit relevantes Forschungsproblem ist. Es gibt vor, worin zentrale Forschungsaufgaben liegen und wie man sich an deren Lösung machen sollte. Paradigmen haben in ihrer Vorbildfunktion einen normativen Charakter. Kuhn beschreibt sie als »allgemein anerkannte wissenschaftliche Leistungen, die für eine gewisse Zeit einer Gemeinschaft von Fachleuten maßgebende Probleme und Lösungen liefern« (Kuhn 1976: 10). So wie in einer Forschungsgemeinschaft ein Problem x durch das Lösungsverfahren z gelöst wurde, so geht man nun an neue Probleme in dem Forschungsgebiet heran. Das heißt, was als Forschungsproblem erkannt und wie es konzipiert wird, worin seine Herausforderung liegt, ist maßgeblich von den geltenden Paradigmen einer Forschungsgemeinschaft geprägt. Das Erkennen und Konzipieren von relevanten Forschungsproblemen gehen mit den Erwartungen einher, diese prinzipiell lösen zu können. Die Lösungsvisionen sind mehr als ein erstes Erkunden, sie sind zielgerichtetes Entwickeln und Testen, d.h. es existieren bestimmte strategische Vorstellungen darüber, wie die Forschungsprobleme zu meistern seien. Mit diesen formierenden Erwartungen und Zuschreibungen bündeln Paradigmen nicht nur das begriffliche und apparative Instrumentarium (Kuhn 1976: 41), sondern ebenso entsprechende Akteur:innen, Ressourcen und Infrastrukturen um sich

(Borup et al. 2006). Ich verstehe hier Paradigmen darüber hinaus als sinnstiftend und formgebend, in dem sie mit typischen Zwecken korrelieren, denen sich eine Forschungsgemeinschaft sinnvollerweise widmen kann. Bei Kuhn stand die Frage nach den Zwecken und Motiven der Forschung weniger im Vordergrund, da er seine Überlungen an der Geschichte der modernen Physik entwickelte, für die er verschiedene Phasen und damit Paradigmen der Forschung ausmachen konnte, die sich aber weitestgehend einem gemeinsamen obersten Zweck verschrieben hatten – die Natur zu erkennen. Im Fall von XAI haben wir es dagegen mit verschiedenen Zwecken zu tun. Meinem Eindruck nach wird dieser Verschiedenheit bloß diskursiv Rechnung getragen, während die Forschungspraxis weitgehend im Modus des normalwissenschaftlichen Rätsellösens der ML-Community verweilt. Mit Kuhn gesprochen, isoliert sich die XAI-Community in dieser Weise von solchen Problemen, »die sich nicht auf die Rätselform reduzieren lassen« (Kuhn 1976: 51), d.h. sowohl Fragestellungen als auch Lösungswege, Methoden, Werkzeuge, theoretische Annahmen, die außerhalb des eigenen Paradigmas liegen, werden abgelehnt, als nicht wichtig erachtet oder für zu schwierig gehalten.

2. Das Paradigma der epistemischen Güte von ML

Gut laufende Maschinen sind nicht erklärungsbedürftig. Aus einer Ingenieurs-Perspektive werden Sie es dann, wenn man Sie optimieren möchte oder wenn ein Fehler auftritt, den man beheben will. Anlass vieler Diskussionen ist dies das Problem des »overfitting«. Dieses bezeichnet die Überanpassung eines ML-Modells zu den Trainingsdaten im Vergleich zu anderen Datensätzen. Liegt eine Überanpassung vor, kann aus einer hohen Treffsicherheit auf den Trainingsdaten nicht auf eine hohe Treffsicherheit für andere Datensätze geschlossen werden. Damit ist die gute Performanz nicht verallgemeinerbar, weil die Maschine nicht tatsächliche Muster, sondern willkürliche gelernt hat. Dieser Effekt lässt sich ebenfalls als Clever-Hans-Effekt bezeichnen (Hernández-Orallo 2019; Lapuschkin et al. 2019), womit Beziehungen gemeint sind, die man fälschlicherweise als Kausalitäten einschätzt, die in Wirklichkeit bloße Korrelationen darstellen. Einen solchen Effekt überhaupt feststellen zu können setzt eine sogenannte ›ground truth‹ voraus, d.h. ein Wissen darüber, welche Muster den Tatsachen entsprechen und welche nicht (Freiesleben/König 2023).[4]

Als Beispiel für dieses Paradigma greife ich den häufig zitierten Artikel von Lapuschkin et al. auf, in dem die Autoren Strategien vorstellen, um ML-Modelle auf

4 Das Festlegen dieser Grundwahrheit ist nicht immer möglich und auch nicht immer gefragt, z.B. wenn es darum geht, neue Zusammenhänge und Muster zu erkennen. Sie praktisch festzulegen ist aufwendig, da die Datensätze entsprechend annotiert werden müssen.

Clever-Hans-Effekte zu prüfen (Lapuschkin et al. 2019). Der Name ›Clever-Hans-Effekt‹ geht auf die Aufdeckung eines Versuchsleiter-Erwartungs-Effekts in der Psychologie des frühen 20. Jhs. zurück. Der Mathematiklehrer Wilhelm von Osten hatte ein Zirkuspferd namens Hans darauf dressiert, richtige Antworten auf beispielsweise einfache Rechenaufgaben zu geben (etwa durch eine Anzahl von Hufschlägen), was er einem staunenden Publikum vorführte. Doch Hans konnte nicht rechnen. Eine Prüfung zeigte, dass das Pferd sein Antwortverhalten an der Körpersprache und anderen Signalen seines Besitzers orientierte. Die richtigen Ergebnisse basierten somit nicht auf einer korrekten Durchführung der Aufgabenstellung, sondern hatten kontingente Ursachen. Im Bereich der KI-Forschung spricht man von einem Kluger-Hans-Effekt:

> »[…] bzw. Clever-Hans-Effekt, wenn in einem Trainingsdatensatz, möglicherweise in versteckter Form, bestimmte Eingangsgrößen vorhanden sind, die mit der richtigen Ausgabe korrelieren, aber wenig mit der Ursache der jeweils adressierten Phänomene zu tun haben.« (Kraus/Ganschow 2022: 39)

Lapuschkin et al. haben einen solchen Effekt bei ML-Modellen nachgewiesen, die beim PASCAL VOC (Visual Object Classification) Wettbewerb gewonnen hatten (Lapuschkin et al. 2019). Es ist nicht trivial, Maschinen so zu konstruieren, dass sie ein Objekt klassifizieren können (Tisch, Stuhl, Tier, Person). Während Menschen dies beiläufig tun (Kaminski 2014), ist eine Objektklassifikation für Maschinen eine solch ausgezeichnete Leistung, dass es in der Informatik üblich ist, hierfür Wettbewerbe zu veranstalten. Dass man im Bereich der ›Computer Vision‹ überhaupt zu vertretbaren Ergebnissen gekommen ist, liegt hauptsächlich an der Vergrößerung der Trainingsdatensätze. In diesem Bereich lässt sich die Performanz-Steigerung der Maschinen aufgrund der gewachsenen Datensätze gut nachvollziehen: In den 1960er Jahren standen Forschenden typischerweise ein Dutzend Bilder zur Verfügung, mit denen sie ihre Maschinen trainieren konnten. In den 1990er Jahren waren es bereits Sätze mit Tausenden von Bildern. Mit Beginn der 2020er Jahre bewegt man sich üblicherweise im Millionenbereich (Crawford 2023: 1368).

Hinter den von 2005/06 bis 2012 organisierten PASCAL VOC Wettbewerben steht das von der EU geförderte Exzellenznetzwerk PASCAL[5] (Everingham et al. 2010; Everingham et al. 2015). Ich werde zunächst die Logik des PASCAL Projektes und der Wettbewerbe beschreiben, um die Dienlichkeit von XAI im Bereich der Bilderkennung zu verorten. Das Exzellenznetzwerk hatte der eigenen Forschungsgemeinschaft zunächst eine Infrastruktur für die Erprobung von ML-Techniken im Bereich der Bilderkennung bereitgestellt. Kern der Infrastruktur war ein frei zugänglicher Datensatz aus Bildern, die annotiert waren. Über eine dazugehörige

5 PASCAL steht für pattern analysis, statistical modelling und computational learning.

Software stellte die Forschergruppe außerdem Standards für die Evaluierung von Algorithmen bereit, die die Bilder aus dem Datensatz für das Lösen bestimmter Aufgaben im Bereich der Bilderkennung nutzen sollten (Everingham et al. 2010). Wichtig war, dass die Bilder annotiert und damit mit einer ›ground truth‹ versehen waren, d.h. es war bekannt und notiert, was auf den Bildern zu sehen ist (Everingham et al. 2015: 98f.). Seit 2006 wurde jährlich ein neuer Datensatz mit annotierter Grundwahrheit bereitgestellt.

Everingham et al. (Everingham et al. 2010) beschreiben, wie sie das Datenset für den Wettbewerb im Jahr 2007 erstellt und kuratiert haben. Welche Bilder aufgenommen und wie diese annotiert werden, beeinflusst auf entscheidende Weise, wie gut die gestellten Forschungsfragen beantwortet werden und die Antwortlösungen verglichen werden können. Die Güte der generierten Datensätze richtete sich nach der formulierten Absicht des Projektes (im Vergleich zu anderen verfügbaren Bild-Datensätzen). Ziel von PASCAL war es, ML-Modelle für ein möglichst breites Spektrum ›natürlicher Bilder‹ trainieren und testen zu können (Everingham et al. 2010: 305), was damals eine neue Herausforderung für die Forschungscommunity war. 2006 hat die Forschergruppe Bilder aus der Microsoft Research Cambridge Datenbank verwendet, die allerdings zu gestellt waren, um der Zielstellung ›natürliche Bilder‹ gerecht zu werden.[6] Seit 2007 haben die Initiatoren deswegen ihren Datensatz über Flickr generiert, wodurch eine Ausgangsbasis an Bildern gegeben war, die nicht alle gleichermaßen unter einer bestimmten Absicht erstellt und hinterlegt sind.

Die Annotation der gewonnenen Bilder geht von folgenden Grundsätzen aus: Jedes Bild wird einer der vorab festgelegten Objektklassen zugeordnet (für 2007 waren diese: »aeroplane, bird, bicycle, boat, bottle, bus, car, cat, chair, cow, dining table, dog, horse, motorbike, person, potted plant, sheep, sofa, train, tv/monitor«; Everingham et al. 2010: 308). Hinzu kommt ein Begrenzungsrahmen (»bounding box«), die den Umfang des Objektes sichtbar eingrenzt (Everingham et al. 2010: 308). Seit 2006 wurden weitere Annotationen eingeführt, die für das Training genutzt werden durften (aber für die Evaluierung nicht notwendig waren): »viewpoint«, »truncation« (wenn der Begrenzungsrahmen nicht das ganze Objekt erfasst, weil auf dem Bild z.B. nur ein Teil einer Person sichtbar ist etc.). Ab 2008 wurde auch ggf. »occluded« angegeben, wenn das Objekt (in Teilen) verdeckt sichtbar ist. Zudem wurden bestimmte Bilder als »difficult« markiert, wenn ihre Eigenschaften es im Vergleich zu anderen schwieriger für Algorithmen machen, Objekte zu erkennen (Sind alle Tiere auf dem Bild Kühe oder ist eines ein Schaf?). Diese Annotation sollte konsistent, akkurat und vollständig sein, weswegen die Annotator:innen geschult und der

6 Der Microsoft Research Cambridge-Datensatz war mit der Absicht gebaut worden, bestimmte Objektklassen möglichst gut abzubilden. Entsprechend zeigen die Bilder i.d.R bestimmte Objekttypen in zentrierter, ausgeleuchteter, zentraler Position. Sie haben damit wenig Varianz für die Zwecke des PASCAL Projektes.

Prozess des Annotierens überwacht und geprüft wurde. Annotieren ist ein sehr aufwendiger und je nachdem, von wem man diese Arbeit durchführen lässt, auch kostspieliger Vorgang.[7] Die Forschungsgruppe wollte die Annotationsarbeit zunächst vollständig auslagern, über die Amazon-Plattform »Mechanical Turk«, über die Arbeitskräfte günstig für kleinteilige Arbeit der digitalen Ökonomie zugänglich gemacht werden.[8] Für die hohen Ansprüche an die Annotation reichte es aber nicht aus (Konsistenz, Akkuratheit, Vollständigkeit), so dass nur Label für die Frage eingekauft wurden, ob eine Objektklasse zu sehen ist oder nicht (presence/absence), wodurch dennoch erheblich Zeit und Kosten gespart wurden (Everingham et al. 2015: 101).

Zu dem Projekt gehörte, neben der Bereitstellung der Forschungsinfrastruktur, ein jährlicher Wettbewerb, der PASCAL VOC Wettbewerb. Diese Ergebnisse wurden auf einem jährlichen Workshop zur Diskussion gestellt. Um am Wettbewerb teilnehmen zu können, mussten die Forscher:innen Algorithmen entwickeln, die zu Beginn (2005/06–2008) mindestens eine der zwei folgenden »principle challenges« lösen sollten: (1) Klassifizierung: »For each of twenty object classes, predict the presence/absence of at least one object of that class in a test image«;[9] (2) Detektion: »For each of the twenty classes, predict the bounding boxes of each object of that class in a test image (if any), with associated real-valued confidence« (Everingham et al. 2010: 304). Diese Kernfragen wurden um zwei zusätzliche Fragen ergänzt, die man aber nicht angehen musste, um am Wettbewerb teilzunehmen: (a) Segmentierungs-Probe: »For each test image, predict the object class of each pixel«, und (b) Personen-Gliederungs-Probe: »For each ›person‹ object in a test image (if any), detect the person, predicting the bounding box of the person, the presence/absence of parts (heads/hands/feet), and the bounding boxes of those parts« (Everingham et al. 2010: 305).[10] Der Wettbewerb bestand aus zwei Phasen: einer Trainings- und einer Testphase. Für das Training konnten die annotierten Datensets verwendet werden, für das Testen wurden un-annotierte Bilder herausgegeben (also Bilder, die eigentlich annotiert waren, aber deren Annotation absichtlich für die Wettbewerber entfernt

7 »For example, annotation of the VOC2008 dataset required around 700 person hours.« (Everingham et al. 2010: 336)

8 Typische Arbeiten sind: transkribieren, Objekte klassifizieren, Rückmeldungen zu Webseiten geben, Onlineinhalte umschreiben (Ipeirotis 2010). Die über die Plattformlogik forcierten Arbeitsbedingungen stehen häufig in der Kritik (vgl. exemplarisch Ellmer 2015).

9 Diese Klassifizierungsaufgabe läuft darauf hinaus, die Pixel richtig zu zuordnen: »to which class does each pixel belong?« (Everingham et al. 2015: 99).

10 Ab 2009 bis 2012 wurde die Segmentierungsaufgabe zur dritten Kernfrage gemacht. Der Test, wie gut Algorithmen bestimmte Körperteile und den Aufbau von menschlichen Körpern erkennen können (bounding boxes), blieb Zusatzaufgabe. Ab 2010 kam die Klassifizierung von Handlungen als weitere Zusatzaufgabe hinzu (Everingham et al. 2015: 100).

wurde), so dass Training und Test auf (logisch) verschiedenen Datensätzen erfolgten. Die Annotationen des Test-Datensets wurden absichtlich bis zum Abschluss des Wettbewerbs nicht publik gemacht, um das Phänomen des ›overfitting‹ vorzubeugen.[11]

Die Aufgaben dieser Wettbewerbe lassen sich als leitende Forschungsfragen der Forschungscommunity zu dieser Zeit verstehen. Die Bereitstellung der »ground truth« über die Annotation stellt sicher, dass man für diese Fragen die richtigen Antworten kennt. Entsprechend ließ sich nicht nur die performative Güte jedes eingereichten Algorithmus bewerten, sondern diese waren untereinander vergleichbar. Hiermit wurde einschätzbar, welche algorithmischen Methoden für welche Aufgaben zielführend sind. Die Ergebnisse aus der Serie von Wettbewerben und Workshops stellen zusammen eine wichtige ›Benchmark‹ für den Bereich der Computer Vision dar (Lapuschkin et al. 2019; Samek/Müller 2019; Hernández-Orallo 2019).

Das Problem des Clever-Hans-Effekts betrifft die Güte von ML-Modellen und damit zunächst die Frage ihrer epistemischen Evaluation. Wie zuverlässig sind die Aussagen, die man mit dem ML-Modell gewinnen kann? Zur Evaluierung von ML-Systemen gibt es verschiedene Gütekriterien bzw. statistische Evaluationsmetriken. Wichtig ist, dass sich die Güte der Systeme auf das Zusammenspiel des trainierten ML-Modells und den gegebenen Daten bezieht. Anders als bei konventionellen Algorithmen ist die Güte dieser algorithmischen Systeme deswegen entscheidend von den Daten abhängig. Hierbei sind Trainingsdaten, Validierungsdaten und der Testdatensatz zu unterscheiden.

Um die Rolle von XAI für die Evaluation der epistemischen Güte von ML-Systemen besser einschätzen zu können, vergleiche ich diese mit einem simplen Maß für die Fehlerquote, der *accuracy* (ACC).[12] Bei einer Klassifikationsaufgabe gibt das Maß der Akkuratheit (ACC) an, wie viele Bilder das System richtig klassifiziert hat.

$$ACC = \frac{Anzahl der korrekt klassifizierten Bilder}{Gesamtzahl der zu klassifizierenden Bilder}$$

Eine Akkuratheit von 100% sagt aus, dass das System jedes Bild im vorliegenden Datensatz richtig klassifiziert hat. Dies wäre die bestmögliche *Performanz*. Bei einer simplen Klassifikationsaufgabe mit nur zwei möglichen Antworten (Ist auf dem

11 Die Initiatoren gehen davon aus, durch den geringen Zeitraum für das Testen zu unterbinden, dass die Teilnehmenden ihre Testdaten händisch annotieren.

12 »Accuracy plays a role especially for data whose factual correctness can be conclusively determined and whose meaning is not ambivalent.« (Mohammed et al. 2024: 2)

Bild ein Pferd zu sehen – ja oder nein?), stellt eine Akkuratheit von 50% eine schlechte Performanz dar, da dies einer bloß zufälligen Zuordnung gleichkommt. Um die Akkuratheit eines Systems mit dieser Formel berechnen zu können, muss man wissen, welche Antworten für jedes gegebene Bild die richtige ist (dies war im PASCAL VOC Wettbewerb der Fall, auch wenn die Klassifikation hier komplexer war). Außerdem muss man die jeweiligen Ergebnisse des zu evaluierenden Systems kennen. Für die Berechnung der Akkuratheit ist demnach keine Einsicht in die Black-Box nötig. Die Art und Weise, wie das Ergebnis gewonnen wurde spielt keine Rolle. Auch ein Zirkuspferd wie Hans könnte demnach eine hohe *accuracy* aufweisen ohne Rechnen zu können. Im Gegensatz zu Gütekriterien für die Fehleranfälligkeit nutzt man erklärbare KI, um einen Einblick in die Frage zu erhalten, wie ein System zu seinem (richtigen) Ergebnis gekommen ist. Dies ist deswegen relevant, weil eine gute Performanz aus falschen Ursachen auf $Datensatz_1$ zu einer schlechten Performanz auf einem $Datensatz_2$ führen kann. Es geht um die Generalisierbarkeit der Nutzung der gewonnenen ML-Modelle über die Trainingsdaten (und ggf. Validierungsdaten) hinaus.

Lapuschkin et al. untersuchten zwei der erfolgreich am Wettbewerb teilgenommen Modelle für die Kategorie Objektklassifikationen (Lapuschkin et al. 2019): das Fisher Vectors (FV) Modell und ein vortrainiertes Deep Neuronal Network (DNN), welches die Autorengruppe für ihre Zwecke auf dem PASCAL VOC Datensatz von 2007 ›feingetuned‹ haben. Beide Modelle bewiesen »excellent state-of-the-art test set accuracy on categories, such as ›person‹, ›train‹, ›car‹, or ›horse‹ of this benchmark« auf (Lapuschkin et al. 2019: 4). Um eine Einsicht darin zu erhalten, wie die Modelle zu diesen guten Leistungen gekommen sind, setzten die Autoren die Methode der »Layer-wise relevance propagation« ein, in der sogenannte Heatmaps visualisieren, welche Zonen eines Bildes (Pixel) besonders einflussreich für die Klassifizierung des Modells sind (›Entscheidung‹). Die Heatmap für das DNN-Modell hebt die Zone hervor, auf der tatsächlich jeweils das Pferd auf den Bildern zu sehen ist. Dies spricht dafür, dass dieses Modell die Pferde aufgrund der richtigen ›Ursachen‹ korrekt klassifiziert. Die Heatmap des FV-Modells hingegen hebt eine Zone unten links auf den Bildern hervor, auf dem nicht die Pferde, sondern ein »source tag« zu sehen ist.[13] Aufgrund dieses Befundes sind die Forschenden die Pferdebilder ›händisch‹, d.h. mit menschlichen Augen durchgegangen und konnten bestätigen, dass sich diese Herkunftsangabe auf allen Pferdebildern in dem Datensatz befindet. Hiernach stellten sie den Verdacht auf, »the FV model has ›overfitted‹ the PASCAL VOC dataset by relying mainly on the easily identifiable source tag, which incidentally correlates with the true features, a clear case of ›Clever Hans‹ behavior« (Lapuschkin

13 Es ist ein schöner historischer Zufall, dass der Clever-Hans-Effekt hier bei der Klassifikation von Pferdebildern auftrat.

et al. 2019: 4). Um ihren Verdacht zu erhärten, haben die Autoren die Herkunftsangabe künstlich aus den Pferde-Bildern entfernt. Während das DNN-Modell gleich performte wie zuvor, büßte das FV-Modell signifikant an Performanz ein. Mehr noch, fügt man die Herkunftsangabe künstlich zu Auto-Bildern hinzu, klassifiziert das FV-Modell diese dann als Pferde: »a clearly invalid decision« (Lapuschkin et al. 2019: 4). Es liegt nahe, dass das trainierte FV-Modell mit hoher Wahrscheinlichkeit alle Bilder als Pferde klassifiziert, die das gleiche »resource tag« aufweisen. Mit diesem Beispiel geben Lapuschkin et al. zu bedenken, sich nicht ausschließlich auf gute Performanzergbnisse zu verlassen, sondern weitere epistemische Gütekriterien zur Evaluation von ML-Modellen heranzuziehen. Es muss die Frage gestellt werden, wie valide eine hohe Performanz eines Systems auf einem Datensatz$_x$ mit Blick auf andere Datensätze ist. Neben den Pferdebildern zirkulieren weitere eingängige Beispiele für Clever-Hans-Effekte in der Forschungsdiskussion zu XAI, etwa Huskies, die von Wölfen aufgrund von Schnee im Hintergrund der Bilder klassifiziert wurden (Ribeiro et al. 2016), oder auch »boats by the presence of water and trains by the presence of rails in the image« (Samek/Müller 2019; Cremers et al. 2019). Die Klassifikationsleistung aufgrund kontingenter Merkmale wird als Fehleinschätzung eingestuft, weil das korrekte Ergebnis aus falschen Gründen erbracht wird und damit die Klassifikationsleistung nicht auf andere Datensätze übertragbar ist. Wichtig ist hierbei, dass die richtige Klassifizierung in diesen Fällen dadurch klar wird, dass sich die korrekte Zuordnung an unserer (menschlichen) Klassifikation der Bilder orientieren kann. Für uns ist es i.d.R. völlig unproblematisch Wölfe von Huskies usw. zu unterscheiden. Komplizierter wird die Frage der epistemischen Güte von Klassifikationsleistungen, wenn nicht bekannt ist, welche Zuordnung die richtige ist bzw. welche Klassen es überhaupt gibt (Caruana et al. 2015).

Die Übertragbarkeit auf andere Datensätze wirft die Frage auf, wie verlässlich ML-Systeme in der realen Welt eingesetzt werden können (Kraus/Ganschow 2022: 39). Genauer sind zwei Fragen nach der Repräsentativität der Daten zu stellen, zum einen, wie angemessen repräsentieren Trainings- und Validierungsdatensätze die eigentlichen Daten der angedachten Anwendungsfelder, und zum anderen, wie gut repräsentieren die Daten die Wirklichkeit, für die sie stehen sollen. In diesen einfachen Fällen der Bildklassifikation lässt sich mit XAI-Methoden die Einsicht gewinnen, ob die ML-Modelle aus den ›richtigen Ursachen‹ gut performen oder nicht. Sie fügen sich in dieser Perspektive in den Werkzeugkasten ein, der dazu dient, die epistemische Güte von KI-Systemen zu bewerten. Als ein solcher Toolkasten ist XAI von Expert:innen für Expert:innen gemacht. Diese Bemühungen verlaufen im Rahmen dessen, was sich mit Kuhn (Kuhn 1976) als Normalwissenschaft bezeichnen lässt: Es gibt innerhalb der Community eine konsensuelle Problemstellung (Einsicht in die schwarzen Kisten zu erlangen, um deren epistemische Güte besser einschätzen zu können). Mir scheint, der größte Teil der XAI-Forschung arbeitet in diesem Paradigma und löst Rätsel dieses epistemischen Typs. Soll XAI anderen Zwecken dienen,

sollte man den modus operandi wechseln, etwa indem man sich ernsthaft realen Anwendungskontexten und damit diversen Nutzer:innen zuwendet.

3. Das Paradigma der effizienten Nutzung von KI-Systemen

Der zweite Zweck, der die XAI-Forschung (diskursiv) entscheidend motiviert, ist die Herausforderung, KI-Systeme in der Anwendung effizient handhaben zu können. Mit dem Beispiel des vom US-Verteidigungsministerium geförderten und über die »Defense Advanced Research Projects Agency« (DARPA) aufgesetzten Forschungsprojekts *XAI* möchte ich demonstrieren, dass diese Zwecksetzung einen Paradigmenwechsel in der Erforschung und Entwicklung von XAI erfordert.

Das DARPA-Projekt stellt eine wichtige Referenz in der Forschungsgemeinschaft dar und hat die Bezeichnung »XAI« maßgeblich verbreitet. Es kann als eine der ersten großflächigen Initiativen einer Regierungsorganisation angesehen werden, das Feld zu rahmen und voranzutreiben (Nannini et al. 2023: 1202; Hoffman et al. 2023). In der Retrospektive der leitenden Köpfe stellt es einen Meilenstein in der XAI-Forschung dar: »The program certainly acted as a catalyst to stimulate XAI research (both inside and outside the program)« (Gunning et al. 2021: 8f.). Das DARPA-Projekt geht über die Motivation von erklärbarer KI durch epistemische Fragestellungen hinaus, da explizit die Perspektive der ›End-User‹ adressiert wurde:

> »The stated goal of explainable artificial intelligence (XAI) was to create a suite of new or modified machine learning techniques that produce explainable models that, when combined with effective explanation techniques, enable end users to understand, appropriately trust, and effectively manage the emerging generation of AI systems.« (Gunning et al. 2021: 1)

Erklärbare KI sollte nicht nur von Expert:innen für Expert:innen entwickelt werden, sondern auch für Primäruser:innen. Das DARPA-Projekt stellt somit einen Perspektivwechsel in der Zwecksetzung von XAI dar, welches die Aufmerksamkeit der Forschung umlenkt, jedenfalls rhetorisch: Es geht um den realen Einsatz von KI-Systemen in diversen Anwendungsfeldern. Man verlässt den geschützten Rahmen der informatischen Laborforschung und Wettbewerbe. Mit den Usern kommt die *Praxis* ins Blickfeld, oder anders gesagt: Man wendet sich rhetorisch zugleich den Nutzer:innen und der Praxis zu.

An dem Projekt, das über vier Jahre lief (2017–2021), waren zwölf Forschergruppen verschiedener US-Amerikanischer Hochschulen und Forschungseinrichtungen beteiligt. Elf dieser Forschungsgruppen hatten einen informatischen Hintergrund (insbesondere aus den Communities des Machine Learning und der Human-

Computer-Interaction (HCI)). Hinzu kam ein Team, das sich der »psychology of explanation« (DARPA 2016: 15) widmen sollte. Für das Gebiet der HCI ist die Hinwendung zu Enduser:innen keineswegs neu (Biran/Cotton 2017); tatsächlich hat sich dieser Bereich der Computer Science herausgebildet, um Computersysteme für Nicht-Expert:innen nutzbar zu machen, etwa im Zuge der Entwicklung und Verbreitung von Minicomputern, dann Personal Computern über die Gestaltung von leichten handhabbaren Schnittstellen wie dem Graphical User Interface, Tastatur und Maus (Dix 2017; Petrick 2020; Harrison et al. 2007). Interessant ist aber, dass sich eine Diskrepanz zwischen der ›informatischen‹ Projektkonzeption und der Perspektive des Teams von Robert R. Hoffman, das für die »psychology of explanation« zuständig war, zeigen lässt. Die ›informatischen‹ Projektkonzeption ziehe ich maßgeblich aus der Ausschreibung des Verteidigungsministeriums zu dem Projekt des *Broad Agency Announcement – Explainable Artificial Intelligence (XAI) – DARPA-BAA-16-53*, vom 10.08.2016 (DARPA 2016). Die DARPA hatte David Gunning als Programmmanager berufen und ihn später zum Direktor des Projektes gemacht, der die wissenschaftliche Ausrichtung verantwortet und Experte in der KI-Forschung ist. Die Ausschreibung wird durch Folien von Gunning flankiert (*Distribution Statement »A«*), mit denen er vermutlich bei interessierten Partnern aus der Industrie und Forschung das Programm des Projektes vorgestellt hat (Gunning 2016). Hinzu kommt eine Publikation zum Ende der ersten Phase (Gunning/Aha 2019), in der Zwischenergebnisse berichtet werden, sowie die Retrospektive nach Abschluss des Projektes (Gunning et al. 2021). Die Beteiligung aus dem HCI-Bereich wird hier in einer bestimmten Weise, nämlich additiv, integriert, während die Gruppe um Hoffman dezidiert eine Sonderrolle zukommt.

Ich interpretiere die Diskrepanz zwischen Hoffmans Team und der informatischen Projektkonzeption so, dass letztere zwar rhetorisch eine Hinwendung zu realen Anwendungskontexten und Enduser:innen formuliert, praktisch jedoch im Modus der informatischen Normalwissenschaft operierte.[14] Hoffmans Team hingegen war vermutlich in einer ambivalenten Rolle: einerseits galt es die zugewiesenen Aufgaben zu erfüllen, andererseits hätte dieses Team die Hinwendung zur Praxis sicherlich anders aufgezogen. Die Differenz der Perspektiven von Gunning und Hoffman dient mir als Demonstration dazu, dass es sinnvoll ist, die Hinwendung zu User:innen und zur Praxis als einen Paradigmenwechsel in der XAI-Forschung zu verstehen, mit dem die Problemstellung, für die XAI eine gute Lösung sein will, neu zu fassen ist.

14 Tatsächlich kommentiere ich nicht die Forschungspraxis oder -ergebnisse, sondern die Konzeption des XAI-Projektes. Das heißt, ich Vergleiche die informatische Selbstbeschreibung von Gunning und wechselnden Ko-Autor:innen mit der von Hoffmans Team.

3.1 Informatische Projektkonzeption

Ich stelle hier die Motivation für das Projekt, die gestellten Forschungsfragen und daran die gebundene Konzeption des Untersuchungsgegenstands heraus sowie die anvisierte Zusammenarbeit bzw. Arbeitsteilung der Forschungsgruppen.

3.1.1 Motivation

Warum setzte die DARPA ein Projekt zu erklärbarer KI auf und warum im Jahr 2016? Das Kernmotiv, welches in den genannten Quellen zu finden ist, ist zweiseitig: Die eine Seite besteht aus einer zweifachen Charakterisierung der ML-Technik, zunächst deren jüngster Erfolg: »Dramatic success in machine learning has led to an explosion of new AI capabilites« (DARPA 2016: 5). Dieser Erfolg wird allein auf die Technologie bezogen (ML); kein Wort über die entscheidenden Randbedingungen (Big Data, Rechenkapazitäten, Verfügbarkeit von Services und Tools, Einbettung und Handhabe der Systeme). Zudem wird, wie in dem eingangs zitierten Lehrbuch zu XAI, das Forschungsgebiet der erklärbaren KI allein aus einem Manko der jüngeren Technologie hergeleitet: »These systems offer tremendous benefits, but their effectiveness will be limited by the machine's inability to explain its decisions and actions to human users« (DARPA 2016: 5). Die informatische Perspektive fokussiert einseitig auf das technische System.

Die andere Seite des Kernmotivs besteht in der Zielstellung der *effektiven Nutzbarkeit* von KI durch Laien:

> »Defense Advanced Research Projects Agency (DARPA) formulated the explainable artificial intelligence (XAI) program in 2015 with the goal to enable end users to better understand, trust, and effectively manage artificially intelligent systems.« (Gunning et al. 2021: 1)

Die Rede davon, dass Nutzer:innen den neuen KI-Systemen »appropriately trust, and effectively manage« (Gunning/Aha 2019: 44) können sollen, findet sich immer wieder in den Quellen zum Projekt (Gunning et al. 2019) und kann als übergeordnete Zielstellung aufgefasst werden. Damit ist die Hinwendung zu Usern/der Praxis prominent platziert.

Warum aber gerade das Verteidigungsministerium hohe Summen in die XAI-Forschung investieren wollte, wird nicht weiter erläutert. Bekannterweise ist die DARPA und das US-Militär generell eine der kontinuierlichen Förderer der Computertechnologien in den USA (Norberg 1996; Edwards 1997; Mahoney 2011), aber welches Anliegen konkret durch die Forschung zur erklärbaren KI befriedigt werden soll, bleibt äußerst vage:

> »The issue [XAI zu entwickeln – SA] is especially important for the Department of Defense (DoD), which is facing challenges that demand the development for more intelligent, autonomous, and symbiotic systems. Explainable AI will be essential if users are to understand, appropriately trust, and effectively manage this incoming generation of artificial intelligent partners.« (DARPA 2016: 5)

Die Ausschreibung artikuliert einen allgemeinen Bedarf an technischem Fortschritt, der an Erklärbarkeit gekoppelt wird ohne über konkrete Nutzungsweisen oder Gründe zu sprechen. Problematisch ist hier nicht die mangelnde Zwecksetzung (Freiesleben/König 2023), sondern die *mangelnde Konkretisierung* der Zwecke hinsichtlich ›user‹ und ›context‹. Der unterstellte allgemeine Bedarf des technischen Fortschritts, der XAI quasi naturwüchsig umfasst, wird sodann gekoppelt an die Hinwendung zum ›User‹ – was wiederum abstrakt bleibt. Mit der Hinwendung zum User und zur Praxis wird somit einerseits ein neuer Gegenstand des Forschungsinteresses formal instituiert, der anderseits bemerkenswert vage bleibt. Inhaltlich findet sich eine zaghafte Konkretisierung von Usern/Praxis in zwei Richtungen: Zum einen werden zwei spezifische Anwendungsfälle genannt, die als ›Use-cases‹ in dem Forschungsprojekt fungieren sollten, zum anderen wird auf viele weitere Anwendungsfelder pauschal verwiesen.

Der erste Anwendungsfall ist ein »intelligence analyst who receives recommendations from a big data analytics system«. Um seinen Beruf gut auszuüben, so die Zuschreibung, muss sie diese Empfehlung nachvollziehen können (Gunning et al. 2021: 2). Dieser Fall stehe für ein bekanntes Problem in der Praxis:

> »The data analytics challenge was motivated by a common problem: intelligence analysts are presented with decisions and recommendations from big data analytics algorithms and must decide which to report as supporting evidence in their analyses and which to pursue further. These algorithms often produce false alarms that must be pruned and are subject to concept drift. Furthermore, these algorithms often make recommendations that the analyst must assess to determine whether the evidence supports or contradicts their hypotheses. Effective explanations will help confront these issues.« (Gunning/Aha 2019: 45f.)

Diese Fallkonkretisierung vermittelt eine erste Idee eines Anwendungsszenarios, sagt jedoch nicht präzise, wann in welcher Form Erklärungen des Systems wie genau für die Analystin hilfreich wären. Der Erklärungsbedarf bleibt undefiniert. Allerdings prägt die Beschreibung die Vorstellung darüber, in welcher Form Erklärungen vorkommen sollen: nämlich *als fehlende Information zu […]*, im Prozess der Entscheidungsfindung der Analystin. Die Aufgabe von XAI besteht somit darin, die notwendigen Informationen verfügbar zu halten und auf verständliche Weise zu

übermitteln. XAI wird hier als Informationsverarbeitungs- und Vermittlungsaufgabe gedacht, die sich kontextunabhängig formulieren lässt.

Das zweite Beispiel stammt aus dem Bereich der militärischen Anwendung von Drohnen in Kampfschauplätzen. Das Militär setzt Drohnen (halb-automatisierte Agenten) ein, um Sachen für Truppen zu transportieren, Informationen einzuholen, Ziele zu identifizieren oder auch abzuschießen. Interessanterweise wird mit Bezug auf diesen zweiten Anwendungsfall von der KI als einem Partner gesprochen, der die Fähigkeiten von Soldat:innen ergänzt:

> »The autonomy challenge was motivated by the need to effectively manage AI partners. For example, the Department of Defense seeks semiautonomous systems to augment warfighter capabilities. Operators will need to understand how these behave so they can determine how and when to best use them in future missions. Effective explanations will better enable such determinations.« (Gunning/Aha 2019: 46)

Was in dieser Umschreibung des zweiten Fallbeispiels ›effectively manage‹ genau bedeuten soll, bleibt offen. Hierbei geht es weniger um fehlende Informationen und deren Vermittlung, sondern darum, das Systemverhalten in pragmatischen Zusammenhängen einschätzen zu können, z.B. die Flugrouten einer teilautomatisierten Drohne.

Anstatt auf die Verschiedenheit beider Anwendungsfälle einzugehen und zu überlegen, welche XAI-Tools für welche Konstellationen und Zwecke geeignet sein könnten, findet sich in den genannten Quellen eine gegenteilige Strategie: Man setzt auf Generalisierung der Anwendungsfälle. Dies passiert, indem weitere Fälle beispielhaft genannt werden – »Consider, for example, a doctor needing to explain a diagnosis to a fellow doctor, a patient, or a medical review board« (Gunning et al. 2021: 8) – und indem ganze Anwendungs*felder* ins Spiel gebracht werden.

Auf diese Weise entsteht der Eindruck einer unspezifischen Inwertsetzung von XAI für Anwendungen von KI, als seien Erklärungen immerzu hilfreich, gewollt oder gar notwendig. Zudem zeigt es, dass die Verschiedenheit der Anwendungskonstellationen in der Konzeption des DARPA-Projektes keine oder nur eine sehr marginale Rolle spielt. Auch die in der Ausschreibung eingeführten Use-Cases spielten in der Durchführung des Projektes keine prägnante Rolle, sondern man kam zu dem Entschluss »that it would be more valuable to explore a variety of approaches across a breadth of domains« (Gunning et al. 2021: 4). Man sah es nicht als zentral für die Forschung an, sich auf bestimmte Beispiele in ihrer Spezifizität festzulegen und einzulassen. Entgegen der proklamierten Hinwendung zu ›den usern‹ denkt man nicht von der Praxis, sondern von der Technologie aus.

So undifferenziert wie der Bezug zu Praxiskontexten ist, so ist er es auch gegenüber den intendierten oder potenziellen Nutzer:innen. Man redet allgemein von den

›Usern‹: »Explainable AI will be essential if users are to understand, appropriately trust, and effectively manage these artificially intelligent partners« (Gunning/Aha 2019: 44). Diese generalisierte Sicht auf intendierte und potenzielle Nutzer:innen von XAI wird indirekt leicht eingeschränkt, indem man den Untersuchungsgegenstand User/Praxis auf solche Anwendungstypen einschränkt, in denen professionelle Entscheidungsträger:innen durch KI-Systeme unterstützt werden:

> »The target of XAI is an end user who depends on decisions, recommendations, or actions produced by an AI system, and therefor needs to understand the rationale for the system's decisions.« (DAPRA 2016: 6)

Es geht nicht um den privaten Bereich, sondern Arbeitswelten. Es geht nicht um Betroffene der KI-assistierten Entscheidungen (Klient:innen, Patient:innen, Datensubjekte), sondern um die Entscheidungsträger:innen. Es geht aber auch nicht um organisationale Strukturen oder institutionelle Einbettungen der Systeme (z.B. die Befehlskette im Militär mit seinen klar geregelten Hierarchien oder die Ebene von Management und Direktion, die in Unternehmen oder Kliniken es zuerst verantworten die fraglichen KI-Systeme überhaupt einzukaufen und anzuwenden). DARPA wählt allein die Mikroperspektive der ›human-computer-interaction‹. Es geht ebenfalls nicht darum, zu vergleichen, ob die Berufstätigen ihre Arbeit mit oder ohne KI-Unterstützung besser ausführen. Die Optimierung der Arbeitsprozesse durch KI-Systeme wird nicht in Frage gestellt. Es geht allein darum mit XAI sicherzustellen, dass die neuen Assistenzfunktionen sinnvoll eingesetzt werden können. Die Handhabbarkeit der Systeme durch ihre Nutzer:innen wird so zum Faktor der Optimierbarkeit von Arbeitsabläufen durch die neue Technologie. Auf diese Weise wird das Projekt in den Rahmen einer abstrakten technischen Fortschrittserzählung gesetzt, was alternative Konzeptionen, etwa die des Ausprobierens, gerade verdeckt. Es geht bei XAI nicht darum, in der Kollaboration verschiedener Interessensvertreter:innen zu testen, in welchen Fällen, welche KI, welche Entscheidungsträger:innen sinnvoll unterstützen kann und inwiefern für wen, wozu und in welcher Art Erklärungen dabei hilfreich oder auch notwendig sein mögen. Die Notwendigkeit der jüngeren KI-Systeme wird schlicht gesetzt und mit ihnen und ihrer Charakterisierung als undurchsichtig die generalisierte Dienlichkeit von XAI. Durch diese undifferenzierte Bezugnahme auf User/Praxis entsteht eine generalisierte Suggestion: Ohne XAI können Entscheidungsträger:innen die neuen, effektiven KI-Systeme nicht sinnvoll nutzen und wären damit vom technischen Fortschritt ausgeschlossen – was es, so der weitere Subtext, zu verhindern gilt.

Dafür, dass die Anwendungsfelder und Fälle in der informatischen Konzeption als hinreichend ähnlich erachtet werden, spricht ebenfalls der Verweis in der DARPA-Ausschreibung auf eine Studie von Kulesza et al. zum »explanatory debugging« (Kulesza et al. 2015), welche dort als Vorbild für die Userorientierte Forschung im

DARPA-Projekt angepriesen wird. Dies ist bemerkenswert, weil die zitierte Studie einen bestimmten Praxisbereich in den Blick nimmt – solche bereits im Einsatz befindlichen Empfehlungssysteme, die Texte klassifizieren müssen, z.B. Spam-Filter für E-Mail-Programme oder die Auswahl von News Feeds (Kulesza et al. 2015: 128). Diese Studie als Vorbild für den militärischen Nutzen von XAI zu nehmen ist deswegen bemerkenswert, weil Anwendungen wie Spam Filter typischerweise anders als militärische Anwendungen nicht als moralisch, politisch oder sozial brisante Kontexte gelten. In der Risikoklassifikation des AI ACTs der EU (European Commission 2022) gelten Spam Filter als das Beispiel für eine Anwendung der Kategorie ›low risks‹. Aus einer Perspektive, die Risikoklassifikationen bzw. die soziale, ethische, politische, legale Sensitivität von Kontexten ernst nimmt, wäre es zumindest fragwürdig, inwiefern das Spam Filter-Beispiel überhaupt als Vorbild für sensitivere Kontexte dienen kann. Eine solche Überlegung scheint in der DARPA-Konzeption keine Rolle zu spielen. Kulesza et al. wählten ihr Beispiel aus technischen und forschungspragmatischen Gründen:

> »We chose text classification because (1) many real-world systems require it (e.g., spam filtering, news recommendation, serving relevant ads, search result ranking, etc.) and (2) it can be evaluated with documents about common topics (e.g., popular sports), allowing a large population of participants for our evaluation.« (Kulesza et al. 2015: 128)

Man könnte überlegen, ob diese Auswahlkriterien, die strategisch sinnvollsten sind, um die effektive Handhabbarkeit von KI-Systemen im Bereich des Militärs oder auch der Medizin, in denen Entscheidungen stark in das Leben von Menschen eingreifen, zu untersuchen. Mir scheint, die undifferenzierte Sicht auf User:innen und Kontexte verfehlt es, überhaupt Fragen der Übertragbarkeit von Studien mit Demonstratoren oder Prototypen oder andere empirisch gewonnene Einsichten präzise adressieren zu können. Die Präsentation der Projektergebnisse als Liste von Stichpunkten (»key takeaways«) verstärkt den Eindruck, als würde man sich entweder für die Übertragbarkeit von Einsichten auf andere Kontexte schlicht nicht interessieren oder als ginge man davon aus, dass diese Einsichten allgemein gültig seien (Gunning et al. 2019; Gunning et al. 2021). Ich demonstriere, wie unplausibel und wenig aussagekräftig die Befunde dadurch werden, an den ersten beiden gelisteten Befunden:

(i) »Users prefer systems that provide decisions with explanations over systems that provide only decisions. (Supported by 11 experiments across performer teams.)« (Gunning et al. 2021: 8)

(ii) »In order for explanations to improve user task performance, the task must be difficult enough that the AI explanation helps (PARC, UT Dallas).« (Gunning et al. 2021: 8)

Es ist fragwürdig, ob (i) allgemein zutrifft. Hoffman et al. widersprechen der Generalisierung: »Explanations are not needed all the time […]« (Hoffman et al. 2023: 241). Befund (ii) ist entweder trivial (wenn nichts erklärungsbedürftig erscheint, braucht es keine Erklärung) oder zu unspezifisch: bezogen auf welche Aufgabentypen, in welchen Kontexten und für welche Berufstätigen trifft dies zu? Was waren die konkreten Bedingungen der entsprechenden Studie (im Labor oder in der Praxis?) und aus welchen Gründen hält man diese Befunde übertragbar und auf was genau? Die Projektkonzeption ließ auf diese Weise wenig Platz für einen echten Paradigmenwechsel, für die man erstens die Problemstellung hätte überdenken müssen, um dann zweitens nach angemessenen Lösungsstrategien zu suchen. Die Organisation der Forschung bot darüber hinaus wenig Raum für die von der informatischen-normalwissenschaftlichen Perspektive abweichende Alternative von Hoffmans Team.

3.1.2 Organisation der Forschung

Das DARPA-Forschungsprojekt sollte drei Forschungsfragen verfolgen: »(1) how to produce more explainable models, (2) how to design explanation interfaces, (3) how to understand the psychological requirements for effective explanations« (Gunning/Aha 2019: 45). Die ersten beiden Fragen waren Aufgabe der elf technischen Forschungsgruppen, aus der ML- und HCI-Community, mit einer Ausrichtung auf Fragen des Interface-Designs.[15] Die dritte Frage lag in der Hand des psychologisch-orientierten Teams von Hoffman. Das Projekt war in zwei Phasen gegliedert. In der ersten Phase (18 Monate) sollten technische Demonstratoren entwickelt werden; in der zweiten Phase (30 Monate) sollten diese getestet werden. Am Ende sollten Prototypen herauskommen, die in einem open source XAI-Toolkit der Öffentlichkeit zur Verfügung gestellt werden.[16] Organisatorisch betrachtet, sah man drei Bereiche vor (Technical Areas (TA)): einen technischen Bereich (TA1), der die Fragen (1) und (2) umfasst, den Bereich der psychologischen Erklärung (TA2), der für die dritte Forschungsfrage zuständig war. Hinzu kam die Evaluation (TA3), welche vom U.S. Naval Reserach Laboratory (ein gemeinsames Forschungslabor der US Navy und des US Marine Corps) unter der Leitung von Eric Vorm verantwortet und organisiert wurde.

Ursprünglich gingen die Beteiligten, laut ihrer veröffentlichten Retrospektive, davon aus, dass die von den elf Teams entwickelten Erklärungstechniken durch

15 Hier HCI im Sinne des kognitiven Paradigmas (Harrison et al. 2007): Es geht um effektive Vermittlung Kanäle, Signale, Informationsdichte.

16 Das Toolkit kann man sich unter diesem Link herunterladen: https://xaitk.org/.

die Gestaltung der Schnittstellen nutzerfreundlich gemacht werden können. Mit den Nutzer:innen wird folglich die Frage der Schnittstellengestaltung zentral. Damit zerfällt die Aufgabe, nutzergerechte Erklärungen zu entwickeln, in zwei technische Komponenten; erstens ein »explainable model« und zweitens das »explainable interface« (Gunning/Aha 2019; Gunning et al. 2021). Diese Arbeitsteilung der Forschung ist von Seiten der Technik gedacht. Eigenschaften des technischen Systems (Opazität, Komplexität) bestimmen den Erklärungsbedarf – Entwicklung von erklärbaren Modellen. Sind diese einmal vorhanden, gilt es sie nutzergerecht zu vermitteln. Beide Schritte stehen additiv zueinander. User sind für die Schnittstellengestaltung relevant, nicht aber für die Frage des Erklärungsbedarfs (siehe kritisch hierzu z.B. Rohlfing et al. 2020). Hinzu kommt, dass die Endnutzer:innen negativ definiert sind, als nicht-Experten. Während die Expert:innen die (neuen) Erklärungstechniken verstehen, braucht es für die Endnutzer:innen die Schnittstelle. Ihr nicht-Expert:innen-Sein ergibt den Vermittlungsbedarf. Damit wird das Zuschneiden auf die ›user‹ auf eine Frage des fehlenden Wissens und der Modalitätsformen verengt, in denen das fehlende Wissen präsentiert werden soll.

Es wird in den Quellen nicht deutlich, ob man sich eine Integration der Ergebnisse des psychologischen Teams in die Entwicklung der neuen XAI-Tools vorstellte. Klar ist, diese Ergebnisse wurden als wichtig für die Evaluation angesehen und damit der Frage, wie man die Effektivität von XAI messen kann. Das psychologische Team sollte hierfür die Grundlage bieten und in diesem Sinn den anderen Projekten assistieren:

> »The program structure anticipated the need for a grounded psychological understanding of explanation. One team was selected to summarize current psychological theories of explanation to assist the XAI developers and the evaluation team.« (Gunning et al. 2021: 3f.)

Die vorhergesehene Hauptaufgabe des psychologischen Teams war »understanding the psychology of explanation by summarizing, extending and applying psychological theories of explanation« (Gunning et al. 2021: 2). Sie sollten die psychologische Fachliteratur zum Erklären überblicken und systematisieren (siehe hierzu den Report: Mueller et al. 2019), und auf dieser Erkenntnisbasis dann ein psychologisch informiertes »computation model« des Erklärens entwickeln, welches sie sodann anhand der Evaluationsergebnisse der XAI-Entwickler:innen validieren sollten (Gunning et al. 2021: 4). In der Retrospektive hat es sich als zu hochgegriffen herausgestellt, ein formalisiertes psychologisches Modell des Erklärens zu erstellen, stattdessen habe das Team um Hoffman beschreibende Modelle erstellt (Gunning et al. 2021: 4). Bemerkenswerterweise lässt sich aus den Publikationen ein Missverständnis darüber erkennen, worin das ›psychologische Modell‹ besteht, welches Hoffmans Team erstellen sollte. Was Gunning et al. (Gunning et al. 2021) als Ergeb-

nis der Arbeit des psychologischen Teams ausgeben (siehe Abbildung 2), markieren Hoffman et al. (Hoffman et al. 2023) als informatische Vorannahme, mit dem das DARPA Projekt gestartet sei (siehe Abbildung 3). Ihr eigentliches psychologisches Modell weisen sie abgrenzend hierzu als Weiterentwicklung zu dieser anfänglichen Annahme aus (siehe Abbildung 4).

3.2 Psychologische Konzeption

Aus der Retrospektive des ›psychologischen‹ Teams um Hoffman lassen sich weitere Diskrepanzen gegenüber der informatischen Perspektive aufzeigen. Zum Beispiel stellen Hoffman et al. zwölf Prinzipien zusammen, welche aus dem DARPA-Projekt bzw. der von diesem stimulierten XAI-Forschung, hervorgegangen sind (Hoffman et al. 2023). Diese zwölf Prinzipien sind weitaus sensibler für die Diversität der Nutzer:innen und Fragen der Relevanz von Erklärungen, als es in der informatischen Perspektive zum Ausdruck kommt. Auch wenn sich die Darstellung der Einsichten aus dem DARPA-Projekt des psychologischen Teams in Teilen so liest, als wären sie neu gewonnen, spricht vieles dafür, dass das psychologische Team die Erforschung von XAI für Nutzer:innen von Beginn an anders konzipiert hätte. Allein der fachliche Hintergrund von Hoffman, der Experte in den Bereichen des ›Cognitive Systems Engineering‹, des ›Human-Centered Computing‹ sowie der ›Human Factors‹-Forschung ist, spricht dafür, dass dem ›psychologischen‹ Team schon vor Projektbeginn klar war, »one explanation does not fit all« (Sokol/Flach 2020). Während es bei Gunning et al. (2021: 8) so klingt, als sei diese Einsicht eine der wesentlichen Lernschritte des Projektes gewesen – »different user types require different types of explanation« –, plädieren Clancey und Hoffman (2021) z. B. dafür, Einsichten wie diese aus der Forschung zu *Intelligent Tutoring Systems* für die XAI-Forschung fruchtbar zu machen. Aus letzterem Bereich seien Erkenntnisse aus mehreren Jahrzehnten Forschung zu gewinnen.

Während die informatische Perspektive von Seiten der Technik gedacht ist, denkt die psychologische HCI-Perspektive von ›den usern‹ aus.[17] Letzteres impliziert wenigstens drei zentrale Forschungsfragen, die in der informatischen Perspektive nicht zur Geltung kommen:

17 Mit der fachlichen Expertise von Hoffman scheint sein Team in der Forschung von den ›users‹ oder ›humans‹ ausgegangen zu sein und nicht von typischen psychologischen Ausgangspunkten wie kognitiven Prozessen oder Persönlichkeitsmerkmalen. Ich danke Ingrid Scharlau für den Hinweis, dass gegenüber solchen psychologischen Ausgangspunkten die Rede von ›dem User‹ einen ziemlich groben Forschungsgegenstand konstruiert, den man in der klassischen Psychologie so nicht findet.

1. Wann sind Erklärungen für Nutzer:innen überhaupt relevant?
2. Wozu dienen Nutzer:innen Erklärungen?
3. Wie geht man mit der Diversität von Kontexten und Nutzer:innen um?

Die informatische Sicht leitet den Erklärungsbedarf aus den Eigenschaften des technischen Systems ab (Komplexität, Opazität) während dieser Perspektivwechsel dazu einlädt, die Bedarfe von Nutzer:innen in verschiedenen Praxiskontexten zu erkunden. Wer so fragt, braucht ein anderes Wissen über Nutzer:innen und Kontexte. Es reicht dann nicht aus User Studies allein für die Evaluation der gewonnenen Werkzeuge einzubeziehen, denn dieses Wissens sollte das Forschungsdesgin von Anfang an informieren:

> »The design of XAI systems must be fully informed by a psychological model based on empirical evidence of what happens when people try to explain complex systems to other people and what happens as people try to reason out how a complex system works […].« (Hoffman et al. 2022: 366)

Indem die DARPA-Konzeption User studies allein für die Evaluation angesetzt hat, weist sie dem Wissen über Nutzer:innen und Kontexten sowie dem Wissen *von* User:innen einen bestimmten Platz zu: es betrifft allein die Effektivität der entwickelten XAI-Tools (Gunning et al. 2021: 4).[18] In der Entwicklung dieser Tools spielt dieses Wissen keine Rolle. Auch diese Platzierung des User-Wissens spricht für die Auffassung, dass die Erklärungen unabhängig von den konkreten, kontextuellen Bedarfen identifizierbar sind. Sie ergeben sich aus Mängeln bzw. Eigenschaften der KI-Systeme allein. In dieser Auffassung kommt das Selbstverständnis des Paradigmas der epistemischen Güte von ML deutlich zum Tragen, in denen es allein darum geht, die KI-Modelle für andere Expert:innen einsichtig zu machen und aufgrund der Homogenität dieser Expertengruppe die Verschiedenheit anderer Nutzer:innen nicht als relevante Kategorie erachtet wird.

Die informatische Perspektive verpasst es so, überhaupt nach der Bedeutung von Erklärung aus Sicht von User:innen in verschiedenen Kontexten zu fragen. Sie entwickelt keine Vorstellung über die Kontextabhängigkeit von Erklärprozessen. Zum Beispiel scheint allein die Tatsache, dass Nutzer:innen in ihren Arbeitswelten häufig mit multiplen Aufgaben und Zielen konfrontiert sind, in denen sich die KI-Systeme mit ihren Empfehlungen (besser oder schlechter) einfügen, einen praktischen Unterschied zu machen, wie Erklärungen effektiv sein können. In der XAI-Forschung geht man stattdessen von einer Vereinfachung aus: »Much XAI research

18 User studies waren ein zentraler Bestandteil der Evaluationsphase, mit insgesamt 12.700 Teilnehmenden, wovon »1900 supervised« und »10 800 unsupervised participants« waren, die z.B. über Amazons Mechanical Turk gewonnen wurden (Gunning et al. 2021: 7).

has assumed a one-person, one-task problem situation« (Hoffman et al. 2023: 243). Solche Vereinfachungen können forschungspragmatisch sinnvoll sein, man sollte sich jedoch über die Differenz zu den eigentlich angedachten Nutzungskontexten im Klaren sein. Um über diese gesättigteren Vorstellungen zu gewinnen, gibt es verschiedene Strategien, wie den Einbezug von Domänenexpert:innen oder anderen Stakeholdern im Sinne eines partizipativen Forschungsdesigns (Simon 2017; Dignum 2019; van der Hoven/Manders-Huits 2020; Friedman/Nissenbaum 1996) oder einer Befragung und Beobachtung von diesen in realen Anwendungskontexten oder im Labor. Zwar scheint es jüngst ein größeres Bewusstsein in der XAI-Forschung für die Notwendigkeit dieser Art von Forschung zu geben (Ribera/Lapedriza García 2019; Langer et al. 2021; Cabitza et al. 2023; Capel/Brereton 2023; Kim et al. 2024) – wenn man mit XAI den Zweck verfolgen will, für User:innen in verschiedenen Kontexten dienliche Erklärungen anzubieten – doch nach wie vor scheint es hierzu wenig empirische Untersuchungen zu geben (so der Befund von Langer et al. 2021). Wang et al. betonen, dass sich mit diesem Perspektivwechsel weitere Forschungsfragen anschließen – zum Beispiel wie sich die Relevanz von Erklärungen für User:innen überhaupt validieren lässt und wie man mit dem Unterschied zwischen Normen guter Erklärungen (aus der Argumentationstheorie und Logik) und dem, wie Personen tatsächlich anderen Dinge erklären, umgehen sollte (Wang et al. 2019: 601). Sollte sich XAI eher an der Norm einer guten Erklärung orientieren oder eher am Vorbild realer Erklärungen?

Die Kernfrage, die sich mit der ernsthaften Hinwendung zu User:innen und der Praxis stellen muss, ist die nach der Relevanz von Erklärungen.[19] Denkt man von der Technik aus, besteht der Bedarf an Erklärung scheinbar allgemein betrachtet, nämlich so lange wie die Technik opak und komplex ist und diese Charakteristika der Technik als erklärungsbedürftig angesehen werden. Nur wenn man diese Zwecksetzung von XAI, die von Eigenschaften der Technik ausgeht, unreflektiert auf eine andere Zwecksetzung von XAI überträgt – nämlich KI-Systeme in der Praxis effektiv nutzen zu können – entsteht der Eindruck einer generalisierten Inwertsetzung von XAI, die von Seiten der User:in und der Praxis gedacht nicht haltbar ist. Es ist sonach geboten, beide Zwecksetzungen voneinander zu unterscheiden: wir haben es hier mit verschiedenen Paradigmen von XAI zu tun.

Um die Frage nach der Relevanz von Erklärungen zu erforschen, schlägt das Team um Hoffman vor, von ›Triggern‹ zu sprechen, die an verschiedenen Stellen im Gebrauch von KI-Systemen auftauchen können. Typische Trigger sind Überra-

19 Was die Frage angeht, ob es überhaupt eine Erklärung in einer bestimmten Situation braucht, könnte man weiter pragmatisch-kontextuelle Faktoren (Relevanz) von individuell-kognitiven Faktoren (Angemessenheit, Sinnhaftigkeit) unterscheiden. Für diesen Hinweis danke ich Katharina Rohlfing.

schungen oder der Bruch mit Erwartungen (Hoffman et al. 2023). Zu diesem Schluss kommen ebenfalls de Graaf und Malle:

> »In short, people try to explain any given behavior (in self or other) if either (a) they themselves wonder why the behavior occurred or (b) they expect that someone else wonders why the behavior occurred.« (de Graaf/Malle 2017: 22)

Dieser psychologische Befund lässt sich gut an einen Kerngedanken der phänomenologischen Technikphilosophie anschließen. Erklärungsbedürftig ist das nichtselbstverständliche; das, was mit Vertrautem bricht. Wichtig ist hier, dass unser eingeübter Umgang mit Technik ebenso zum Bereich des Selbstverständlichen gehört wie alles andere auch. Technik ist in diesem Sinne Teil der Lebenswelt (als Inbegriff des Selbstverständlichen) und damit aus phänomenologischer Perspektive gerade das, was wir nicht (bewusst) zu verstehen brauchen, um sie nutzen zu können (Blumenberg 1981; Kaminski 2010). So wie wir keine Fachphysiker:innen sein brauchen, um eine Straßenbahn sinnvoll zu nutzen, liegt es ebenso nicht auf der Hand, wie und warum z.B. eine Einsicht, ähnlich der von Heatmaps, für die Bilderkennung notwendig sein sollte, um das Verhalten von Drohnen einschätzen zu können. Dem modernen Großstädter genügt es »daß er auf das Verhalten des Straßenbahnwagens ›rechnen‹ kann, er orientiert sein Verhalten daran; aber wie man eine Trambahn so herstellt, daß sie sich bewegt, davon weiß er nichts« (Weber 1992: 86). Max Webers Beispiel der Straßenbahn gibt zu bedenken, dass das nötige Verständnis über das Verhalten von Technik von anderer Art ist als das entsprechende Expertenwissen zum Bau dieser Technik.

Vor dem Hintergrund solcher Überlegungen ließen sich weitere Forschungsfragen für die XAI-Community gewinnen. Ein Fragebündel geht in die Richtung zu konzipieren, worin das Unerwartete eigentlich besteht: Sind es Abweichungen von sozialen Normen? Das Unvertraute? Das Neue? Wie zeigt sich dieses jeweils in verschiedenen konkreten Arbeitskontexten? Hieran angeschlossen ließe sich ein weiteres Fragebündel in die Richtung ausbuchstabieren, wofür genau XAI eingesetzt werden soll. Es könnte beispielsweise einen Unterschied machen, ob es darum geht Nutzer:innen bei der Aneignung von Neuem (neuen technischen Hilfsmitteln, neuen Aufgaben, etc.) zur Seite zu stehen (quasi als eine Art bessere Bedienungsanleitung) oder ob es darum geht Fälle, die von den Üblichkeiten abweichen oder schwerer einzuschätzen sind mit Hilfe von KI und ggf. Erklärungen zu dieser besser nachvollziehen zu können (Lebovitz et al. 2022). Weiter macht es einen Unterschied ob XAI Angebote machen soll, um etwas besser zu verstehen, oder ob es letztlich (z.B. im medizinischen oder juristischen Bereich) darum geht, Entscheidungen rechtfertigen zu können (Krishnan 2020).

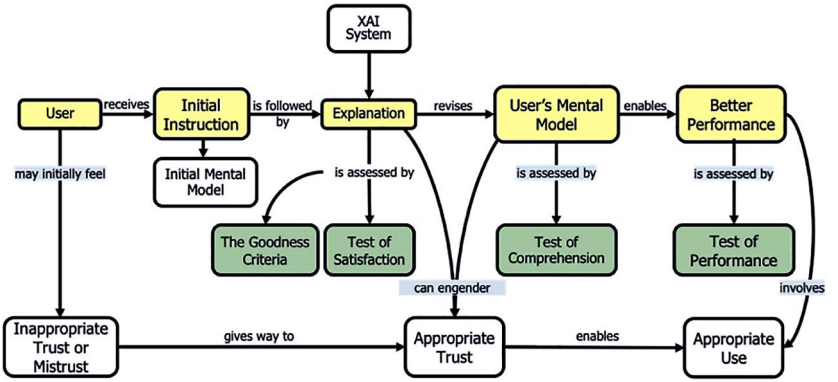
Die Frage nach der Relevanz von Erklärungen aus User-Sicht erfordert somit eine differenzierte Sicht auf die instrumentellen Zwecksetzungen von XAI und da-

mit ebenfalls auf die verschiedenen Kontexte und Nutzer:innen. Die praktische Notwendigkeit dieser Differenzierungen ergibt sich nicht in der Perspektive, die von der Technik aus den Bedarf an Erklärungen ableitet.

Diese verschiedenen Sichtweisen auf XAI implizieren außerdem verschiedene Vorstellungen darüber, wie die Technik und die Nutzer:innen zusammenspielen sollen. Diese Vorstellungen zum Verhältnis von ›user‹ und XAI hängen wiederum mit der Zwecksetzung von XAI zusammen, d.h. mit den Zielvisionen wie eine wünschenswerte Zusammenarbeit von XAI und Nutzer:innen aussehen würde. Diese Fragen hängen mit der Aufgabenstellung im DARPA-Projekt zusammen und fordern das psychologische Team, ein psychologisches Modell des Erklärens zu entwickeln. Die Darstellung der Ergebnisse dieser Modellerstellung demonstriert deutlich den Unterschied zwischen der informatischen und der psychologischen Perspektive im DARPA Projekt, denn das, was Gunning et al. (2021) als Ergebnis der Arbeit von Hoffmans Team herausstellen und als psychologisches Modell des Erklärens titulieren (Abbildung 2) weisen Hoffman et al. (2023) dezidiert als eben nicht-psychologisches, sondern informatisches Modell des Erklärens zurück – mit dem das DARPA-Projekt gestartet ist: Dieses Modell ist kein Ergebnis der Forschungsarbeit, sondern eine Explikation von Vorannahmen. Hoffman et al. üben harsche Kritik an diesem Modell, weil es den Prozess des Erklärens in der Logik einer Fütterung (›spoon feeding‹) modelliere (Abbildung 3): »The explanation is generated and then delivered, (ideally) to good effect.« (Hoffman et al. 2023: 239). Dieses Fütter-Modell impliziert eine Reihe von Unterstellungen. User werden hier als passive Empfänger von Erklärungen vorgestellt, die keinen aktiven Beitrag zum Erklärprozess liefern (siehe kritisch dazu Rohlfing et al. 2021): »the ›spoon feeding‹ paradigm is blind to the fact that users engage in a motivated, deliberative attempt to make sense of the AI system and any explanatory material that may be presented.« (Hoffman et al. 2023: 239)

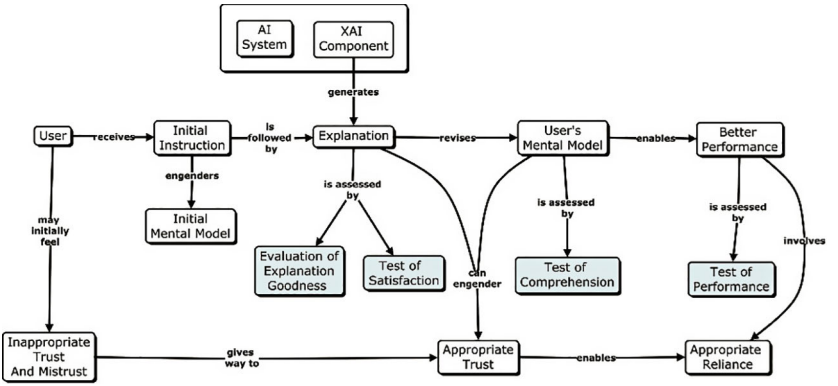
Die Logik des informatischen Modells basiert auf der Annahme einer simplen Informationsübertragung von einem Sender (XAI-System) zu einem Empfänger (user) durch bestimmte Kanäle (Schnittstellengestaltung). In dieser Vorstellung liegt der aktive Part beim XAI-System. Ob die Übertragung gelingt, dafür ist die Güte der Kanäle ausschlaggebend. Indem Hoffman et al. diese Vorstellung als »spoon feeding paradigm« ausweisen (Hoffman et al. 2023: 239), heben sie die unterstellte Passivität der User:in drastisch hervor, denn bei einer typischen Fütterung (von Haustieren, Babys oder pflegebedürftigen Personen) gehört es allein zur Rolle des Fütterungssubjekt den Mund zur passenden Zeit zu öffnen und wieder zu schließen. Der aktive Beitrag zur Verdauung des Futters liegt allein im Kauen und Schlucken – der eigentliche Verdauungsvorgang läuft dank funktionierender Organe wie automatisch ab. Die ganze Musik beim Füttern liegt auf der Seite der Fütternden. Die Gefütterten haben in der Regel wenig mitzubestimmen über das, womit sie gefüttert werden (Ist es gut für mich, brauche ich das?).

Abbildung 2: Darstellung der vermeintlichen Ergebnisse des psychologischen Teams nach Gunning et al.



Quelle: Gunning et al. 2021: 5

Abbildung 3: Rückweisung der informatischen Vorannahmen zum Verhältnis XAI-user (›spoon feeding model‹)



Quelle: Hoffman et al. 2023

Sie nehmen nur indirekt darauf Einfluss, wie viel sie bekommen. Versteht man XAI als Fütter-Aufgabe, treffen also die Fütternden (Systemdesigner:innen) allein alle Relevanzentscheidungen, die User:innen hingegen werden infantilisiert.[20]

20 Die Praxis des Fütterns lässt sich sehr wohl im familiären Bereich oder in der Pflege als eine Situation sozialer Aushandlungen verstehen, in der die Gefütterten keineswegs rein passiv sein müssen. Mir scheint aber die Wahl der Metapher von Hoffman darauf abheben zu wollen, dass User:innen hier als rein passive Gefäße modelliert sind. Diese Wahl spielt vermut-

Die Vorstellungen zum Verhältnis von User:in und XAI lassen sich darauf beziehen, welche Zielvisionen für die XAI-Entwicklung eigentlich angesetzt wird. Wie sehe eine optimale XAI in der Nutzung aus? Aus der informatischen Konzeption des DARPA-Projektes lassen sich hierzu nur indirekt Vermutungen aufstellen. Da diese Konzeption von der Technik (allein) her denkt, wäre die optimale XAI eine perfektionierte Technik. Worin die Perfektion der Technik liegen könnte, dafür scheinen mir unterschwellig zwei Ideale eine Rolle zu spielen: Das Ideal der Selbstevidenz und das Ideal der vollständigen Automatisierung.

Beim ersten Ideal der Selbstevidenz ist die KI nicht (mehr) fragwürdig bzw. sind die angebotenen Erklärungen so klar, dass keine Fragen übrigbleiben und man eine vergleichbare Situation hat wie bei einer vollkommen transparenten KI. Die Problemstellung (opake und komplexe technische Systeme) löst sich in dieser Vision quasi auf, ergo gibt es keinen Bedarf mehr an XAI bzw. hat XAI ihre genuine Aufgabe optimal erfüllt.[21] Die zweite Vision ist das Spiegelbild der Selbstevidenz für XAI. Eine vollkommen automatisierte XAI ist nicht angewiesen auf das Zusammenspiel mit ›usern‹ oder Kontexten, sie schafft es allein aus sich heraus flexibel jeden Erklärungsbedarf zu befriedigen. In beiden Zielvorstellungen spielen Nutzer:innen und die Kollaboration dieser mit der XAI keine tragende Rolle. Sie passen zu der Ignoranz gegenüber der Diversität von ›usern‹ und Anwendungsfeldern. In dieser Sicht erfüllt die perfekte Technik die Funktion, menschliche Arbeit zu substituieren.

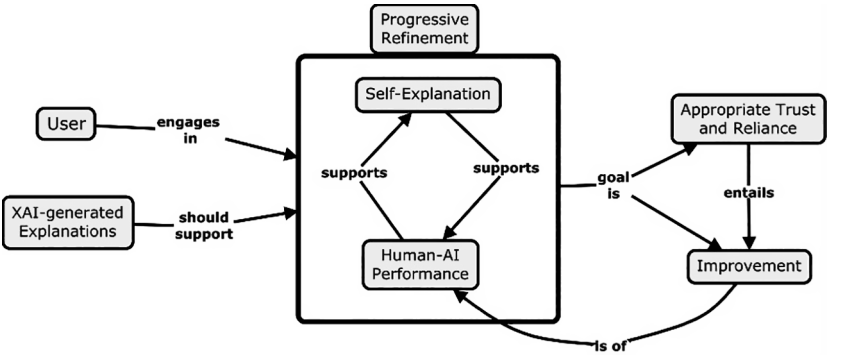
Das Team um Hoffman bietet in diesem Punkt ebenfalls eine alternative Sicht an: bei ihnen geht es nicht um Substitution, sondern Augmentation. Nicht die vollständige Automatisierung (oder die vollständige Auflösung des Erklärungsbedarfs im Sinne vollkommener Transparenz) ist das Ziel, sondern Kollaborationen zwischen Nutzer:innen und KI-Systemen. Dies betonen sie mit ihrem Prinzip »All Explanations involve Self-Explanation«. Es solle bei XAI nicht darum gehen, durch die automatisierten Erklärungen Mängel der Technologie zu beheben (ihre Opazität, ihre Komplexität), sondern darum Nutzer:innen in die Lage zu versetzen, sich das Fragwürdig-Gewordene mit Hilfe der XAI *selbst zu erklären*. Dementsprechend sollte eine oberste Design-Maxime wie folgt lauten: »Explain unto others in such a way as to help them explain to themselves« (Hoffman et al. 2023: 238). Die Ausgaben des XAI-Systems werden diesem Prinzip nach als Material und Hilfestellung für den Erklärungsprozess gesehen, der kollaborativ zwischen den Nutzer:innen und dem System stattfindet und in dem die User:innen einen aktiven Part spielen: Sie müssen in die Lage versetzt werden, sich das Verhalten des Systems mit Bezug auf das,

lich auf ein oft kritisiertes Modell des Lernens an, in dem Lernen ebenfalls mit rein passiver Nahrungsaufnahme verglichen wird.

21 Dieses verheißt jedenfalls das Versprechen der Transparenz. Dass diese allein praktisch nicht automatisch sinnvolle Anhaltspunkte liefert, wurde häufig diskutiert (vgl. Vogelmann 2019; Stamboliev 2023).

was für sie in einem spezifischen Kontext relevant ist, selbst erklären zu können. Der Output von XAI-Systemen (»representations«) darf also nicht mit der Erklärung selbst verwechselt werden, sondern assistiert den Erklärungsprozess (Hoffman et al. 2023). Das XAI-System erhält hierbei die Rolle eines Assistenten, denn letztlich geht es um das Vermögen der Nutzer:innen. Für diese Zielvision hat das Team ihr psychologisches Modell von XAI als Kollaboration erstellt (Abbildung 4).

Abbildung 4: Selbst-Darstellung der Ergebnisse des ›psychologischen‹ Teams: XAI als Kollaboration



Quelle: Hoffman et al. 2023

Bereits 2013 hatte sich Hoffman in Ko-Autorschaft in »The Seven Deadly Myths of ›Autonomous Systems‹« kritisch gegenüber eindimensionalen und überzogenen Konzeptionen der Autonomisierung von Technik geäußert (Bradshaw et al. 2013). Der springende Punkt hierbei ist, dass sie Autonomie nicht als Eigenschaft technischer Systeme missverstanden wissen wollen, sondern als Attribut des soziotechnischen Systems, d.h. dem Zusammenspiel von Mensch und Technik in spezifischen Kontexten mit ihren Bedingungen. Insofern Autonomie als Zielstellung für die technische Entwicklung angesehen wird, sollte man diese Zielstellung nicht einseitig auf die Eigenschaften des technischen Systems beziehen. Vielmehr sei das eigentliche Ziel effektive Teams von Menschen und Maschinen/Software in bestimmten Kontexten zu schaffen; es geht also darum die Autonomie dieses soziotechnischen Systems zu steigern. Der Kontextbezug ist dabei nicht außer Acht zu lassen. Autonomie ist auf Handlungssituationen bezogen: »Functions can't be automated effectively in isolation from an understanding of the task, the goals, and the context« (Bradshaw et al. 2013: 56). Im Design technischer Systeme sind, entweder explizit oder implizit, notwendigerweise Annahmen über die Kontexte der Anwendungen, ihre Nützlichkeit und beschränkte Bedingungen enthalten – Funktionen lassen sich

gar nicht anders konstruieren und denken (Bradshaw et al. 2013: 57; Vermaas 2010; Lenk 1982). Diesen Überlegungen folgend lässt sich vermuten, die informatische Konzeption des DARPA-Projektes, die offenkundig nur oberflächlich über Kontexte und User:innen nachgedacht hat, hat stillschweigend den Kontext der informatischen Forschung und den User-Typ des ML-Experten (eben in einer defizitären Variante charakterisiert durch mangelndes Wissen) auf das Forschungsthema XAI für Laien übertragen. Doch so wenig wie man Autonomie als eine diskrete, separierbare Komponente eines Systems missverstehen sollte, so sollte man auch Erklärbarkeit des Systemgeschehens konzipieren als »capability of the larger system enabled by the integration of human and machine abilities« (Bradshaw et al. 2013: 56). Als Vermögen (»capability«) ist die Verwirklichung von Autonomie bzw. Erklärbarkeit dann auf kontextuelle Bedingungen angewiesen. Ihre Relevanz und Angemessenheit, ergibt sich über die kontextuellen Zwecksetzungen des Handelns. Den Gedanken, nicht technische Eigenschaften, sondern das Team Mensch-Maschine, in den Fokus zu nehmen, stellen auch Miller et al. an den Anfang ihrer Einführung zum *Special Issue on Explainable AI*, welches von den Forschenden des DARPA-Projektes initiiert wurde: »AI can be seen as one integrative part of a cognitive work system and its broader social or organizational context.« (Miller et al. 2022)

3.3 Hinwendung zur Praxis als Paradigmenwechsel

Der Einbezug von Nutzer:innen und Kontexten ist wichtig und kann methodisch und theoretisch vielfältig angegangen werden. Die Arbeit des Teams von Hoffman ist nur ein Beispiel hierfür, das deswegen interessant ist, weil es im Kontext des DARPA-Projektes konzipiert wurde. Allein schon die hier demonstrierten Unterschiede zwischen den anfänglichen informatischen Vorstellungen von der Forschung an der ›psychology of explanation‹ und dem, was Hoffmans Team veröffentlicht hat, sprechen dafür, dass wir es bei der Hinwendung zu Nutzer:innen und zu Kontexten mit einem Paradigmenwechsel gegenüber der normalwissenschaftlichen Forschung der ML-Community im Paradigma der epistemischen Güte von ML zu tun haben. Ich plädiere dafür, diesen Paradigmenwechsel ernst zu nehmen. Wir haben es mit einer anderen Problemkonzeption zu tun als derjenigen, die das Paradigma der epistemischen Güte von ML sinnvoll anleitet.

Die Problemkonzeptionen sollten sich aus den Zwecksetzungen von XAI herleiten, über die sich die Community ernsthafter Gedanken machen sollte (Krishnan 2020; Freiesleben/König 2023). Mir geht es nicht darum eine ›human-centered‹ (Capel/Brereton 2023) gegenüber einer ›technology-centered‹ XAI auszuspielen, sondern darum, die Verschiedenheit dieser, mit dem jeweiligen Fokus auf sinnvoll einhergehende Zwecksetzungen zu beachten (Gunning et al. 2019). Will man sich ernsthaft mit der Güte von XAI für Laien in realen Anwendungsfällen beschäftigen, sollte man die Forschung entsprechend anders organisieren als es im Paradigma der epis-

temischen Güte von ML der Fall ist. Dies fängt bei der Gewichtung der beteiligten Disziplinen an, geht über die Erstellung der leitenden Forschungsfragen und Reflexion der leitenden Vorannahmen, bis hin zur Frage des Einbezugs von Nutzer:innen und der Kontextabhängigkeit von Erklärungen hinaus.

4. Das Paradigma der Vereinbarkeit mit Grundwerten und Normen

Mein drittes Paradigma, der Vereinbarkeit von KI mit Grundwerten und Normen, stellt erneut einen Perspektivwechsel dar. Es geht um ›ethical and social concerns‹, die insbesondere als Auseinandersetzung um ›bias‹, ›fairness‹ und ›accountability‹ von KI diskutiert werden. Genährt aus Sorgen, epistemischen und normativen Unsicherheiten im Umgang mit KI sowie dem vorauseilenden Gehorsam der Tech-Industrie (Lepri et al. 2018; Tworek 2019; Floridi 2021), welche beschwört, *AI for the social good* zu entwickeln, ist ein genuiner Diskursraum entstanden, der unter dem Namen *AI Ethics* firmiert. Dieser Diskurs konstituiert sich über das abstrakte gemeinsame Interesse, KI-Systeme gesellschaftsfähig zu machen und folgt größtenteils der Leitidee, *AI Ethics* als Selbstregulierung einzusetzen (Alpsancar 2023). Für Unternehmen geht es um ihr Selbstverständnis und Ansehen, Akzeptanzfragen und die Konformität mit rechtlichen Vorgaben. Die Politik will KI als Schlüsseltechnologie für wirtschaftlichen Erfolg ihrer Nationen fördern und sucht nach einem passenden regulativen Rahmen (Nannini et al. 2023). *AI Ethics* ist ein anwendungsorientiertes Forschungsgebiet vielfältiger Expertisen, welches ebenso in die Zuständigkeit der Computer Science und ihr angrenzende Disziplinen fällt, wie in die Produktentwicklung großer Tech-Konzerne. Zu diesem Diskursraum zählt ebenfalls eine in der breiteren Öffentlichkeit geführte Debatte um ›high profilic cases‹, über die in Blogs, Tech-Magazinen oder auch Buchpublikationen diskutiert wird (O'Neil 2016; Eubanks 2017; Benjamin 2019; Crawford 2021). Parallel zu der breiteren öffentlichen Debatte entstanden über die letzten Jahre zahlreiche ethische und politische Initiativen, die das Thema der Vereinbarkeit von KI mit gesellschaftlichen Grundwerten auf ihre Agenda setzten (Hagendorff 2020; Nannini et al. 2023). Im Jahr 2019 waren bereits mehr als 80 ethische Richtlinien oder KI-Kodizes öffentlich zugänglich (Jobin et al. 2019; Morley et al. 2020), die von Industrieverbänden wie der IEEE oder ACM, von Unternehmen wie IBM oder Google, oder von staatlichen Institutionen initiiert wurden, wie etwa die High Level Expert Group (2019) der Europäischen Kommission.

Dieser Diskurs adressiert zum einen bestimmte Einsatzgebiete, nämlich »socially significant and morally weighty contexts« (Walmsley 2021: 585; Floridi et al. 2018), in denen die KI-gestützten Entscheidungen einen erheblichen Einfluss auf das Leben derjenigen haben, die von diesen Entscheidungen betroffen sind, z.B. dem Finanzwesen (Zarsky 2016; Pfeiffer et al. 2023), dem Personalwesen (Starke

et al. 2022; Strohmeier 2020), dem Justizwesen (Corbett-Davies et al. 2017; Fortes 2020), dem Gesundheitswesen (Marabelli et al. 2018; López-Martínez et al. 2020) oder bei anderen staatlichen Hoheitsaufgaben (Allhutter et al. 2020). Zum anderen stehen Produkte und Services bekannter Tech-Unternehmen im Zentrum der Aufmerksamkeit, deren Einsatz diskriminierende Effekte hervorbrachte. Klassifikationsfehlern von Bilderkennungsprogrammen, die im geschützten Rahmen der laborähnlichen Wettbewerbe der ML-Community primär epistemisch von Bedeutung sind, kommen hier gesellschaftliches Gewicht zu, wenn farbige Personen durch die Google Photo App als ›gorilla‹ kategorisiert werden (Kasperkevic 2015), das Konzentrationslager in Dachau auf dem von Yahoo bereitgestelltem Foto-Dienst Flickr als ›gym‹ ausgewiesen wird (Hern 2018) oder in Facebooks Daily Mail Video Nutzer:innen, die Videos mit schwarzen Personen angeschaut hatten, ›weitere‹ Videos über Primaten empfohlen bekommen (AIAAIC 2021). Durch einen solchen »steady stream of examples and anecdotes of problematic biases, prejudices and other errors that have been automated and reinforced« (Walmsley 2021: 585), sei insgesamt eine große Skepsis gegenüber KI entstanden (Miller 2019: 1; Gilpin et al. 2018).

AI Ethics funktioniert größtenteils als Resonanzraum dieser Sorgen und Debatten. *Ethics* wird dabei hauptsächlich in attributiver Bedeutung verwendet: *ethische KI* ist gute KI. Synonym zu ethischer KI spricht man auch von einer KI, die ›sozialen Werten‹ gerecht wird. Die gesellschaftliche Verträglichkeit von KI ist für viele eine Aufgabe des *Engineerings*. Beispielhaft hierfür steht das Buch von Kearns und Roth, *The Ethical Algorithm* (Kearns/Roth 2019). Die beiden Informatiker verbinden ihre universitäre Forschung zum »socially aware algorithm design« mit einer Beratungstätigkeit für große Tech-Unternehmen wie Amazon, Apple und Facebook. Eine der größten Konferenzen dieser wachsenden Community ist die *ACM Conference on Fairness, Accountability, and Transparency (ACM FAccT)*.[22] Die Tech-Branche sponsert in diesem Themengebiet nicht nur einzelne Forscher (Roth und Kearns sind z.B. Amazon Scholars) und wissenschaftliche Tagungen wie die FAccT, sondern stellt auch sogenannte *Ethicists* ein, die in Unternehmen die gesellschaftliche Verträglichkeit der Produkte im Entwicklungsprozess sicherstellen sollen (Metcalf et al. 2019).

Der Diskursraum und seine zugehörigen Praktiken des Engineerings, des Ausrichtens, Sorgens und Debattierens trägt durchaus Züge von dem, was Petra Gehring für den Bereich der Bioethik als Realexperiment der Rechtspolitik

22 Die American Computing Machinery ist der größte nordamerikanische Fachverband der Computer Science und hat 2019 die Schirmherrschaft über die Konferenzreihe übernommen, die 2018 zunächst als FAT* Conference gestartet war. Fairness, Accountability und Transparency waren schon zuvor auf mehreren Workshops ein Thema. Auf größeren Konferenzen der Computer Science unterstreicht aber die Zusammenlegung zu einer eigenen Konferenz die gewachsene Bedeutung des Themengebietes.

beschrieben hat. Wie im Fall der Bioethik hallt der Ruf nach *AI Ethics* nicht nur durch die Massenmedien, Bildungseinrichtungen, Forschungsbetriebe und ganze Industrien, sondern auch in den »Vorhöfen und Foren legislativ tätiger oder Gesetzgebung zumindest erwägender Politik« (Gehring 2016: 144). Ebenso wie im Fall der Bioethik zeigt sich AI Ethics praktisch als ein »Mittelding aus Sachverständigeneinlassung, Meinung und Entscheidungsvorschlag, Beiträge zu Genehmigungsverfahren zu leisten, zur freiwilligen Selbstverwaltung von Körperschaften, Verbänden, Forschungseinrichtungen, zur parlamentarischen Arbeit und zum Regierungshandeln« (Gehring 2016: 146). Man kann AI Ethics als Mittel sehen, härtere Regulierungen zu umgehen (Floridi 2021), als Zwischenlösung auf dem Weg zu Regulierungen (Robles Carillo 2020) oder auch als Konkretisierungshilfe für die Rechtssprechung (Surden 2020). Der Rechtspolitik mag KI-Ethik ähnlich wie die Bioethik als Testfeld dienen, auf dem Normierungen abgewägt und deren Akzeptanz von Seiten verschiedener gesellschaftlicher Akteur:innen vorgeführt werden kann (Gehring 2016). Das Verhältnis zur Regulierung freilich ist komplex (Hilgendorf 2020), es sollte jedoch deutlich sein, dass man sich mit dem Paradigma des ›value alignment‹ auf einer politischen Spielwiese tummelt.

Der Diskurs der *AI Ethics* ist freilich breiter als XAI. Jedoch wird letzteres als ein entscheidendes Mittel angesehen, um die Gesellschaftsverträglichkeit von KI zu gewährleisten. Im Umkehrschluss erhält die XAI-Forschung eine ethische und soziale Zwecksetzung: XAI dient der Sicherung des ›value alignment‹ von KI. Diese dritte instrumentale Zwecksetzung beginnt interessanterweise bei der gleichen Problemkonzeption, wie es für die technisch-methodische XAI-Forschung typisch ist. Als Kern des Problems wird die schwärze von Kisten angesehen:

> »Such technologies are frequently described as a ›black box‹, capable of producing powerful results, but with little ability on the part of their creators to understand exactly how and why they make the decisions they do.« (Hern 2018)

Entsprechend schreibt man die ungewollten Nebeneffekte des Einsatzes von KI, die rassistischen und sexistischen Effekte, pauschal ›den Algorithmen‹ zu (Kasperkevic 2015). Die argumentative Verkettung verläuft hier wie folgt: die neuen ML-Systeme sind mächtig, aber opak. Um Fragen nach Fairness, Bias und Verantwortbarkeit adressieren zu können, brauche es mehr Erklärbarkeit und Transparenz dieser Systeme: »These explanations are important to ensure algorithmic fairness, identify potential bias/problems in the training data, and to ensure that the algorithms perform as expected« (Gilpin et al. 2018). Eine Einschränkung erhält diese allgemeine Valorisierung von XAI durch den Blick auf bestimmte Anwendungsbereiche, nämlich solche, in denen das menschliche Unvermögen die Maschine zu verstehen, zum gesellschaftlichen Problem wird, insbesondere dann, wenn Anwender:innen von KI-Systemen das Verhalten der Maschine gegenüber anderen rechtferti-

gen müssen. Konkreter wird die ethische-soziale Dienlichkeits-Zuschreibung selten. Ebenso vage bleibt die Zuschreibung dahingehend, unter welchen Bedingungen XAI überhaupt wie, für wen und für was hilfreich sein kann:

> »AI applications can have great societal impact, improving our societies and building a better world. Explainable AI can facilitate our greater adoption of AI applications by empowering us to address important issues like fairness, bias, verifiability, safety, and accountability.« (Kamath/Liu 2021: 10f.)

In den ethischen Richtlinien zu KI finden sich ähnliche pauschale Zuschreibungen dieser Dienlichkeit. Exemplarisch sei hier aus der *Recommendation on the Ethics of Artificial Intelligence* der UNESCO zitiert, die im November 2021 veröffentlicht und von allen 193 Mitgliedstaaten angenommenen wurde:

> »Transparency and explainability relate closely to adequate responsibility and accountability measures, as well as to the trustworthiness of AI systems.« (UNESCO 2021: 22)

Wie mehrere Meta-Reviews dieser ethischen KI-Richtlinien gezeigt haben, wird *explainability* (oder ein damit verwandtes Konzept wie *transparency* oder *interpretability*) als zentral für die KI-Entwicklung gesetzt (Hagendorff 2020; Morley et al. 2020; Jobin et al. 2019). Auch wenn Erklärbarkeit in diesen Richtlinien häufig selbst als Prinzip oder Grundwert deklariert wird, ist sie doch i.d.R als eine Art proto-ethischer Faktor angesehen, d.h. Erklärbarkeit wird über ihre Dienlichkeit für die Bewahrung und/oder Förderung der anderen ethischen Grundwerte valorisiert (Tsamados et al. 2022).

Es bleibt allerdings unklar, ob XAI allgemein als dienlich angesehen wird oder nur für bestimmte Fälle (welche?). Wird Erklärbarkeit als notwendige oder gar hinreichende oder doch nur als förderliche Bedingung für den Schutz und die Förderung der ethischen und sozialen Grundwerte verstanden? Zudem bleibt auszuhandeln, was unter den gesetzten Werten überhaupt jeweils zu verstehen ist, sowohl den ethischen und sozialen Werten auf der einen Seite als auch Erklärbarkeit/Transparenz auf der anderen Seite. Höherstufig wäre außerdem zu klären, an welchen Kriterien oder Maßstäben sich die Festlegung dieser Punkte überhaupt orientieren und wer hierfür zuständig sein soll. Meinem Eindruck nach geht derzeit zu wenig Energie in die Klärung dieser Punkte, während zugleich sehr viel in die (technische) Umsetzung des ›value alignment‹ investiert wird. Durch diese Diskrepanz klafft eine Lücke zwischen den Prinzipien der Richtlinien und der Praxis des ›Engineering‹ auf. Die rhetorisch-diskursive omnipräsente Inwertsetzung von XAI und die zahlreichen XAI-Techniken stehen einander unvermittelt gegenüber.

Auf die klaffende Lücke zwischen der Deklaration von »high-level principles« (Hickok 2021: 41) und der Umsetzung dieser Prinzipien in die Praxis wurde mehrfach hingewiesen. Häufig zieht man als Konsequenz aus dieser Diskrepanz den Schluss, ethische Richtlinien seien nichts weiter als zahnlose Tiger und AI Ethics seien gar insgesamt unnütz (Schwartz 2004; Popescu 2016; McNamara et al. 2018; Rességuier/Rodrigues 2020; Munn 2022). Die Richtlinien und weiteren Bemühungen in dem Bereich dienten höchstens der Tech-Industrie dazu, den Anschein einer Auseinandersetzung mit gesellschaftlichen Fragen zu wahren. Mit diesem *Ethics-Washing* versuche man härtere rechtliche Regulierungen vorzubeugen (Floridi 2021: 620; Wagner 2018; Yeung et al. 2020; Bietti 2020) und sich gegen tieferbohrende Fragen der Öffentlichkeit zu immunisieren (Mittelstadt 2019: 501).

Doch aus der berechtigten Sorge vor einer Verflachung der Ethik durch unternehmerisches Marketing (Bietti 2020) lässt sich noch ein alternativer Schluss ziehen: dass die eigentliche ethische Arbeit erst nach dem Setzen von Grundwerten beginnt. Man sollte nicht missverstehen, was ethische Kodizes überhaupt leisten können und was nicht. Im besten Fall dienen sie der Verständigung über und Kodifizierung von grundlegenden Prinzipien und Werten, denen ein Kollektiv seine Handlungen und Entscheidungen gegenüber verpflichtet. Diese Dokumente können jedoch nur ein kleiner Baustein einer Auseinandersetzung mit möglichen ethischen und sozialen Konsequenzen von technologischen Entwicklungen sein. Dass sie aus sich heraus nicht unbedingt eine hinreichende Motivation darstellen, sich de facto moralisch zu verhalten, ist klar und ein altbekanntes Problem, da typischerweise verschiedene Interessen in Konflikt miteinander geraten (Fehige/Wessels 1998). Folglich entstehen neue Arbeitspakete und weiterer Klärungsbedarf: Welche Pflichten sollte wer gegenüber wem haben? Welche Anreize lassen sich für welche Fälle schaffen? Welche Motive greifen? Mit welchen Interessen steht ein ethisches ›value alignment‹ im Konflikt? Welche fördert es? Welche übersieht es?

Zu dem Motivationsbedarf kommt eine Interpretationsaufgabe. Werte sind per se unterbestimmt und daher interpretationsbedürftig. Die eigentliche Herausforderung liegt somit darin, diese zusätzlichen Aufgaben institutionell zu integrieren und dies auf eine je nach Fall/Bereich angemessene und rechtfertigbare Weise. Für die Übersetzung von Grundwerten in technische Anforderungen im Designprozess gibt es in der Fachliteratur verschiedene Heuristiken aus dem Bereich des Value-Sensitive Designs sowie des Responsible Research and Innovation (van de Poel 2016; Simon 2017; Hallensleben et al. 2020). Zudem bietet eine ganze Reihe von »frameworks, tools, and checklists« (Hickok 2021: 41) eine Orientierungshilfe für die Prozesse dieser Umsetzung, die spezifischer darauf eingehen, welche Stakeholder und Akteur:innen eigentlich wann und für welche Fragen einbezogen werden sollten. Wichtig ist es den Status dieser theoretischen Heuristiken zu beachten. Diese sind Reflexionswerkzeuge, sie bieten Hilfe zur Selbsthilfe. Sie liefern Orientierungsmittel, mit denen man sich selbst in einem gegebenen

Fall eine Orientierung verschaffen kann (Hubig 2007). Der Form nach handelt es sich nicht um algorithmisches Wissen, welches sich mit geschickten technischen Mitteln automatisieren lässt. Ebenso wenig sollten die Kodizes mit Kochrezepten verwechselt werden, die dann gut funktionieren, wenn die Kochenden auf bewährte Routinen und Mittel zurückgreifen können. Die Herausforderung für eine ethische Orientierung von KI besteht gerade darin, nur in begrenztem Maße auf tradiertes Wissen und Erfahrungen zurückgreifen zu können. Die Richtlinien sollten eher mit Kompass und Karte verglichen werden, die auf hoher See (wo es sonst relativ wenig äußere Orientierung gibt) helfen können, seinen Weg zu finden. Sie geben dabei das Ziel der Reise nicht vor, sie präferieren keinen bestimmten Pfad, aber sie können die eigenen Entscheidungen unterstützen.

Mit diesen Überlegungen sollte klar geworden sein, dass sich die Dienlichkeit von XAI-Techniken für ethische und soziale Zwecke nur im Zusammenhang mit Antworten auf die oben genannten Fragen seriös konkretisieren lässt. In diesem Sinne schlage ich vor, die Dienlichkeit von XAI für ethische und soziale Zwecke als einen echten Paradigmenwechsel zu verstehen, der ganz andere Anforderungen stellt, als es die Routinen, Werkzeuge und Ansätze der Normalwissenschaft der XAI-Forschung erahnen lassen. Hinzu kommt, dass nicht nur die Angemessenheit erprobter Mittel und Ansätze zur Diskussion stehen sollte, sondern ebenfalls die Frage der Spielregeln: Wann braucht es überhaupt eine ethische Ausrichtung von KI-Systemen, wann nicht? Worin besteht ein guter Weg, ein KI-System an ethischen und sozialen Grundwerten zu orientieren? Wer sollte hierzu eine konkrete Vorstellung entwickeln? Wer sollte sie umsetzen? Woran kann man erkennen, ob diese Umsetzung gelungen ist? Nur im Licht der Antworten auf diese Fragen lässt sich erkennen, ob XAI in einem bestimmten Fall nützlich oder gar notwendig sein kann oder nicht. Folgend demonstriere ich an Beispielen wie verwickelt diese Fragen mit der Einschätzbarkeit über die Güte von XAI für ethische und soziale Zwecke sind.

4.1 Minimierung von Diskriminierungsrisiken

Bei der Frage nach der Vereinbarkeit von KI mit gesellschaftlichen Grundwerten und Normen geht es offenkundig um normative Fragen. Diese werden in ihrer Normativität überwiegend zu wenig beachtet. Ob etwas normativ gesehen richtig oder falsch ist, ist eine Frage der Anerkennung, nicht der Erkenntnis. Es gehört zum Selbstverständnis von Demokratien, dass normative Fragen nicht dogmatisch vorgegeben werden, sondern gesellschaftlich ausgehandelt werden können. Die meisten normativen Fragen sind i.d.R. nicht thematisch; Gesellschaften ziehen ihre normative Orientierung aus ihrer Kultur, ihrer Tradition, bestehenden Moral- und Rechtssystemen usw. Interessant ist nun, dass mit dem Einsatz von KI-Systemen eine Reihe spezifischer Aushandlungsfragen im Raum stehen. Dabei können Fälle

in ihrem normativen Gewicht und ihrer Komplexität sehr verschieden sein. Ein für solche Einschätzungsfragen simples Beispiel ist der ›rassistische Seifenspender‹, auf den der Softwareentwickler Chukwuemeka Afigbo in einem Twitter-Post vom 16.08.2017 aufmerksam gemacht hatte. Aufgrund der Einstellung des Lichtsensors spendete der Seifenspender farbigen Händen keine Seife, wohl aber weißen Papiertüchern. Afigbo kommentierte: »If you have ever had a problem grasping the importance of diversity in tech and its impact on society, watch this video« (Afigbo 2017). Eine Diskriminierung im sozialen Sinne liegt vor, wenn ungerechtfertigterweise Gleiche ungleich behandelt werden oder Ungleiche gleich (Beck et al. 2019). Entscheidend ist, dass diese Diskriminierung nicht gerechtfertigt ist. Viele Antidiskriminierungsgesetzte verbieten eine Diskriminierung aufgrund geschützter Merkmale wie Rasse, Alter, Geschlecht, Sexualität und Religion (Kolleck/Orwat 2020). Dass in diesem Fall eine solche ungerechtfertigte Ungleichbehandlung vorliegt, dürfte unstrittig sein. Wir können davon ausgehen, dass es nicht im Interesse der Hersteller und Betreiber dieses Seifenspenders lag, bestimmte Personengruppen vom Gebrauch des Seifenspenders auszuschließen. Entsprechend lässt sich dieser Fall als Beispiel für eine Fehleinstellung der Hardware ansehen, der unbeabsichtigte negative Effekte hervorbrachte. Diese hätten leicht verhindert werden können, hätte man den Seifenspender ausreichend getestet. Wie Krishnan argumentiert, steht dieser Fall für ein KI-induziertes Diskriminierungsproblem, für dessen Abhilfe keine XAI-Techniken gebraucht werden (Krishnan 2020).[23] Da sogenannte *biases* an sehr vielen Stellen aus verschiedenen Gründen und Ursachen im Lebenszyklus von KI-Systemen auftreten können, scheint es insgesamt plausibel, dass verschiedene Strategien helfen können, um Diskriminierungsrisiken zu minimieren und zu managen (Klier 2024).

Fallspezifische Antworten auf folgende Fragen können diese Einschätzung informieren: (1) Bestehen Diskriminierungsrisiken? Gegenüber wem? (2) Welche Interessen haben die beteiligten Stakeholder? (3) Bestehen Werte- und/oder Interessenskonflikte? (4) Wie kann wer von diesen Risiken und Interessen Kenntnis erlangen? (5) Welche Mittel der Risikobewältigung und des Umgangs mit Konflikten stehen zur Verfügung und können hier (von wem) als angemessen angesehen werden?

4.2 Entscheidungen nachvollziehbar machen?

Wie komplex diese Fragen werden können, zeigt der viel diskutierte Fall um den Einsatz des *Correctional Offender Management Profiling for Alternative Sanctions* (COM-

23 Wie plausibel diese Einschätzung ist, hängt davon ab, was man unter Erklärbarkeit versteht oder was XAI-Techniken leisten können. Für das Testen, wann der Seifenspender Seife gibt, braucht man die Kiste sicherlich nicht zu öffnen, aber freilich braucht es eine gewisse Kenntnis von dem System, um die Einstellungen zu berichtigen.

PAS) Systems, das in den USA im Justizwesen zum Einsatz kommt. Die jüngere Debatte hatte eine investigative Recherche von ProPublica ausgelöst, die die Berechnungen von COMPAS als rassistisch eingeschätzt haben (Angwin et al. 2016). Die Software von Northpointe Inc. (heute Equivant) liefert Vorhersagen zur Risikoeinschätzung und wird als Unterstützung benutzt, wenn Bewährungsstrafen festgelegt werden, die Höhe der Kaution bestimmt wird oder auch andere Urteile von Richter:innen gefällt werden. Hierzu ordnet COMPAS die Verurteilten in verschiedene Risikogruppen ein (hohes, mittleres, geringes) bzgl. ihrer Rückfallwahrscheinlichkeit und nutzt dazu sozioökonomische Daten, Daten zur familiären Situation sowie Persönlichkeitsdaten, z.B. Stresstoleranz (Kolleck/Orwat 2020). Die Daten wurden teils aus den Akten, teils aus Befragungen bezogen. Rasse/Hautfarbe werden nicht als Daten einbezogen, es besteht jedoch der Verdacht, dass diese über sogenannte Proxyvariablen (z.B. den Wohnsitz oder das Alter) dennoch in der Berechnung des Risiko-Scores einen gewichtigen Unterschied machen.

An die Veröffentlichung des Berichts von ProPublica schloss sich eine breite Debatte an. Flores et al. bemängelten statistische Fehler und ein mangelndes Verständnis für Daten auf Seiten der investigativen Journalist:innen und behaupteten, dass ›der Algorithmus‹ gegenüber Schwarzen und Weißen gleich kalibriert sei (Flores et al. 2016). Corbett-Davies et al. betonten, dass es einen trade-off im Systemdesign zwischen einer Optimierung auf das Kriterium der öffentlichen Sicherheit und Fairnesskriterien gibt, die auf eine faire Behandlung von Individuen unabhängig ihrer Zugehörigkeit zu bestimmten Gruppen (z.B. Rasse) setzt (Corbett-Davies et al. 2017). Hamilton ergänzte die Diskussion um den Aspekt, dass Frauen von COMPAS hinsichtlich der Rückfallwahrscheinlichkeit benachteiligt würden, da das System ihnen ein zu hohes Risiko zuweisen würde (Hamilton 2019). Insgesamt führte diese Debatte zu verschiedenen Einsichten. Zunächst, dass die verschiedenen Bewertungen von COMPAS auf verschiedenen Definitionen und damit Kriterien von Fairness zurückzuführen sind (Hedden 2021). Des Weiteren, dass sich algorithmische Systeme jeweils nur auf ein Fairness-Kriterium ausrichten lassen und nicht mehreren zugleich gerecht werden können (Chouldechova/Roth 2020; Corbett-Davies et al. 2023). Damit reicht es nicht einfach aus, zu zeigen, dass bestimmte KI-Systeme fair sind, sondern man muss spezifizieren, auf welches Fairness-Kriterium die Systeme eingestellt sind und sollte begründen, warum man dieses Kriterium für diesen Fall für angemessen hält. Folglich entstehen Anerkennungsfragen und es stellt sich die Frage, wer diese normative Festlegung eigentlich zu treffen hat bzw. welche Akteur:innen hieran beteiligt sein sollten.

Die Debatte hat außerdem gezeigt, wie wichtig die institutionelle Einbettung bei der Bewertung dieser Fragen ist. Rudin et al. argumentieren grundsätzlich gegen den Einsatz proprietärer Software im Justizwesen, denn das Geschäftsgeheimnis verhindere hier, dass Betroffene oder unabhängige Dritte bzw. die Gesellschaft überhaupt eine Bewertung der Software vornehmen können (Rudin et al. 2020). Da-

mit wechselt die Diskussion auf eine andere Ebene: Ist es für staatliche Institutionen bzw. ist in bestimmten Anwendungsbereichen der Gebrauch von proprietärer Software überhaupt legitim, wenn durch diese tief in das Leben von Individuen eingegriffen wird? Rudin et al. schlagen vor, transparente Systeme zu nutzen, die von unabhängigen Expertengruppen hinsichtlich geprüft werden können, z.B. hinsichtlich verschiedener Fairnesskriterien (Rudin et al. 2020). Zur Diskussion steht damit, ob und wem ein Anspruch darauf zukommen sollte, KI-Systeme in bestimmten Anwendungsbereichen prüfen und nachvollziehen zu können (Alfrink et al. 2023). Es hängt von der Einschätzung ab, was genau es heißen soll, dass Entscheidungen nachvollziehbar sind bzw., dass KI-Systeme überprüfbar sind und welche Mittel man an dieser Stelle für angemessen hält. Auch steht zur Frage, ob ein Einsatz proprietärer Software wie COMPAS mit dem Prinzip des ordnungsgemäßen Verfahrens (›due process‹) im Einklang steht.

Walmsley argumentiert, dass aus Sicht der Betroffenen ein entscheidender Unterschied zwischen der Klassifikation durch COMPAS und einer Klassifikation, die von Anwält:innen, Richter:innen oder Zeug:innen vorgenommen wird, besteht, denn sie können COMPAS nicht »cross-examine […] in the same way they could with a human witness« (Walmsley 2021: 592). Es geht also nicht nur darum, die Software in bestimmter Weise überhaupt nachvollziehbar zu machen (entweder, indem man nur transparente Software für bestimmte Fälle nutzt, oder durch XAI), sondern auch um die Frage, für wen Software in welcher Weise und in welchem Umfang nachvollziehbar sein muss (Burrell 2016). Die meisten Angeklagten werden kein Expertenwissen des Machine Learning aufweisen. Was kann es heißen, dass die Entscheidungen für sie dennoch nachvollziehbar sind? In wessen Hand legt man diese Übersetzungsarbeit? Gehört sie künftig zum Beruf der Anwältin? Oder sollte man Angeklagten eine Art Recht auf Beratung durch KI-Expert:innen zugestehen? Wenn ja, woher kommt diese Expertise und wer bezahlt sie? Die COMPAS-Debatte zeigt deutlich, dass man bei bestimmten Anwendungsbereichen zunächst eine ganze Reihe von Fragen der institutionellen Einbettung von KI klären sollte, bevor man sinnvoll einschätzen kann, welche XAI-Techniken für diese Fälle überhaupt zuverlässig sind.

4.3 Unternehmerische Techniken des Value Alignments

Es sind viele Fälle denkbar, bei denen es wie im Seifenspender-Beispiel ausreicht, die Systeme vor Gebrauch hinreichend zu testen, um Diskriminierungsrisiken zu minimieren. Andere Fälle, vielleicht all solche Anwendungen, die im Rahmen des EU AI Acts in die Gruppe der hohen Risiken gehören, ähneln eher dem Beispiel von COMPAS. Da hier die Anwendung von KI sowohl individuelle Grundrechte als auch das Gemeinwohl tangiert, spricht vieles dafür, höhere (regulative) Anforderungen

für ihren Einsatz zu stellen, etwa die Forderung nach Nachvollziehbarkeit der Entscheidungen.

Darüber hinaus gibt es wenigstens eine weitere Gruppe von Fällen, bei denen Stereotype reproduziert werden und Diskriminierungsrisiken zu verzeichnen sind, die jedoch nicht per se in die Gruppe der hohen Risiko-Anwendungen gehören. Hierzu gehören all jene Social Media und Onlineangebote, in denen KI-gestützt Inhalte moderiert oder generiert werden oder in andere Entscheidungen in Kommunikation, Information und Konsum eingreifen. Diese Fälle scheinen mir normativ höhere Unsicherheiten mit sich zu bringen, weil weniger klar ist, welche politische, soziale und moralische Brisanz man ihnen zuschreiben soll. Wäre es wünschenswert, dass der Gesetzgeber auch für diese Fälle bestimmte Auflagen vorsieht? Aus welchen guten Gründen? Ungeachtet dieser normativen Unsicherheiten haben Tech-Unternehmen längst diverse technische Strategien der Handhabe etabliert. Diese Fälle scheinen mir weitestgehend unter dem Radar der XAI-Community zu fliegen (Finke et al. 2022). Sie sind aber allein darum schon aufschlussreich, da man es hiermit, aus technischer Sicht, mit dem gleichen Problem wie bei COMPAS zu tun hat: »The problem of achieving agreement between our true preferences and the objective we put into the machine is called the *value alignment problem*: the values or objectives put into the machine must be aligned with those of the human« (Russell/Norvig 2021, Hervorh. im Original). Wenn man alle Fälle, vom Seifenspender über COMPAS bis zur Content Moderation als Problem des ›value alignment‹ konzeptualisiert, liegt es nahe, die in diesem Bereich etablierten technischen Lösungen auch für die anderen Fälle einzusetzen.

Eine simple technische Lösung für unerwünschte Nebeneffekte sind eingebaute Filter. Drei Jahre, nachdem der Post von Alciné zur Missklassifikation von Google Photos viral ging, berichtete das Tech Magazine *Wired* über einen von ihnen durchgeführten Test, der darauf schließen ließ, dass Google bestimmte Kategorien aus ihrem Klassifikationssystem schlicht gelöscht hat, um der rassistischen Zuordnung Herr zu werden. *Wired* hatte mit 40.000 Fotos die Klassifizierung getestet, darunter Fotos von Affen, die aber nicht als solche kategorisiert wurden. Google habe darauf bestätigt, die Kategorien ›gorilla‹, ›chimp‹, ›chimpanzee‹ und ›monkey‹ aus ihrem System gelöscht zu haben (Simonite 2018; Vincent 2018).

Filterfunktionen gibt es in verschiedenen Komplexitätsstufen. Eine weitere Strategie ist eine bestimmte Form der Mensch-Maschine-Arbeitsteilung, wofür das Einrichten des Large-Language-Models von OpenAI, auf dem ChatGPT-4 basiert, illustrativ ist. Informationen hierzu finden sich auf der Webseite von OpenAI sowie in dem *Technical Report* zu ChatGPT-4. OpenAI hat sich selbst das Ziel gestellt, seine Chatbots in Einklang mit drei Werten – den drei H's – zu entwickeln: Die Ausgaben der Bots sollten »honest«, »helpful« und »harmless« sein (Open AI 2023). Ehrlichkeit (»honesty«) ist begrifflich eine missverständliche Wahl, da sie sich auf den inneren Zustand eines Subjektes bezieht. Sachlich gemeint ist, dass die Aussagen

korrekt/wahr sein sollen (Heitzinger/Woltran 2024: 143). In seinem technischen Bericht gibt OpenAI zwei Schritte an, in denen sie das Large-Language-Modell erstellt haben: ein »pre-training« und ein »fine-tuning« des Modells. Aus beiden Lernwegen wurde dann das eigentliche Modell gebildet (Open AI 2023). Das pre-training fand automatisiert statt (überwachtes Lernen eines artifiziellen neuronalen Netzes), auf einer sehr großen Datenmenge, die zum Teil aus frei zugänglichen Internetquellen und andererseits aus lizenzierten Datenquellen stammen (OpenAI et al. 2024). Auf diesem Wege wurde das Modell in die Lage gebracht, zu bestimmen, welches Wort (token) sinnvollerweise auf ein vorheriges Wort erscheinen soll – die statistische Grundlage dieser Generierung von Texten. Der zweite Schritt dient dem Zweck des ›value-alignment‹. ChatGPT soll nur solche Text generieren, die mit ›unseren‹ Grundwerten vereinbar sind. Der Schritt des Fine-Tunings ist folglich eine Kontrollmaßnahme:

> »Then, we ›fine-tune‹ these models on a more narrow dataset that we carefully generate with human reviewers who follow guidelines that we provide them. Since we cannot predict all the possible inputs that future users may put into our system, we do not write detailed instructions for every input that ChatGPT will encounter. Instead, we outline a few categories in the guidelines that our reviewers use to review and rate possible model outputs for a range of example inputs. Then, while they are in use, the models generalize from this reviewer feedback in order to respond to a wide array of specific inputs provided by a given user.« (Open AI 2023)

Da die Daten, auf denen das Modell trainiert wurde, unerwünschte Inhalte beinhalten (›toxic content‹), lernt das Modell ebendiese (z.B. Beschimpfungen, diskriminierende Aussagen, Verleumdungen des Holocausts). Ohne die manuell-maschinelle Kontrollmaßnahme würde das Modell diese Inhalte ungefiltert ausgeben, wenn es die Token-Wahrscheinlichkeiten nahe legen würde. Die Kontrollmaßnahme, von Unternehmensseite auch als Sicherheitsmaßnahme verstanden (›do no harm‹), baut auf einer spezifischen Arbeitsteilung von Maschine und Mensch auf. Da die Maschine nicht allein beurteilen kann, ob generierter Text mit den drei H's vereinbar ist oder nicht, braucht es eine Beurteilung durch Menschen. Da die Menschen aber nicht den ganzen Datensatz prüfen können und praktisch nicht jeder denkbare Prompt vorab präventiv eingehegt werden kann, haben die ›human reviewers‹ die Prüfung nur auf einer vergleichsweise viel geringeren Datenmenge durchgeführt, damit diese Prüfung überhaupt für menschliche Arbeitskräfte (auch für große Gruppen) durchführbar ist. Die ›humans‹ waren dafür zuständig, vier vom Modell gebildete Antworten auf einen Prompt in eine Reihenfolge zu bringen (von der besten zur schlechtesten). Dieses Ranking wurde dann als Feedback in das Modell zurückgegeben, wodurch dieses seine Zuordnungen von Antworten zu Prompts

weiter optimierte, nach der Strategie des »reinforcement learning from human feedback (RLHF)« (Heitzinger/Woltran 2024: 143).

Für den Prüfvorgang waren den Arbeitskräften Instruktionen vorgegeben, worauf sie achten sollten, z. B. auf die Konformität mit geltendem Gesetz: »do not complete requests for illegal content« (Open AI 2023). Sodann gab es auch ›höherstufige‹ Anweisungen wie »avoid taking a position on controversial topics« (Open AI 2023). Außerdem wurden wöchentliche Treffen des OpenAI Management mit den Prüfer:innen anberaumt, um fragwürdige Fälle und offene Fragen zu klären: »This iterative feedback process is how we train the model to be better and better over time« (Open AI 2023).

Mittlerweile ist ein eigener Geschäftsbereich dadurch entstanden, KI so zu kalibrieren, dass sie sich nicht unerwünscht ›verhält‹. An diese etablierte Praxis lassen sich zahlreiche normative Fragen anschließen. OpenAI wirft in ihrem Blog selbst die Frage auf, wer eigentlich die Werte, Instruktionen und Standards bestimmen sollte, an denen diese Systeme orientiert werden (Open AI 2023). Es wäre auch zu diskutieren, ob durch solche normative Festlegungen privater Akteur:innen die kulturelle Hegemonie des Westens in bestimmten Bereichen des Internets weiter gefestigt wird (Goffi et al. 2021; Siapera 2022; Shahid/Vashistha 2023).

Dazu gesellen sich Fragen zu den Arbeitsbedingungen der menschlichen KI-Unterstützer:innen (Perrigo 2023; Hagendorff 2022); und dies umso mehr, weil die Arbeit der Einhegung dieser Software wohl auf Dauer stattfinden muss, wie die zahlreichen Beispiele des »Jailbreaks« zeigen, in denen die normativen Regeln des Systems umgegangen wurden (Xie et al. 2023). Hinzu kommen weitere Fragen zur Vereinbarkeit von Software dieser Art mit anderen hier nicht diskutierten gesellschaftlichen Grundwerten, wie Umweltschutz und Nachhaltigkeit (George et al. 2023; Khowaja et al. 2024). Hier ließe sich diskutieren, ob der hohe CO_2-Ausstoß gepaart mit dem Potential von Large-Language-Models, wie ChatGPT-4, in diversen elektronischen Geräten und Services als KI-Aufrüstung verbaut zu werden, nicht mit dem öffentlichen Interesse einhergeht, über die Umweltkosten dieses technischen Fortschritts aufgeklärt zu werden.

5. Über die Verschiedenheit der Zwecksetzungen

Die XAI-Forschung täte gute daran, ernsthaft über den Zweck nachzudenken, für den XAI entwickelt und optimiert wird (Krishnan 2020; Colaner 2022; Alpsancar et al. 2024). Freiesleben und König (2023) zufolge, kommen die meisten Beiträge in der XAI-Forschung ohne eine Zwecksetzung ihrer Tools aus, sie reflektieren nicht auf die Dienlichkeit ihrer Technologie, sondern kümmern sich abstrakt um deren Optimierung. Ich vermute, dieser Eindruck trifft zu einem großen Teil zu, er übersieht allerdings, dass sehr wohl Zwecke genannt und Dienlichkeiten gesetzt werden

– dies passiert allerdings erstens pauschal und generalisierend und zweitens eher rein diskursiv, d.h. ohne forschungspraktischen Bezug auf die Arbeit an den XAI-Technologien. Entsprechend rege ich nicht nur dazu an, über die Zwecke zu reflektieren und eine konkrete Zwecksetzung anzugeben, sondern ebenso über die kategoriale und forschungspraktische Verschiedenheit der Zwecksetzungen Klarheit zu gewinnen.

Im Paradigma der epistemischen Güte von ML ist der Sinn des Forschens innerhalb der ML-Community lokalisiert. Hier entwickeln einige Expert:innen für andere Expert:innen bessere Einsichten in bestimmte formale Zusammenhänge von ML-Systemen, um diese optimieren zu können und Fehler zu beheben. Die Relevanz der Erklärungen ergibt sich hier aus der Forschung der Community selbst: Welche Systeme sollen in welcher Hinsicht besser verstanden werden. Die Evaluation der XAI-Techniken und deren Maßstäbe ergeben sich aus dieser Zwecksetzung und sollten sich den üblichen methodischen Standards fügen.[24]

Im zweiten Paradigma verändert sich offenkundig die Zusammensetzung der Akteur:innen und Nutzer:innen, und Kontexte werden Teil des Forschungsgegenstandes, die gemäß der methodischen Standards aus der Psychologie, Sozial- und Kulturwissenschaft erforscht werden sollten. Der Methodenkasten erweitert sich disziplinär und die interdisziplinäre Zusammenarbeit wird zum Thema. Außerdem mag es von Bedeutung sein, in einigen Fällen partizipative Verfahren zu integrieren. Die Relevanz und Angemessenheit von Erklärungen lässt sich in diesem Paradigma nicht (allein) aus technischen Überlegungen gewinnen; es braucht folglich einen echten Perspektivwechsel. Mit meinem Kommentar zum XAI-Forschungsprojekt der DARPA wollte ich demonstrieren, dass es wichtig ist, sich über strukturelle Fragen bezüglich der Zuständigkeit von Forschungsfragen und Aufgaben Gedanken zu machen. Dies betrifft ebenfalls die Gewichtung der Beteiligung verschiedener Expertisen, also eine forschungspolitische Frage. Ein Klärungsbedarf besteht auch in konzeptioneller Hinsicht. Die von mir herausgestellte Differenz der ›informatischen‹ und der ›psychologischen‹ Perspektive im XAI-Projekt zeigt, dass praktische Unterschiede darin bestehen, wie man den Erklärungsprozess modelliert. Für eine erfolgreiche interdisziplinäre Zusammenarbeit scheint es hilfreich, sich über die Zielvision zu verständigen. Es ist etwas vollkommen anderes, XAI-Techniken mit

24 Nach Freiesleben und König (2023) ist die Community noch dabei diese auszubilden und befindet sich deswegen noch nicht in einem paradigmatischen Zustand der normalwissenschaftlichen Forschung im Sinne Kuhns. Dies mag der Fall sein, betrifft aber den Kerngedanken meines Arguments nicht – dass die normalwissenschaftliche Forschung eines Paradigmas nicht ohne weiteres auf die normalwissenschaftliche Forschung eines anderen Paradigmas übertragen werden kann und, dass wir es in der XAI-Welt mit wenigstens drei grundsätzlich zu unterscheidenden Paradigmen zu tun haben.

Blick auf das Endziel der Automatisierung zu entwickeln (unsere Erklärungen werden irgendwann so gut, dass alles selbst-evident ist, und dann brauchen wir keine User:innen für die Zwecke der XAI) oder, ob man die Techniken daraufhin optimiert, dass Teams von User:innen und KI-Systemen kollaborieren sollen. Diese Zielvorstellungen implizieren ein anderes Gewicht der Kontextualität des Gebrauchs von KI. Während er im Automatisierungs-Leitbild zu vernachlässigen ist, denn die ideale Automatisierung ist eben unabhängig von Kontexten (so die Ideologie), ist es in der Kollaborations-Vision gerade entscheidend, von welchen Kontexten und konkreten Konstellationen man ausgeht. Je nach Zielvision gilt es andere forschungspraktische Fragen zu berücksichtigen.

Der Diskurs der AI Ethics und das Problem des ›value alignment‹ konstituieren meiner Ansicht nach ein weiteres, drittes Paradigma des Forschens und Entwickelns von XAI, für das im Sinne Kuhns die erprobten Instrumente, Methoden und theoretischen Ansätze der anderen beiden Paradigmen nicht geeignet sein können – weil wir es hier mit einem anderen Typus von Problem zu tun haben, der andere Konzeptionen von Lösungsräumen und -strategien erfordert. Dieses Paradigma ist auch darum besonders, viel weniger im Bereich der Forschung verortet zu sein als ›in der Gesellschaft‹ – in Industrie, Politik, Recht und der darauf bezogenen Forschung und Entwicklung. Hier ist es nicht eine methodische Option, echte Kontexte und Nutzer:innen, z.B. über Feldstudien, Interviews oder partizipative Verfahren, in die (inter-)disziplinäre Forschung einzubeziehen, sondern die Bearbeitung des Problems findet primär in anderen gesellschaftlichen Bereichen als der Wissenschaft statt. Dementsprechend haben wir es in diesem Paradigma noch einmal mit einer grundverschiedenen Konstellation von Akteur:innen und Strukturen, sodann Machtverhältnissen zu tun.

Macht man sich die Verschiedenheit dieser drei Paradigmen bewusst, sollte es überraschen, dass de facto relativ ähnliche Lösungsstrategien des Engineerings für diese verschiedenen Probleme ausprobiert werden und ›laufen‹. Es wäre an der Zeit, das Engagement und die Expertise der Ingenieurswissenschaften, Data Science und Informatik mit anderen Expertisen so zu kombinieren, dass sich der alte Fehler eines ›technological fix‹ (de Bruijn et al. 2022) nicht im Bereich der Entwicklung und Optimierung von XAI wiederholt. Stattdessen könnte man das Feld der XAI diversifizieren, das (informelle) Wissen verschiedener sozialer Stakeholder einbeziehen, eine Debatte über gesellschaftliche Nützlichkeit von XAI und KI anregen und die mit ihrer Nutzung einhergehenden Risiken und Chancen bedacht abwägen, wobei man auch reflektieren sollte, warum, wie und für wen die (vermeintlichen) Zwecke von XAI und KI profitabel/nützlich sind.

Literatur

Adadi, A.; Berrada, M. (2018): Peeking inside the black-box. A survey on explainable artificial intelligence (XAI), in: *IEEE access*, 6, 52138–52160.

Afigbo, C. (2017): Post from @nke_ise, in: Twitter/X: 16.08.2017. [https://twitter.com/nke_ise/status/897756900753891328] (Zugriff: 28.12.2023).

AIAAIC (2021): Facebook labels black men ›primates‹. [https://www.aiaaic.org/aiaaic-repository/ai-algorithmic-and-automation-incidents/facebook-labels-black-men-primates] (Zugriff: 28.12.2023).

Alfrink, K.; Keller, I.; Kortuem, G.; Doorn, N. (2023): Contestable AI by design. Towards a Framework, in: *Minds and Machines*, 33(4), 613–639.

Allhutter, D.; Mager, A.; Cech, F.; Fischer, F.; Grill, G. (2020): Der AMS-Algorithmus. Eine Soziotechnische Analyse des Arbeitsmarktchancen-Assistenz-Systems (AMAS). Endbericht, Wien: Institut für Technikfolgen-Abschätzung (ITA).

Alpsancar, S. (2023): What is AI Ethics? Ethics as means of self-regulation and the need for critical reflection, in: International Conference on Computer Ethics, 1(1), 1–17. [https://soremo.library.iit.edu/index.php/CEPE2023/article/view/227].

Alpsancar, S.; Matzner, T.; Philippi, M. (2024): Unpacking the purposes of explainable AI, in: Arias-Olivia, M.; Pelegrin-Borondo, J.; Murata, K.; Palma, A.M.L.; Ollé Sensé, M. (Hg.), Smart Ethics in the Digital World. Proceedings of the ETHICOMP 2024. 21st International Conference on the Ethical and Social Impacts of ICT, Logroño: Universidad de La Rioja, 31–35. [https://dialnet.unirioja.es/descarga/articulo/9326091.pdf].

Angwin, J.; Larson, J.; Mattu, S.; Kirchner, L. (2016): Machine Bias. There's software used across the country to predict future criminals. And it's biased against blacks, in: ProPublica. [https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing] (Zugriff: 15.06.2024).

Beck, S.; Grunwald, A.; Jacob, K.; Matzner, T. (2019): Künstliche Intelligenz und Diskriminierung. Herausforderungen und Lösungsansätze (Whitepaper), München: Plattform Lernende Systeme.

Bellon, J.; Gransche, B.; Nähr-Wagener, S. (Hg.) (2022): Soziale Angemessenheit. Forschung zu Kulturtechniken des Verhaltens, Wiesbaden: Springer.

Benjamin, R. (2019): Race After Technology. Abolitionist Tools for the New Jim Code, Cambridge/Medford (MA): Polity.

Bietti, E. (2020): From Ethics Washing to Ethics Bashing. A View on Tech Ethics from within Moral Philosophy, in: FAT* '20. Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, New York: Association for Computing Machinery, 210–219.

Biran, O.; Cotton, C. (2017): Explanation and justification in machine learning. A survey, in: *IJCAI-17 Workshop on Explainable AI (XAI)*, 8(1), 8–13.

Blumenberg, H. (1981): Technisierung und Lebenswelt unter Aspekten der Phänomenologie, in: Ders., Wirklichkeiten in denen wir leben. Aufsätze und eine Rede, Stuttgart: Reclam, 7–54.

Borup, M.; Brown, N.; Konrad, K.; Van Lente, H. (2006): The sociology of expectations in science and technology, in: *Technology analysis & strategic management*, 18(3/4), 285–298.

Bradshaw, J.M.; Hoffman, R.R.; Woods, D.D.; Johnson, M. (2013): The Seven Deadly Myths of Autonomous Systems, in: *IEEE Intelligent Systems*, 28(3), 54–61.

Burrell, J. (2016): How the machine thinks. Understanding opacity in machine learning algorithms, in: *Big Data & Society*, 3(1). [doi.org/10.1177/2053951715622512].

Cabitza, F.; Campagner, A.; Malgieri, G.; Natali, C.; Schneeberger, D.; Stoeger, K.; Holzinger, A. (2023): Quod erat demonstrandum? Towards a typology of the concept of explanation for the design of explainable AI, in: *Expert Systems with Applications*, 213, 1–16.

Capel, T.; Brereton, M. (2023): What is Human-Centered about Human-Centered AI? A Map of the Research Landscape, in: CHI '23. Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, New York: Association for Computing Machinery, 1–23.

Caruana, R.; Lou, Y.; Gehrke, J.; Koch, P.; Sturm, M.; Elhadad, N. (2015): Intelligible Models for HealthCare. Predicting Pneumonia Risk and Hospital 30-Day Readmission, in: KDD '15. Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York: Association for Computing Machinery, 1721–1730.

Cath, C.; Wachter, S.; Mittelstadt, B.; Taddeo, M.; Floridi, L. (2018): Artificial intelligence and the ›good society‹. The US, EU, and UK approach, in: *Science and engineering ethics*, 24, 505–528.

Chakrabarti, R.; Sanyal, K. (2020): Towards a ›Responsible AI‹. Can India take the lead?, in: *South Asia Economic Journal*, 21(1), 158–177.

Chouldechova, A.; Roth, A. (2020): A snapshot of the frontiers of fairness in machine learning, in: *Communication of the ACM*, 63(5), 82–89.

Clancey, W.J.; Hoffman, R.R. (2021): Methods and standards for research on explainable artificial intelligence. Lessons from intelligent tutoring systems, in: *Applied AI Letters*, 2(4), 1–8.

Colaner, N. (2022): Is explainable artificial intelligence intrinsically valuable?, in: *AI & Society*, 37(1), 231–238.

Corbett-Davies, S.; Pierson, E.; Feller, A.; Goel, S.; Huq, A. (2017): Algorithmic Decision Making and the Cost of Fairness, in: KDD '17. Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York: Association for Computing Machinery, 797–806.

Corbett-Davies, S.; Gaebler, J.D.; Nilforoshan, H.; Shroff, R.; Goel, S. (2023): The Measure and Mismeasure of Fairness, in: *Journal of Machine Learning Research*, 24(312), 1–117.

Crawford, K. (2021): Atlas of AI. Power, Politics, and the Planetary Costs of Artificial Intelligence, New Haven/London: Yale University Press.

Crawford, K. (2023): Archeologies of Datasets, in: *The American Historical Review*, 128(3), 1368–1371.

Cremers, A.B.; Englander, A.; Gabriel, M.; Hecker, D.; Mock, M.; Poretschkin, M.; Rosenzweig, J.; Rostalski, F.; Sicking, J.; Volmer, J. et al. (2019): Trustworthy Use of Artificial Intelligence. Priorities from a Philosophical, Ethical, Legal and Technological Viewpoint as a Basis for Certification of Artificial Intelligence, Sankt Augustin: Fraunhofer Institute for Intelligent Analysis.

Defense Advanced Research Projects Agency (DARPA) (2016): Broad Agency Announcement. Explainable Artificial Intelligence (XAI), 1–52. [https://www.darpa.mil/program/explainable-artificial-intelligence] (Zugriff: 24.04.2024).

de Graaf, M.M.; Malle, B.F. (2017): How people explain action (and autonomous intelligent systems should too), in: AAAI Fall Symposium Series 2017, Washington [DC]: AAAI Press,19-26.

de Bruijn, H.; Warnier, M.; Janssen, M. (2022): The perils and pitfalls of explainable AI. Strategies for explaining algorithmic decision-making, in: *Government information quarterly*, 39(2), 1–8.

Dignum, V. (2019): Responsible Artificial Intelligence. How to Develop and Use AI in a Responsible Way, Cham: Springer.

Dix, A. (2017): Human–computer interaction, foundations and new paradigms, in: *Journal of Visual Languages Computing*, 42, 122–134.

Doshi-Velez, F.; Kim, B. (2017): Towards a rigorous science of interpretable machine learning, in: arXiv. [https://arxiv.org/abs/1702.08608] (Zugriff: 11.06.2024).

Edwards, P.N. (1997): The Closed World. Computers and the Politics of Discourse in Cold War America, Cambridge (MA): The MIT Press.

Ellmer, M. (2015): Digitale Arbeitsteilung. Amazon Mechanical Turks sozial konstruierte Designmuster und die Steuerung von Human-Computation-Arbeit, in: *Momentum Quarterly*, 4(3), 174–186.

Eubanks, V. (2017): Automating inequality. How high-tech tools profile, police, and punish the poor, New York: St. Martin's Press.

European Commission (2022): Regulatory framework proposal on artificial intelligence. [https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai] (Zugriff: 15.06.2024).

Everingham, M.; Eslami, S.M.A.; Van Gool, L.; Williams, C.K.I.; Winn, J.; Zisserman, A. (2015): The Pascal Visual Object Classes Challenge. A Retrospective, in: *International Journal of Computer Vision*, 111(1), 98–136.

Everingham, M.; Van Gool, L.; Williams, C.K.I.; Winn, J.; Zisserman, A. (2010): The Pascal Visual Object Classes (VOC) Challenge, in: *International Journal of Computer Vision*, 88(2), 303–338.

Fehige, C.; Wessels, U. (1998): Preferences. An introduction, in: Fehige, C.; Wessels, U. (Hg.), Preferences, Berlin/New York: de Gruyter, xx-xliii.

Finke, J.; Horwath, I.; Matzner, T.; Schulz, C. (2022): (De)Coding Social Practice in the Field of XAI. Towards a Co-constructive Framework of Explanations and Understanding Between Lay Users and Algorithmic Systems, in: Degen, H.; Ntoa, S. (Hg.), Artificial Intelligence in HCI, Cham: Springer, 149–160.

Flores, A.W.; Bechtel, K.; Lowenkamp, C.T. (2016): False Positives, False Negatives, and False Analyses. A Rejoinder to ›Machine Bias. There's Software Used Across the Country to Predict Future Criminals. And It's Biased Against Blacks‹, in: *Federal Probation*, 80(2), 38–46.

Floridi, L. (2021): The End of an Era. From Self-Regulation to Hard Law for the Digital Industry, in: *Philosophy & Technology*, 34(4), 619–622.

Floridi, L.; Cowls, J.; Beltrametti, M.; Chatila, R.; Chazerand, P.; Dignum, V.; Luetge, C.; Madelin, R.; Pagallo, U.; Rossi, F. et al. (2018): AI4People – An Ethical Framework for a Good AI Society. Opportunities, Risks, Principles, and Recommendations, in: *Minds and Machines*, 28(4), 689–707.

Fortes, P.R.B. (2020): Paths to digital justice. Judicial robots, algorithmic decision-making, and due process, in: *Asian Journal of Law and Society*, 7(3), 453–469.

Freiesleben, T.; König, G. (2023): Dear XAI community, we need to talk! Fundamental misconceptions in current XAI research, in: Longo, L. (Hg.), Explainable Artificial Intelligence. First World Conference, Cham: Springer, 48–65.

Friedman, B.; Nissenbaum, H. (1996): Bias in computer systems, in: *ACM Transactions on Information Systems (TOIS)*, 14(3), 330–347.

Gehring, P. (2016): Ethik als Realexperiment von Rechtspolitik. Zum Dreiecksverhältnis von Bioethik, Recht und Politik, in: *Jahrbuch für Wissenschaft und Ethik*, 20(1), 143–162.

George, A.S.; George, A.H.; Martin, A.G. (2023): The Environmental Impact of AI. A Case Study of Water Consumption by Chat GPT, in: *Partners Universal International Innovation Journal*, 1(2), 97–104.

Gilpin, L.H.; Bau, D.; Yuan, B.Z.; Bajwa, A.; Specter, M.; Kagal, L. (2018): Explaining Explanations. An Approach to Evaluating Interpretability of Machine Learning, in: arXiv. [https://arxiv.org/abs/1806.00069] (Zugriff 14.06.2024).

Goffi, E.R.; Colin, L.; Belouali, S. (2021): Ethical Assessment of AI Cannot Ignore Cultural Pluralism. A Call for Broader Perspective on AI Ethic, in: *Arribat-International Journal of Human Rights Published by CNDH Morocco*, 1(2), 151–175.

Goodman, B.; Flaxman, S. (2017): European Union regulations on algorithmic decision-making and a »right to explanation«, in: *AI magazine*, 38(3), 50–57.

Guidotti, R.; Monreale, A.; Ruggieri, S.; Turini, F.; Giannotti, F.; Pedreschi, D. (2018): A survey of methods for explaining black box models, in: *ACM computing surveys (CSUR)*, 51(5), 1–42.

Gunning, D. (2016): Explainable Artificial Intelligence (XAI). DARPA/I2O. [https://sites.cc.gatech.edu/~alanwags/DLAI2016/(Gunning)%20IJCAI-16%20DLAI%20WS.pdf] (Zugriff: 14.05.2024).

Gunning, D.; Aha, D. (2019): DARPA's Explainable Artificial Intelligence (XAI) Program, in: *AI Magazine*, 40(2), 44–58.

Gunning, D.; Stefik, M.; Choi, J.; Miller, T.; Stumpf, S.; Yang, G.-Z. (2019): XAI – Explainable artificial intelligence, in: *Science Robotics*, 4(37). [doi.org/10.1126/scirobotics.aay7120].

Gunning, D.; Vorm, E.; Wang, J.Y.; Turek, M. (2021): DARPA's explainable AI (XAI) program. A retrospective, in: *Applied AI Letters*, 2, 1–11.

Gyevnar, B.; Ferguson, N.; Schafer, B. (2023): Bridging the transparency gap. What can explainable AI learn from the AI Act?, in: Gal, K.; Nowé, A.; Nalepa, G.J.; Fairstein, R.; Rădulescu, R. (Hg.), ECAI 2023. 26th European Conference on Artificial Intelligence, Amsterdam u.a.: IOS Press, 964–971.

Hagendorff, T. (2020): The Ethics of AI Ethics. An Evaluation of Guidelines, in: *Minds and Machines*, 30(1), 99–120.

Hagendorff, T. (2022): Blind Spots in AI Ethics, in: *AI and Ethics*, 2(4), 851–867.

Hallensleben, S.; Hustedt, C.; Fetic, L.; Fleischer, T.; Grünke, P.; Hagendorff, T.; Hauer, M.; Hauschke, A.; Heesen, J.; Herrmann, M. et al. (2020): From Principles to Practice. An interdisciplinary framework to operationalise AI ethics (Technischer Bericht), Gütersloh: Bertelsmann Stiftung.

Hamilton, M. (2019): The sexist algorithm, in: *Behavioral sciences & the law*, 37(2), 145–157.

Harrison, S.; Tatar, D.; Sengers, P. (2007): The three paradigms of HCI, in: Alt. Chi. Session at the SIGCHI Conference on human factors in computing systems San Jose, California, USA, New York: Association for Computing Machinery, 1–18.

Hedden, B. (2021): On statistical criteria of algorithmic fairness, in: *Philosophy and Public Affairs*, 49(2), 209–231.

Heitzinger, C.; Woltran, S. (2024): A Short Introduction to Artificial Intelligence. Methods, Success Stories, and Current Limitations, in: Werthner, H.; Ghezzi, C.; Kramer, J.; Nida-Rümelin, J.; Nuseibeh, B.; Prem, E.; Stanger, A. (Hg.), Introduction to Digital Humanism. A Textbook, Cham: Springer, 135–149.

Hern, A. (2018): Google's solution to accidental algorithmic racism. Ban gorillas, in: *The Guardian*, 12.01.2018. [https://www.theguardian.com/technology/2018/jan/12/google-racism-ban-gorilla-black-people?CMP=share_btn_url] (Zugriff: 28.12.2023).

Hernández-Orallo, J. (2019): Gazing into Clever Hans machines, in: *Nature Machine Intelligence*, 1(4), 172–173.

Hickok, M. (2021): Lessons learned from AI ethics principles for future actions, in: *AI and Ethics*, 1(1), 41–47.

High Level Expert Group (2019): A definition of AI. Main capabilities and disciplines. [https://digital-strategy.ec.europa.eu/en/library/definition-artificial-intelligence-main-capabilities-and-scientific-disciplines] (Zugriff: 15.06.2024).

Hilgendorf, E. (2020): Robotik, Künstliche Intelligenz, Ethik und Recht. Neue Grundlagenfragen des Technikrechts, in: Hentschel, A.; Hornung, G.; Jandt, S. (Hg.), Mensch – Technik – Umwelt. Verantwortung für eine sozialverträgliche Zukunft, Baden-Baden: Nomos, 545–564.

Hoffman, R.R.; Miller, T.; Clancey, W.J. (2022): Psychology and AI at a Crossroads. How Might Complex Systems Explain Themselves?, in: *American Journal of Psychology*, 135(4), 365–378.

Hoffman, R.R.; Miller, T.; Klein, G.; Mueller, S.T.; Clancey, W.J. (2023): Increasing the Value of XAI for Users. A Psychological Perspective, in: *KI-Künstliche Intelligenz*, 37, 237–247.

Hu, M. (2020): Cambridge Analytica's black box, in: *Big Data & Society*, 7(2), 1–6.

Hubig, C. (2007): Die Kunst des Möglichen II. Grundlinien einer dialektischen Philosophie der Technik, Band 2. Ethik der Technik als provisorische Moral, Bielefeld: transcript.

Hüllermeier, E. (2020): Towards Analogy-Based Explanations in Machine Learning, in: Torra, V.; Narukawa; Y.; Nin, J.; Agell, N. (Hg.), Modeling Decisions for Artificial Intelligence, Cham: Springer, 205–217.

Ipeirotis, P.G. (2010): Analyzing the Amazon Mechanical Turk Marketplace, in: *XRDS: Crossroads, The ACM magazine for students*, 17(2), 16–21.

Jobin, A.; Ienca, M.; Vayena, E. (2019): Artificial Intelligence. The global landscape of AI ethics guidelines, in: *Nature Machine Intelligence*, 1, 389–399.

Kamath, U.; Liu, J. (2021): Explainable Artificial Intelligence. An Introduction to Interpretable Machine Learning, Cham: Springer.

Kaminski, A. (2010): Technik als Erwartung. Grundzüge einer allgemeinen Technikphilosophie, Bielefeld: transcript.

Kaminski, A. (2014): Sein und »als«. Notizen zu einer Denkfigur in Heideggers Werk, in: *Filozofija i društvo*, 25(4), 21–28.

Kaminski, M.E. (2019): The right to explanation, explained, in: *Berkeley Technology Law Journal*, 34(1), 189–218.

Kasperkevic, J. (2015): Google says sorry for racist auto-tag in photo app, in: *The Guardian*, 01.07.2015. [https://www.theguardian.com/technology/2015/jul/01/google-sorry-racist-auto-tag-photo-app?CMP=share_btn_url] (Zugriff: 05.06.2024).

Kearns, M.; Roth, A. (2019): The Ethical Algorithm. The Science of Socially Aware Algorithm Design, New York: Oxford University Press.

Khowaja, S.A.; Khuwaja, P.; Dev, K.; Wang, W.; Nkenyereye, L. (2024): ChatGPT Needs SPADE (Sustainability, PrivAcy, Digital divide, and Ethics) Evaluation. A Review, in: *Cognitive Computation*. [doi.org/10.1007/s12559-024-10285-1].

Kim, M.; Kim, S.; Kim, J.; Song, T.-J.; Kim, Y. (2024): Do stakeholder needs differ? Designing stakeholder-tailored Explainable Artificial Intelligence (XAI) interfaces, in: *International Journal of Human-Computer Studies*, 181, 1–12.

Klier, M. (2024): Grundlagen zu Bias & Fairness in KI-Systemen. [https://bias-and-fairness-in-ai-systems.de/grundlagen/] (Zugriff: 31.05.2024).

Kolleck, A.; Orwat, C. (2020): Mögliche Diskriminierung durch algorithmische Entscheidungssysteme und maschinelles Lernen. Ein Überblick, Berlin: Büro für Technikfolgen-Abschätzung beim Deutschen Bundestag (TAB).

Kraus, T.; Ganschow, L. (2022): Anwendungen und Lösungsansätze erklärbarer Künstlicher Intelligenz, in: Hartmann, E.A. (Hg.), Digitalisierung souverän gestalten II, Berlin/Heidelberg: Springer, 38–50.

Krishnan, M. (2020): Against interpretability. A critical examination of the interpretability problem in machine learning, in: *Philosophy & Technology*, 33(3), 487–502.

Kuhn, T.S. (1976[1962]): Die Struktur wissenschaftlicher Revolutionen. Bd. 2, Frankfurt a.M.: Suhrkamp.

Kulesza, T.; Burnett, M.; Wong, W.-K.; Stumpf, S. (2015): Principles of Explanatory Debugging to Personalize Interactive Machine Learning, in: IUI '15. Proceedings of the 20th International Conference on Intelligent User Interfaces, New York: Association for Computing Machinery, 126–137.

Langer, M.; Oster, D.; Speith, T.; Hermanns, H.; Kästner, L.; Schmidt, E.; Sesing, A.; Baum, K. (2021): What do we want from Explainable Artificial Intelligence (XAI)? A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research, in: *Artificial Intelligence*, 296. [10.1016/j.artint.2021.103473].

Lapuschkin, S.; Wäldchen, S.; Binder, A.; Montavon, G.; Samek, W.; Müller, K.-R. (2019): Unmasking Clever Hans predictors and assessing what machines really learn, in: *Nature Communications*, 10(1), 1–8.

Laux, J.; Wachter, S.; Mittelstadt, B. (2023): Three Pathways for Standardisation and Ethical Disclosure by Default under the European Union Artificial Intelligence Act, in: *Computer Law & Security Review*, 53, 1–11.

Lebovitz, S.; Lifshitz-Assaf, H.; Levina, N. (2022): To Engage or Not to Engage with AI for Critical Judgments. How Professionals Deal with Opacity When Using AI for Medical Diagnosis, in: *Organization Science*, 33(1), 126–148.

Lenk, H. (1982): Zur Sozialphilosophie der Technik, Frankfurt a.M.: Suhrkamp.

Lepri, B.; Oliver, N.; Letouzé, E.; Pentland, A.; Vinck, P. (2018): Fair, Transparent, and Accountable Algorithmic Decision-making Processes, in: *Philosophy & Technology*, 31(4), 611–627.

Lipton, Z.C. (2018): The Mythos of Model Interpretability. In machine learning, the concept of interpretability is both important and slippery, in: *Queue*, 16(3), 31–57.

López-Martínez, F.; Núñez-Valdez, E.R.; García-Díaz, V.; Bursac, Z. (2020): A case study for a big data and machine learning platform to improve medical decision support in population health management, in: *Algorithms*, 13(4), 1–19.

Mahoney, M.S. (2011): Histories of computing, Cambridge (MA): Harvard University Press.

Marabelli, M.; Newell, S.; Page, X. (2018): Algorithmic Decision-Making in the US Healthcare Industry. [https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3262379] (Zugriff: 15.06.2024).

Markus, A.F.; Kors, J.A.; Rijnbeek, P.R. (2021): The role of explainability in creating trustworthy artificial intelligence for health care. A comprehensive survey of the terminology, design choices, and evaluation strategies, in: *Journal of biomedical informatics*, 113, 1–11.

McNamara, A.; Smith, J.; Murphy-Hill, E. (2018): Does ACM's code of ethics change ethical decision making in software development?, in: ESEC/FSE 2018. Proceedings of the 2018 26th ACM joint meeting on European software engineering conference and symposium on the foundations of software engineering, New York: Association for Computing Machinery, 729–733.

Meske, C.; Abedin, B.; Klier, M.; Rabhi, F. (2022): Explainable and responsible artificial intelligence, in: *Electronic Markets*, 32(4), 2103–2106.

Metcalf, J.; Moss, E.; boyd, d. (2019): Owning Ethics. Corporate logics, Silicon Valley, and the Institutionalization of Ethics, in: *Social Research. An International Quarterly*, 86(2), 449–476.

Miller, T. (2019): Explanation in artificial intelligence. Insights from the social sciences, in: *Artificial Intelligence*, 267, 1–38.

Miller, T.; Hoffman, R.R.; Amir, O.; Holzinger, A. (2022): Special issue on Explainable Artificial Intelligence (XAI), in: *Artificial Intelligence*, 307. [10.1016/j.artint.2022.103705].

Mittelstadt, B.D. (2019): Principles alone cannot guarantee ethical AI, in: *Nature Machine Intelligence*, 1, 501–507.

Mittelstadt, B.; Russell, C.; Wachter, S. (2019): Explaining Explanations in AI, in: FAT* '19. Proceedings of the Conference on Fairness, Accountability, and Transparency, New York: Association for Computing Machinery, 279–288.

Mohammed, S.; Brandner, L.T.; Burtscher, F.; Hallensleben, S.; Harmouch, H.; Hauschke, A.; Heesen, J.; Hildebrandt, S.; Hirsbrunner, S.D.; Keselj, J. et al. (2024): A Data Quality Glossary. [https://zenodo.org/records/10474880] (Zugriff: 14.06.2024).

Morley, J.; Floridi, L.; Kinsey, L.; Elhalal, A. (2020): From What to How. An Initial Review of Publicly Available AI Ethics Tools, Methods and Research to Translate Principles into Practices, in: *Science and Engineering Ethics*, 26, 2141–2168.

Mueller, S.T.; Hoffman, R.R.; Clancey, W.J.; Emrey, A.; Klein, G. (2019): Explanation in Human-AI Systems. A Literature Meta-Review, Synopsis of Key Ideas and Pu-

blications, and Bibliography for Explainable AI. [https://apps.dtic.mil/sti/citations/AD1073994] (Zugriff: 17.06.2024).

Munn, L. (2022): The uselessness of AI ethics, in: *AI and Ethics*, 3, 869–877.

Nannini, L.; Balayn, A.; Smith, A.L. (2023): Explainability in AI Policies. A Critical Review of Communications, Reports, Regulations, and Standards in the EU, US, and UK, in: FAccT '23. Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, New York: Association for Computing Machinery, 1198–1212.

Norberg, A. (1996): Changing computing. The computing community and DARPA, in: *IEEE Annals of the History of Computing*, 18(2), 40–53.

O'Neil, C. (2016): Weapons of math destruction. How big data increases inequality and threatens democracy, New York: Crown.

Open AI. (2023): How should AI systems behave, and who should decide? [https://openai.com/index/how-should-ai-systems-behave/] (Zugriff: 17.06.2024).

OpenAI; Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F.L.; Almeida, D.; Altenschmidt, J.; Altman, S. et al. (2024): GPT-4 Technical Report, in: arXiv. [https://arxiv.org/abs/2303.08774] (Zugriff: 17.06.2024).

Panigutti, C.; Hamon, R.; Hupont, I.; Fernandez Llorca, D.; Fano Yela, D.; Junklewitz, H.; Scalzo, S.; Mazzini, G.; Sanchez, I.; Soler Garrido, J. et al. (2023): The role of explainable AI in the context of the AI Act, in: FAccT '23. Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, New York: Association for Computing Machinery, 1139–1150.

Perrigo, B. (2023): Exclusive. OpenAI Used Kenyan Workers on Less Than $2 Per Hour to Make ChatGPT Less Toxic, in: *Time Magazine*, 18.01.2023. [https://time.com/6247678/openai-chatgpt-kenya-workers/] (Zugriff: 10.06.2024).

Petrick, E.R. (2020): Building the Black Box. Cyberneticians and Complex Systems, in: *Science, Technology, & Human Values*, 45(4), 575–595.

Pfeiffer, J.; Gutschow, J.; Haas, C.; Möslein, F.; Maspfuhl, O.; Borgers, F.; Alpsancar, S. (2023): Algorithmic Fairness in AI, in: Business & Information Systems Engineering, 65, 209–222.

Popescu, A.-I. (2016): In brief. Pros and Cons of corporate codes of conduct. *Journal of Public Administration, Finance and Law*, 9(9), 125–130.

Ras, G.; van Gerven, M.; Haselager, P. (2018): Explanation methods in deep learning. Users, values, concerns and challenge, in: Escalante, H.J.; Escalera, S.; Guyon, I.; Baró, X.; Güçlütürk, Y.; Güçlü, U.; van Gerven, M. (Hg.), Explainable and Interpretable Models in Computer Vision and Machine Learning, Cham: Springer, 19–36.

Rességuier, A.; Rodrigues, R. (2020): AI ethics should not remain toothless! A call to bring back the teeth of ethics, in: *Big Data & Society*, 7(2), 1–5.

Ribeiro, M.T.; Singh, S.; Guestrin, C. (2016): »Why Should I Trust You?«. Explaining the Predictions of Any Classifier, in: KDD '16. Proceedings of the 22nd ACM

SIGKDD International Conference on Knowledge Discovery and Data Mining, New York: Association for Computing Machinery, 1135–1144.

Ribera, M.; Lapedriza García, À. (2019): Can we do better explanations? A proposal of user-centered explainable AI, in: IUI Workshops 2. [https://api.semanticscholar.org/CorpusID:84832474] (Zugriff: 17.06.2024).

Roberts, H.; Cowls, J.; Morley, J.; Taddeo, M.; Wang, V.; Floridi, L. (2021): The Chinese approach to artificial intelligence. An analysis of policy, ethics, and regulation, in: *AI & society*, 36, 59–77.

Robles Carrillo, M. (2020): Artificial intelligence. From ethics to law, in: *Telecommunications Policy*, 44(6), 1–16.

Rohlfing, K.J.; Leonardi, G.; Nomikou, I.; Rączaszek-Leonardi, J.; Hüllermeier, E. (2020): Multimodal Turn-Taking. Motivations, Methodological Challenges, and Novel Approaches, in: *IEEE Transactions on Cognitive and Developmental Systems*, 13(2), 260–271.

Rohlfing, K.J.; Cimiano, P.; Scharlau, I.; Matzner, T.; Buhl, H.M.; Buschmeier, H.; Esposito, E.; Grimminger, A.; Hammer, B.; Häb-Umbach, R. et al. (2021): Explanation as a Social Practice. Toward a Conceptual Framework for the Social Design of AI Systems, in: *IEEE Transactions on Cognitive and Developmental Systems*, 13(3), 717–728.

Rudin, C. (2019): Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, in: *Nature Machine Intelligence*, 1(5), 206–215.

Rudin, C.; Wang, C.; Coker, B. (2020): The Age of Secrecy and Unfairness in Recidivism Prediction, in: *Harvard Data Science Review*, 2(1), 1–53.

Russell, S.; Norvig, P. (4. Auflage 2021): Artificial Intelligence. A modern approach, London: Pearson.

Samek, W.; Müller, K.-R. (2019): Towards Explainable Artificial Intelligence, in: Samek, W.; Montavon, G.; Vedaldi, A.; Hansen, L.K.; Müller, K.-R. (Hg.), Explainable AI. Interpreting, Explaining and Visualizing Deep Learning, Cham: Springer, 5–22.

Samek, W.; Wiegand, T.; Müller, K.-R. (2017): Explainable artificial intelligence. Understanding, visualizing and interpreting deep learning models, in: arXiv. [https://arxiv.org/abs/1708.08296] (Zugriff: 11.06.2024).

Schwartz, M.S. (2004): Effective corporate codes of ethics. Perceptions of code users, in: *Journal of business ethics*, 55, 321–341.

Shahid, F.; Vashistha, A. (2023): Decolonizing Content Moderation. Does Uniform Global Community Standard Resemble Utopian Equality or Western Power Hegemony?, in: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, 1–18.

Siapera, E. (2022): AI content moderation, racism and (de) coloniality, in: *International journal of bullying prevention*, 4(1), 55–65.

Simon, J. (2017): Value Sensitive Design and Responsible Research and Innovation, in: Hansson, S.O. (Hg.), The Ethics of Technology. Methods and Approaches, London/New York: Rowman & Littlefield, 219–236.

Simonite, T. (2018): When It Comes to Gorillas, Google Photos Remains Blind, in: *Wired*, 11.01.2018. [https://www.wired.com/story/when-it-comes-to-gorillas-google-photos-remains-blind/] (Zugriff: 28.12.2023).

Sokol, K.; Flach, P. (2020): One Explanation Does Not Fit All, in: *KI-Künstliche Intelligenz*, 34(2), 235–250.

Sovrano, F.; Sapienza, S.; Palmirani, M.; Vitali, F. (2022): Metrics, explainability and the European AI act proposal, in: *J*, 5(1), 126–138.

Stamboliev, E. (2023): Proposing a Postcritical AI Literacy. Why We Should Worry Less about Algorithmic Transparency and More about Citizen Empowerment, in: *Media Theory*, 7(1), 202–232.

Starke, C.; Baleis, J.; Keller, B.; Marcinkowski, F. (2022): Fairness perceptions of algorithmic decision-making. A systematic review of the empirical literature, in: *Big Data & Society*, 9(2), 1–16.

Strohmeier, S. (2020): Algorithmic decision making in HRM, in: *Encyclopedia of electronic HRM*, 1, 54–59.

Surden, H. (2020): Ethics of AI Law. Basic Questions, in: Dubber, M.D.; Pasquale, F.; Das, S. (Hg.), The Oxford Handbook of Ethics of AI, New York: Oxford University Press, 719–736.

Tsamados, A.; Aggarwal, N.; Cowls, J.; Morley, J.; Roberts, H.; Taddeo, M.; Floridi, L. (2022): The ethics of algorithms. Key problems and solutions, in: *Ai & Society*, 37(1), 215–230.

Tworek, H. (2019): Social Media Councils, in: Owen, T.; Docquir, P.F.; Donovan J.; Etlinger, S.; Fay, R.; Girard, M.; Gorwa, R.; Kimmelman, G.; Klonick, K.; McDonald, S.M. et al. (Hg.), Models for Platform Governance. A CIGI Essay Series, Waterloo: Centre for International Governance, 97–102.

UNESCO (2021): Recommendation on the Ethics of Artificial Intelligence. [https://unesdoc.unesco.org/ark:/48223/pf0000381137] (Zugriff: 7.06.2023).

van de Poel, I. (2016): An Ethical Framework for Evaluating Experimental Technology, in: *Science and Engineering Ethics*, 22(3), 667–686.

van der Hoven, J.; Manders-Huits, N. (2020): Value-sensitive design, in: Olsen, J.K.B.; Pedersen, S.A.; Hendricks, V.F. (Hg.), The Ethics of Information Technologies, London: Routledge, 329–332.

Vermaas, P.E. (2010): Focussing Philosophy of Engineering. Analyses of Technical Functions and Beyond, in: van de Poel, I.; Goldberg D. (Hg.), Philosophy and Engineering. An Emerging Agenda, Dordrecht u.a.: Springer Netherlands, 61–73.

Vilone, G.; Longo, L. (2021): Notions of explainability and evaluation approaches for explainable artificial intelligence, in: *Information Fusion*, 76, 89–106.

Vincent, J. (2018): Google ›fixed‹ its racist algorithm by removing gorillas from its image-labeling tech, in: *The Verge*, 12.01.2018. [https://www.theverge.com/2018/1/12/16882408/google-racist-gorillas-photo-recognitionalgorithm-ai] (Zugriff: 28.12.2023).

Vogelmann, F. (2019): Transparency's Trap. Problems of an Unquestioned Norm, in: Berger, S.; Owetschkin, D. (Hg.), Contested Transparencies, Social Movements and the Public Sphere. Multi-Disciplinary Perspectives, Cham: Springer, 35–54.

Wagner, B. (2018): Ethics As An Escape From Regulation. From ›Ethics-Washing‹ To Ethics-Shopping?, in: Bayamlioglu, E.; Baraliuc, I.; Janssens, L.A.W.; Hildebrandt, M. (Hg.), BEING PROFILED:COGITAS ERGO SUM. 10 Years of Profiling the European Citizen, Amsterdam: Amsterdam University Press, 84–89.

Walmsley, J. (2021): Artificial intelligence and the value of transparency, in: *AI & Society*, 36(2), 585–595.

Wang, D.; Yang, Q.; Abdul, A.; Lim, B.Y. (2019): Designing Theory-Driven User-Centric Explainable AI, in: CHI '19. Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, New York: Association for Computing Machinery, 1–15.

Weber, M. (1992): Wissenschaft als Beruf, in: Ders., Wissenschaft als Beruf 1917/1919. Politik als Beruf 1919 [Studienausgabe der Max Weber-Gesamtausgabe, Band I/17], Tübingen: Mohr Siebeck, 71–112.

Xie, Y.; Yi, J.; Shao, J.; Curl, J.; Lyu, L.; Chen, Q.; Xie, X.; Wu, F. (2023): Defending ChatGPT against jailbreak attack via self-reminders, in: *Nature Machine Intelligence*, 5(12), 1486–1496.

Yeung, K.; Howes, A.; Pogrebna, G. (2020): AI Governance by Human Rights–Centered Design, Deliberation, and Oversight. An End to Ethics Washing, in: Dubber, M.D.; Pasquale, F.; Das, S. (Hg.), The Oxford Handbook of Ethics of AI, New York: Oxford University Press, 76–106.

Zarsky, T. (2016): The trouble with algorithmic decisions. An analytic road map to examine efficiency and fairness in automated and opaque decision making, in: *Science, Technology, & Human Values*, 41(1), 118–132.

Verteilte Anerkennung

Wie künstliche Intelligenz die Theorie der Anerkennung verändert

Natalia Juchniewicz

Abstract: *The article argues that, based on contributions from social philosophy, the relationships between humans and AI and the effects of AI on social life can also be conceptualized with the necessary differentiation. This is done in continuation of discussion approaches that refer to A. Honneth's theory of recognition. It is shown how this theory makes it possible to include non-human persons as ›partial persons‹ and in emotionally important relationships; and why a (classical) argument from reciprocity and consciousness is not a necessary element of the recognition relationship itself. Recognition theory, for its part, must be expanded to include collective intelligence and distributed action. However, this also leads to a new challenge for ethics and social research: that of distributed responsibility.*

Keywords: *artificial intelligence (AI); recognition; partial persons; distributed responsibility; distributed recognition*

Bei den Blickwinkeln, unter denen soziale Beziehungen im 21. Jahrhundert analysiert werden, kommt der Digitalisierung der Technologien, die wir nutzen, und den entsprechenden sozialen Praktiken zweifellos eine zentrale Rolle zu. Der zunehmenden Bedeutung künstlicher Intelligenz (KI), die häufig Bedenken in normativ sensiblen Bereichen, im Zusammenhang mit Entscheidungsprozessen und vorgeschlagenen Lösungen hervorruft, ist dabei ein besonderes Gewicht zuzumessen. »Almost every day, the news media report on achievements of AI helping to overcome a great variety of real-world problems.« (Peeters et al. 2021: 217) Sosehr KI es beschleunigt, verschiedene Probleme zu diagnostizieren, und sosehr sie schnellere Entscheidungsfindung durch die Analyse von Big Data und *machine computing* ermöglicht, gilt es doch zugleich sich dessen bewusst zu sein, dass die Bestimmung dessen, was gerecht, akzeptabel oder förderlich für das Wohlergehen der Menschen

ist, nicht allein auf der Bewertung von Daten beruhen sollte. Dies darf schon allein deswegen nicht außer Acht kommen, weil zahlreiche Aspekte des menschlichen Daseins über die messbaren Indikatoren hinausgehen, die KI zur Verfügung stehen. Auch im Hinblick auf das soziale Leben kann KI, wenn sie ihrer Technik gemäß unsere Identitäten auf der Grundlage *vergangener* Aktivitäten bestimmt und definiert, uns nur auf Vorbestimmtes beschränken und ließe wenig Raum für menschliche Veränderungen oder die Begegnung mit Neuem aus dem Internet. Was KI nur vermag, ist, aufgrund früherer Nutzeraktivitäten etwas zu projizieren, und basiert dabei auf der Voraussetzung, dass die menschliche Identität fest (van Dijck 2013) und vorhersehbar ist.

> »With their predictive capabilities and relentless nudging, ubiquitous but imperceptible, AI systems can shape our choices and actions easily and quietly. This is not necessarily detrimental. […] It may foster social interaction and cooperation. [But it] may also exert its influencing power beyond our wishes or understanding, undermining our control on the environment, societies, and ultimately our choices, projects, identities, and lives.« (Taddeo/Floridi 2018: 752)

Mein Fokus in diesem Artikel liegt darauf, welche Antworten aus der Sozialphilosophie abzuleiten sind, um den Veränderungen und Herausforderungen Rechnung zu tragen, die in unseren intersubjektiven Beziehungen aufgrund der zunehmenden Präsenz künstlicher Intelligenz in unseren alltäglichen Aktivitäten sich ergeben. Im Letzten zielen die vorgestellten Überlegungen darauf ab, die umfassend gewordenen Auswirkungen des zunehmenden Einflusses und der Präsenz von KI in unserer gesellschaftlichen Welt zu erforschen. Dazu suche ich zu erläutern, warum das Konzept der *Anerkennung* eine entscheidende Rolle auch bei der Reflexion der Beziehung zwischen Mensch und KI spielen sollte und wie gegenwärtige Theorien der Anerkennung nicht-menschliche Entitäten wie ›KI‹ in ihren Rahmen einbeziehen könnten.

Unzweifelhaft scheint, dass die zunehmende Präsenz von KI verschiedenste Bedenken und Hoffnungen hervorruft, Verschiedenstes, das mit unserer Privatsphäre, unseren Beschäftigungen, Entscheidungsprozessen, der Erkenntnis von Strukturen (›Muster‹) und der Vorausschau zukünftiger Ereignisse zusammenhängt. All dies beeinflusst in der Tat, wie wir uns als Individuen wahrnehmen, und es wirkt sich nicht weniger auf die Anerkennung aus, die wir in sozialen Beziehungen erfahren. Unsere alltäglichen Interaktionen mit künstlicher Intelligenz, die durch verschiedene Technologien wie Browser, soziale Medien und Chatbots an uns kommt, prägen denn viele normative Konsequenzen innerhalb unseres Lebens. Von einfachen Algorithmen bis hin zum *machine learning* und *deep learning*, diese Technologien formen zahlreiche Gegebenheiten, mit denen wir alltäglich zu tun haben. So sind wir, und das ist ein entscheidender Punkt, bei der Konstruktion unseres Selbstgefühls nun heute erheblich auf die Vermittlung durch Elemente von KI angewiesen. Dies führt

zu einem nicht zu unterschätzenden Aspekt des sozialen Problems der Anerkennung – dass unsere erfahrene Anerkennung nun wesentlich durch die Beziehungen zwischen Menschen und *KI* geprägt werden könnte. Diese Veränderungen auf jeden Fall mitreflektieren zu müssen, gilt umso mehr, als Individuen nicht unbedingt bemerken, dass KI keine menschliche Person ist, da sie menschenähnliches Verhalten zeigen kann (nicht nur den Turing-Test besteht, sondern auch in verschiedenen sozialen Situationen zufrieden stellende Interaktionen liefert; s. Borenstein/Pearson 2010).

Der nachfolgende Artikel ist in drei Abschnitte gegliedert. Im ersten werde ich zunächst allgemein die soziale Bedeutung von KI erklären und dabei den aktuellen Ansatz, wie die aus den Beziehungen zwischen Mensch und KI sich ergebenden sozialen Probleme bestimmt werden können, vorblickend unter Verwendung der Sprache der Anerkennungstheorie erläutern. Im zweiten Abschnitt werde ich die Theorie der Anerkennung einführen und mich insbesondere auf ihre ausgereifteste Form fokussieren, wie sie sich von Axel Honneth entwickelt findet. Näheres zum Potenzial dieses Ansatzes lässt sich sodann durch Analyse der Kommentare von Arto Laitinen und Heikki Ikäheimo gewinnen; beider Einsichten heben wichtige Probleme und Begriffe hervor, die in Bezug auf Erörterungen von KI und deren Rolle in den menschlichen Praktiken verwendet werden können. Laitinens Perspektive ermöglicht eine Diskussion über verschiedene Formen der Anerkennung, einschließlich eines weitreichenden Verständnisses, das einbefassend verschiedenartige Objekte veranschlagen kann. Ikäheimos Sichtweise hilft zu erkennen, dass auch schon generell in verschiedenen Dimensionen der Anerkennung eine Wechselseitigkeit der Anerkennung nicht immer als Voraussetzung steht. Davon ausgehend kann ein Ansatz dahin entwickelt werden, bei dem statt des einen menschlichen Parts in Anerkennungsrelationen nun eine Instanz von KI stünde. Konkret lässt sich so in einem nächsten Schritt darlegen, wie das auch in KI-Interaktionen zentral bestehende Problem, welches Bewusstsein die Subjekte der Anerkennung haben oder entwickeln, angegangen werden kann. Dazu werde ich dem Vorschlag von Nolen Gertz über ›den unerlaubten Schritt‹ (»illicit move«) in der Hegelschen Tradition der Anerkennung folgen. Im dritten und letzten Abschnitt schließlich werde ich einen Ansatz vorschlagen, der auf Studien zu dem Neuen in dem Verhältnis von Mensch und KI angewandt werden kann. Dabei stütze ich mich das Konzept der »partial persons« (Hirvonen 2017), was auf KI Anwendung finden kann, das generative Modell der Anerkennung (Laitinen) – Anerkennung als performative Formung bzw. Veränderung von Normativem und Sozialem – und werde mich der Betonung der Notwendigkeit anschließen, einen neuen Ansatz zur Beschreibung von verteilter Agency und von Verantwortung in den Relationen zwischen Mensch und KI zu entwickeln (Peeters et al. 2021; Taddeo/Floridi 2018). Im Ganzen werde ich dafür argumentieren, dass die Entwicklung verantwortlicher Haltungen zu den Wirklichkeiten wie Möglichkeiten von KI und ihren sich entwickelnden Formen nicht ausschließlich auf KI selbst fo-

kussieren sollte – auf ihre Gestaltung, die Algorithmen oder die Nutzung –, sondern auch auf die Menschen. Sind doch die Menschen nicht nur Nutzer/Anwender der Technologie, sondern Individuen, die ihr Empfinden von Wert, sozialer Anerkennung und Respekt in ihren Beziehungen damit gestalten.

Wenn es um die Anerkennung zwischen Menschen und KI geht, gibt es zwei mögliche Perspektiven zu bedenken. Die erste konzentriert sich auf die Analyse von Anerkennungstheorien und Argumente, die es möglich machen, KI und die Beziehung dazu in ihren Rahmen einzubeziehen. Die zweite Perspektive beinhaltet die Analyse von KI selbst und ihrer Einwirkungen im Sozialen, um zu bestimmen, wann und warum ein Problem von Anerkennung entsteht (s. Jacobs 2024). Die erste Perspektive lässt sich theoretisch und die zweite empirisch nennen. Im vorliegenden Artikel wird mein Schwerpunkt hauptsächlich auf der theoretischen Perspektive liegen. Ich werde versuchen zu zeigen, dass innerhalb der bestehenden Anerkennungstheorien ein Raum vorhanden ist, der es uns ermöglicht, auch über KI zu diskutieren und neue Aspekte der Anerkennung zu beleuchten, die berücksichtigt werden müssen.

1. KI, Gesellschaft und Gesichtserkennungstechnologien

Im Jahr 2005 prognostizierte Ray Kurzweil, dass an einem bestimmten, nicht allzu fernen Zeitpunkt Maschinen leistungsfähiger sein würden als menschliche Intelligenz und an dem wir, um effektiver denken zu können, mit den Maschinen verschmelzen müssen. Bekannt wurde dies unter dem zum Schlagwort gewordenen ›Singularität‹, mit dem der Einschnitt in allem bisherigen menschheitlichen Geschehen, von da an das Eintreten in ein fundamental neues Seinsstadium gekennzeichnet sein sollte. (Kurzweil 2005) Dass diese Vision einer allgemeinen künstlichen Intelligenz verwirklicht ist, ist immer noch eher spekulativ als auf echten Technologien basierend. Doch ist offenkundig, dass die zeitgenössische KI bereits eine eminente Rolle spielt in verschiedenen Bereichen menschlicher Entscheidungspraxis, wie zum Beispiel in allem, was mit Logistik zu tun hat, bei Versicherungen, im Gesundheitswesen und auch in der Bildung (Coeckelbergh 2020: 3). Zusätzlich hat sich der Gedanke, künstliche Intelligenz in eine spezifische Technologie zu integrieren, wie etwa in eine Robotermaschine oder maschinelle Assistenz, um menschliche Tätigungen und Verhalten zu verbessern, zum Konzept des Internets der Dinge (*Internet of Things*) entwickelt. Dieser Wandel zielt auf die Schaffung eines Netzwerks von jeweiligen, miteinander verbundenen technologischen Umgebungsfeldern, die Verbindung über das Internet haben können. Diese Perspektive hat den Einsatz von KI erweitert, die nun in allen digitalen Technologien implementiert werden kann, genauer gesagt, in allen Geräten mit Internetzugang (und dies nimmt inzwischen schon lange den bei weitem höchsten Anteil des Datenauf-

kommens im Internet ein). So nimmt künstliche Intelligenz gegenwärtig zahlreiche Formen an, und unsere Interaktionen und Beziehungen zu ihr sind, auch wo im Einzelnen nicht erkannt, zu integralen Bestandteilen unseres täglichen Lebens geworden.

In der vorhandenen Literatur wurde der soziale Einfluss von KI hauptsächlich aus einer normativen Perspektive analysiert. Der Fokus liegt dabei auf den möglichen Folgen. Es wurden sowohl die verbundenen Hoffnungen als auch die im Allgemeineren erkennbaren Risiken dessen, dass KI in unseren Entscheidungsprozessen mit enthalten ist, herausgehoben, unter Letzterem die Beschränkung der persönlichen Autonomie, die Reproduktion von Ungleichheit, Diskriminierung, Rassismus, die Destabilisierung der Demokratie, die Förderung von Polarisierung und nicht zuletzt die ökologischen Umweltfolgen usw. (Avnoon et al. 2023: 2). In KI stecken Potenziale, soziale Ungerechtigkeiten gerade zu perpetuieren, indem sie Stereotypen und bestehende Diskriminierung verstärkt und gleichzeitig neue Formen epistemischer Ungerechtigkeit hervorruft. (Rafanelli 2022) Was dabei jedoch in der Forschung bislang weniger häufig untersucht wird, sind die gewissen zur Theorie der Anerkennung gehörenden Aspekte, mit denen diese Bewertungsdimensionen offenkundig in engem Zusammenhang stehen. Dazu zunächst ein erster Aufriss.

So enormen Einfluss künstliche Intelligenz auf die Gesellschaft hat, die Bewertung ihrer Wirkung ist ambivalent. KI könnte uns helfen zu verstehen, wer wir sind, aber auch menschliche Fähigkeiten abwerten und die autonome Selbstverwirklichung verringern; sie kann das Spektrum der Dinge, die wir tun können, erweitern und menschliches Handeln ausdehnen, aber auch dazu führen, dass menschliche Verantwortung dabei wegfällt; KI kann uns dabei helfen, mehr zu erreichen (individuell und kollektiv), aber auch die menschliche Kontrolle über verschiedene Prozesse reduzieren; und schließlich kann sie Einfluss nehmen auf die gesellschaftliche Zusammengehörigkeit, indem sie menschliche Interaktionen mit anderen Menschen und der Welt weiterentwickelt, aber sie kann auch die Selbstbestimmung des Menschen untergraben. (Vgl. Floridi et al. 2018) Für die Ziele meiner Erörterung ist es interessant, dass bei der Suche nach einer Sprache, um in der Konstatierung dieser ambivalenten Bewertungen unsere sozialen Beziehungen zu KI zu beschreiben, auf Lösungen aus der Ethik und Bioethik zurückgegriffen wird (vgl. ebd.) und höchstens in geringerem Maße auf etwas aus der Sozialphilosophie. Wie im Folgenden gezeigt, sollte jedoch die Theorie der Anerkennung eine bedeutende Rolle in unserer aktuellen Diskussion über KI spielen. Denn sie ist ein weitreichendes Konzept, das es ermöglicht, über Phänomene wie das Selbstverständnis in vermittelten Beziehungen (mit verschiedenen Menschen, aber auch Technologien) und Selbstachtung im Zeitalter der zunehmenden Rolle von KI im Rechtlichen und der Rechtsmacht von Verhältnissen und in ökonomischen Markt-Beziehungen zu sprechen. Integral einbefasst ist darin, dass die Präsenz von KI in unserem täglichen Leben sowohl bewusst sein kann, wie wenn wir Entwicklungen auf dem Markt be-

obachten, die von aufstrebenden Technologien beeinflusst werden, als auch unbewusst, wenn wir unsere Webbrowser nutzen und unbewusst unser Selbstwertgefühl durch KI-Algorithmen gestalten, d.h. gestalten lassen (zum Beispiel, wenn jemand unter Fettleibigkeit leidet und nach einem geeigneten Arzt online sucht, wird dieser Mensch dabei oder danach aller Wahrscheinlichkeit nach Benachrichtigungen über neue Diäten und gesunde Lebensstile erhalten und der private Nachrichten-Feed wird sich mit Bildern von trainierten Athleten füllen – was potenziell die Bemühungen zur Selbstfürsorge gerade untergraben wird).

Ein aktuelles Problemfeld mag die Tragweite, die die Perspektive auf Anerkennung haben könnte, illustrieren. Es bringt die konkrete Frage einer ›Politik der Anerkennung‹ ein. Rosalie A. Waelen schlägt vor, dass die sozialen und politischen Auswirkungen von KI im Licht der Anerkennungstheorie von Ch. Taylor verstanden werden können, da diese die universellen Prinzipien der Gleichheit und Würde für alle Bürger ansetzt und doch zugleich die Bedeutung dessen betont, deren Besonderheit in Unterschieden zwischen Individuen und Gruppen Rechnung zu tragen. (Waelen 2022: 217) Dies spielt eine entscheidende Rolle etwa auch bei der Bewertung der sozialen und normativen Konsequenzen von Technologien, die für Waelen das empirische Beispiel sind, das von ihr dazu untersucht wurde – die Technologien, die für die Gesichtserkennung entwickelt wurden. Da Anerkennung (qua Erkennen) in ihrem unmittelbarsten Sinne »Identifizieren« oder »Kategorisieren« bedeutet, vollziehen Technologien, die auf der Identifikation oder Kategorisierung menschlicher Gesichter basieren, buchstäblich das (An)Erkennen von Menschen. Selbst wenn KI lediglich die Gesichtsmerkmale einer Person mit bestimmten Mustern abgleicht, sind die sozialen und normativen Auswirkungen bedeutend.

Waelen unterscheidet dabei drei Formen der Fehleinschätzung (»misrecognition«), die auftreten können, wenn Gesichtserkennungstechnologien eingesetzt werden, um Menschen zu identifizieren. Die erste ist die umfassende Fehlidentifikation. Sie tritt auf, wenn eine Person durch KI aufgrund von ungenauen Daten mit etwas oder jemandem fälschlich identifiziert wird. Ein Beispiel dafür könnte eine falsche Identifizierung einer Person als Verdächtiger durch die Strafverfolgungsbehörden sein, was strukturell in besonderem Maße dunkelhäutige Menschen betrifft, da die KI dunkelhäutige Menschen nicht richtig erkennt. Dies ist nicht nur in Bezug auf eine falsche Darstellung von Menschen mit dunkler Haut beleidigend. Sondern es hat auch praktische Auswirkungen, da für sie das Risiko steigt, fälschlicherweise als jemand identifiziert zu werden, der ein Verbrechen begangen hat (Waelen 2022: 218). Die zweite Form ist die Fehlkategorisierung im Einzelnen. Sie tritt auf, wenn die KI verschiedene Aspekte menschlicher Gesichter als konkrete Verhaltensweisen oder Emotionen interpretiert. Eine solche (An)Erkennung impliziert, dass menschliche Gesichter Emotionen grundsätzlich in ähnlicher Weise zeigen, sozusagen anthropologisch universell, was es der KI ermöglicht, Individuen basierend auf ihren Gesichtsausdrücken als ›traurig‹, ›glücklich‹ oder ›zwinkernd‹ zu kategorisie-

ren. Das Letzte, das ›Zwinkern‹, kann als überzeugende Illustration dafür stehen, wie eine KI Asiaten möglicherweise fälschlich als ›zwinkernd‹ kategorisiert, weil ihre Gesichtsmerkmale von denen abweichen, die weiße westliche Personen gemeinhin haben. »Miscategorization implies that people have elements of their identity misunderstood or ignored, or worse, that they are treated wrongly because of certain elements of their identity.« (Waelen 2022: 219) Die dritte Form der Fehleinschätzung liegt in der Unfähigkeit, die subjektive Identität zu erkennen. Ihre Bedeutung ergibt sich aus der Tatsache, dass in KI-Gesichtserkennungstechnologien eine Person ausschließlich basierend auf ihrem Gesicht interpretiert wird, das in einer stark polaren Matrix beurteilt wird. Unsere Gesichter müssen vorgegebenen Kategorisierungen entsprechen. Das Problem entsteht besonders für Personen, deren äußeres Erscheinungsbild nicht-binär ist, für die, die sich als nicht-binär identifizieren, oder für solche mit gemischtethnischem Hintergrund (Waelen 2022: 220). In all solchen Fällen können potenzielle Fehlklassifizierungen von Individuen erhebliche schädliche Auswirkungen haben und ihr Selbstwertgefühl, das sie aus dem entwickeln, was die Gesellschaft ihnen entgegenbringt – in Gestalt ihrer eingesetzten KI –, negativ beeinflussen. Dies unterstreicht, welche Bedeutung eine nuanciertere Analyse von Anerkennungstheorien bei Untersuchung der Beziehungen zwischen Mensch und KI und ihrer sozialen Folgen haben wird. (S. auch Waelen/Wieczorek 2022)

Die Theorie der Anerkennung kann, wie ich zeigen werde, dazu beitragen, die Fragen zu verstehen, die sorgfältig berücksichtigt werden müssen, um eine verantwortungsbewusste Implementierung von KI in die Gesellschaft zu ermöglichen, da KI die Anerkennung und ebenso Fehlbeurteilung von Individuen beeinflusst. Dies beinhaltet ein ganzes Untersuchungsprogramm. Im Folgenden werde ich dazu zunächst die Theorie von Axel Honneth heranziehen und erläutern, wie verschiedene Aspekte dieser Theorie zur Erörterung der Beziehungen zwischen Mensch und KI beitragen können. Anschließend werde ich die Theorie von Laitinen vorstellen, die Konzepte liefert, die helfen können, die Anwendung der Anerkennungstheorie in diesem spezifischen Kontext zu strukturieren. In Hinsicht auf die beiden Schlüsselprobleme im Zusammenhang mit der Erweiterung der Anerkennungstheorie auf KI, nämlich Gegenseitigkeit und Bewusstsein, wird sich sodann eine Perspektive von Ikäheimo zur Theorie der Anerkennung als fruchtbar erweisen, die es in Anschlag zu bringen erlaubt, dass Gegenseitigkeit nicht immer in allen Formen der Anerkennung die Voraussetzung ist. Schließlich wird ein letzter Unterteil das Problem des Bewusstseins anhand von Gertz' Interpretation dieser Frage thematisieren.

2. Anerkennungstheorie

Anerkennung besteht in der Fähigkeit, sich selbst als Subjekt nur durch Interaktion mit einem anderen Subjekt zu verstehen, d.h. durch Intersubjektivität. (Haber-

mas 1968) In diesem Sinne liefert Anerkennung eine starke Grundlage für Identität und ist psychologisch entscheidend. (Taylor 1992; Honneth 1994) Der Prozess der Anerkennung basiert auf einer Identifizierung, worin bestimmte Merkmale durch eine*n Anerkenner*in qualifiziert werden, die das anerkannte Individuum zu einem *Subjekt normativer Überlegungen* machen. Sobald wir bestimmte Merkmale, die ein Wesen beispielsweise als bewusst definieren, identifizieren, ist es unvermeidlich, die Implikationen dieses Merkmals auch auf praktische Aspekte zu erstrecken, wie freien Willen und Autonomie. Daher ist Anerkennung auch aus normativer Sicht wichtig, was sich in Ethik, Sozial- und politischer Theorie widerspiegelt (Kloc-Konkołowicz 2015). Anerkennung kann auch durch ihre gegenteilige Position, die fehlende oder versagte Anerkennung, diagnostiziert werden. Diese tritt auf, wenn ein Individuum bedeutsame Merkmale besitzt, die es beispielsweise zu einem autonomen Wesen machen, diese Merkmale jedoch von anderen nicht anerkannt werden. Fehlende oder versagte Anerkennung führt oft zu einem Kampf um Anerkennung und kann Grundlage für soziale und politische Bewegungen sein (Taylor 1992; Butler 1987).

Ein wichtiger Aspekt der Anerkennung, für einige Forscher sogar der entscheidende (Brandom 2007), ist die Annahme der Gegenseitigkeit. Anerkannt zu werden bedeutet – dies auf den ersten Blick –, eine Verständigung zu gewinnen über beiderseitig verschiedene normative Wesentlichkeiten oder Bedürfnisse, dabei unter der Voraussetzung, dass wir diese auf Gegenseitigkeit gewährleisten können. Was darin jedoch nicht so klar ist, ist das, ob Gegenseitigkeit notwendiges Element jedweder Formen der Anerkennung ist, oder ob dies nur der Anerkennung in ihrem Ideal entspricht. Wie in den folgenden Teilen erläutert, gibt es im Konkreten menschlicher Sozialexistenz verschiedene Formen der Anerkennung, von denen durchaus nicht alle auf Gegenseitigkeit beruhen. Zum Beispiel Eltern, die den Bedürfnissen ihrer Kinder gerecht zu werden suchen, erwarten von diesen nicht unbedingt Gegenseitigkeit, jedenfalls nicht in der idealen Form elterlicher Liebe; sie anerkennen ihre Kinder allem voran als Wesen, die normativ gerechtfertigt eine Fürsorge erwarten (dürfen). Ein weiteres Problem, das mit dem Punkt der Gegenseitigkeit aufgeworfen ist, betrifft das, was auch in den Beziehungen zwischen Mensch und KI oft als entscheidend betrachtet wird, nämlich das Bewusstsein beim Anerkennenden wie Anerkannten. Kann jemand (oder allgemeiner: eine Subjekt-Verkörperung) Anerkanntsein haben, ohne sich dessen bewusst zu sein? Ist es dann eine echte Anerkennung? Ich werde argumentieren, dass Anerkennungstheorien in diesem Aspekt von Gegenseitigkeit und Bewusstsein nicht streng festgelegt sind – und dadurch die Möglichkeit eröffnen, den Bereich der anerkannten Objekte auch auf künstliche Intelligenz auszudehnen.

2.1 Die Theorie von Axel Honneth im Kontext von KI

Eine der einflussreichsten Interpretationen der Anerkennungstheorie wurde von Axel Honneth vorgelegt. Im Anschluss an Hegel und G.H. Mead kennzeichnet Honneth, dass es drei Formen der Anerkennung gibt: Liebe, Recht und Solidarität (verstanden als Sittlichkeit oder Wertschätzung; Honneth 1994). Liebe ist notwendig, um elementar Selbstvertrauen aufzubauen, und sollte in einer lebensübergreifenden Verbundenheit von ›Familie‹ vermittelt werden, in der die Bedürfnisse der Menschen, insbesondere der Kinder, Gegenstand der Fürsorge der Eltern sind. Durch Liebe kann eine Person das grundlegende Gefühl für sich selbst und ihre körperliche Integrität aufbauen. Recht ist ein Ausdruck des kognitiven Respekts der Menschen vor einem menschlichen Wesen, das in der Lage ist, sich moralisch verantwortungsbewusst zu verhalten. Zu wissen, dass jemand als eine Person nicht nur von der eigenen Familie anerkannt wird, sondern auch ein Subjekt des Rechts ist, stärkt unter allen Beteiligten das Selbstwertgefühl und das Gefühl der Autonomie, was Grundlage für die Rechte im Rechtssystem bildet. Solidarität schließlich wird mit anderen Menschen geteilt, wenn wir ihren Beitrag zu einer Gesellschaft, der auf ihren Werten, ihrem Beruf, ihrem Engagement usw. basiert, hochschätzen. Dies bedeutet, dass eine Person aufgrund ihrer Bedeutung für die Gesellschaft soziale Anerkennung erreichen kann. – Nach Honneth können all diese Formen der Anerkennung, insofern sie fundamental sind, auch ihre Gegenpositionen haben, wie Missbrauch oder Vergewaltigung, Verweigerung von Rechten, oder in der Gesellschaftsgemeinschaft Ausschluss und Herabwürdigung oder Beleidigung. Dieses Negative sollte ebenfalls berücksichtigt werden, wenn es die soziale Bedeutung der Anerkennung zu analysieren gilt.

Alle diese drei Formen der Anerkennung: Liebe, Recht und Solidarität, waren in Honneths Theorie auf empirische Forschung gegründet (hauptsächlich Psychologie/Psychoanalyse und Sozialwissenschaften). Das hat die Diskussion über Anerkennung in der Bedeutung von Anerkennungsprozessen für Menschen gestärkt, aber auch eingeschränkt. Doch ist offenkundig, dass es, wie in der Forschung über verschiedene Formen menschlicher Interaktion und menschlichen Tuns sich erwiesen hat, unterschiedliche Objekte menschlicher Fürsorge, Achtung oder Wertschätzung gibt, die nicht alle zwangsläufig menschenförmig sein müssen. Das lässt sich unschwer illustrieren.

Die Erfahrung zu lieben, wohl eine der komplexesten Emotionen der menschlichen Natur, kann sich auf Gott, die Natur, Tiere oder sogar auf verschiedene Tätigkeiten richten (z.B. wenn wir Liebe zu unserer Arbeit ausdrücken). Wenn Liebesempfindungen etwas sind, das bei der Entstehung von Selbstvertrauen, größerer Selbstbestimmung oder einem Gefühl der Sicherheit aufkommt, können sie auch durch ein Technisches zustande kommen. – Pflegeroboter dienen als bemerkenswertes Beispiel für solches Technische. Entwickelt wurden sie, um

Einzelpersonen in Krankenhäusern oder Pflegeheimen zu unterstützen, Medikamente zu verabreichen, bei der Mobilität zu helfen oder Gespräche zu führen, wenn menschliche Pflegekräfte beschäftigt sind. Forschungsergebnisse zeigen, dass solche Beziehungen zu Robotern genauso befriedigend sein können wie diejenigen, die mit einem menschlichen Gegenüber entstehen, und starke emotionale Bindungen wachsen lassen. Dies kann gleichwohl nicht bedeuten, dass dies nicht noch ethisch überdacht werden müsste. (Vallor 2011) Gerade weil Menschen starke Beziehungen zu künstlichen Intelligenzen aufbauen können, die ihr eigenes Verhalten beeinflussen, ist es so wichtig, die verschiedenen Ebenen der Anerkennung und des Mangels an Anerkennung in solchen Beziehungen zu berücksichtigen (Cappuccio et al. 2019). Solchen emotionalen Aspekten bei künstlicher Intelligenz wird heute auch für die Gestaltung von sozialen Robotern stark Rechnung getragen. »This places strong demands on social robotics to build robots that understand others' actions, intentions, and emotions and show emotions themselves; that know when to listen to the human or act on its own preferences; that develop social competence, can keep up a normal conversation, form social relationships, learn from experience, and perhaps have a personality.« (Brinck/Balkenius 2020: 54; s. auch Breazeal 2002; Dautenhahn 2007) Diese Fähigkeit der künstlichen Intelligenz, menschliches Verhalten zu imitieren, wirft die Frage auf, ob die Nutzer von z.B. sozialen Robotern von den Entwicklern dieser Technologien subjektiv und mit Respekt vor ihrer Würde behandelt werden oder ob sie einer emotionalen Manipulation erliegen, die ethisch fragwürdig ist.

Ebenfalls in der Literatur diskutiert in Bezug auf Anerkennung wurde das Problem der Rechte, und es gibt viele Lösungsvorschläge, die auf KI angewendet werden können, um sie zu einem Rechtssubjekt zu machen. (Rodrigues 2020) Rechte auf KI auszuweiten, kann durch das, wie die Theorie der Anerkennung für Gruppen und Gruppenidentitäten angewandt wurde, gerechtfertigt werden. O. Hirvonen stellt hierzu fest: »In our everyday social practices we grant the status of personhood to any agent that is [capable of an addressive performance towards us]« (Hirvonen 2017: 147), was bedeutet, dass wir eine Zurechenbarkeit und Verantwortlichkeit zuschreiben und dies auf den kommunizierenden Agenten übertragen können. Personenhaftigkeit kann dabei nach Hirvonen in der Linie der Theorie der Anerkennung von Hegel und Honneth gefasst sein, und dementsprechend muss, um respektvoll behandelt zu werden, das gegeben sein, dass eine Entität ihrerseits in der Lage ist, abstraktes, moralisches Denken zu entwickeln, dass sie über sprachliche Fähigkeiten verfügt, dass sie selbstbewusst und dass sie rational ist. Dies könnte in gewisser Weise auf solche Formen von KI wie Chat-GPT oder fortgeschrittenere Roboter zutreffen, die normativen Erwartungen ähnlich folgen wie Menschen dem Gesetz. Hirvonen schlägt in seiner Argumentation für die Personenhaftigkeit von Gruppen vor, dass wir dort von »partial persons« sprechen können – Entitäten, die nicht alle

Kriterien erfüllen, die wir auf Menschen anwenden, aber dennoch in einigen Dimensionen der Anerkennung als Personen betrachtet werden können.

Das Beispiel von Gruppen, wie von Hirvonen dargestellt, verdeutlicht allerdings zugleich den Unterschied in den Dimensionen von Liebe und Recht in Bezug auf Prozesse von Anerkennung. Wir lieben keine Gruppen als abstrakte Entitäten, sondern lieben vielmehr bestimmte Individuen innerhalb dieser Gruppen (ansonsten die Gruppe vielleicht in einem übertragenen Sinne, wenn sie die Verkörperung einer von uns geliebten Idee ist). Dennoch sollten Gruppen, auch ohne Liebe, als Entitäten mit Rechten anerkannt werden, und dass man ebenso in Verhältnissen der Solidarität mit ihnen stehen kann. Ähnlich dann stellt sich das Problem aber auch bei KI (s. Gunkel 2014). Kann KI im Sinne des Vorschlags von Hirvonen »teilweise« die Kriterien der Personenhaftigkeit erfüllen? Die Antwort könnte bejahend ausfallen, denn verschiedene Formen von KI kommunizieren, lernen und können sich an Diskussionen über Werte beteiligen.

Solidarität schließlich erweist sich als einer der herausforderndsten Aspekte der Anerkennung in den Beziehungen zwischen Mensch und KI. Denn bei Solidarität spielt in einem gewissen Maß das mit, ein Verständnis für oder eine Identifikation mit dem zu haben, in welchen Kämpfen die unterschiedlichen Gruppen oder Individuen stehen. Gemäß Honneths Beschreibung ist Solidarität erreichbar, wenn Menschen ihre Verbundenheit erkennen und die vielfältigen Rollen und Fähigkeiten wertschätzen, die für die Gesellschaft wesentlich sind. Die Frage ist also: Kann man Solidarität gegenüber KI empfinden? Unter Bezugnahme auf die Forschung von Singer (Singer 2010) schlägt Nolen Gertz diesbezüglich vor, dass in bestimmten Situationen starke Impulse zur Projektion und zur Identifikation in der Tat zu einem Gefühl der Solidarität auch mit einer Technologie führen können: die Projektion menschlicher Funktionen, die für die Gesellschaft wichtig sind, und die persönliche Identifikation mit diesen Funktionen. Ein Beispiel für dieses Verhalten könne in der Art und Weise beobachtet werden, wie Soldaten Roboter auf dem Schlachtfeld behandeln, wo scheinbar irrationale Handlungen, wie das Risiko eigenen menschlichen Lebens, um einen Roboter zu retten, auftreten. Die Erklärung für dieses Verhalten läge genau im Gefühl der Solidarität – Solidarität in einer mit dem technischen Artefakt gemeinsamen Schicksalssituation. Die Soldaten nehmen den Roboter als ›Kameraden‹ wahr und sehen darin ein Spiegelbild von sich selbst. Folglich reagieren sie in einer Weise, die mit dem Verhalten übereinstimmt, das sie ihrerseits von allen anderen Individuen (und den eingesetzten Maschinen, auf die sie sich verlassen) im Krieg erwarten. (Gertz 2018)

Die Beispiele zu Liebe, Recht und Solidarität, obwohl kurz dargestellt, liefern starke Gründe, um das Thema der Anerkennung auch in der zeitgenössischen Philosophie der Technologie, insbesondere im Zusammenhang mit KI anzugehen. Wir stehen vor dem Erfordernis, unser genaues Verständnis von Selbstvertrauen, Selbstachtung und Selbstwertgefühl in unserer gegenwärtigen modernen Gesell-

schaft zu erforschen, seit diese Vorstellungen nicht mehr nur von Interaktionen zwischen Menschen geprägt werden, sondern auch dem Einfluss von Interaktionen zwischen Mensch und KI unterstehen. Um dazu einen Ansatz vorzuschlagen, der eine Diskussion über KI im Kontext der Anerkennung ermöglicht, möchte ich die Theorie von Arto Laitinen einbringen.

2.2 Die Theorie von Laitinen und verschiedene Formen der Anerkennung

Laitinen diagnostiziert, dass es verschiedene Dimensionen des Konzepts der Anerkennung gibt, die berücksichtigt werden sollten. Das beginnt damit, dass Anerkennung eindimensional oder dagegen multidimensional verstanden werden kann. Eindimensionale Anerkennung ist zum Beispiel in der Care-Ethik gegeben, bei der in der Struktur ›A anerkennt B als Z‹ dieses ›Z‹ konstant ist (beispielsweise in der Beziehung zwischen Arzt und Patient), während sich das ›Z‹ in der multidimensionalen Anerkennung je nach Veränderung des Status der Person ändern kann (Laitinen 2002: 464f.). Zweitens kann Anerkennung eine praktische oder symbolische Bedeutung haben. Wenn jemand anerkannt wird, weil er jemand ist oder bestimmte Eigenschaften hat, bedeutet dies, dass er*sie ›entsprechend behandelt werden sollte‹, was dabei im Konkreten von verschiedenen praktischen und symbolischen Kontexten abhängt. Drittens aber kann Anerkennung auch als ein entweder weites Konzept verstanden werden, bei dem verschiedene Situationen Verschiedenes, ihr Genüge zu tun, auslösen können, oder dagegen als strikte Auffassung, bei der die Regeln der Anerkennung in der Gesellschaft bereits existieren und wir die Bedeutung, was in einem Fall die Anerkennung ist, den verschiedenen eintretenden Situationalitäten als solchen zuordnen.

Diese verschiedenen Ebenen der Anerkennung führen nach Laitinen zu zwei allgemeinen Modellen der Anerkennung: einem generativen Modell und einem responsiven Modell. Im generativen Modell wird Anerkennung als *Handlung* betrachtet – eine Handlung, die normative Gründe etabliert und die soziale Realität effektiv verändert oder, mit anderen Worten, als *performativer Akt*, der Veränderungen in der Ontologie der Realität bewirkt. Das generative Modell sieht Anerkennung als einen Prozess, der, letztlich mit jeder Ereignung seines Akts, neue Attribute von anerkannten Objekten oder Subjekten enthüllt. Auf der anderen Seite das responsive Modell geht davon aus, dass innerhalb der Gesellschaft bereits normative Gründe für die Gewährung von Anerkennung an verschiedene Objekte oder Subjekte existieren. Es ist ein Modell der Zuteilung. Dieses Modell legt nahe, dass der Akt der Anerkennung eine Form des ›Benennens‹ oder ›Zuschreibens‹ ist, bei dem Rechte, Eigenschaften, Bedürfnisse usw. der rechtmäßig anerkannten Entität zugewiesen werden. Diese Prozesse verändern entsprechend die Realität nicht von sich aus, indem sie neue Bedeutungen von Anerkennung einführen würden, vielmehr werden sie allenfalls den Bereich der Objekte und Subjekte erweitern, die anerkannt werden sollten, also

was unter die gesellschaftlich gegebenen Bedeutungen fällt bzw. der betreffenden Zuschreibung zuteil werden kann.

Die Frage, die sich im Zusammenhang mit KI ergibt, lautet dann: Können *Objekte* ›anerkannt‹ werden, und kann ein Objekt eine Person in einem Sinne ›anerkennen‹, der über die (dem Objekt mitgegebene) bloße technische Funktionalität hinausgeht? – Laitinen weist darauf, dass es vier Arten von Objekten gibt, die einen Wert tragen und die es rechtfertigen, die Reaktionen, die sie basierend auf Gründen und Normen hervorrufen, jeweils zu analysieren. Diese Objekte sind 1. instrumentell wertvolle Objekte; 2. intrinsisch wertvolle Objekte; 3. lebende Wesen, die sich (in ihren Vitalprozessen) entfalten können, aber nicht unterscheiden können zwischen Verletzung und Beleidigung; 4. selbstbewusste Akteure, die in der Lage sind, Überzeugungen über die Überzeugungen anderer und über sich selbst zu bilden. (Laitinen 2002: 466) Wie zu sehen, erfüllt nur die letzte Art von Objekt die Kriterien, die normalerweise an Personen in einem vollen Sinne dieses Begriffs gestellt werden, und Laitinen betont, »[that] only persons can be recognizers« (Laitinen 2002: 465). Er macht jedoch keine solche Einschränkung für die oder das Anerkannte, die, wie aus der obigen Typologie ersichtlich, auch einen im weiteren Sinne objektmäßigen Charakter haben könnten. Allerdings gilt es zu bedenken, dass Technologien, wenn wir die oben erwähnte Analyse von Waelen berücksichtigen, nicht nur das Potenzial haben, Menschen (an)zuerkennen, sondern sie auch falsch anerkennen können, was zu normativen Konsequenzen führt. (S. auch Brinck/Balkenius 2020)

Man könnte jedoch anführen, dass eine einseitige Form der Anerkennung, die auf Objekte gerichtet ist oder die durch sie umgekehrt Menschen gezollt ist, nicht mit einem angemessenen Verständnis von Anerkennung übereinstimmt. Dafür gäbe es zwei Gründe. Zum einen, dass KI keine Intentionalität besitzt, um die Werte oder Identitäten, die von verschiedenen Individuen repräsentiert werden, zu schätzen – alles ist für sie gewissermaßen gleich, ein Faktum ihrer Umwelt. Und zum andern behandeln Menschen, außer in begrenzten Fällen, KI eben nicht gleichwertig mit anderen Menschen. – In den nächsten beiden Unterabschnitten werde ich argumentieren, dass die Annahme von Gegenseitigkeit der Anerkennung und Bewusstsein beim Anerkennenden/Anerkannten kein notwendiges Element der Theorie der Anerkennung ist. Was damit aber nicht berührt ist, ist, dass trotzdem die KI-Technologie, die auf begrenzter Anerkennung basiert, möglicherweise nicht unseren Erwartungen an ›wahre‹ Anerkennung entspricht. Deren normative Implikationen sind und bleiben weiterhin bedeutend.

2.3 Die Theorie von Ikäheimo und das Problem der Gegenseitigkeit der Anerkennung

H. Ikäheimo geht für die von ihm erörterten Fragen zurück auf die Hegelsche Tradition. Er hebt die Bedeutung von drei Dimensionen hervor, um gemäß dieser Tradi-

tion zu verstehen, was eine Person und ihre sozialen Prozesse ausmacht: Singularität (Einzigartigkeit), Autonomie (Universalität) und Partikularität. (Ikäheimo 2002) In Bezug auf die Singularität werden wir als Individuen mit einzigartigen Eigenschaften und unserer »suchness« anerkannt, wobei der Anerkennende sein Verhältnis zu uns an unserem Glück und Wohlbefinden ausrichtet. Als autonomen Wesen können uns Rechte gewährt werden, die dabei als Bedingung auf bestimmten universellen Werten basieren, die von allen Individuen geteilt werden. Als besondere Wesen schließlich wünschen wir uns, uns von anderen zu unterscheiden, und suchen Anerkennung für diese Partikularität. In diesem Sinne verkörpern wir, in unserem Uns-Unterscheiden, einen sozial geschätzten Wert. Ikäheimo argumentiert überzeugend, dass man diese drei Dimensionen bei der Anwendung auf Honneths Kategorisierung der Formen der Anerkennung, beinhaltend Liebe, Recht und Solidarität, berücksichtigt finden kann oder aber auch nicht.

Im Kontext der Liebe werden wir als einzigartige Wesen anerkannt, aber nicht ausschließlich als völlig autonom dabei (da wir immer noch auf andere angewiesen sind, um unsere Bedürfnisse zu erfüllen, was im Fall von Kindern deutlich wird) und auch nicht ausschließlich in unserer besonderen Partikularität (da von Eltern erwartet wird, dass sie ihre Kinder lieben, selbst wenn sie nicht einzigartig sind). Im Raum der Rechte werden alle Dimensionen unseres Menschseins berücksichtigt. Wir werden als Menschen mit unseren inhärenten Bedürfnissen anerkannt, ebenso als autonome Wesen, die zu verantwortungsbewusstem Verhalten fähig sind, und nicht zuletzt als einzigartige Individuen, insbesondere wenn wir eine Minderheit oder Ethnie repräsentieren, die um Anerkennung kämpft. Schließlich Solidarität, die wiederum einen speziellen Fokus hat, nicht alle Dimensionen gleicherweise in sich einbefasst. Sie zielt in erster Linie auf unsere Besonderheit und betont unsere einzigartigen Fähigkeiten, die Werte, die wir verkörpern, und unseren Beitrag zur Gesellschaft.

In Ikäheimos Interpretation der Theorie der Anerkennung liegt der Schwerpunkt darauf, verschiedene Identitätsdimensionen auf verschiedene Momente der Anerkennung anzuwenden. Anstatt Liebe, Rechte und Solidarität als separate Analysebereiche zu betrachten, könne die Theorie der Anerkennung aus der dreifachen Perspektive von Einzigartigkeit, Autonomie und Besonderheit dargelegt werden. Wenn jedoch die Anerkennung durch diese Konzepte untersucht wird, ist es nicht mehr immer notwendig, dass klassischerweise eine *Gegenseitigkeit* der Anerkennung stattfindet.

Die Singularität, die beinhaltet, ein menschliches Wesen in seiner*ihrer »suchness« wahrzunehmen, wird in den Bereichen Liebe und Rechte erfüllt. Diese Formen der Anerkennung betonen den Wert einer Person als Person, ohne dabei das, dass sie bestimmte spezifische Eigenschaften besitzen muss. Schon bei der Singularität wird deutlich, dass nicht immer eine gegenseitige Anerkennung erforderlich ist. Ich kann als Mensch, als ein menschliches Wesen, von meiner Familie oder durch das

Rechtssystem anerkannt werden, auch ohne dass ich mir dessen bewusst bin. Ich kann Menschenrechte haben, ohne vollständig zu verstehen, wie sie im internationalen Recht formuliert sind. Darüber hinaus sollte ich selbst dann, wenn ich gegen das gehandelt habe, was mir umgekehrt entgegengebracht wird, also wenn ich ein Verbrechen begangen habe und bestraft werde, immer noch als Mensch in meiner wesentlichen Natur behandelt werden. Dies verdeutlicht, dass die willentliche Anerkennung, andere »as such« anzuerkennen, nicht unbedingt eine Voraussetzung ist. Singularität der Person als Form der Anerkennung kann gegenseitig sein, kann aber auch ohne Gegenseitigkeit existieren.

Autonomie, wie sie auf dem universellen Verständnis basiert, dass jeder Mensch Vernunft und Würde besitzt, ist der umfassendste Aspekt und ist möglicherweise das einzige, in dem Gegenseitigkeit wirklich Berücksichtigung findet. Als autonomes Wesen kann ich Unabhängigkeit von anderen erreichen, weil sie es mir garantieren, und im Gegenzug behandele ich sie als ebenso autonom. Autonomie ist eng mit Verantwortung und Selbstachtung verbunden; verlässlich real verwirklicht kann sie nur durch das Rahmenwerk der Rechte werden.

Menschen in ihrer Partikularität zu verstehen, die sich auf die Einzigartigkeit von Fähigkeiten, Verhalten, dabei Individuen und Gruppen bezieht, manifestiert sich durch Schätzung, die sowohl in Rechten als auch in Solidarität zu finden ist. Diese Perspektive entspricht den theoretischen Grundlagen der ›Politik der Differenz‹ (Taylor 1992) und dient als Grundlage für die Wertschätzung von Individuen aufgrund ihres Beitrags zur Gesellschaft. Auch hier gilt es zu beachten, dass bei dem, jemanden aufgrund bestimmter Fähigkeiten zu schätzen, Gegenseitigkeit nicht unbedingt erforderlich ist. Entsprechend unrealistisch ist, zu erwarten, dass es eine gegenseitige Solidarität mit jedem Individuum, jedem Beruf und jeder Facette des Lebens der Menschen gäbe.

Anerkennung also ist nicht immer gegenseitig. Damit stellt sich die Folgefrage: Ist es hingegen notwendig anzunehmen, dass jedes Subjekt oder Objekt der Anerkennung dabei *bewusst* ist? – Um diese Frage zu beantworten, legt es sich nahe, sich noch einmal dem Urtext, Hegels *Phänomenologie des Geistes* zuwenden. Hegels Grundmodell beschreibt die gegenseitige Anerkennung von zwei Selbstbewusstseinen – Selbstbewusstwerdung, die von einem anderen Selbstbewusstsein anerkannt wird – und impliziert dabei, dass Gegenseitigkeit diesem Prozess konstitutiv innewohnt (das also, was wie gesehen nicht immer wesenhaft erforderlich ist). Aber, geht Hegel davon aus, dass jedes solche Selbst tatsächlich Bewusstsein haben muss? Und haben wir irgendwelche Evidenzen, dies anzunehmen? N. Gertz hat dafür einen tiefgehenden Ansatz vorgeschlagen.

2.4 Das Problem des Bewusstseins: ein Ansatz von Gertz

Gertz schlägt vor, das Problem der Anerkennung so zu untersuchen, dass man die Möglichkeit in Betracht zieht, dass auch Technologie in die Position des Anderen in der Anerkennungsbeziehung gestellt wird. (Gertz 2018) Diese alternative Perspektive auf Anerkennung führt zu neuen Implikationen für unser Verständnis dieses Konzepts. »The actual moments of the encounter as described by Hegel [in the *Phenomenology of Spirit* – NJ]«, so Gertz, »leave open the possibility that the other need to be a consciousness, but only appear to be a consciousness« (Gertz 2018: 142). Wie Gertz argumentiert, klassifiziert das Selbstbewusstsein, im Moment der Begegnung mit »dem Anderen«, diesen *als* »Selbstbewusstsein«, durch *Projektion* – die Projektion seiner eigenen Wünsche und Bedürfnisse (Gertz 2018: 143); und begrenzt damit eigentlich die Bedeutung der Anerkennung auf die Bedingungen seiner eigenen Selbstkenntnis. Die ›Lösung‹ für diese Beschränkung, Selbstbewusstsein als die einseitig ausgehende Projektion eines Bewusstseins zu behandeln, besteht nichtsdestoweniger im bekannten ›Kampf auf Leben und Tod‹, der den Unterschied in den Wünschen und Bedürfnissen enthüllt.

Es ist jedoch ein attraktiver Gedanke, im Einklang mit dem, was Gertz den »illicit move« jener Projektion nennt (Gertz 2018: 142), gegenläufigerweise zu erwägen, dass bei Übertragung der Anerkennungsbeziehung vom Bewusstsein auf Technologie der anfängliche Moment der Begegnung zweier Selbst im ›Kampf um Anerkennung‹ eben kein Wissen über den Anderen erfordert, sondern nur über uns selbst. Dies lässt die Möglichkeit zu, einem Objekt Bewusstsein *zuzuschreiben*, das sich so verhält, als ob es bewusst wäre, auch wenn es dies nicht ist. Zwar könnte man argumentieren, dass ohne wahres Bewusstsein beim Anderen die ›echte‹ Anerkennung in dieser Beziehung fehle. Aber ist dies immer der Fall? Wenn man zum Beispiel mit einem Chatbot in sozialen Medien spricht, ohne zu wissen, dass es keine echte Person ist, ›gibt‹ es dann keine Anerkennung? Ebenso, wenn man einem Instagram-Profil einer KI-generierten Person folgt, ohne den Unterschied zu echten Individuen zu erkennen, wirkt sich dies etwa auf das Konzept der Anerkennung aus?

Wie darauf zu antworten wäre, ließe sich im Anschluss an Waelen lernen. Waelen folgt Laitinen und führt einen Ansatz der angemessenen Berücksichtigung (»adequate regard«) ein, der sich ausschließlich auf die Haltung von A gegenüber B stützt, nicht jedoch das Umgekehrte. (Waelen 2022; s. Laitinen 2010) Wenn ich etwas als jemanden anerkenne und es versäume, zwischen einer Technologie und einem lebendigen Subjekt zu unterscheiden, ist dies zweifellos ein Fehler, indem ich der in der Technologie wirkenden KI menschliche Qualitäten zuschreibe. Doch kann dasselbe über die praktischen Konsequenzen dieser Anerkennung gesagt werden? »If we follow the adequate regard account […], it does not matter if the person considers the technology to be capable of recognizing them. All the matters under this understanding of recognition is the effect the system has on the person's

self-development.« (Waelen 2022: 221) KI in falscher Weise als Person anzuerkennen, wirkt sich zwar immer noch darauf aus, wie der Anerkennende sich selbst in Bezug auf die KI wahrnimmt und sie effektiv anstelle von Beziehungen zu echten Personen nimmt. Ansonsten aber gilt: »Mutual [sic!] recognition has been achieved when one individual shows that its behavior can be influenced by the other.« (Brinck/Balkenius 2020: 65) Wie deutlich, basiert das Argument hier auf dem Einfluss auf das Verhalten einer (menschlichen) Person, was ausreichend ist, um von Anerkennung zu sprechen, und was eben auch durch solche Technologien wie KI ermöglicht werden könnte. – Das hätte nun weitreichende theoriekonzeptionelle Folgen. Nach meiner Überzeugung ist es notwendig, dafür nicht nur bestehende Theorien der Anerkennung zu reformulieren, um dieses in der Sozialphilosophie so fruchtbare Konzept angemessen auch im Hinblick auf die Realität der intersubjektiven Beziehungen, die Menschen mit KI aufbauen, zu fassen. Sondern es muss auch konzeptionell mit Erkenntnissen vonseiten der Philosophie der Technik erweitert werden. Dazu ein letzter Abschnitt.

3. Neue Herausforderungen für eine KI mit einbeziehende Sozialphilosophie

Im gegenwärtigen Diskurs der den Entwicklungen und Veränderungen durch KI gewidmeten Philosophie hat sich tendenziell die Neigung zu zwei gegensätzlichen Lagern gebildet: entweder eine technologiezentrierte Perspektive zu betonen oder eine menschenzentrierte. (Peeters et al. 2021) Erstere legt den Schwerpunkt auf die fortschreitenden Fähigkeiten von KI – Fähigkeiten, die nicht nur das Potenzial haben, Menschen in rechnerischen Aufgaben zu ersetzen, sondern auch als eine Alternative in Situationen dienen könnten, in denen Skepsis bezüglich der menschlichen Entscheidungsfindung besteht, wie beispielsweise in Fragen der Gerechtigkeit (Peeters et al. 2021: 220). Die letztere Perspektive lenkt demgegenüber die Aufmerksamkeit auf die Probleme, die sich aus der KI selbst ergeben, wie etwa Bedenken, wie es Menschen dabei ergeht, oder die sich abzeichnende Zukunft in ethischer Hinsicht zu beurteilen. Dies sind Probleme, die KI aus sich heraus, KI allein nicht lösen kann. Sie erfordern vielmehr, dass es verantwortungsbewusste menschliche Entscheidungsfindung auf verschiedenen Ebenen moralischer, rechtlicher und sozialer Komplexität gibt.

Angesichts der fundamentalen Divergenz zwischen den Perspektiven führen Peeters et al. einen dritten Ansatz ein, den sie als den der kollektiven Intelligenz umreißen (»Collective Intelligence«). Das zentrale Argument dieses Ansatzes besteht darin, dem Rechnung zu tragen, dass ›Intelligenz‹, ob menschlich oder künstlich, nicht isoliert existiert. Um Prozesse und Leistungen von Intelligenz im Einklang mit dem, was ›Intelligenz‹ wesensmäßig ausmacht, tatsächlich umfassend zu ana-

lysieren, ist es notwendig, sie auf Ebene von Gruppierungen und Vernetzungen zu untersuchen – und sich dann heute auch auf das Zusammenwirken zwischen Menschen und Maschinen zu konzentrieren (Peeters et al. 2021: 222). Nach meinem Dafürhalten ist es dabei in Folge auch entscheidend, nicht nur die kognitive Dimension des Menschen im Kontext dieser Kollektivität zu betrachten, sondern ebenso die sozialen Aspekte. Allerdings ergibt sich bei solcher Betrachtung eminent die Frage nach der Verantwortung.

Taddeo und Floridi argumentieren, dass Entscheidungen, die auf KI basieren, das Ergebnis von verteiltem Handeln sind: Interaktionen zwischen verschiedenen menschlichen und nicht-menschlichen Akteuren wie Designern, Entwicklern, aber auch Software und Hardware verschiedener Technologien. Dies würde zu *verteilter Verantwortung* (»distributed responsibility«) führen. (Taddeo/Floridi 2018) Das wäre für die Ethik schon grundlegend ein neues Phänomen; es gäbe nicht mehr das eine Subjekt einer betreffenden Verantwortung. Ethik war traditionell auf individuelle Handlungen und Absichten ausgerichtet, jedoch nicht auf kollektive und aufgeteilte. Diese neue Herausforderung für die Ethik führt zu einer Idee der Verantwortung, die diese nicht an die Intentionen der Menschen koppelt – ob jemand etwas so ›gewollt‹ hat bei seinem*ihrem Tun oder Unterlassen –, sondern vielmehr berücksichtigt, was etwas als die Ergebnisse (und dabei nicht immer beabsichtigte) eines betreffenden Konglomerats von Menschlichem und Technologischem ist. Und seit es dazu gekommen ist, dass KI eine ›lernende‹ Technologie ist, können nicht alle Ergebnisse dieses Prozesses überhaupt vorhergesagt werden. Nichtsdestoweniger aber erfordert es Nachänderungen, wenn die KI auf mit einem Bias behafteten Daten trainiert wird oder wurde. (S. Burrell 2016; Jacobsen 2023) – Dies beschreibt eigentlich nur, was ohnehin alltäglich geschieht, nur die Theorie hat dem bisher noch mit einseitigen Modellen nicht entsprochen. Fast jeden Tag entscheiden wir alle, welche Technologie wir haben möchten, welche Praktiken, die durch Technologie angeregt werden, wir akzeptieren und welche davon geändert werden müssen, sind wir doch Nutzer verschiedener Technologien, und ein Teil dieser verteilten Verantwortung liegt allenthalben auch in unseren Händen.

Das Konzept der verteilten Ethik hat Auswirkungen auf das Verständnis von Anerkennung, das, sosehr in ihm Anerkennung als ein Prozess und als sich im Laufe der Zeit entwickelnd gefasst, traditionell die Prozesse beschrieben hat, die mit der Entwicklung von Selbstwissen, persönlicher und gruppenbezogener Identität, dem normativen Rahmen von sozialen Konflikten usw. verbunden sind. Die Anerkennungstheorie hat konkrete Konzepte bereitgestellt, um die verschiedenen Momente innerhalb dieser Prozesse aufzuspüren und zu verstehen. Die zwei wesentlichsten Punkte, die heute an die bisherigen Horizonte rühren, sind zum einen das, dass positive Dimension der Anerkennung und negative Dimension der Missachtung nicht mehr so klar in ihrer Bedeutung sind, wenn, wie zunehmend eingetreten, der Prozess der Anerkennung *verteilt* ist und dies über das (Zwischen-)Menschliche

hinausgeht. Und zum anderen eben das, dass der Anerkennende wie der Anerkannte nicht unbedingt menschlich und darin ›volles‹ Personsubjekt sein müssen, sondern Hirvonen gemäß eine »partial person« sein können. Daher stellt sich nicht nur die Frage, ob KI als (teilweise) Person anerkannt werden kann, KI als »der Andere« im Kampf um Anerkennung, und wie sich dies tatsächlich auf Liebe, Recht und Solidarität auswirken würde, sondern auch, was es für ein menschliches Wesen bedeutet, von KI ›anerkannt‹ zu werden.

Die Fragen reichen bis in die Struktur des Konzepts der Anerkennung. Auf KI als solche wird in der Regel wohl das responsive Modell der Anerkennung (s.o.) angewendet, das auf der Zuweisung von anerkannten Merkmalen beruht. Die KI muss dabei dann in einer Weise gestaltet sein, dass Benutzer*innen (Menschen) sie als Technologie (an)erkennen können, die konkrete Bedürfnisse erfüllt, oder als fähig dazu, Muster im menschlichen Verhalten zu erkennen. Im Falle von Menschen jedoch, insbesondere in ihren Beziehungen zu KI, könnte das generative Modell der Anerkennung (s.o.) eindeutig sachhaltiger sein, auch wenn noch nicht ganz klar ist, wie sich soziale Beziehungen, menschliches Verhalten und die emotionalen und intimen Dimensionen des menschlichen Lebens unter dem Einfluss von KI verändern können. Die in diesem Artikel verwendeten Beispiele, um den Einfluss von KI auf unsere täglichen Interaktionen zu veranschaulichen, enthüllen auch die Potenziale, wie KI uns Menschen als Träger der Anerkennung wahrscheinlich verändern könnte. KI wird schon jetzt zunehmend nicht mehr nur als Werkzeug betrachtet, das menschlichen Zielen dient, sondern auch als Partner oder Freund (s. Brinck/Balkenius 2020: 54), der die Art und Weise beeinflusst, wie Individuen *sich* im Akt der Anerkennung wahrnehmen. Dies ließe sich als ein Bereich aufkommender neuer Formen der Anerkennung verstehen, die der empirischen Forschung bedürfen.

Was sich festhalten lässt

Die große und einflussreiche Leistung der Theorie der Anerkennung war, verschiedene Formen des Aufbaus menschlicher Identität, des Selbstbewusstseins, der Gegenseitigkeit und der Intersubjektivität darzulegen. Solange sie traditionell verschiedene Phänomene im sozialen Leben explizierte, wie den Kampf um Anerkennung, verschiedene Formen sozialer Fehlzuweisungen und die Notwendigkeit, dass Menschen sich emotional, in ihren Rechten und im ökonomischen Wohlergehen entwickeln (können), wurde sie hauptsächlich auf Beziehungen unter Menschen in Ansatz gebracht. Mit den sich ändernden technologischen Bedingungen des sozialen Lebens und der zunehmenden Rolle nicht-menschlicher Akteure (Latour 2005) und nicht-menschlicher Handlungen (Bowden 2015) fordert das Feld der Beziehungen von Menschen und KI dabei eine weitere Forschung unter dem Gesichtspunkt der Theorie.

Ich habe dafür argumentiert, dass Anerkennung dabei nicht allein durch die instrumentelle Fähigkeit der KI verstanden werden kann, verschiedene menschliche Merkmale anzuerkennen bzw. fehlzuinterpretieren, und auch nicht schon dann, wenn sie soziale Hintergründe erkennen kann. Die Beziehungen von Menschen und KI sollten vielmehr konsequent im Kontext der Sozialphilosophie analysiert werden (s. Jacobs 2024). Das war die Motivation dabei, Honneths Theorie der Anerkennung einzuführen und sie in der Folge durch Interpretationen wie von Laitinen und Ikäheimo zu erweitern. Auf dieser Basis lassen sich gezeigterweise verschiedene Formen der Anerkennung (Liebe, Recht und Solidarität) und mögliche Fehleinschätzungen, verschiedene Analyseebenen (Singularität, Autonomie und Partikularität) und Modelle (ein responsives Modell und ein generatives Modell) identifizieren. Die dabei zu gewinnenden Differenzierungen erhöhen stark das Potenzial der Theorie. Die Untersuchung all dieser Aspekte könnte in diesem Sinne dazu beitragen, einen neuen Ansatz zur Beschreibung der Beziehungen von Menschen und KI zu entwickeln.

Neben der theoretischen Analyse auch empirische Beispiele für die Argumente einzubringen, stärkt dies noch weiter. Das zeigte sich hier an den eingeführten Konzepten der »partial persons«, der verteilten Verantwortung und der kollektiven Intelligenz. Mit ihnen ist man nicht mehr an Denkschemata gefesselt, die eindeutige – auch qualitativ eindeutige – Grenzen zwischen rein menschlichen und rein künstlichen Aktivitäten ansetzen. Von dem aus spricht vieles heute dafür, dass die Theorie der Anerkennung auch einer Neudeutung bedarf und zu einer *verteilten Anerkennung* werden muss. Erst so wird sie den sich ändernden Bedingungen des sozialen Lebens gerecht werden. Statt des responsiven Modells könnte in der Gesellschaft im Zeitalter der künstlichen Intelligenz das generative Modell der Anerkennung die größere Erkenntnis bieten: Anerkennung als performativer Akt, als eine Handlung, die normative Gründe etabliert und die soziale Realität effektiv verändert.

Literatur

Avnoon, N.; Kotliar, D.M.; Rivnai-Bahir, S. (2023): Contextualizing the ethics of algorithms. A socio-professional approach, in: *New Media & Society* [https://doi.org/10.1177/14614448221145728].

Bowden, S. (2015): Human and Nonhuman Agency in Deleuze, in: Roffe, J.; Stark, H. (Hg.), Deleuze and the Non/Human, London: Palgrave Macmillan, 60–80.

Brandom, R.B. (2007): The structure of desire and recognition. Self-consciousness and self-constitution, in: *Philosophy & Social Criticism*, 33(1), 125–148.

Breazeal, C. (2002): Designing sociable robots, Cambridge (MA): The MIT Press.

Brinck, I.; Balkenius, C. (2020): Mutual Recognition in Human-Robot Interaction. A Deflationary Account, in: *Philosophy & Technology*, 33, 53–70.

Butler, J. (1987): Subjects of Desire. Hegelian Reflections in Twentieth-century France, New York: Columbia University Press.

Cappuccio, M.; Peeters, A.; McDonald, W. (2020): Sympathy for Dolores. Moral Consideration for Robots Based on Virtue and Recognition, in: *Philosophy & Technology*, 33, 9–31.

Coeckelbergh, M. (2020): AI Ethics, Cambridge (MA): The MIT Press.

Dautenhahn, K. (2007): Socially intelligent robots. Dimensions of human–robot interaction, in: *Philosophical Transactions of the Royal Society B: Biological Sciences*, 362(1480), 679–704.

Floridi, L.; Cowls, J.; Beltrametti, M.; Chatila, R.; Chazerand, P.; Dignum, V.; Luetge, C.; Madelin, R.; Pagallo, U.; Rossi, F.; Schafer, B.; Valcke, P.; Vayena, E. (2018): AI4People – An Ethical Framework for a Good AI Society. Opportunities, Risks, Principles, and Recommendations, in: *Minds and Machines*, 28(4), 689–707.

Gertz, N. (2018): Hegel, the Struggle for Recognition, and Robots, in: *Techné. Research in Philosophy and Technology*, 22(2), 138–157.

Gunkel, D.J. (2014): A Vindication of the Rights of Machines, in: *Philosophy and Technology*, 27(1), 113–132.

Habermas, J. (1968): Arbeit und Interaktion. Bemerkungen zu Hegels Jenenser ›Philosophie des Geistes‹, in: Ders., Technik und Wissenschaft als ›Ideologie‹, Frankfurt a.M.: Suhrkamp, 9–47.

Hirvonen, O. (2017): Groups as Persons? A Suggestion for a Hegelian Turn, in: *Journal of Social Ontology*, 3(2), 143–165.

Honneth, A. (1994): Kampf um Anerkennung. Zur moralischen Grammatik sozialer Konflikte, Frankfurt a.M.: Suhrkamp.

Ikäheimo, H. (2002): On the Genus and Species of Recognition, in: *Inquiry*, 45(4), 447–462.

Jacobs, K. (2024): Roboter gegen Einsamkeit? Zur Reproduktionsdynamik falscher und mangelnder Anerkennung durch »soziale« KI, in: Adophi, R.; Alpsancar, S.; Hahn, S.; Kettner, M. (Hg.), Philosophische Digitalisierungsforschung. Verantwortung, Verständigung, Vernunft, Macht, Bielefeld: transcript, 219–256.

Kloc-Konkołowicz, J. (2015): Anerkennung als Verpflichtung. Klassische Konzepte der Anerkennung und ihre Bedeutung für die aktuelle Debatte, Würzburg: Königshausen & Neumann.

Kurzweil, R. (2005): The Singularity is Near. When Humans Transcend Biology, New York: Viking Press.

Laitinen, A. (2002): Interpersonal Recognition. A Response to Value or a Precondition of Personhood?, in: *Inquiry*, 45(4), 463–478.

Laitinen, A. (2010): On the scope of ›recognition‹. The role of adequate regard and mutuality, in: Schmidt am Busch, H.-C.; Zurn, C.F. (Hg.), The Philosophy of Recognition. Historical and Contemporary Perspectives, Lanham: Rowman & Littlefield, 319–342.

Latour, B. (2005): Reassembling the social. An introduction to actor network theory, Oxford: Oxford University Press.

Peeters, M.M.M.; van Diggelen, J.; van den Bosch, K.; Bronkhorst, A.; Neerincx, M.A.; Schraagen, J.M.; Raaijmakers, S. (2021): Hybrid collective intelligence in a human–AI society, in: *Ai & Society*, 36(1), 217–238.

Rafanelli, L.M. (2022): Justice, injustice, and artificial intelligence. Lessons from political theory and philosophy, in: *Big Data & Society*, 9(1) [https://doi.org/10.1177/20539517221080676].

Rodrigues, R. (2020): Legal and human rights issues of AI. Gaps, challenges and vulnerabilities, in: *Journal of Responsible Technology*, 4(3), 100005.

Singer, P.W. (2010): Wired for War, New York: The Penguin Press.

Taddeo, M.; Floridi, L. (2018): How AI can be a force for good. An ethical framework will help to harness the potential of AI while keeping humans in control, in: *Science*, 361(6404), 751–752.

Taylor, C. (1992): The politics of recognition, in: Gutmann, A. (Hg.), Multiculturalism. Examining the Politics of Recognition, New Jersey: Princeton University Press, 25–73.

Vallor, S. (2011): Carebots and Caregivers. Sustaining the Ethical Ideal of Care in the Twenty-First Century, in: *Philosophy & Technology*, 24, 251–268.

Van Dijck, J. (2013): ›You have one identity‹. Performing the self on Facebook and LinkedIn, in: *Media, Culture & Society*, 35(2), 199–215.

Waelen, R.A. (2022): The struggle for recognition in the age of facial recognition technology, in: *AI and Ethics*, 3(1), 215–222.

Waelen, R.; Wieczorek, M. (2022): The Struggle for AI's Recognition. Understanding the Normative Implications of Gender Bias in AI with Honneth's Theory of Recognition, in: *Philosophy & Technology*, 35(2), article 53.

Algorithmenkritik und die Suche nach dem »Außen«

Tobias Matzner

Abstract: *This article reflects on critiques of algorithms in recent years, i.e. texts that critically engage with algorithms from an ethical or political perspective. It shows that many of these works derive their normativity from an »outside« of the algorithmic. For example, an algorithmic classification is criticised on the basis of a comparison with a non-algorithmic, human-made classification. Three forms of »outside« are presented. (1) The »outside« as a site of instrumental control over algorithms. Two further forms of the »outside« are found in theories that reject an instrumental understanding of technology: (2) The unpredictable, which marks an inherent limit of the algorithmic, from which some authors draw political consequences. (3) A form of social interaction that is not algorithmically (or non-technically) mediated and that is opposed to the algorithmically mediated form. This article critiques all three forms and proposes a different critique of algorithms, one that does not seek an »outside« but rather begins »within« the algorithmic itself, drawing its normativity from the difference between algorithms or algorithmically mediated situations themselves.*

Keywords: *criticism of algorithms; digital ethics; unpredictability; normativity; philosophy of technology*

In den letzten Jahren wurden eine Vielzahl von Texten veröffentlicht, die sich kritisch mit der Ethik oder Politik von Algorithmen und künstlicher Intelligenz auseinandersetzen. Dabei kommen ganz unterschiedliche Begriffe und Vorstellungen von Algorithmen – und Technologie allgemein – sowie deren ethischer oder politischer Kontextualisierung zum Tragen. In diesem Text geht es um eine argumentative Figur, die viele dieser Texte trotz der Differenzen eint: Die Kritik an einer algorithmischen Praxis oder einer algorithmischen Anwendung entsteht durch den Bezug zu einem »Außen« der Algorithmen. So wird z. B. eine algorithmische Klassifikation kritisiert basierend auf dem Vergleich mit einer nicht-algorithmischen Klassifikation, also einer Klassifikation durch Menschen. Eine ganze Reihe von Begriffsschöpfungen nach dem Muster »algorithmic X« lassen sich finden: »algorithmic gover-

nance« (Rouvroy 2013), »algorithmic cultures« (Seyfert/Roberge 2016), »algorithmic war« (Amoore 2009) und viele mehr. Oft wird hier das »algorithmic X« kritisiert, indem es auf eine nicht-algorithmische Form von X bezogen wird. Es werden jedoch nicht nur solche Differenzen zwischen algorithmischen und nicht-algorithmischen Varianten des Gleichen bemüht. Eine noch grundlegendere Form des »Außen« stammt aus Überlegungen, was Algorithmen per se nicht möglich oder zugänglich ist.

Zur Begriffsschärfung und Definition dessen, was Algorithmen sind und was sie ausmacht, ist der Bezug zu beiden Formen des »Außen« sicher richtig und wichtig. Hier geht es jedoch um etwas anderes: Die Differenz zwischen Algorithmen und dessen »Außen« wird zur Quelle der politischen Bewertung von Algorithmen.

Im Folgenden werde ich drei verschiedene Formen dieser politischen oder ethisch-politischen Argumentation über ein »Außen« der Algorithmen vorstellen. Erstens das »Außen« als Ort der Kontrolle über Algorithmen, zweitens das »Außen« als das nicht-Berechenbare oder Unberechenbare und drittens, das »Außen« als die Form sozialer Interaktion, die durch Algorithmen gerade bedroht ist. Für jeden Punkt werde ich einige Probleme anführen, die dann zum Schluss zeigen lassen, dass die hier vorgestellte Kritik von außen durch eine Kritik »innerhalb« des Algorithmischen ergänzt werden muss. Diese speist sich aus unterschiedlichen Formen und Konfigurationen von Algorithmen. Statt der Differenz zwischen Algorithmen und einem Außen betrachtet diese Kritik also die Differenz zwischen unterschiedlichen Formen oder Konfigurationen von Algorithmen.

1. Kontrolle über Algorithmen

Die erste Variante des Arguments über ein Außen speist sich aus einem liberalen oder auch instrumentellen Technikverständnis. Demzufolge wird Technik zum Problem, wenn sie nicht mehr von hinreichend autonomen Subjekten kontrolliert wird und in ihren Konsequenzen beherrschbar bleibt. Technik ist dieser Vorstellung nach ein Objekt, das einem Subjekt als Instrument für seine Ziele dienen soll. Die Politik der Technik wird also durch das Subjekt und dessen Anliegen definiert, und beides ist der Technik äußerlich gegenübergestellt. In anderen Worten kann dieselbe Technik in den Händen des einen Subjekts ein hilfreiches Werkzeug sein, in denen eines anderen ein Mittel für Missbrauch oder Gewalt. Das politische relevante »Außen« der Algorithmen sind also die Menschen, die sie nutzen, aber auch von ihrer Nutzung betroffen sind. Das instrumentelle Technikverständnis ist hier nun aber nicht deskriptiv, sondern normativ zu verstehen. Algorithmen bergen demzufolge die Gefahr, sich genau dieser Kontrolle und damit einer verantwortbaren Politik zu entziehen.

Diese Vorstellung ist sehr wirkmächtig in Ethikleitlinien und Policy-Empfehlungen. Die dort geforderten Werte wie Transparenz oder Einwilligung haben zum Ziel, den menschlichen Nutzer*innen oder Betroffenen diese Kontrolle zu sichern oder wiederzugeben. Umgekehrt speist sich Kritik dann daraus, dass diese Kontrolle über die Techniknutzung und das Verstehen ihrer Konsequenzen gerade bei Algorithmen oft nicht mehr möglich sei. Gängige Motive sind hier die sogenannte »Blackbox« der Algorithmen, womit der Zugang zu Quellcode oder Designentscheidungen gemeint ist, der lediglich Programmierer*innen möglich ist (Kitchin 2017; Matzner 2017). Im Kontext von künstlicher Intelligenz wird auch angeführt, dass nicht einmal die Entwickler*innen selbst verstehen könnten, wie die Technik wirklich funktioniert (Rohlfing et al. 2020). Ähnliche Argumente finden sich auch in Bezug auf Trainingsdaten von Machine Learning Algorithmen, deren Quelle und Verarbeitung aus Sicht der Anwendung und der Betroffenen nicht nachvollziehbar ist (Mühlhoff 2019). Eine verwandte Argumentationsweise findet sich auch in der Kritik des sogenannten Überwachungskapitalismus zum Beispiel bei Shoshanna Zuboff (Zuboff 2019). Sie argumentiert, dass aus der Tätigkeit von Menschen online digitale Produkte entwickelt werden und somit diese ohne deren Wissen als »Sklaven« arbeiteten (Zuboff 2016). In all diesen Argumenten ist die selbstbestimmte Nutzung der Technologie beziehungsweise deren Unmöglichkeit die Quelle der Kritik – noch nicht einmal die bewusste Aufgabe der Selbstbestimmung ist nach Zuboff und anderen noch möglich. Die Nutzung vieler digitaler Technologien erfolgt, ohne die Konsequenzen kennen zu können.

Diese Gegenüberstellung von menschlichem Subjekt und technischem Objekt und das ihr zugrundeliegende instrumentelle Technikverständnis werden in weiten Bereichen der Technikphilosophie, der Medienwissenschaft und der Science and Technology Studies kritisch gesehen. Hier wird betont, dass menschliche Subjektposition immer schon in einer Art und Weise durch Technik geformt ist, die von Menschen nicht komplett nachvollzogen oder durchdrungen werden kann. Zum Beispiel hat N. Kathrine Hayles in ihrer Geschichte der Kybernetik dargestellt, wie letztere einerseits als liberales Projekt verstanden werden muss, das aus dem Versuch entstand, neue technische Nutzungsmöglichkeiten zu erschließen. Andererseits hat sie aber das liberale Subjekt radikal in Frage gestellt, wenn es als komplexe Form von Informations- und Kommunikationsschleifen Tieren oder sogar Maschinen ontologisch gleichgestellt wird (Hayles 1999). In einflussreichen Teilen der Technikphilosophie wird Technik grundlegend so verstanden, dass ihr Sinn genau darin besteht, Menschen Handlungen zu ermöglichen, die sie genau nicht selbst durchführen können oder nicht mehr selbst durchführen können müssen. Das heißt, ihr Sinn besteht gerade darin, gewisse Dinge nicht zu wissen oder können zu müssen (Krämer 2021). Aus politischer Sicht lässt sich hinzufügen, dass dieses Technikverständnis einem individualistischen Politikverständnis entspricht (Matzner 2019b). Daraus folgt unter anderem, dass die Verantwortung für die Tech-

nologie reduziert wird auf die Verantwortung für die Techniknutzung durch das individuelle Subjekt. Gerade im Kontext von Algorithmen ist es aber nicht möglich, wirklich von verantwortungsvoller Nutzung zu sprechen, weil die Technik schlicht und ergreifend nicht überschaut werden kann. Selbst informatisch hervorragend gebildete Menschen können nur erahnen, was mit ihren Daten geschieht, wenn sie beispielsweise eine Webseite aufrufen. Die informierte Einwilligung welche durch Cookie-Warnungen oder das Akzeptieren von Lizenzen suggeriert wird, ist somit eigentlich nicht möglich (Matzner 2024: 147ff.). Statt diese Unmöglichkeit zu bemängeln und zu versuchen, sie dennoch herzustellen, wäre die Alternative, die Verantwortung gleich bei den Betreibern oder den Entwicklern zu suchen und nicht mehr beim Individuum selbst. Unter dem Stichwort »Responsibilisierung« wird diskutiert, wie durch solche Maßnahmen, die vordergründig so klingen, als wollten sie die Autonomie der Subjekte durch mehr Kontrolle vergrößern, diese eigentlich einschränken, indem sie sie mit einer nicht zu lösen Aufgabe überfordern (Matzner et al. 2016). Zudem stimmt die implizierte Zuschreibung, dass die Normativität einer Technik aus der Nutzung kommt, spätestens im Fall von Algorithmen in aller Deutlichkeit nicht mehr. Wenn z.B. ein Unternehmen einen Algorithmus zur Bewertung von Bewerbungen kauft, der ein Bias aufweist, dann ist das nicht dem Unternehmen oder seinen Angestellten zuzuschreiben. Das heißt, sowohl aus technisch praktischer Sicht ist es nicht möglich, eine volle Kontrolle auszuüben, wie es auch aus normativer Sicht nicht angemessen ist, diese Aufgabe den Nutzer*innen alleine zu überlassen.

Abstrakt gesprochen lassen sich diese Kritiken am instrumentalen Technikverständnis darin zusammenfassen, dass hier das Subjekt und damit die Quelle der Kritik der Technik rein äußerlich ist. Darüber hinaus ist auch das mit dieser Kritik verbundene Verständnis von Subjekten in vielen Aspekten von kritischer Theorie bis Systemtheorie kritisiert worden. Da all diese Punkte in der Literatur ausführlich diskutiert sind, belasse ich es in diesem Text mit dieser kurzen Erwähnung und gehe zur nächsten Form über, in der Algorithmen durch ein »Außen« kritisiert werden.

2. Das Unberechenbare

Eine zweite Gruppe von Theorien versammle ich hier unter dem Begriff des Unberechenbaren. Diese arbeiten nicht mit einer Gegenüberstellung von Mensch und Technik. Vielmehr finden sich hier Theorien, die der Technik sogar einen recht großen Einfluss auf menschliche Subjekte zumessen. Das Außen des Algorithmischen besteht hier viel mehr in einer prinzipiellen Unverfügbarkeit für Algorithmen. Diese ist unabhängig von deren Nutzung oder Einsatzkontext.

Genauer müssen in Bezug auf den Begriff des Unberechenbaren zwei Theorielinien unterschieden werden. Die eine bezieht sich auf Erkenntnisse der theoreti-

schen Informatik und Mathematik, in denen es um die Grenzen der Berechenbarkeit geht. Hier geht es also darum zu klären, wie Berechenbarkeit mathematisch gefasst werden kann und was in dieser Hinsicht überhaupt berechenbar ist. Wichtige Grundlagentexte der Informatik haben gezeigt, dass es tatsächlich Dinge gibt, die von Algorithmen oder, um genauer zu sein, von deren mathematischem Modell wie der sogenannten Turingmaschine oder formalen Kalkülen nicht berechenbar sind (Turing 1937; Church 1936; für eine Zusammenfassung siehe Matzner 2024: 15ff.). In dieser Tradition finden sich beispielsweise die Arbeiten von Luciana Parisi, die sich allerdings nicht direkt auf Alan Turing stützt, sondern auf den etwas obskuren Mathematiker Gregory Chaitin und seine Arbeit zum Unberechenbaren (Parisi 2016). Dennoch bezieht sich auch diese Theorie auf die grundsätzliche, formal definierte Berechenbarkeit. Hier ist im Kontext aktueller Anwendung von Algorithmen zu betonen, dass sich Berechenbarkeit in diesen Texten immer auf deterministische Berechenbarkeit bezieht, das heißt auf das garantierte Erreichen des richtigen Ergebnisses. Maschinelles Lernen ist aber genau aus dem Versuch entstanden, mit diesen Grenzen der Berechenbarkeit umzugehen, indem eben nicht deterministisch gerechnet wird. Stattdessen werden statistische, heuristische oder andere näherungsweise Rechenverfahren genutzt. Bei all diesen wird das Ergebnis nicht garantiert gefunden, sondern nur mit einer hinreichenden Wahrscheinlichkeit und hinreichender Genauigkeit erreicht. Eine Kritik, die sich daraus speist, dass Algorithmen, weil sie maschinell rechnen, gewisse Dinge per se nicht können, verfehlt also gerade Algorithmen im Sinne von maschinellem Lernen, der künstlichen Intelligenz oder vielen anderen zeitgenössischen Anwendungen, wie z.B. Optimierungsverfahren, die alle nicht deterministisch rechnen und sich somit Dingen, die Algorithmen tatsächlich deterministisch nicht können, dennoch anzunähern.

Diese erste Theorielinie des Unberechenbaren als ein Außen des Algorithmischen werde ich hier also nicht weiter verfolgen. Stattdessen beschäftige ich mich mit der zweiten Form des Unberechenbaren. Diese bezieht sich nicht auf mathematische oder maschinelle Berechenbarkeit. Stattdessen geht es um die prinzipiellen Möglichkeiten von Algorithmen in ihrem aktuellen Einsatzkontext und insbesondere um die Möglichkeiten des maschinellen Lernens. Diese Form des Unberechenbaren wurde vor allem von Louise Amoore in ihrem Buch Cloud Ethics ausbuchstabiert (Amoore 2020). Sie untersucht dort verschiedene algorithmische Verfahren der Prädiktion im Bereich der Sicherheit, der Genanalyse und andere mehr. Sie zeichnen sich alle genau dadurch aus, unsichere, probabilistische oder sogar die Zukunft betreffende Aussagen zu suchen. Für Amoore sind die Ausgaben von Algorithmen aber nicht nur deshalb unsicher, weil sie approximativ oder prädiktiv sind. Denn nicht einmal die im Rahmen der Statistik möglichen Garantien einer Aussage hält Amoore für gerechtfertigt.

Sie betrachtet Algorithmen nicht als Rechnung, sondern als automatisierten Akt des Schreibens. Algorithmen, egal wie sie genau mathematisch arbeiten, liefern Zei-

chen, welche eine Aussage über die Welt beinhalten. Diese Aussagen von Algorithmen sind für Amoore nun prinzipiell unsicher, weil sie so unsicher sind, wie es jeder Akt des Schreibens ist. Hier bezieht sich Amoore auf Foucaults Überlegungen zur Autorschaft (Foucault 1988) sowie auf die Theorie von Jacques Derrida zur Schrift. Derzufolge kann jede Art des Schreibens nur versuchen, seine vermeintlich sichere Bedeutung einzuholen, diese aber nie letztendlich erreichen (Derrida 2001). Algorithmen sind als Akte des Schreibens – wenn auch des maschinellen Schreibens – in diesem Sinne genauso unsicher; oder genauer gesagt, die Bedeutung des algorithmischen Schreibens kann nie abgeschlossen werden. Es ist genau diese prinzipielle Unabgeschlossenheit von Schrift, die Amoore als das Unberechenbare bezeichnet.

Damit richtet sie sich erst einmal gegen Versuche, die Ethik von Algorithmen durch Zugang oder Transparenz zu lösen:

> »The calls for access to the source code of algorithms are a means to try to read how a likelihood ratio is written. Yet, scrutinizing the code itself would reveal nothing of the racialized, prejudicial outputs of the algorithm that are written via training data, validation data, and the feedback loops of experiments in the field. Indeed, the Github platform is itself a form of distributed and iterative writing in which multiple developers contribute to the rewriting and editing of software.« (Amoore 2020: 97)

Damit richtet sich Amoore auch gegen die im ersten Abschnitt diskutierte Strategie, Subjekten gegenüber der Technik Kontrolle zu geben – in diesem Fall eben durch Einsicht in den Quellcode als die entscheidende Instanz, wo Kontrolle über die Technik ausgeübt würde. Neben dem Code wird der Output von Algorithmen des maschinellen Lernens aber durch viele Faktoren bestimmt: Trainingsdaten, die Feedbackloops der Trainingsprozesse. Und auch der Code selbst ist nicht eine zentrale Instanz sondern oft verteilt und in verschiedenen Formen der Zugänglichkeit reguliert (Matzner 2024: 89ff.), wofür Amoore hier als Beispiel die Codeverwaltungsplattform GitHub angibt. All diese Faktoren sind aber nicht nur eine Quelle von großer Komplexität für Amoore. Sie stellen in vielerlei Hinsicht einen Bezug zu anderen sozialen, kulturellen, und ökonomischen Prozessen her und sind damit eine Quelle von Kontingenz. Selbst wenn also hypothetischerweise der Aufwand betrieben werden könnte, der nötig ist, um der Komplexität algorithmischer Prozesse gerecht zu werden, könnten diese nicht durchdrungen werden. Sie sind aufgrund der Kontingenzen in den Prozessen selbst kontingent und damit prinzipiell unabgeschlossen. Genau so, wie ein Text laut Foucault nicht auf einen Autor oder dessen Intentionen zurückzuführen ist, sondern seine Bedeutung durch vielerlei Aspekte erhält, die insbesondere auch in Bezügen zu anderen Texten entsteht, so lässt sich auch keine definitive Quelle der Bedeutung algorithmischer Ausgaben finden. Amoore betrachtet Algorithmen als technischen Schreibprozess, deren Bedeutung aber ebenso

relational und offen zu verstehen ist. »The algorithm iterates beyond the moment of its inscription, distributing the writing through multiple characters, from the training data to the back propagation of errors.« (Amoore 2020: 100) Damit, so Amoore weiter, sind Algorithmen gerade nichts Besonderes:

> »But my point is that we should remember that algorithms are not unique, or somehow ›outside the text.‹ All acts of writing and reading, whatever the text, necessarily confront illegibility and the impossibility of reading. We do not need new resources or technologies to prize open the unreadable algorithm. Instead we could begin from that unreadability as the condition of all engagements with text. It is something familiar to us all as social and political theorists.« (Amoore 2020: 100)

Das ethische Problem besteht folglich vielmehr darin, zu denken, Algorithmen wären etwas Besonderes. Diese Besonderheit hat viele Formen: Algorithmen als eine objektive Instanz, frei von sozialen oder kulturellen Einflüssen, eine Quelle von Berechenbarkeit in der Unberechenbarkeit des Lebens, ja sogar eine Produktionsweise neuen Wissens. All diese Vorstellungen haben gemein, dass Algorithmen als etwas gesehen werden, das klare, lesbare, eindeutige Ergebnisse produziert. Für Amoore ist das Unberechenbare, das sie auch das Unlesebare oder Unattribuierbare (unattributable) nennt, also nicht das Problem, wie oben im ersten Abschnitt diskutiert. Vielmehr ist es die Quelle ihrer Ethik. Dieser geht es einerseits darum, diese Offenheit anzuerkennen, was große Konsequenzen für die Frage hat, welchen Wert algorithmischen Entscheidungen beigemessen werden. Das bedeutet für sie nicht, dass diese keinen Wert hätten. Aber sie müssen eben mit der Vorsicht und Kompetenz sozial und politisch kundiger Menschen »gelesen« werden, wie sie im obigen Zitat schreibt. Andererseits geht es auch darum, diese Offenheit zu wahren. Denn nur in ihr findet sich auch die Möglichkeit für Veränderung und Widerstand.

> »In place of an ethics that seeks to make an illegible algorithm legible to the world, a cloud ethics recognizes the nonclosure in all forms of writing and pushes the fabulation of the algorithm beyond what can currently be read of its logic.« (Amoore 2020: 160)

Die hier angesprochene »fabulation«, so konstatiert Amoore, sei bereits am Werk. Denn die Informatik hätte kein Problem damit, »Fabeln« über ihre Technik zu erzählen, z.B. dass sie eine Möglichkeit maschineller Vernunft beinhalte. Auch Algorithmen selbst fabulierten: »they invent a people, write them into being as a curious body of correlated attributes, grouped into clusters derived from data that are themselves fabulatory devices« (Amoore 2020: 158). Diese als Lüge oder Ideologie zu enttarnen, verschließt sich aber aufgrund Amoores hier ausgeführtem Verständnis.

Stattdessen ist die Herausforderung einer Ethik, die Unabgeschlossenheit zu nutzen, und den technikdeterministischen oder vorurteilsbehafteten Fabulationen andere – eben ethische – beizufügen.

Dieses Betonen einer prinzipiellen Möglichkeit von Alternativen einer prinzipiellen Offenheit ist ein erster wichtiger Schritt für jede ethische Beschäftigung mit Algorithmen. Dies ist gerade in einem Kontext relevant, der sehr stark von Optimierungslogiken geprägt ist. Dennoch muss eine ethische Beschäftigung mit Algorithmen auch über dieses prinzipielle Moment der Offenheit hinaus zu konkreten Alternativen finden können. Es geht in einer Ethik der Algorithmen auch um die Orientierung der Entwicklung anderer oder besserer Verfahren. Auch dabei ist es wichtig, die prinzipielle Unabgeschlossenheit mitzudenken. Und nicht der Verlockung zu verfallen, zu denken, jetzt die richtige Lösung zu entwickeln. Was aber die ethische Fabulation gegenüber etwa derjenigen der Informatik besser macht, bleibt aus Sicht von Amoores Theorie unklar. Die Umsetzung jeglicher Fabulation von Algorithmen wird notwendigerweise auf genau die Kontingenzen stoßen, die Amore ja auch analysiert: Abhängigkeit von Daten, von Code und dessen Möglichkeiten, von materiellen Bedingungen des Algorithmischen (Matzner 2024: 125). Mit diesen muss sich dann im Konkreten und aufgrund von konkreten Kritiken der bestehenden algorithmischen Anwendungen auseinandergesetzt werden. Das bedeutet also nicht, dass Amoores Ethik ihren Zweck verfehlt, aber sie muss durch eine Kritik innerhalb des Algorithmischen ergänzt werden.

3. Der Verlust des Außen

Die dritte Form von Kritik argumentiert in Bezug auf den Verlust des Außen. Auch hier geht es um ein prinzipiell Offenes, Unberechenbares. Wenn aber Amoore der Meinung ist, dass gerade auch Algorithmen immer in Bezug zu einem Unabgeschlossenen und Unberechenbaren stehen, so liegt in dieser Form der Kritik die Betonung darauf, dass Algorithmen einer eigenen technischen Logik folgten. Das Offene und Unberechenbare wird auf Seiten der Menschen verordnet und gerade nicht bei den Algorithmen. Die Gefahr besteht dann folglich nicht darin, die prinzipielle Offenheit des Algorithmischen zu vergessen oder zu verleugnen, wie Amoore dies schreibt. Stattdessen geht es darum, dass die abgeschlossene Logik des Algorithmischen auf die zwischenmenschliche Offenheit übergreift und diese tatsächlich verdrängt oder zerstört. Diese Möglichkeit von Algorithmen, menschliche Aktion oder Interaktion zu determinieren, wird damit begründet, dass sich die algorithmische Logik dadurch auszeichne, bestimmte menschliche Dispositionen anzusprechen.

Solch ein Argument findet sich z.B. in der Beschreibung der vielzitierten Idee der »Filterblase«. Dieser zufolge kann die algorithmische Auswahl der Inhalte di-

gitaler Medien ein menschliches Grundbedürfnis ansprechen, Dinge, die wir mögen oder die unseren Positionen ähnlich sind, weniger kritisch zu betrachten als andere (Pariser 2011). Im Gegensatz zu der im ersten Abschnitt diskutierten Form der Kritik, die auf autonome Technikkontrolle aufsetzt, wird hier also mit einem Menschenbild operiert, das anerkennt, dass Menschen nicht nur rationale autonome Wesen sind. In der Theorie der Filterblase geht es noch darum, diese Autonomie dennoch so weit wie möglich wieder herzustellen und damit sich gegen den algorithmischen Einfluss zu verwehren. Andere Theorien, die mit dem Verlust des Außen argumentieren, gehen hier weiter und beziehen sich auf ein Menschenbild, das sich von dem liberalen autonomen Subjekt deutlich unterscheidet. In prägnante Form bringt diese Argumentation zum Beispiel Antoinette Rouvroy in ihrer Konzeption der algorithmischen Gouvernementalität. Sie erklärt ganz explizit: »I do not intend to rehabilitate the autonomous, unitary, perfectly intentional and rational subject, the fundamental unit of liberalism. As for the ›subject‹ or the ›person‹, I hypothesise that there has never been anything to be nostalgic about.« (Rouvroy 2013: 157) Stattdessen bezieht sie sich auf Louis Althusser, Judith Butler und Jacques Derrida, um das Subjekt zu beschreiben, um das es ihr geht: »These ›pragmatic‹ accounts understand the ›self‹ as a process rather than a phenomenon, a process happening between individuals, in a space that both presupposes and constitutes ›the common‹.« (Rouvroy 2013: 158) Algorithmen bedrohen ihr zufolge also genau diesen Prozess, »the inactual, potential dimensions of human existence, its dimensions of virtuality, the conditional mode of what people ›could‹ do, their potency or agency«. (Rouvroy 2013: 158)

Diese Bedrohung wird möglich, weil Algorithmen die Architekturen und Umgebungen, in denen Menschen handeln, kontrollieren.

> »Algotithmic governmentality thus exhibits a new strategy of uncertainty management consisting in minimising the uncertainty associated with human agency: the capacity humans have to do or not to do all they are physically capable of. Effected through the reconfiguration of informational and physical architectures and/or environments within which certain things become impossible or unthinkable, and throwing alerts or stimuli producing reflex responses rather than interpretation and reflection, it affects individuals in their agency […].« (Rouvroy 2013: 155)

Im Gegensatz etwa zu Foucaultianischen Ansätzen, die Algorithmen als neue Form der Subjektivierung (Matzner 2017) oder Biopolitik beschreiben (Cheney-Lippold 2011), geht es laut Rouvroy gar nicht mehr um einen Einfluss auf Subjekte. Algorithmische Gouvernementalität muss sich nicht um Subjekte kümmern, weil ihre Absichten, Intentionen und dergleichen gar nicht wichtig sind. Das ist erst einmal eine treffende Beschreibung vieler datengetriebenen Anwendungen. Hier wird da-

mit operiert, kleinste Datenspuren, die aus der Sicht eines Subjektes völlig belanglos sein mögen, mittels Mustererkennung zu verknüpfen. Was zählt, ist nicht eine umfassende digitale Erfassung eines Subjekts, sondern die Sammlung von möglichst vielen »Signalen« unterschiedlicher Subjekte, die das zu operationalisierende Ziel beschreiben. Wenn es also beispielsweise darum geht, potentielle Kund*innen einer Versicherung zu bewerten, geht es nur darum, was in der Breite der Daten z.B. für die Wahrscheinlichkeit eines Schadensfalls spricht. Der Rest des Subjekts ist egal, oder wie Rouvroy schreibt »does not matter«. Die Idee eines digitalen Abbilds des Subjekts, eines »data double« (Lyon 2014), das dann mit dem realen Subjekt kritisch verglichen werden könnte, zielt also an der Realität vorbei.

Wenn Algorithmen kein Subjekt mehr erfassen wollen, versuchen sie auch keines als Subjekt zu beeinflussen. Stattdessen zielten Algorithmen auf Umgebungen und Architekturen, in denen sich Subjekte vermeintlich frei bewegen könnten. Empfehlungssysteme wählen etwa im Hintergrund mediale Inhalte oder Produkte von Onlineshops (inklusive deren Preise) aus, während das für die Kund*innen eben »der Inhalt« des Mediums oder der Suche im Shop ist. Sie werden nicht als Subjekt adressiert, die Inhalte sind einfach da. Dass andere ganz andere Auswahlen bekommen und an welchem Merkmal der eigenen Person das liegt, ist wenn überhaupt nur durch aufwändige Vergleiche nachvollziehbar. Rouvroy hatte beim Schreiben des Textes sicher auch Anwendungen im Blick, die geplant waren, aber immer noch nicht richtig funktionieren. Biometrische Kontrollen öffnen den einen fast magisch Türen, während sie andere diskret aussortieren. Intelligente Videoüberwachung kontrolliert im Hintergrund Passagierflüsse in der U-Bahn und lässt störende Personen entfernen. All diesen Konfigurationen von Umgebungen ist gemein, dass Rouvroy sie mit behavioristischem Vokabular beschreibt. Sie verändern Handlungsmöglichkeiten, nicht durch Ansprache von Subjekten, sondern arbeiten »through the reconfiguration of informational and physical architectures and/or environments within which certain things become impossible or unthinkable, and throwing alerts or stimuli producing reflex responses« – wie oben zitiert. Rouvroy operiert hier mit Denkfiguren, die an die frühe Kybernetik erinnern. Hier wurde das liberale Menschenbild durch die Idee herausgefordert, auch der Mensch sei nur eine Feedbackmaschine. Gleichzeitig war die Kybernetik aber auch ein liberales Projekt, das technischen Fortschritt verfolgte. So schreibt N. Kathrine Hayles über den Gründer der Kybernetik, Norbert Wiener:

> »Throughout his major writings, he struggled to reconcile the tradition of liberalism with the new cybernetic paradigm he was in the process of creating. When I think of him, I imagine him laboring mightily to construct the mirror of the cyborg. He stands proudly before this product of his reflection, urging us to look into it so that we can see ourselves as control-communication devices, differing in no substantial regard from our mechanical siblings. Then he happens to glance over

> his shoulder, sees himself as a cyborg, and makes a horrified withdrawal.« (Hayles 1999: 87)

Ganz ähnlich wird auch für Rouvroy der Mensch gegenüber dem Algorithmus zu einem Reiz-Reaktionswesen, nur dass hier nicht mehr das liberale Subjekt bedroht ist, sondern die zwischenmenschliche Subjektkonstitution. Was hier verloren geht, beschreibt sie selbst als »Außen«:

> »How do we find an ›outside‹, an excess of the world over reality, a space of recalcitrance from which to gain solidity and to practise critique? [… We] should realise that the fundamental stake – what has to be preserved as a resource antecedent to both the ›subject‹ and sociality, as excess of the world over the algorithmic reality, is ›the common‹; this ›in between‹, this space of common appearance (comparution) within which we are mutually addressed to each other.« (Rouvroy 2013: 160)

Es ist nicht ganz klar, welchen ontologischen Status das menschliche Subjekt in dieser Argumentation hat. Auf der einen Seite könnte diese ganz ähnlich wie die Kritik Amoores gelesen werden. Dann ginge es also darum, dass es im menschlichen Handeln eine prinzipielle Offenheit und konstitutives Aufeinanderbezogen-sein gibt, ein »exzess«, welcher durch algorithmische Anwendungen verdeckt oder ignoriert wird. (Auch wenn der relevante Unterschied zu Amoore bliebe, dass dieses Außen eine rein menschliche Sache ist.) An den Stellen, wo Rouvroy aber von Reiz und Reflex spricht, klingt es so, als sähe sie tatsächlich die Möglichkeit, dass dieses Außen, d.h. menschliche Sozialität durch Algorithmen verschwindet; dass der Mensch eben als Reizreaktionstier anfällig ist für diese Form von Steuerung, in der dann der inhärente Exzess menschlicher Interaktion nicht mehr nur verdeckt oder ignoriert wird, sondern tatsächlich zerstört ist.

Folgt man der zweiten Lesart, so wäre das eine ziemlich technikdeterministische Sicht. Algorithmen könnten dann quasi menschliche Sozialität mit behavioristischer Steuerung überschreiben. Diese Sicht übersieht, dass die Algorithmen nicht so klar und deterministisch funktionieren, wie versprochen. Gerade die unbemerkte, unbewusste Manipulation durch Algorithmen ist ein oft diskutiertes Risiko, das sich aber in der Realität als deutlich komplexer herausstellt, als Menschen nur den richtigen Reiz vorzuwerfen. Hier muss die Kritik also darauf achten, den überzogenen Ansprüchen der Anbieter und Entwickler solcher Software nicht einfach zu folgen. Ein exemplarischer Fall wären die Versprechen von Cambridge Analytica, sogar wichtige Wahlen manipulieren zu können, die sich im Nachhinein als wenig haltbar herausgestellt haben (Denham 2020). Zumindest muss gesehen werden, dass algorithmische Veränderungen von Umgebungen und Handlungsmöglichkeiten nicht einfach Menschen treffen, sondern sozial und kulturell situierte Personen, die entsprechend unterschiedlich auf die algorithmischen Ak-

tivitäten reagieren oder unterschiedlich von ihnen betroffen sind (Matzner 2019a). Diese Differenzen zwischen unterschiedlichen Auswirkungen desselben Algorithmus auf sozio-kulturell verschieden situierte Menschen, bräuchte eben dann eine Kritik von »innen«, das heißt basierend auf der Differenz zwischen verschiedenen sozio-technischen Konfigurationen. Ebenso müssen dann auch die verschiedenen Auswirkungen diverser Algorithmen, die dann nicht mehr alle in dasselbe einfache Reiz-Reaktions-Schema passen, in den Blick genommen werden.

Auch wenn nach der ersten Lesart die Offenheit zwischenmenschlicher Interaktion als prinzipiell gegeben gesehen wird, die dann durch Algorithmen verdeckt oder ignoriert wird, bleibt das Argument in dieser Form bei einer Gegenüberstellung von menschlicher Handlung und technischer Konfiguration von Umgebungen. Für eine ethisch-politische Bewertung fehlt auch hier eine soziale Kontextualisierung der jeweils betroffenen menschlichen Subjekte. Diese läge vor dem theoretischen Hintergrund von Rouvroys Arbeiten sehr nahe. Ihre Quellen wir Althusser, Butler, oder Boltanski haben alle auf unterschiedliche Weise versucht, soziale, ökonomische und kulturelle Differenzen nicht nur in Bezug auf gesellschaftliche Strukturen, sondern auch auf der individuellen Ebene zu fassen. Dass diese Analyse hier fehlt, liegt also nicht am theoretischen Design, sondern eher an einer anderen Interessenlage – vielleicht auch an dem Eindruck, den kybernetische oder behavioristische Beschreibungen vom Verschwinden des Menschen machen können.

4. Schluss

In diesem Text wurden drei unterschiedliche Formen präsentiert, in denen Algorithmen »von außen« kritisiert werden. Die erste Form der Kritik, eine Gegenüberstellung von Mensch und Technik basierend auf einem instrumentellen Technikverständnis, ist schon aus technikphilosophischer Sicht – also noch vor der Frage nach Kritik – vielfältig kritisiert worden. Sie taucht hier dennoch auf, weil sie nach wie vor viele konkrete – und dann auch normative – Vorhaben zum Umgang mit Technik informiert: Gesetze, Data Literacy Workshops, Ethikrichtlinien und mehr. Entsprechend gilt es zu fragen, wie diese Vorhaben zu einem technikthеoretisch angemesseneren Begriff von Technik gebracht werden können.

Für die Frage nach einer Kritik von Algorithmen relevanter sind aber die beiden anderen Formen der Kritik. Beide liefern wichtige Inspirationen, müssen aber durch eine kontextualisierte Kritik »von innen« ergänzt – aber nicht ersetzt – werden. Das »Unberechenbare« als ein »Außen« der Algorithmen zeichnet sich im Gegensatz zu den beiden anderen hier diskutierten Formen dadurch aus, dass es nicht mit einem Gegensatz zwischen Menschen und Algorithmen arbeitet, sondern prinzipielle Unverfügbarkeiten für die Technologie an sich sucht. Es wurde gezeigt, dass die grundlegenden Arbeiten der Informatik, in der Form, wie sie in der aktuellen

Kritik aufgenommen werden, das heißt ohne Einbeziehung der Erkenntnisse zu probabilistischem und statistischem Rechnen, nicht viel zu aktuellen Anwendungen des Maschinellen Lernens und Ähnlichem sagen können. Mit Amoore wurde aber gezeigt, dass die Figur des Unberechenbaren dennoch relevant werden kann. Indem sie Algorithmen als eine Form des Schreibens auffasst, umgeht sie die Einschränkungen, die engere, auf bestimmte Eigenschaften von Informationstechnik ausgerichtete Reflexionen haben. Durch ihre Analyse des Schreibens mit Foucault und Derrida umgeht sie auch die Problematik, die Programmierung als entscheidenden Ort des Schreibens von Algorithmen überzubewerten – oder die Fetischisierung von Code wie Wendy Chun diese Sicht prominent kritisiert hat (Chun 2008). Stattdessen schreiben die Algorithmen selbst und sind damit aber der prinzipiellen Unabgeschlossenheit jeglichen Schreibens ausgesetzt. Kritisch arbeitet diese Unabgeschlossenheit vor allem negativ: gegen eine zu technikdeterministische Sicht, gegen die vermeintliche Objektivität von Algorithmen, gegen die Überbetonung der Macht von Programmierern. Wie diese prinzipielle Unabgeschlossenheit aber jeweils zu füllen ist, wird damit nicht beantwortet. Für Amoores ethnographisch inspirierte Arbeitsweise ist das kein Problem, sie liefert diese Inhalte aus der Empirie. Hier müsste aber nicht nur empirisch, sondern auch normativ angesetzt werden, was dann aber ohne Einbezug der konkreten Anwendung und Situation der Subjekte nicht zu leisten ist.

Rouvroy und ähnlich argumentierende Autor*innen kehren zurück zu einem Gegensatz zwischen Mensch und Technik. Während aber die Seite menschlicher Subjekte hier komplex ko-konstitutiv gedacht ist, wirkt die Technik deterministisch-reduktionistisch. Rouvroy wählt dazu selbst den Begriff des »data behaviorism«. Damit wird implizit auch die sozialtheoretische Komplexität reduziert. Während Menschen in vielerlei Hinsicht aufeinander bezogen sind, und damit sowohl in Machtverhältnissen stehen, als auch die Möglichkeit eines Exzesses, eines Aufbrechens, einer Veränderung besteht, wird diese Differenzierung in beide Richtungen eingeebnet. Gegenüber der Technik wird der Mensch zu einem Reiz-Reaktions-Wesen. Die Bedrohungen, die Rouvroy hier konkret im Kopf hat, sind sicher real, bedürfen aber zweierlei Präzisierung. Inzwischen wurde deutlich, dass Technik nicht alle Menschen gleich bedroht, die vermeintlich basal-behavioristische Wirkweise hat selbst sozio-kulturelle Indizes. Und zweitens funktioniert die Technik nicht so glatt und deterministisch, wie sie hier wirkt. In Bezug auf diesen Punkt wäre es spannend, die zweite Form der Kritik, die eine Unabgeschlossenheit der Technik thematisiert, mit der dritten, welche die Offenheit menschlicher Interaktion betont, zu verbinden. Davon könnte die Amooresche Form der Kritik umgekehrt ein stärker zwischen Technik und Menschen ko-konstitutives und damit auch an sozio-politischer Normativität reichhaltigeres Technikverständnis bekommen.

Literatur

Amoore, L. (2009): Algorithmic War. Everyday Geographies of the War on Terror, in: *Antipode*, 41(1), 49–69.

Amoore, L. (2020): Cloud ethics. Algorithms and the attributes of ourselves and others, Durham: Duke University Press.

Bruns, A. (2019): Are Filter Bubbles Real?, Cambridge: Polity Press.

Cheney-Lippold, J. (2011): A New Algorithmic Identity. Soft Biopolitics and the Modulation of Control, in: *Theory, Culture & Society*, 28(6), 164–181.

Chun, W.H.K. (2008): On »sourcery,« or code as fetish, in: *Configurations*, 16(3), 299–324.

Church, A. (1936): An Unsolvable Problem of Elementary Number Theory, in: *American Journal of Mathematics*, 58(2), 345–363.

Denham, E. (2020): Letter RE. ICO investigation into use of personal information and political influence, 2020. [https://ico.org.uk/media/action-weve-taken/2618383/20201002_ico-o-ed-l-rtl-0181_to-julian-knight-mp.pdf] (Zugriff: 29.05.2024).

Derrida, J. (2001): Writing and difference, London: Routledge.

Foucault, M. (1988): Was ist ein Autor?, in: Ders. (Hg.), Schriften zur Literatur, Frankfurt a.M.: Suhrkamp, 7–31.

Hayles, N.K. (1999): How We Became Posthuman, Chicago: University of Chicago Press.

Kitchin, R. (2017): Thinking Critically about and Researching Algorithms, in: *Information, Communication & Society*, 20(1), 14–29.

Krämer, S. (2021): Digitalism as a Cultural Technique. From Alphanumerics to AI, in: *Goethe.De.*, 2021. [https://www.goethe.de/prj/k40/en/eth/dig.html] (Zugriff: 26.05.2024).

Lyon, D. (2014): Surveillance, Snowden, and Big Data. Capacities, Consequences, Critique, in: *Big Data & Society*, 1(2), 1–13.

Matzner, T. (2017): Opening Black Boxes Is Not Enough. Data-Based Surveillance In Discipline and Punish And Today, in: *Foucault Studies*, 23, 27–45.

Matzner, T. (2019a): Plural, situated subjects in the critique of artificial intelligence, in: Sudmann, A. (Hg.), The Democratization of Artificial Intelligence, Bielefeld: transcript, 109–121.

Matzner, T. (2019b): The Human Is Dead – Long Live the Algorithm! Human-Algorithmic Ensembles and Liberal Subjectivity, in: *Theory, Culture & Society*, 36(2), 123–144.

Matzner, T. (2022): Algorithms as Complementary Abstractions, in: *New Media & Society*, 26(4). [https://doi.org/10.1177/14614448221078604].

Matzner, T. (2024): Algorithms. Technology, Culture, Politics, Abingdon/Oxon: Routledge.

Matzner, T.; Masur, P.K.; Ochs, C.; von Pape, T. (2016): Do-It-Yourself Data Protection – Empowerment or Burden?, in: Gutwirth, S.; Leenes, R.; De Hert, P. (Hg.), Data Protection on the Move, Dordrecht: Springer, 277–305.

Mühlhoff, R. (2019): Menschengestützte Künstliche Intelligenz. Über die soziotechnischen Voraussetzungen von »deep learning«, in: *Zeitschrift für Medienwissenschaft*, 11(2), 56–64.

Pariser, E. (2011): The Filter Bubble. What the Internet Is Hiding from You, London: Viking.

Parisi, L. (2016): Automated Thinking and the Limits of Reason, in: *Cultural Studies ↔ Critical Methodologies*, 16(5), 471–481.

Rohlfing, K.J.; Cimiano, P.; Scharlau, I.; Matzner, T.; Buhl, H.M.; Buschmeier, H.; Esposito, E. et al. (2020): Explanation as a social practice. Toward a conceptual framework for the social design of AI systems, in: *IEEE Transactions on Cognitive and Developmental Systems*, 13(3), 717–728.

Rouvroy, A. (2013): The end(s) of critique. data-behaviourism vs. due-process, in: De Vries, K.; Hildebrandt, M. (Hg.), Privacy, Due Process and the Computational Turn. The Philosophy of Law Meets the Philosophy of Technology, London: Routledge, 143–167.

Seyfert, R.; Roberge, J. (Hg.) (2016): Algorithmic Cultures. Essays on Meaning, Performance and New Technologies, London/New York: Routledge.

Turing, A.M. (1937): On Computable Numbers, with an Application to the Entscheidungsproblem, in: *Proceedings of the London Mathematical Society s2*, 42(1), 230–265.

Zuboff, S. (2016): Wie wir Googles Sklaven wurden, in: *FAZ.NET*, 05.03.2016. [https://www.faz.net/aktuell/feuilleton/debatten/die-digital-debatte/shoshana-zuboff-googles-ueberwachungskapitalismus-14101816.html] (Zugriff: 26.05.2024).

Zuboff, S. (2019): The Age of Surveillance Capitalism. The Fight for the Future at the New Frontier of Power, London: Profile Books.

II
Verständigungsverhältnisse

Redefreiheit, Digitalisierung und die Rolle der Philosophie*

Micha Werner

Abstract: *The ongoing digital transformation of almost all areas of human action and agency calls for a readjustment of the norms that regulate these practices. For example, the digitisation of communicative practices poses new challenges to their functioning. This paper explains some of these challenges and argues that they cannot be met by a normative framework that focuses mainly on defensive (free speech and property) rights. In the context of mediated digital communication, the application of such a framework may even have paradoxical consequences. Accordingly, this paper argues for a broader system of communicative rights, understood as rights to meaningful participation in well-functioning communicative practices.*

Keywords: *communicative rights; freedom of speech; digital transformation; digital media; libertarianism*

Die digitale Transformation fast aller menschlichen Handlungsbereiche legt oft die Neuaneignung und ggf. Neujustierung der Normen nahe, die diese Praxen regulieren. Welchen Beitrag auch die Philosophie dafür leisten kann, soll hier am Beispiel der Redefreiheit illustriert werden. Dazu wird erst an das spezifische Thema herangeführt (1), dann an klassische Begründungen der Rede- und Äußerungsfreiheit erinnert (2) und gezeigt, inwiefern der digitale Kommunikationswandel die mit dem Zeitalter der Massenmedien einsetzende Verschiebung des kommunikativen Flaschenhalses von den Äußerungs- zu den Rezeptionschancen ins Extrem treibt (3). Daher kann die Anwendung klassisch liberaler, primär abwehrrechtlicher Konzepte der Redefreiheit auf digital gewandelte Redekontexte etwa im Hinblick auf die Stellung der sogenannten Intermediäre zu Problemen führen (4), deren Lösung

* Für vielfältige Anregungen und Verbesserungen danke ich Klaus Beck, Stefanie Averbeck-Lietz, Tim Kirchner, Kira Boots, den Teilnehmer:innen des Kolloquiums Praktische Philosophie und den Herausgeber:innen dieses Bandes.

nicht allein von der Selbstverantwortung privatwirtschaftlicher Akteure zu erhoffen ist (5). Welche Rolle die Philosophie bei der Bewältigung solcher Probleme spielen kann, hängt selbst von substantiellen sozialphilosophischen Überzeugungen ab (6). In Abgrenzung von Karsten Webers libertärem Verständnis digitaler Redefreiheit als Informationsverfügungsfreiheit (7) wird dafür plädiert, den abwehrrechtlichen Kern der Redefreiheit durch Teilhaberechte (auf faire Kommunikationschancen) und Partizipationsrechte (auf die Mitgestaltung von Kommunikationsstrukturen) zu ergänzen. Hilfreiche Beiträge zur Konkretisierung dieser kommunikativen Rechte im Kontext eines komplexen Mediensystems und Gefüges funktional ausdifferenzierter Teilöffentlichkeiten kann die Philosophie nur als Partnerin im interdisziplinären Diskurs mit anderen Disziplinen leisten (8).

1. Digitaler Kommunikationswandel

Immer deutlicher wird, dass Digitalisierung und Vernetzung einen kulturellen Umbruch bezeichnen, vergleichbar mit der Durchsetzung des Buchdrucks, wenn nicht gar der Schriftkultur. Im Kern meint Digitalisierung nur die Übersetzung von Informationen verschiedenen Typs (von Texten, Noten, Messergebnissen, Bildern, Film- oder Tonaufnahmen und so weiter) in eine auf Ziffern (*digits*) basierende Form. Revolutionäres Potential entfaltet sie dank der Verwendung eines binären Formats, das es ermöglicht, jene Informationen mittels effizienter Technologien (v.a. elektronisch, magnetisch oder optisch) zu speichern, über vernetzte Systeme zu übermitteln, zu durchsuchen oder noch weitergehend ›maschinell‹ (z.B. mittels selbstlernender neuronaler Netzwerke) zu be- und verarbeiten. Wer wie ich in den 1960er Jahren geboren wurde, hat miterlebt, wie rasant und wie tiefgreifend jene Technologien nahezu alle Bereiche menschlicher Lebenspraxis durchdrungen und transformiert haben: Institutionen, Praktiken und Methoden der Produktion, des Konsums und des ökonomischen Austauschs ebenso wie solche der Forschung, der wissenschaftlichen, privaten oder politischen Kommunikation, der politischen Herrschaftsorganisation, der Personenüberwachung, der Kriegsführung, der kulturellen Selbstverständigung, der Literatur-, Kunst- und Musikproduktion, der religiösen Vergemeinschaftung, des Kochens, der Freizeitgestaltung, der Partnerwahl und der individuellen Selbstsorge und Selbstkontrolle bis hin zu Praktiken der Meditation oder Steuerung des Schlafverhaltens.

Nun gehört zum normativen Selbstbild moderner Gesellschaften, dass sie die von ihnen selbst hervorgebrachten Transformationsprozesse nicht einfach als quasi-naturwüchsiges Schicksal über sich ergehen lassen, sondern vielmehr zum Gegenstand kritischer Reflexion und bewusster Gestaltung machen. Teilt man dieses Ideal, steht man vor der Aufgabe, diese Reflexions- und Gestaltungaufgaben im Ausgang von den bestehenden Strukturen gesellschaftlicher Arbeitsteilung zu organi-

sieren. Nimmt man an, dass die Philosophie auch heute noch aufgerufen ist, Beiträge zur argumentativen Kritik menschlicher Lebensvollzüge und ihrer sozialen und kulturellen Objektivierungen zu leisten, wird man folgern, dass der digitale Wandel auch Ansatzpunkte für philosophische Betätigung bietet. Allerdings ist es eine offene Frage, ob und wie es gerade einer grundlagenkritischen und daher oft vorsichtig-tastenden Disziplin wie der Philosophie gelingen kann, weder von der Vielgestaltigkeit noch von der Geschwindigkeit der digitalen Revolution überwältigt zu werden.

Sucht man die Antwort in der Priorisierung von Themenbereichen nach dem Grad ihrer Dringlichkeit, liegt es nahe, dort anzusetzen, wo die Voraussetzungen und Ressourcen der gesellschaftlichen Selbstverständigung, Selbstkritik und Selbstbestimmung ihrerseits von den digitalen Transformationsprozessen erfasst werden. Das ist etwa dort der Fall, wo KI-basierte Systeme unsere bisherigen Auffassungen von Autonomie, Verantwortlichkeit, Zurechenbarkeit oder Autor:innenschaft in Frage stellen, und ebenso dort, wo der digitale Wandel die gesellschaftlichen Verständigungsverhältnisse umwälzt (Hahn/Langenohl 2017) – und damit in jene Sphäre eingreift, in der die kritische Reflexion auf diese Umwälzung selbst erfolgen muss. Die folgenden Überlegungen konzentrieren sich auf den zweiten Themenbereich, den digitalen Wandel[1] der Strukturen öffentlicher Kommunikation.

Würdigen wir zunächst die neuen Möglichkeiten, die durch die Entwicklung digitaler Kommunikationstechnologien eröffnet werden: Diese Technologien ermöglichen interaktive (auch synchron-interaktive) und »selbstvermittelte« Kommunikationsprozesse zwischen Nutzer:innen, die grundsätzlich über den gesamten Globus verteilt sein können (Schönhagen 2004). Die Digitalisierung hat die Kosten der Informationsverbreitung radikal vermindert[2] und damit potentiell – hier sind die Befunde weniger eindeutig (Deibert/Villeneuve 2004; Chang/Lin 2020) – umgekehrt den Aufwand staatlicher Zensurmaßnahmen erhöht. Grundsätzlich erschwert auch die Dynamik digitaltechnologischer Entwicklungen und bereits die Tatsache, dass neben den alten Medien neue Alternativangebote entstanden sind, die Konsolidierung von Einflussmonopolen. Schließlich eröffnen die Verbindung von Text-, Bild- und Tonübertragung und Technologien der künstlichen Datenverarbeitung gänzlich neuartige und faszinierende Möglichkeiten zur Gestaltung von Kommunikationsprozessen, etwa durch neue Formen der Textauswertung, der

1 Der gängige Begriff des »digitalen Wandels« wird der Kürze halber auch hier verwandt; er darf aber nicht im Sinne eines technologischen Determinismus missverstanden werden. Was er bezeichnet, sind nämlich Veränderungen, die durch die Entwicklung digitaler Technologien zwar allererst ermöglicht, aber nicht auch schon durch sie allein erzwungen werden, sondern stets durch zusätzliche (kulturelle, sozioökonomische…) Faktoren zu erklären sind.

2 Jedenfalls für diejenigen, die überhaupt Netzzugang haben; vgl. Riehm/Krings 2006.

Datenpräsentation, der Musikvermittlung und der multimedialen Verschränkung von Inhalten, durch neue Assistenztechnologien für Personen mit auditiven oder visuellen Einschränkungen, durch digitale Wörterbücher und Übersetzungssysteme, durch personalisierte Such- und Vorschlagsalgorithmen und jüngst auch durch *chatbots* und andere Anwendungen generativer KI.

Den zahlreichen neuen Möglichkeiten digitaler Kommunikation entsprechen aber auch neue Herausforderungen für die verantwortliche Gestaltung von Kommunikationsprozessen. Diese Herausforderungen betreffen zum einen die möglichen Hebel, an denen Gestaltungsversuche überhaupt ansetzen können. Schwierigkeiten der Regulierung digitaler Netzkommunikation, die durch deren transnationalen Charakter und/oder durch VPN's, Overlay-Netzwerke und starke Verschlüsselung geschaffen werden, mag man im Hinblick auf Zensurversuche autoritärer Regierungen begrüßen. Sie erschweren aber auch potentiell legitime gesellschaftliche Regulierungen etwa im Interesse des Jugendschutzes, zum Schutz der Privatsphäre, zur Verfolgung von Drogen-, Waffen- oder Menschenhandel oder der Einhegung von Hassrede (Bromell 2022; Frischlich 2022). Zum anderen erfordert die digitale Transformation von Kommunikationsprozessen auch eine kritische Überprüfung der Normen und Ideale, die für die Regulierung und die Bewertung dieser Prozesse einschlägig sind.

Soweit die Wissenschaft Beiträge zur kritischen Reflexion dieses Wandels leisten kann, ist wesentlich die Kommunikations- und Medienwissenschaft angesprochen, die die Beschreibung, Erklärung und Deutung von Medien, Kommunikationsstrukturen, -mustern und -praktiken als einen ihrer zentralen Forschungsgegenstände betrachtet (Alm et al. 2022). Im Hinblick auf ihre (vielfältigen) Fragestellungen, Deutungsperspektiven, Methoden und Erklärungsansätze (Schweiger/Beck 2019) steht sie im Austausch mit anderen Wissenschaften (Sutter/Mehler 2010: 7ff.), darunter auch der Philosophie. Welche spezifischen Beiträge man sich dabei von der Philosophie erhofft, hängt wesentlich vom Philosophieverständnis ab. Relativ unkontrovers (und entsprechend vage) ist die Auffassung, dass Philosophie eine Rolle bei der Interpretation, Klärung und Begründung begrifflicher, methodischer und normativer Grundlagen sinnvoller und vernünftiger Lebensvollzüge und ihrer Objektivierungen spielen kann.[3] Aus dieser allgemeinen und vagen Bestimmung folgt bereits, dass eine Zusammenarbeit zwischen Kommunikationswissenschaft und Philosophie grundsätzlich auf zwei verschiedenen Ebenen angesiedelt

3 Strittiger ist, welche Methoden der Philosophie dabei zur Verfügung stehen und inwieweit diese sich von den Methoden anderer Wissenschaften unterscheiden. Entsprechend ist auch strittig, wie stark oder schwach, kontextrelativ oder voraussetzungsvoll mittels philosophischer Methoden einlösbare Geltungsansprüche sein können. Für ein Spektrum jüngerer Angebote vgl. Ragland/Heidt 2001.

sein kann, da sowohl Kommunikationspraxen selbst als auch die auf sie reflektierende Kommunikationswissenschaft als Gegenstände philosophischer Reflexion in Frage kommen.

Entsprechend kann Philosophie zum einen aus sprachphilosophischem, logischem, kulturphilosophischem, sozialphilosophischem oder kommunikationsethischem Erkenntnisinteresse ihren Blick auf Kommunikationsakte, -praxen, -formen oder -strukturen richten. In dieser Hinsicht ist sie Partnerin der Kommunikationswissenschaft, die teils überlappende, teils ergänzende Projekte verfolgt. So werden sich, da menschliche Kommunikation grundsätzlich eine normregulierte Praxis ist, sowohl die Kommunikationswissenschaft als auch die Philosophie für Beschaffenheit, Funktion und Wandel dieser Normen interessieren. Erstere wird dies jedoch primär in der Perspektive einer empirisch beschreibenden, deutenden und erklärenden Disziplin tun (s. vergleichend Averbeck-Lietz 2017), die Philosophie dagegen primär im Interesse einer normativen Rekonstruktion oder Kritik. Da einerseits auch die kommunikationswissenschaftliche Deutung und Erklärung sozialer Institutionen, Praxen und Akte die Frage der objektiven Gültigkeit der diese Institutionen, Praxen und Akte stützenden Werte oder Normen in Betracht zieht (Zillich et al. 2016) oder aus methodologischen Gründen ziehen muss (Habermas 1981) und da andererseits die philosophischen Bemühungen um die Begründung von Werten oder Normen nicht indifferent gegenüber deren realer sozialer Bedeutung und Einbettung sein können, ist eine Zusammenarbeit beider Disziplinen für beide potentiell profitabel, wenn nicht sogar unvermeidlich.

Zum anderen kann Philosophie auch aus wissenschaftstheoretischem Interesse auf die Kommunikationswissenschaft als Forschungsgegenstand reflektieren. In dieser Hinsicht kann sie sowohl als methodische Hilfsdisziplin der Kommunikationswissenschaft auftreten und deren Bemühungen um eine Reflexion auf das eigene disziplinäre Selbstverständnis (Geise et al. 2021a; Geise et al. 2021b) unterstützen als auch übergreifende wissenschaftsphilosophische Ziele verfolgen. Als verbliebener Rumpf einer früheren Universalwissenschaft, aus der heraus sich bis in jüngste Zeit immer wieder Einzelwissenschaften ausdifferenziert haben, kann Philosophie auch eine Rolle bei der Übersetzung von Beiträgen aus verschiedenen Einzeldisziplinen spielen, oder, in der Rolle einer bewussten Einzelwissenschafts-Dilettantin, Anstöße für potentiell fruchtbare Kooperationen liefern.

Im Folgenden möchte ich den ersten Bereich möglicher Zusammenarbeit zwischen Philosophie und Kommunikationswissenschaft bei der Reflexion auf den digitalen Wandel exemplarisch illustrieren. Am Beispiel der Redefreiheit möchte ich zeigen, dass der digitale Wandel neue Fragen hinsichtlich der Bedeutung, Institutionalisierung und Rechtfertigung kommunikativer Grundnormen aufwirft, zu deren Klärung auch die Philosophie beitragen kann.

2. Kommunikationsgrundrechte

Von vorrangiger Bedeutung für die Reflexion des digitalen Kommunikationswandels ist aus normativer Sicht die Frage, inwieweit er Menschen- oder grundlegende Bürger:innenrechte betrifft. In der Literatur finden sich unterschiedliche Konzepte einschlägiger Rechte, und zwar in der Regel Freiheitsrechte wie Äußerungs- und Redefreiheit(en), Informationsfreiheit(en) oder Kommunikationsfreiheit(en). Diese Kommunikations(grund)rechte weisen großflächige Überlappungen auf, sind aber nicht deckungsgleich. Rede- und Äußerungsfreiheit werden insbesondere in der angelsächsischen Diskussion oft austauschbar verwandt, weil das Konzept von »Rede« juristisch so weit interpretiert wird, dass es auch nicht-symbolische Ausdrucksformen wie Musik oder abstrakte Kunst einschließt (Tushnet/Chen/Blocher 2017). Die Formulierung in Art. 5, Abs. 1 GG schützt zugleich mit der Äußerungsfreiheit auch die »Pressefreiheit und die Freiheit der Berichterstattung durch Rundfunk und Film«. Das Konzept der »Informationsfreiheit« geht einerseits deutlich über das Konzept der Äußerungsfreiheit hinaus, insofern es zum einen auch exklusive private Verfügungsrechte über Informationen (also auch Rechte auf Nicht-Äußerung) und zum anderen auch Rechte auf Informationszugang (ebenfalls durch Art. 5, Abs. 1 GG geschützt) einschließt (Weber 2009). Anderseits scheint es auf nicht-symbolische Ausdruckshandlungen nicht unmittelbar anwendbar. Das Konzept der Kommunikationsfreiheit(en) (Beck 2021) bzw. -grundrechte (Sell 2017; Asscher 2002) stellt individuellen Abwehr-, Verfügungs- oder Zugangsrechten typischerweise auch partizipative Rechte auf Teilhabe an und Mitgestaltung von gemeinschaftlichen Praxen und Institutionen der Kommunikation zur Seite.

Rechte auf Äußerungs- und Redefreiheiten sind in den Rechtssystemen der westlichen Demokratien am stärksten verankert. Als *moralische* Rechte reichen sie – wie alle moralischen Rechte – genau so weit wie die Gründe, die für sie angeführt werden können. Klassischerweise werden als zentrale Argumentationstypen vor allem wahrheitsbasierte (Mill 1977), autonomie- (Scanlon 1972; Brison 1998) und demokratiebasierte Argumente (Meiklejohn 1948; Schauer 1983; Sunstein 1993) angeführt (häufig, wie schon bei Mill, miteinander verflochten); daneben (oder wiederum damit vermittelt) finden sich u.a. auch würde-, selbstentfaltungs-, toleranz-, glücks- oder tugendbasierte Argumente (zum Überblick vgl. Barendt 2007: 1ff.; Greenawalt 1989; Schauer 1982: 3ff.). Für die Überzeugung, dass wir Redefreiheiten schützen sollten, sprechen also Annahmen wie die, dass diese Freiheiten nötig sind für eine funktionierende Demokratie, oder für die kooperative Wahrheitssuche, oder dass sie unmittelbar als Aspekte ihrerseits begründeter, noch allgemeinerer Rechte oder Grundfreiheiten zu verstehen sind, etwa auf Individualität, Selbstbestimmung und freie Selbstentfaltung. Diese Annahmen speisen

dann Argumente, die teils konsequentialistisch und gemeinwohlorientiert und teils nicht-konsequentialistisch und teils auch ›absolutistisch‹ ausbuchstabiert werden.[4]

Diese Argumente – und ebenso die spezifischen Formulierungen der Rede-, Presse- oder Informationsfreiheit, die in den Rechtssystemen liberaler Demokratien verankert sind, sowie die Traditionen ihrer Interpretation und Abwägung gegen andere Grundrechte – sind jedoch wesentlich noch im Kontext von Kommunikationsstrukturen entwickelt worden, die sich von denen des digitalen Zeitalters deutlich unterscheiden. Es ist daher zu fragen, inwieweit der digitale Kommunikationswandel Auswirkungen auf unser Verständnis, auf die Stichhaltigkeit von Begründungen oder den relativen Stellenwert der Redefreiheit hat oder haben sollte.

3. Digitale Revolution der Kommunikationsökonomie

Betrachten wir zunächst (in stark stilisierter Darstellung) einige Merkmale der Öffentlichkeit im 18. und 19. Jahrhundert. Information war vergleichsweise knapp. Bei deutlich niedrigerem und zudem noch ungleicher verteiltem sozialen Wohlstand waren die Kosten der Verbreitung geschriebener Texte deutlich höher als heute. Chancen auf die aktive und passive Teilnahme an gesellschaftlichen Verständigungsprozessen waren durch ökonomisch bedingte Grenzen des Bildungszugangs beschränkt. Als Mittel der gezielten Minderung oder Unterdrückung von Kommunikationschancen kam vor allem die direkte Unterdrückung der Vervielfältigung oder Verbreitung von Äußerungen in Frage – durch Zensur, die bei der Druckerpresse ansetzt. Die nachträgliche Unterdrückung der Rezeption bereits in Umlauf gebrachter Äußerungen durch lesekundige und um Neuigkeiten aktiv bemühte Bürger:innen war vergleichsweise mühsam. Auf indirekte Weise lässt das der Vorbericht in Büchners *Hessischem Landboten* [1834] erkennen, der Leser:innen vor den Konsequenzen möglicher Zensurverletzung schützen will:

> »Dieses Blatt soll dem hessischen Lande die Wahrheit melden, aber wer die Wahrheit sagt, wird gehenkt, ja sogar der, welcher die Wahrheit liest, wird durch meineidige Richter vielleicht gestraft. Darum haben die, welchen dies Blatt zukommt, Folgendes zu beobachten:

4 Die zentrale Rolle, die John St. Mills Überlegungen noch in der gegenwärtigen Diskussion spielt (sichtbar etwa in van Mill 2021), erklärt sich vermutlich aus dem ›hybriden‹ Charakter seiner Position: Indem Mill einerseits beispielsweise den Nutzen fürs gemeinschaftliche Erkenntnisstreben, andererseits das grundlegende Individualrecht auf Handlungsfreiheit betont, das nur bei drohendem Schaden für Andere einzuschränken ist, verschränkt er Elemente, deren Verhältnis zueinander Raum für Interpretationen und unterschiedliche Schwerpunktsetzungen lässt.

1. Sie müssen das Blatt sorgfältig außerhalb ihres Hauses vor der Polizei verwahren;
2. sie dürfen es nur an treue Freunde mitteilen;
3. denen, welchen sie nicht trauen, wie sich selbst, dürfen sie es nur heimlich hinlegen;
4. würde das Blatt dennoch bei einem gefunden, der es gelesen hat, so muss er gestehen, dass er es eben dem Kreisrat habe bringen wollen;
5. wer das Blatt nicht gelesen hat, wenn man es bei ihm findet, der ist natürlich ohne Schuld.« (Büchner 2016: 5)

Unterstellte man eine hinreichend funktionierende Konkurrenz von Buch- und Zeitungsverlagen, Vereinsblättern, Parteizeitungen und Flugschriften, war Zensur primär von der Staatsmacht zu fürchten. Der Schutz privater Eigentumsrechte und Gewerbefreiheiten, insbesondere von Zeitungs- und Buchverlagen, harmoniert unter jener – mehr oder weniger kontrafaktischen[5] – Unterstellung funktionierender Meinungskonkurrenz mit dem Schutz der Kommunikationschancen wenigstens des gebildeten Teils der Gesellschaft. In diesem Rahmen war es für die Gebildeten in gewisser Hinsicht naheliegend, Kommunikationsgrundrechte wesentlich als Abwehrrechte gegen staatliche Einmischung zu verstehen, das heißt als Abwehrrechte auf freie Meinungsäußerung, auf Pressefreiheit und die freie Verfügung von Verleger:innen über ihr Privateigentum. Sind solche Rechte sozial etabliert, lassen sie sich in einem zweiten Schritt gegebenenfalls noch sozialstaatlich ergänzen, etwa durch Anspruchsrechte auf Bildungszugang und eine informationelle Grundversorgung (Kubicek 1996; Weber 2005: 183).

Schon mit dem Aufkommen der (elektronischen) Massenmedien haben sich die technischen und ökonomischen Rahmenbedingungen jedoch grundlegend gewandelt. Erst recht unterscheiden sich die Rahmenbedingungen digitaler Kommunikation von denen des 18. und 19. Jahrhunderts. Für diejenigen Personen, die über Netzzugang verfügen, tendieren nun die Kosten der globalen Informationsverbreitung gegen Null. Entsprechend schwillt das Volumen der Redebeiträge an und – dies ist entscheidend – verlagert sich der die Kommunikationschancen einzelner Sprecher:innen beschränkende Flaschenhals von den Chancen auf die *Äußerung und Verbreitung* der eigenen Rede zu den *Rezeptions*chancen (Franck 1998; Goldhaber 1997; Beck/Schweiger 2001). Knapp und kostbar sind nun vor allem die Chancen darauf,

5 Um den Eindruck zu vermeiden, einer »ungerechtfertigten Idealisierung« (Habermas 1990: 15; vgl. 34 ff.) der vor-elektronischen bürgerlichen Öffentlichkeit Vorschub zu leisten sei daran erinnert, dass die vorliegenden Bemerkungen nicht als Beitrag zur Geschichte der Medienökonomie gedacht sind. Sie sollen erklären, in welchem Kontext, die Fokussierung auf eine Redefreiheit, die als Abwehrrecht gegenüber dem Staat verstanden wird, plausibel erscheint. Dieser Kontext kann durchaus auch idealisierende Aspekte zeitgenössischer Selbstdeutungen beinhalten.

dass die eigene Stimme irgendwie durch das betäubende Brausen des Meinungsmeers hindurchdringt, indem beispielsweise einzelne *tweets*, *blogs*, Websites oder Videos einen vorderen Platz in Suchergebnissen finden oder von *social media*-Algorithmen anderen Nutzer:innen vorgeschlagen werden; eine Situation, die häufig mit dem Label der »Aufmerksamkeitsökonomie« bezeichnet wird.

Tim Wu, der als Rechtswissenschaftler den Begriff der Netzneutralität geprägt hat, hält diese Verschiebung des kommunikativen Flaschenhalses für den gravierendsten Aspekt des Wandels der Kommunikationsumgebung. Nicht länger sei die Rede selbst knapp, sondern die Aufmerksamkeit der Hörer:innen. Entsprechend gehen laut Wu auch die neuen Bedrohungen für den öffentlichen Diskurs von diesen veränderten Bedingungen aus.[6] Im Rahmen der Netzkommunikation liegt nämlich ein lohnender Ansatzpunkt für die zielgerichtete Manipulation von Kommunikationsprozessen in der gezielten Beeinflussung der *Chancen auf die Auffindbarkeit* von in gewisser Weise bereits ›verbreiteten‹ oder ›geteilten‹ Redebeiträgen durch potentielle Rezipient:innen. In anderen Worten: Entsprechende Strategien setzen nicht mehr an der Quelle an, sondern sozusagen bei Eingriffen am oder im Leitungsnetz, und zwar (denkt man beispielsweise an personalisierte Newsfeeds) sozusagen bis hin zum individuellen Wasserhahn.

Ein ebenso alltäglich-triviales wie prägnantes Beispiel für solche Strategien, das aus dem Bereich der kommerziellen Kommunikation stammt, liegt in dem Versuch von Online-Handelsplattformen, über ein umfangreiches Netz zunächst unverdächtiger ›Partner‹-Webseiten möglichst viele Suchanfragen potentieller Kund:innen auf die eigenen Server zu lenken. Informationen über die auf alternativen Plattformen verfügbaren Angebote bleiben dann zwar prinzipiell verfügbar; sie lassen sich jedoch auf diese Weise unter der Flut eigener ›suchmaschinenoptimierter‹ Beiträge begraben.[7] Analoge Strategien werden auch im Bereich der politischen Rede verwandt. Der gut dokumentierte Betrieb von Jewgeni W. Prigoschins Sankt Petersburger Trollfarm (Bastos/Farkas 2019; Dawson/Innes 2019) macht(e) sich

6 »The most important change in the expressive environment can be boiled down to one idea: it is no longer speech itself that is scarce, but the attention of listeners. Emerging threats to public discourse take advantage of this change.« (Wu 2018: 548)

7 Anreize für die Partnerseiten, auf das eigene Angebot zu verlinken, können dank der Nachverfolgbarkeit des Nutzer:innenverhaltens durch die ökonomische Beteiligung an den generierten Transaktionen oder Seitenaufrufen generiert werden. Eine Amazon-Informationsseite für Partnershops und -websites erläutert: »Mit Tracking-IDs können Partner die Performance verschiedener Websites oder Werbestrategien analysieren und gleichzeitig Werbekostenerstattungen […] akkumulieren. Beispielsweise könnte der Partner partnerID-21 mithilfe der Tracking-ID partnerID-1-21 die Aktivitäten in seinem Shop verfolgen und mithilfe der Tracking-ID partnerID-2-21 darüber hinaus die Aktivitäten seines Newsletters messen. Unter jeder Tracking-ID werden Werbekostenerstattungen für sein Partnerkonto partnerID-21 generiert.« (Amazon 2023)

ebenso wie die im Auftrag der chinesischen Staatspropaganda tätige sogenannte ›freiwillige 50 Cent-Armee‹ (Han 2018: 101ff., 152ff.) die günstigen Kosten der Informationsverbreitung zunutze, um Informationskanäle mit den gewünschten Inhalten zu fluten. Auch wenn das Ziel, konkurrierende Informationen vollständig unsichtbar zu machen, dabei nicht immer verfolgt oder gar tatsächlich erreicht wird, lässt sich durch die breite Streuung gewünschter Inhalte die Informationsumgebung potentiell in einer aus Sicht der jeweiligen Akteure gewünschten Weise beeinflussen; etwa indem die wahrgenommene Glaubwürdigkeit der verbreiteten Inhalte durch wiederholte Rezeption (»truth effect«: Dechêne et al. 2010) erhöht oder ein bestimmtes Meinungsklima simuliert wird, um Konformitätseffekte zu erzielen, oder um Unsicherheit und Misstrauen gegenüber Institutionen zu schüren (Bennett/Livingston 2018). Hendricks und Vestergaard formulieren prägnant:

> »Traditionally, tyrants' strategy has been to keep the people's information level at an absolute minimum as means of exerting power. Through censorship and punishment, oppressors could deprive people of the sources of information considered problematic. […] However, in the age of information, a similar propagandistic effect may be obtained by […] drowning […] citizens, voters, and media in misinformation and noise. This [is] realized without censorship and suppression of freedom of speech.« (Hendricks/Vestergaard 2019: xiii)

Entsprechende Strategien lassen sich noch weiter rationalisieren, seit sich auch die Textgenerierung digitalen Systemen übertragen lässt (Milmo/Hern 2023). Trollfarmen können durch die Nutzung von generativer KI für die Texterstellung sowohl ihre Kosten senken als auch Gefahren durch *whistleblower* oder andere Folgen menschlicher Unkalkulierbarkeit mindern. Wenn die digitale Automatisierung neben Gestaltung und Austausch auch die Produktion von Informationen erfasst, ist die von Wu beschriebene Umwälzung der Kommunikationsökonomie von einer Knappheits- zu einer Überflussökonomie an einen Extrempunkt gelangt. Knapp und kostbar bleiben dabei weiterhin – und, relativ betrachtet, mehr denn je – die zur Verfügung stehende Zeit, Aufmerksamkeit und Verarbeitungskapazität von Kommunikationspartner:innen.

Das Label »Aufmerksamkeitsökonomie« benennt die praktisch relevanten Aspekte und Konsequenzen dieser Situation allerdings nur in einer eigentümlich selektiven und perspektivisch verzerrten Weise. Die durch den Begriff hervorgehobene Knappheit der Chancen auf das Rezipiertwerden von verlautbarten Äußerungen erscheint salient primär aus der Perspektive von (potentiellen) von *Informationsanbieter:innen bzw. Sprecher:innen*, die mit ihren Mitteilungen eine mehr oder weniger spezifizierte Gruppe möglicher Rezipient:innen erreichen möchten. Die Kombination aus (Des-)Informationsüberfluss einerseits, begrenzten Rezeptions- und Verarbeitungsressourcen andererseits, bringt jedoch für die *Informati-*

onssuchenden bzw. Rezipient:innen ebenso große Herausforderungen mit sich. Diese Herausforderungen werden durch den Begriff der »Aufmerksamkeitsökonomie« eher verdeckt. Denn knapp und kostbar ist aus der Rezipient:innenperspektive ja keineswegs die eigene Aufmerksamkeit für inhaltlich beliebige Kommunikationsbeiträge. Vielmehr sind es die eigenen Chancen, sich in der Überfülle des Informationsangebots sinnvoll zu orientieren, Beiträge einzuordnen und genau diejenigen Inhalte zu selektieren, die am Maßstab der aufgeklärten eigenen Interessen und Werte tatsächlich rezeptionswürdig sind. Kostbar und knapp sind aus ihrer Sicht also vor allem die Ressourcen für eine qualifizierte, zuverlässige oder wenigstens vertrauenswürdige oder auf transparenten Kriterien basierende Einordnung und Selektion von Kommunikationsbeiträgen.

Im Vergleich zu vordigitalen Kommunikationsstrukturen deutlich verändert haben sich schließlich auch die Technologien und sozioökonomischen Institutionen, die die Koordination der Perspektiven der (verschiedenen Gruppen von) Sprecher:innen und Rezipient:innen leisten und die das Spielfeld prägen, auf dem der Ausgleich ihrer teils komplementären und teils konkurrierenden Interessen ausgehandelt wird. Die vor-elektronischen Orientierungs- und Selektionsmechanismen des Verlagswesens, des kritischen Journalismus und der Kaffeehausgespräche sind zwar keineswegs verschwunden, in den vergangenen Jahrzehnten jedoch erheblich zurückgedrängt worden. Ablesen lässt sich dies beispielsweise am veränderten Zahlenverhältnis zwischen Journalist:innen und PR-Expert:innen, die im Auftrag von Unternehmen, Regierungsinstitutionen, Parteien oder sonstigen Interessengruppen strategischen Einfluss auf die Informationsumgebung nehmen. Der Kommunikationswissenschaftler Stephan Ruß-Mohl hält mit Verweis auf Greenslade (Greenslade 2014) fest:

> »Mit der Digitalisierung verschiebt sich beschleunigt die Machtbalance zwischen Journalismus und PR: Kam in den [19]80er Jahren in den USA statistisch noch auf einen PR-Experten jeweils ein Journalist, so wurde daraus bis zum Jahr 2008 jeweils eine Übermacht von vier bis fünf PR-Experten pro Journalist […].« (Ruß-Mohl 2022)

Seit 2008 hat sich das Verhältnis noch weiter verschoben; 2018 kamen auf jede:n Journalist:in bereits sechs PR-Professionals, wobei letztere im Durchschnitt auch etwa ein Drittel mehr verdienten als die Journalist:innen (Schneider 2018).

Gänzlich neu ist im Vergleich mit der vordigitalen Kommunikationsumgebung die Rolle der sogenannten *Informationsintermediäre* (kurz Intermediäre). Die Rezeptionschancen digitaler Kommunikationsbeiträge liegen vielfach in der Hand – genauer gesagt: in der Verfügung der als Betriebsgeheimnis gehüteten Computeralgorithmen – dieser meist privatwirtschaftlich organisierten Akteure. Einige von ihnen bilden in ihrer jeweiligen Domäne zumindest regionale De-facto-Monopole oder Oli-

gopole. So lag der Marktanteil von *Alphabets Google* bei Suchmaschinen im April 2023 bei über 90% (StatCounter); im Bereich der sozialen Kurznachrichten hat *X* (vormals *Twitter*) eine zentrale Stellung inne, im Bereich der sozialen Netzwerke und *social videos* dominieren *Meta Platforms* (zu dem neben *Facebook* auch *Instagram*, *WhatsApp* und *Messenger* gehören), *TikTok* und *YouTube* (letzteres wiederum zu *Google* gehörig), wobei auch größere Konkurrenten wie das zu *Alphabet* gehörige Business-Netzwerk *LinkedIn* oder das zu *Meta* gehörende *Threads* oft anderen Mitgliedern der sogenannten *big five* der IT-Branche (*Alphabet, Amazon, Apple, Meta, Microsoft*) gehören. Die Konzentration im Bereich der Informationsintermediäre ist dabei kein Zufall, sondern wesentlich das Ergebnis von Netzwerkeffekten: So ist der Zugang zu sozialen Netzwerken für Nutzer:innen wie für Werbetreibende typischerweise desto wertvoller, je mehr Mitglieder schon beigetreten sind.

Die präzedenzlose Rolle der Intermediäre ist unter anderem dadurch gekennzeichnet, dass sie zugleich als Kommunikationsplattformen dienen, auf denen Nutzer:innen Informationen austauschen können, Kontaktaufnahmen zwischen Kommunizierenden nahelegen, dabei durch technische Vorgaben wesentlichen Einfluss die Form möglicher Kommunikationsbeiträge ausüben und schließlich vielfach selbst Inhalte darbieten (die teilweise auf eigene weitere digitale oder nicht-digitale Produkte bezogen sind) und/oder diese Inhalte jedenfalls aggregieren oder KI-basiert weiterverwerten (was wiederholt zu heftigen Konflikten über Urheberschafts- und Verwertungsrechte führt) und auch priorisieren. *Last but not* least verfügen sie über nie dagewesene Fähigkeiten der Gewinnung und Auswertung von Nutzer:innendaten und des Monitorings von Kommunikationsprozessen (Beck 2021: 97ff.; Gabriel 2023). Die gewonnenen Einsichten in individuelle Einstellungen, Präferenzen und Persönlichkeitsprofile nutzen sie nicht nur zur personalisierten Priorisierung von Äußerungen der Mitglieder. Sie vermarkten diese Einsichten auch als Ressource für vielfältige Nutzungszwecke, insbesondere an Anzeigenkunden, deren Zahlungen in der Regel die wichtigste, in manchen Fällen sogar die einzige Einnahmequelle der Intermediäre darstellen. Im Kontext eines Plädoyers für Kommunikationsfreiheiten im Sinne ›positiver‹ Ermöglichungsrechte betont Andrew T. Kenyon 2021 die tiefgreifenden Veränderungen, die sich durch die Fähigkeit der Intermediäre ergeben haben, Kommunikationsströme zu steuern und kommunikative Kanäle zwischen verschiedenen Teilnehmer:innen(gruppen) zu erweitern oder zu verengen:

> »[M]ediated speech now is very different than it was ten or twenty years ago. It now includes major internet intermediaries reshaping content creation, circulation and use – and reformulating publics as they go – indeed, perhaps enabling the targeting of speech that is not even public in ways that have generally been understood before […]. That targeting of speech arises through corporate and state surveillance of a form and extent that was impossible in the twentieth century,

and through powerful processes of automation. They are reshaping the circulation of speech and changing who sees what in terms of content; they are arguably changing the ›exposure diversity‹ of recipients.« (Kenyon 2021: 11)

4. Mögliche Folgen digitaler Redefreiheit als Abwehrrecht

Ein quantitativ erheblicher Teil der digitalen Verständigung findet damit in einem technologischen und informationsökonomischen Kontext statt, dessen Strukturen etwa 1791, als das Recht auf *free speech* im Ersten Zusatz zur US-Verfassung etabliert wurde, oder 1859, als John St. Mill in *On Liberty* seine Überlegungen zur Äußerungsfreiheit entwickelt hat, nicht einmal ansatzweise erkennbar waren. Nicht absehbar war daher auch, welche Konsequenzen eine Institutionalisierung kommunikativer Grundfreiheiten, die wesentlich um die Äußerungsfreiheit im Sinne eines Abwehrrechts gegen den Staat gruppiert ist, im Kontext digitaler Verständigungsverhältnisse zeitigen könnte.

Bei der Skizze dieser Konsequenzen möchte ich mich auf die Regulierung in den USA konzentrieren.[8] Interessant ist der US-amerikanische Regelungskontext erstens, weil das Recht auf *free speech* darin einen besonders hohen verfassungsrechtlichen Rang genießt und klar als Abwehrrecht gegen den Staat interpretiert wird, so dass sich die möglichen Konsequenzen eines entsprechenden Verständnisses in besonders klarer Weise illustrieren lassen. Zweitens ist die US-amerikanische Diskussion relevant, weil die auch hier dominierenden Suchmaschinen und sozialen Medien (außer TikTok) dort ihren Hauptsitz haben. Obgleich nationale und europäische Regelungen wie das deutsche Gesetz zur Verbesserung der Rechtsdurchsetzung in sozialen Netzwerken (NetzDG) oder die europäische Datenschutz-Grundverordnung (DSGVO) und demnächst der Digital Services Act (DSA) und AI-Act auch für ausländische Unternehmen gelten, resultieren daraus Durchsetzungsprobleme (exemplarisch Krempl 2023) und stehen lokale Regulierungsbemühungen unter handelspolitischem Druck. Drittens werden bereits in der Zeit vor-digitaler Massenmedien entwickelte Medienregulierungen und sie tragende oder flankierende Konzepte wie das einer ›positiven Medienordnung‹ (Hoffmann-Riem 2002) nicht nur durch digitaltechnologische Entwicklungen herausgefordert (Hoffmann-Riem 2020), sondern auch durch solche der politischen Kultur.

8 Die im vorliegenden Kontext unvermeidlichen Vereinfachungen könnten den falschen Eindruck nahelegen, dass ein gerader Weg von den Auffassungen der ›Verfassungsväter‹ zu der die aktuelle Verfassungsrechtsprechung der USA prägenden Interpretation der Redefreiheit führt. Tatsächlich ist der Weg jedoch keineswegs gerade, und über die Auffassungen der *framers* liegen in Bezug auf das *First Amendment* unzureichende Informationen vor (Graber 1991; Bunker 2012; Smolla 2020; mit Verweis auf mögliche Implikationen für die Diskussion über Netzneutralität Feldman 2017).

Schließlich sind auch aus der politischen Philosophie und Medienethik Plädoyers für die Orientierung am US-amerikanischen Vorbild und für die Konzentration auf eine als ›klassisches und zentrales liberales Recht‹ verstandene Meinungsfreiheit zu vernehmen (Weber 2009: 19). Darauf möchte ich später zurückkommen.

Unter den Konsequenzen der Fokussierung auf eine abwehrrechtlich verstandene Meinungsfreiheit für die US-amerikanische Regulierung digitaler Kommunikation möchte ich folgende drei Aspekte hervorheben:

- Erstens können wirkliche oder potentielle Nutzer:innen privater Plattformen nicht unter Berufung auf das *First Amendment* gegen Maßnahmen der Intermediäre klagen, von denen sie, die Nutzer:innen, ihre Kommunikationschancen beeinträchtigt sehen, da der Verfassungszusatz nur gegen Eingriffe staatlicher Institutionen gerichtet ist. Aus demselben Grund sind auch *internet service providers* nicht durch das *First Amendment* auf die Netzneutralität verpflichtet (Balkin 2009: 429f.).
- Zweitens werden Intermediäre nicht selbst als Sprecher:innen derjenigen Inhalte behandelt, die auf ihren Plattformen geteilt werden. Explizit geregelt ist dies im 1996 erlassenen »Kommunikationsanstandsgesetz« (*Communications Decency Act*), Abschnitt 230 des Bundesgesetzes: »Kein Anbieter oder Nutzer eines interaktiven Computerdienstes soll als Herausgeber oder Sprecher von Informationen behandelt werden, die von einem anderen Inhaltsanbieter zur Verfügung gestellt worden sind.« Entsprechend haften Intermediäre nicht in derjenigen Weise für die von ihnen gehosteten Inhalte wie klassische Medien, Herausgeber:innen oder Journalist:innen, die etwa in Verleumdungsklagen zur Rechenschaft gezogen werden können. Dasselbe Gesetz räumt den Intermediären andererseits explizit das Recht ein, Inhalte zu moderieren, die »obszön, unzüchtig, lüstern, schmutzig, exzessiv gewalttätig, belästigend oder in sonstiger Weise anstößig« sind, und zwar ausdrücklich unabhängig davon, ob diese Inhalte als solche unter den verfassungsrechtlichen Schutz der Äußerungsfreiheit fallen würden. Section 230 ist politisch heftig umstritten,[9] hat aber bis jetzt allen juristischen Auseinandersetzungen standgehalten (Barnes/Zakrzewski 2023).
- Drittens können sich Intermediäre nach verbreiteter Auffassung ihrerseits auf den Schutz der Redefreiheit berufen, um staatliche Eingriffe in ihre Such- oder Vorschlagsalgorithmen oder andere Elemente des technischen Designs ihrer Plattformen abzuwehren. Denn anders als die von Nutzer:innen geteilten Inhalte sind Algorithmen und Plattformdesign ja tatsächlich von den Intermediären selbst gestaltet; und das *First Amendment* schützt nicht nur Äußerungen von Privatpersonen, sondern auch die *corporate free speech* von Unternehmen

9 Aktuell wird u.a. diskutiert, ob Abschnitt 230 die Plattformen auch von Verantwortung für Inhalte freistellt, die KI-generiert sind (Livni/Kessler/Mattu 2023).

(wenngleich in der Literatur nachhaltig strittig ist, ob der Schutzstatus beider Redeformen identisch ist oder sich unterscheiden sollte; für eine expansive Lesart vgl. Redish 2021; kritisch Sunstein 1993). Demnach sind Maßnahmen, durch die Intermediäre die Kommunikationschancen von Nutzer:innen beeinflussen, nicht nur durch wirtschaftliche Freiheiten der Intermediäre gesichert, sondern fallen zusätzlich in die breit definierte Klasse jener Äußerungen, die durch das *First Amendment* vor staatlichen Eingriffen bewahrt werden (vgl. kritisch Whitney/Simpson 2019).

Versteht man Äußerungsfreiheit lediglich als Abwehrrecht gegen den Staat und behandelt man Plattformen wie natürliche Sprecher:innen, sind diese Regulierungen grundsätzlich nachvollziehbar. Aber sind sie der neuartigen Rolle der Intermediäre angemessen? Ist beispielsweise die Freistellung der Intermediäre von Mitverantwortung für die geteilten Inhalte auch noch im Hinblick auf die neuartigen Vermittlungsleistungen personalisierter Vorschlagsalgorithmen plausibel? Wenn sich der kommunikative Flaschenhals von den Äußerungs- zu den Rezeptionschancen verlagert, und wenn Intermediäre zunehmend so verfahren, dass sie aus einer gigantischen Fülle von Netzinhalten diejenigen anhand selbst entwickelter Algorithmen auswählen lassen, die sie Nutzer:innen personalisiert präsentieren; wenn sie gar, basierend auf jenen Netzinhalten, KI-Systeme individuelle Antworten auf Nutzer:innenfragen geben lassen – tun sie dann nicht Ähnliches wie beispielsweise eine Person, die einen individualisierten Brief verfasst, indem sie Passagen aus Zeitungen zusammenklebt, und das Ergebnis jemandem in den Briefkasten schiebt? Sollten wir sie also nicht als Sprecher:in ansehen, die für den Inhalt der algorithmisch selektierten oder gar KI-generierten Antwort verantwortlich ist, ähnlich wie den:die Editor:in eines Collagebriefs? Freilich nutzen Intermediäre digitale Maschinen zur Selektierung und/oder Generierung ihrer personalisierten Botschaften. Da sie die Funktionsweise dieser Maschinen jedoch selbst kontrollieren und gemäß eigenen Intentionen zu optimieren suchen, während sie für Nutzer:innen als solche unkontrollierbar und in aller Regel zudem völlig opak bleibt, scheint dieser Unterschied nur von geringer Bedeutung.

Der Medienrechtler Jonathan Peters hält vor allem die normative Gleichsetzung der Intermediäre mit natürlichen Personen, die vor staatlichen Freiheitseingriffen geschützt werden müssen, für unplausibel. Er vergleicht die Plattformen im Hinblick auf ihren Einfluss auf die Redechancen von Bürger:innen vielmehr ihrerseits mit transnationalen Quasi-Regierungen:

> »As nongovernmental entities, the platforms are generally unconstrained by constitutional limits, including those imposed by the First Amendment. They are mostly free to develop and enforce their content rules and community guidelines as they please. They also have the freedom to decide how to display and prioritize

> their content using algorithms. Their terms-of-use, which operate effectively as a contract with users, empower the platforms to remove forbidden content, to suspend or deactivate user accounts, and otherwise to address content problems. In these ways, social-media platforms act as arbiters of free expression, conducting a form of ›private worldwide speech regulation‹ and developing a de facto jurisprudence. As the legal scholar Jeffrey Rosen put it, ›[The] lawyers at Facebook and Google and Microsoft have more power over the future of […] free expression than any king or president or Supreme Court justice.‹ Rebecca MacKinnon, the researcher and Internet freedom advocate, once wrote that big Internet companies are the ›sovereigns of cyberspace‹.« (Peters 2018; mit Verweis auf Rosen 2011; MacKinnon 2012)

Als Diagnose der aktuellen globalen Situation erscheint Peters' Beschreibung überspitzt formuliert. Das gilt zumal im Hinblick auf die zunehmenden europäischen Regulierungsbemühungen der jüngeren Zeit. Gerald Spindler kommentiert 2022 die Entwicklung »in den letzten zehn Jahren« wie folgt: »Standen zunächst noch die weitgehende Befreiung von jeder Verantwortlichkeit für Plattformen im Vordergrund, hat sich die Einstellung der Gesetzgeber gegenüber den Intermediären dahingehend entwickelt, dass diese als Gatekeeper und Schlüsselinstanzen in der Meinungsbildung wesentlich stärker in die Pflicht genommen werden.« (Spindler 2022: 126) Nationale und europäische Regelungen wie die bereits erwähnten NetzDG, DSGVO, DSA oder AI-Act greifen durchaus in die Souveränität der kommerziellen ›Netzkönige‹ ein. Allerdings verdanken sich diese Regulierungsbemühungen gerade einem anderen normativen Rahmen als dem, dessen Implikationen hier untersucht werden sollen. Zudem wirkt Peters' Diagnose zwar hier und heute überspitzt, aber doch weiterhin relevant, wenn man folgende Punkte bedenkt: Erstens räumen auch die genannten Regelungen den Intermediären durch das Instrument der »regulierten Selbstregulierung« wesentliche Entscheidungskompetenzen ein, was zwar im Hinblick auf die pragmatischen Herausforderungen der Regulierung nahe liegen mag, aber gleichwohl Anlass zu demokratietheoretischen Bedenken, zu Fragen bezüglich der Rechtssicherheit und zu Sorgen im Hinblick auf mögliches *overblocking* gibt (Peukert 2022). Zweitens nehmen diese Regulierungen die Intermediäre nur bei Verstößen gegen besonders scharfkantige Kommunikationsnormen in die Verantwortung, d.h. entweder klar justiziabel sind (Datenschutz, Handel mit illegalen Waren und Dienstleistungen, einzelne Aspekte des Jugendschutzes, Verleumdung…) und/oder zumindest ein besonderes Gefährdungspotential mit sich bringen.

Das verbleibende Tagesgeschäft der Aggregation, Vermittlung und (ggf. auch personalisierten) Priorisierung von Inhalten nach Maßgabe von Kriterien oder Algorithmen, die für die Nutzer:innen typischerweise intransparent sind, wird von den bestehenden Regulierungen ansonsten kaum tangiert. Plattformen, die eine

marktbeherrschende Stellung einnehmen, behalten insoweit den empfindlichen Flaschenhals der digitalen Kommunikationsströme fest im Griff.[10] Dadurch üben sie weiterhin einen bestimmenden Einfluss auf den Verlauf von Kommunikationsströmen und die faktischen Rezeptionschancen aus, ohne jedoch in vollem Umfang jenen Normen unterworfen zu sein, die auf klassischen *gatekeepers* wie Journalist:innen und kritischen Herausgeber:innen lasten. Entsprechend kommentiert auch Giovanni De Gregorio seine detaillierte Rekonstruktion der Stellung der Intermediäre in der Europäischen Union mit den Worten:

> »Despite […] judicial efforts, the challenges raised by online platforms are far from being solved. European courts have extensively addressed the problem of enforcement in the digital age. […] Still, the challenge of content moderation raises constitutional concerns. The increasing active role of online platforms in content moderation questions not only the liability regime of the e-Commerce Directive but also constitutional values such as the protection of fundamental rights and the rule of law.« (De Gregorio 2022: 59)

5. Benevolente Intermediäre?

Hält man Peters' Diagnose insoweit für weiterhin zutreffend, bleibt allerdings weiter die Frage offen, wie besorgniserregend der Einfluss der kommerziellen ›Netzkönige‹ eigentlich ist. Auch der von Peters zitierte Jeffrey Rosen[11] räumt ein, dass gerade ihre Macht den Äußerungsfreiheiten von Bürger:innen nicht selten Schutz gegen »tremendous pressure from repressive countries around the world, and from Western democracies« (Rosen 2012: 1536) geboten habe. Victoria Bacher (2022: 69) weist darauf hin, dass beispielsweise Facebook als Maßnahme der »Unternehmenspolitik« entschieden habe, problematische Posts mit einem Label zu versehen, das sie als *fake news* ausweist, ohne (im Normalfall) in der EU durch Gesetz dazu gezwungen zu sein. Gerade letzterer Hinweis lässt sich freilich (konform mit Bachers Intentionen) auch als Problemanzeige lesen: Wenn es im Belieben der Intermediäre liegt, die von ihnen vermittelten Desinformation als *fake news* zu kennzeichnen, sie zu blockieren, sie unkommentiert weiterzuleiten oder gar gezielt durch Empfehlungsalgorithmen hervorzuheben, erscheint dies legitimationstheoretisch zumindest pro-

10 Deutlichere Einschränkungen wären mit schärferen Regulierungen – Einschränkungen oder effizienten Transparenzverpflichtungen – der Personalisierung bzw. des *microtargeting* verbunden (Bacher 2022); ein Vorschlag für »vergleichsweise geringe Einschränkungen des Microtargeting« (Jaursch 2022) im Bereich der politischen Werbung wird derzeit in der EU beraten (Europäische Kommission 2021).

11 Professor der George Washington Law School und nicht identisch mit Jeffrey A. Rosen, dem gegen Ende der Trump-Präsidentschaft geschäftsführenden Justizminister.

blematisch. Ohne entsprechende regulatorische Zwänge ist nicht einmal sichergestellt, dass derartige Entscheidungen überhaupt regelbasiert getroffen werden und nicht vielmehr als erratische oder den Privatinteressen des Plattformeigners verpflichtete Willkürentscheidungen. Jack M. Balkin, Professor für Verfassungsrecht und insbesondere das First Amendment, stellt 2018 fest:

> »Currently, speech platforms do not govern in the same way that liberal democratic states do. Enforcement of community norms often lacks notice, due process, and transparency. […] Platform operators may behave like absolutist monarchs, who claim to exercise power benevolently, but who make arbitrary exceptions and judgments in governing online speech.« (Balkin 2018: 1196f.)

Die jüngsten Entwicklungen beim sozialen Kurznachrichtendienst *X* scheinen Balkins Beobachtungen nachdrücklich zu bestätigen. Gleichwohl könnte man hoffen, dass etwa die Musk'schen Hemdsärmlichkeiten im Umgang mit Nutzer:innenrechten lediglich Übergangsphänomene sind, weil es letztlich im wohlverstandenen Eigeninteresse der Intermediäre liegt, ihre *Corporate Social Responsibility* auch auf die Integrität der von ihnen beeinflussten Kommunikationsumgebung zu beziehen. Schließlich scheint es ja ein natürliches Anliegen der Plattformen, Nutzer:innen eine optimale *user experience* zu bieten, und (so könnte man vermuten) zu diesem Zweck den realen Kommunikationsbedürfnissen potentieller Nutzer:innen entgegenzukommen. Werden sie ihre Plattformen daher nicht schon aus Eigeninteresse so gestalten, dass sie die Werte, Ziele und Funktionen, die für die Begründung der Redefreiheit in Anspruch genommen werden, befördern, also etwa die Wahrheitsfindung und die individuelle und politische Selbstbestimmung begünstigen? Oder befürchten Mahner:innen wie Balkin (Balkin 2004; Balkin 2009) doch zu Recht, dass unzureichend regulierte Intermediäre die demokratische Kultur gefährden? Sind Plattformen wie *X* bzw. *Twitter* tatsächlich, wie der konservative Kolumnist Brett Stephens 2020 behauptet hat, gezielt dafür »gestaltet zu provozieren und herabzusetzen, dafür, die Selbstgefälligkeit von Unterstützern zu befriedigen und Gegner zum Kochen zu bringen, dafür, den nationalen Diskurs auf die Ebene von Gegrunz und Gegengegrunz zu reduzieren«?[12]

Dies sind, soweit sie überhaupt sinnvoll zu beantworten sind, primär empirische Fragen, deren Klärung nicht der Philosophie zufällt. Einige tentative Überlegungen sind jedoch angezeigt:

Erstens unterstellt die optimistische Lesart einen hinreichend funktionierenden Wettbewerb zwischen Intermediären. Soweit sie im ökonomischen Eigeninter-

12 »…designed for provocations and put-downs; for making supporters feel smug; for making opponents seethe; for reducing national discourse to the level of grunts and counter-grunts« (Stephens 2020).

esse handeln, haben sie anderenfalls nämlich kaum Anreize, realen Nutzer:innenbedürfnissen entgegen zu kommen. Diese Bedingung ist jedoch bestenfalls eingeschränkt erfüllt. Auf die Oligiopolstruktur des Plattformmarkts und die Bedeutung von Netzwerkeffekten wurde bereits hingewiesen. Dass es für Nutzer:innen oft mit prohibitiv hohen Kosten verbunden ist, bislang genutzte Plattformen zu verlassen, lässt sich erneut am Beispiel *Twitter/X* illustrieren.

Zweitens haben die Intermediäre u.a. aufgrund der verfügbaren Nutzerprofile Möglichkeiten der Beeinflussung des Verhaltens von Nutzer:innen, die diese dazu bringen könnte, nicht im Interesse ihrer eigenen aufgeklärten Bedürfnisse zu handeln. Auf die umfangreiche Diskussion über solche potentiell manipulativen Möglichkeiten kann hier allerdings nicht weiter eingegangen werden (vgl. exemplarisch Susser/Rössler/Nissenbaum 2019; Jongepier/Klenk 2022).

Schließlich lässt sich annehmen, dass das Interesse der Plattformbetreiber:innen wesentlich darauf gerichtet sein muss, möglichst nachhaltig möglichst viele möglichst zahlungskräftige Nutzer:innen in einer Weise an die eigene Plattform zu binden, die sie möglichst empfänglich für Angebote von digitalen oder nicht-digitalen Services und Produkten macht, die entweder von den Betreiber:innen selbst oder von Anzeigenkund:innen offeriert werden.[13] Es liegt nicht auf der Hand, dass dieses Interesse auch beispielsweise ein Interesse an der Wahrheit von geteilten Inhalte impliziert, oder daran, dass Funktionen der politischen Meinungsbildung gut erfüllt oder dass Kommunikationschancen fair zugeteilt werden. Mögliche Folgen der Orientierung an Kundenbindung, Kundenengagement und Schaffung eines günstigen Werbeumfelds sind auch von werbefinanzierten Magazinen oder dem Privatfernsehen bekannt – etwa die Präferenz für primär unterhaltsame und materialistische Werthaltungen befördernde Inhalte, die als Kontext für Anzeigen besonders geeignet sind.

Eine wahrscheinliche Spannung zwischen den genannten Anliegen der Intermediäre und den vernünftigen Erwartungen an gesellschaftliche Kommunikationsprozesse kann man schon darin sehen, dass Plattformen *prima facie* ein Interesse an der Maximierung der auf den Plattformen verbrachten Nutzungszeit haben. Das ist nicht unbedingt förderlich für die Kommunikationseffizienz. In jedem Fall ist es eine Herausforderung für die Medienbildung – und, wie zunehmend gut belegt ist, für die psychische Gesundheit (aktuell The U.S. Surgeon General 2023; American

13 Laut Tristan Harris, dem ehemaligen Design-Ethiker bei Google, verfolgen die Plattformen drei Ziele: »1. The engagement goal: to increase usage and to make sure users continue scrolling. 2. The growth goal: to ensure users are coming back and inviting friends that invite even more friends. 3. The advertisement goal: to make sure that while the above two goals are happening, the companies are also making as much money as possible from advertisements.« (The Social Dilemma 2023)

Psychological Association 2023; Riehm et al. 2019) – von Kindern und Jugendlichen (Saunders 2003).

Empirische Studien haben zudem gezeigt, dass Unwahrheiten in sozialen Medien häufiger geteilt werden als wahre Inhalte – vermutlich einfach deshalb, weil unwahre Behauptungen häufiger überraschend und daher interessant erscheinen (Vosoughi/Roy/Aral 2018). Würden sich Intermediäre bei der Priorisierung von Inhalten also ausschließlich am faktischen *user engagement* orientieren, könnte das also gerade auf die Priorisierung von Unwahrheiten hinauslaufen; das direkte Gegenteil von verantwortlichem Journalismus.

Plattformbetreiber:innen haben schließlich auch kein genuines Interesse, psychologischen Verzerrungen wie dem *confirmation bias* entgegenzuwirken, also der bei uns allen feststellbare Neigung, uns vorzugsweise Inhalte anzueignen, die bereits vorhandene Auffassungen bestätigen. Es ist daher plausibel anzunehmen, dass personalisierte Vorschlagsalgorithmen, die lediglich auf die Erhöhung der aus dem bisherigen Verhalten erschlossenen Rezeptionswahrscheinlichkeit zielen, die Polarisierung von Meinungen und affektiven Werthaltungen noch verstärken werden. Laut einer systematischen Metaanalyse, die auf der Auswertung von 121 Studien basiert, zeigen tatsächlich »nearly all experiments […] that social media can further aectively polarize people« (Kubin/von Sikorski 2021: 196).[14] Verhindern ließe sich dieser Polarisierungseffekt nur dadurch, dass bei der Gestaltung der Vorschlagsalgorithmen eigens besondere Vorkehrungen (Vermeulen 2022) zur Erhöhung des Pluralismus vorgeschlagener Inhalte getroffen werden. Solche Vorkehrungen liegen aber – jenseits öffentlichen Drucks oder gesetzlicher Regulierung – nicht unbedingt im Interesse der Plattformen.

Damit ist offenbar noch nicht gezeigt, dass die Nutzung sozialer Medien einen *stärkeren* Polarisierungseffekt hat als die Orientierung in einer Medienlandschaft, die von (ggf. ebenfalls stark polarisierten) traditionellen Medien dominiert ist – etwa kommerziellen Nachrichtensendern, die sich einer bestimmten Klientel verpflichtet fühlen. Deutlich wird aber zumindest, dass die grundsätzliche Offenheit sozialer Medien für unterschiedliche Inhalte sich für die algorithmisch adressierten Nutzer:innen nicht automatisch in eine ›pluralistische, am Gebot der Meinungsvielfalt und Angebotsvielfalt orientierte‹ Medienerfahrung übersetzen wird, auf die etwa der deutsche Medienstaatsvertrag abzielt.

14 Eine der ausgewerteten Studien untersucht beispielsweise im Labor die Wirkung von personalisierten YouTube-Vorschlagsalgorithmen im Vergleich mit künstlich ›de-personalisierten‹ Suchanfragen und kommt zum Ergebnis »that ideological reinforcement […] is heightened by political videos selected by the YouTube recommender algorithm based on participants' own search preferences« (Cho et al. 2020: 14f.). In der Tagespresse werden neben kleineren Experimenten mit YouTube oder TikTok wiederholt einigermaßen dramatische Fallbeispiele persönlicher Radikalisierungsprozesse beschrieben (z.B. Roose 2019).

Neben der Priorisierung von *Inhalten* lassen sich auch die durch das Design digitaler Plattformen vorgegebenen *Formen* der Kommunikation und der Gemeinschaftsbildung daraufhin befragen, wie zuträglich sie einer verantwortlichen, an Wahrheit und vernünftiger Selbstverständigung orientierten Diskussion sind. Zu diesem Fragenkomplex, der zu vielfältigen kommunikationswissenschaftlichen, soziologischen und psychologischen Forschungsbemühungen und -kontroversen Anlass gibt, sind an dieser Stelle nur einige flüchtige Beobachtungen und spekulative Überlegungen möglich.

Zweifellos spielt die Mitgliedschaft und aktive Beteiligung an digitalen Gemeinschaften mittlerweile für viele Menschen eine wesentliche Rolle für die Sozialisierung und Identitätsentwicklung. In der Podcast-Serie *Rabbit Hole* der *New York Times* (Roose et al. 2020) hebt Mark Zuckerberg nachdrücklich das Potential sozialer Medien zur Gemeinschaftsbildung hervor. In nachfolgenden Interviews mit Betroffenen und Angehörigen werden jedoch Fälle berichtet, in denen die Bindung an virtuelle Gemeinschaften andere soziale Kontakte weitgehend verdrängt hat. Manchen Personen mag gerade die digital mögliche Distanz und Anonymität erleichtern, soziale Kontakte zu pflegen. Mutmaßlich lassen sich Selbstbilder im virtuellen Raum leichter formen. Das mag Ausdrucksbereitschaft und Kreativität fördern, aber vielleicht nicht das Bewusstsein, für Äußerungen nötigenfalls auch in der nicht-digitalen Umgebung einstehen zu müssen.

Die spezifische Form, in der sich digitale Gruppen konstituieren, wird jedenfalls wesentlich durch das Design der Plattformen bestimmt, das Verwendungsweisen teils eröffnet und mit unterschiedlichem Nachdruck nahelegt (*affordances*), teils erschwert oder verschließt (*restrictions*). Manche Designelemente erwecken auf frappierende Weise den Eindruck, als wollten die Designer:innen im Interesse des *user engagement* gezielt Merkmale nachbilden, die Freud als typisch für die Teilnehmer:innen nicht-digitaler Massenveranstaltungen behauptet hat: erhöhte Emotionalität, Spontaneität und verminderte Ausdruckshemmung, begrenztes Differenzierungsvermögen, starke Identifikation innerhalb der Gruppe, die typischerweise durch die gemeinsame projektive Identifikation mit Führer:innen (bzw. Influencer:innen) zusammengehalten wird, und nicht zuletzt – zumindest diese Beobachtung deckt sich mit den Befunden der modernen Gruppenpsychologie[15] – die wertgeladene Unterscheidung zwischen In-Group und Out-Group (Freud 1999).

15 Welche der Freud'schen Beschreibungen und Erklärungen des Gruppenverhaltens sich aus heutiger Sicht bestätigen lassen, muss hier offenbleiben; für den Laien, der sich die Mühe macht, Freuds Originaltexte zu lesen, ist immerhin auffällig, dass zentrale Einsichten der »New Crowd Psychology« (Reicher 2022) bereits von Freud formuliert werden, einschließlich wichtiger Einwände an den einseitigen Phänomenbeschreibungen und oberflächlichen Erklärungsversuchen der »Old Crowd Psychology« etwa Gustave LeBons.

Soweit *Social-media*-Kommunikation in Schriftform stattfindet, begünstigt sie Emotionalität und Spontaneität keineswegs. Allerdings bestehen Unterschiede zwischen dem Austausch per *snail mail*, der Briefpartner:innen viel Zeit lässt, ihre Überlegungen reflektiert zu entwickeln und zu gestalten, und der digitalen Sofortkommunikation, die rasche Reaktionen erwarten lässt. Die durch viele Plattformen ermöglichte Anonymität der Teilnehmer:innen spiegelt einen wesentlichen Aspekt der (wahrgenommenen) Rolle von Teilnehmerinnen an lokalen Massenveranstaltungen: Es ist plausibel anzunehmen, dass sie das Gefühl persönlicher Verantwortlichkeit in ähnlicher Weise mindern könnte. Die durch das Design von Plattformen wie *X* vorgegebenen Beschränkungen von Textlängen beschneidet unvermeidlich die Differenzierungsmöglichkeiten. Sie erzwingt Vereinfachungen, wie sie auch für spontane Einwürfe bei Gruppenveranstaltungen charakteristisch sind. (Hierauf bezieht sich vielleicht Stephens' böse Behauptung, *Twitter* reduziere den Diskurs auf das Niveau von Grunzlauten.) Emoticons führen schließlich auch in die schriftliche Kommunikation ein schematisierendes Äquivalent emotionaler Mimik oder Gestik ein und sind das vielleicht klarste Indiz für den gezielten Versuch, die Emotionalität textbasierter Kommunikation zu erhöhen (momentan ist es nur noch mit Mühe möglich, Emoticons auf den Tastaturen mobiler Endgeräte unsichtbar zu machen).

Ein besonders interessantes Designelement sozialer Medien sind die auf manchen Plattformen vorgegebenen *like buttons*, die an die *One-click*-Schaltflächen zum Kauf von Produkten erinnern. Häufig ermöglichen sie keinerlei Differenzierung oder Perspektivierung der durch ihren Gebrauch ausgedrückten Werturteile, sondern verpflichten evaluative Äußerungen auf eine binäre Logik.[16] Zugleich – und dies ist besonders interessant – verknüpfen sie die binären Stellungnahmen und Werturteile mit der auch von außen sichtbaren (vgl. Beck 2021: 113ff.) Sortierung der Nutzer:innen in »Liker:innen« und »Nicht-Liker:innen«, also mit der gleichfalls binären Unterscheidung zwischen *in-group* und *out-group*. Auf der Ebene des technischen Designs ist damit eine Struktur festgeschrieben, die exakt der Logik einer von W. Lance Bennett (Bennett 2012) beschriebenen ›konsumistischen‹ Form von *identity politics* entspricht: Äußerungen von Einschätzungen, Stellungnahmen und Wertentscheidungen werden durch die im Plattformdesign verankerte starre technische Kopplung mit Bekenntnissen zu einer Gruppenzugehörigkeit verwoben.

Schließlich ermöglichen auch die über die *like buttons* generierten Informationen den Plattformbetreiber:innen direkte Einblicke in Haltungen der Nutzer:innen, deren Nutzung für entsprechende Vorschläge erneut die Gruppenidentifikation verstärken und polarisierend wirken mag. Damit bieten diese Kommunikationsumge-

16 Manche Kommentierungssysteme geben wenigstens Zugang zu einer begrenzten Pluralität vordefinierter Bewertungsdimensionen (z.B. »insightful«, »funny« etc.) und/oder ermöglichen Abstufungen.

bungen mutmaßlich keine günstigen sozialen und epistemischen Bedingungen für eine freie kritische Selbsterkundung im Interesse eines autonomen Lebens (Rössler 2017: 133ff.), oder für die unvoreingenommene argumentative Selbstverständigung einer pluralistisch verfassten Öffentlichkeit.

6. Zwischenbemerkung zur Rolle der Philosophie

Diese vorläufigen und fragmentarischen Überlegungen legen zumindest die Vermutung nahe, dass Netzunternehmen ihre Plattformen in Abwesenheit von Regulierungen oder öffentlichem Druck nicht schon aus unternehmerischem Eigeninteresse so gestalten werden, dass die kommunikativen Ziele der Wahrheitsfindung und der kritischen Selbstverständigung einer pluralistischen Öffentlichkeit befördert werden. Wenn das stimmt und wenn, wie zuvor behauptet, die alleinige Fokussierung auf die als Abwehrrecht verstandene Redefreiheit zur Sicherstellung verantwortlicher Plattformgestaltung unzureichend ist, weil sie Intermediäre zu globalen Kommunikationsschiedsrichtern macht, die ihre Regeln in wesentlichem Umfang selbst bestimmen können, lässt sich fragen, ob wir nicht entweder bei der Gestaltung digitaler Kommunikation neben der Redefreiheit andere Kommunikationsgrundrechte stärker ins Spiel bringen und/oder höher gewichten sollten und/oder ob wir bereits die Redefreiheit selbst anders interpretieren sollten, indem wir sie selbst auf das Vermögen zur fairen Partizipation an Verständigungspraxen interpretieren. Bevor diese Frage (in Abschnitt 8) wieder aufgegriffen wird, sind jedoch Komplikationen zu bedenken.

Erstens lässt sich zurückfragen, wer durch das »Wir« eigentlich angesprochen ist. Festlegungen über die Regulierung der digitalen Infrastruktur können in der Demokratie nur von demokratisch legitimierten Gremien getroffen werden. Daraus ergeben sich wiederum Folge- und Unterfragen wie die folgenden: Wie und inwieweit kann und soll die Organisation von Verständigungspraxen (ggf. subsidär) der funktionalen Ausdifferenzierung von Kommunikationspraxen Rechnung tragen? Welche Anforderungen an und Restriktionen für mögliche Regulierungen der Öffentlichkeit ergeben sich aus dem Umstand, dass demokratische Willensbildung ihrerseits eine funktionierende politische Öffentlichkeit voraussetzt? Welche Rolle kann und soll im Kontext der demokratischen Verständigung über Verständigungsverhältnisse die Philosophie spielen?

Zweitens stellt sich die Frage, inwieweit die unterstellten Gestaltungsspielräume für legitime begriffliche und normative Re-Arrangements kommunikativer Grundrechte überhaupt vorhanden sind. Es könnte ja sein, dass sich zur Verteidigung traditioneller Interpretationen liberaler Abwehrrechte so starke Argumente anführen lassen, dass Reformen digitaler Kommunikationsstrukturen – beispielsweise substantielle Eingriffe in die Kompetenzen privatwirtschaftlich organisierter

Intermediäre – von vornherein als illegitim ausscheiden, und zwar gegebenenfalls selbst dann, wenn sich zeigen ließe, dass die bestehenden Strukturen (auch) mit Problemen oder Risiken behaftet sind. Soweit man der Philosophie eine Rolle bei der methodischen Prüfung solcher Grundrechtsargumente zuschreibt, wird man von ihr folglich Beiträge zur Demarkation der Gestaltungsspielräume für Kommunikationsarrangements erwarten.

Damit hängt nun aber die Frage, welche substantiellen Beiträge die Philosophie zur Gestaltung digitaler Verständigungsverhältnisse leisten kann oder leisten sollte, ihrerseits bereits von *substantiellen* normativen Überzeugungen ab: Philosoph:innen, die Meinungsfreiheit als ein vorpolitisch begründetes Abwehrrecht interpretieren, das gegenüber kommunikativen Teilhabe- und Partizipationsrechten strikt vorrangig ist, werden der Philosophie eine zentrale, allerdings rein defensive Rolle zuweisen, die vor allem in der Abwehr demokratischer (Re-)Regulierungsbemühungen einer Kommunikationssphäre besteht, die als *marketplace of ideas* verstanden wird und ebenso wie der ökonomische Marktplatz vor staatlichen Eingriffen möglichst zu bewahren ist. Philosoph:innen, die größere Spielräume legitimer gesellschaftlicher Gestaltung des Kommunikationsraums sehen, werden einerseits ihre professionsgebundene normative Autorität entsprechend bescheidener interpretieren. Andererseits werden sie aber ihre Aufgabe auch in ein weiteres Spektrum von Fragen und Gestaltungsoptionen betreffenden, allerdings tendenziell ›weicheren‹ Beiträgen sehen – in Anregungen oder bedingten Empfehlungen, Beiträgen zur Positions- und Interessenklärung, Problemerhellung oder Normhermeneutik.

7. Redefreiheit als libertäres Informationsverfügungsrecht?

Ein prägnantes Beispiel für die erste Perspektive hat Karsten Weber vor allem in seinem Aufsatz »Die Informationsfreiheit und der Zusammenhang von Abwehr- und Anspruchsrechten« (Weber 2009) geliefert. Nachdem er schon zuvor für »ein unlimitiertes Recht auf freie Rede« (Weber 2007: 39) eingetreten war, unternimmt er hier den Versuch, das Verständnis der Meinungsfreiheit als ›klassisches und zentrales liberales Recht‹ (Weber 2009: 19) im Rahmen eines umfassenderen Konzepts von Informationsfreiheit zu aktualisieren: »[M]it theoretischen Überlegungen aus dem Bereich der liberalen bzw. libertären politischen Philosophie« möchte er zeigen, »dass Informationsfreiheit letztlich immer nur im Sinne einer negativen Freiheit verstanden werden kann und sollte« (Weber 2009: 17f., 21).

In der Einbettung der Meinungsfreiheit in das Rahmenkonzept der Informationsfreiheit, das sowohl die »1. die Freiheit vor Informationseingriffen«, »2. die Freiheit zur Verwendung eigener Informationen« und »3. die Freiheit beim Zugriff auf Informationen« (Weber 2009: 19) umfassen soll, manifestiert sich dabei bereits ein wesentlicher Aspekt der Weber'schen Position: Informationsfreiheiten werden in all

ihren drei Spielarten als private Verfügungsrechte über das Gut ›Information‹ interpretiert und damit nach dem Muster eines Rechts auf Privateigentum modelliert, das in der Nachfolge Lockes und Nozicks als vorpolitisches Recht verstanden wird (Weber 2009: 25).

Insoweit ist denn auch Webers Schlussfolgerung nicht überraschend, dass der Austausch von Verfügungsrechten über Informationen oder Daten aufgrund privatrechtlicher Verträge (etwa Nutzungsvereinbarungen zwischen Plattformanbieter:innen und Nutzer:innen) unproblematisch ist und gänzlich der freien Selbstverantwortung der Vertragspartner:innen anheimgestellt werden kann (Weber 2009: 25) und dass darüber hinaus für diejenigen, die nicht Eigentümer:innen bestimmter Information sind, keinerlei Anspruchsrechte auf Zugang zu diesen Information geltend gemacht werden können, aber auch umgekehrt niemandem verwehrt werden darf, eigene Informationen für andere freizugeben, also beispielsweise unter eine *Open Source*-Lizenz zu stellen (Weber 2009: 26ff.). Auch die Meinungsfreiheit soll anscheinend ganz auf das Prinzip zurückgeführt werden, »dass jede Person selbst mit ihren Daten und Informationen tun und lassen können soll, was sie will« (Weber 2009: 17). »Freie Rede impliziert« daher »dass niemand daran gehindert werden darf, seine Meinung frei zu äußern, aber auch nicht mehr – freie Rede ist ein negatives Recht bzw. ein Abwehrrecht« (Weber 2009: 27).

Sollten aus der Institutionalisierung dieser abwehrrechtlich interpretierten Redefreiheit im Kontext digitaler Kommunikationsverhältnisse problematische Konsequenzen etwa im Hinblick auf unsere Bemühungen der Wahrheitsfindung, der individuellen und gemeinschaftlichen Selbstbestimmung hervorgehen (wie in vorangehenden Abschnitten dieses Texts vermutet), böte dies aus Webers Sicht keine zureichende Rechtfertigung für den Versuch einer Umgestaltung der Kommunikationsstrukturen. Denn mögliche Ansprüche auf die Realisierung individueller oder gemeinschaftlicher Werte sind nach seiner Auffassung nicht schwerwiegend genug, um Eingriffe in die als vorrangig betrachteten negativen Rechte zu begründen: »Positive Rechte in Bezug auf Informationen schränken elementare negative Rechte zu stark ein und stellen somit illegitime Eingriffe in das Leben der Betroffenen dar« (Weber 2009: 21). Entsprechend folgert Weber, »dass es Aufgabe staatlicher Institutionen sein muss« das Abwehrrecht auf Informationsfreiheit

> »durch entsprechende Maßnahmen zu schützen […] bzw. alles zu unterlassen, was dieses Recht verletzen könnte. Es folgt aber auch, dass es nicht die Aufgabe staatlicher Institutionen sein kann und darf, zu Umverteilungsmaßnahmen zu greifen, um ein irgendwie geartetes positives Recht im Zusammenhang mit Informationen zu realisieren. Zumindest in Bezug auf den Umgang mit Informationen ist die Konsequenz daraus, dass sich staatliche Institutionen soweit als nur irgendwie möglich aus dem Leben der Menschen zurückziehen.« (Weber 2009: 28)

Wäre das richtig, gäbe es für eine Neujustierung von Kommunikationsgrundrechten etwa als Reaktion auf den digitalen Medienwandel also gar keinen Spielraum.

Gegen Webers Überlegungen lassen sich allerdings auf unterschiedlichen Ebenen Einwände erheben. Der zurückhaltendste Einwand würde die normativen Grundlagen der Weber'schen Position weitestgehend akzeptieren, aber auf blinde Flecken seiner Situationsbeschreibung hinweisen, die, wie die mögliche Abhängigkeit von marktbeherrschenden Plattformen und die resultierenden Einschränkungen der Vertragsfreiheit, eher für eine sozusagen ordoliberalistische Organisation des Kommunikationsraums denn für einen radikalliberalen Nachtwächterstaat sprechen. Eine radikale Entgegnung bestünde in der Kritik des libertären Rahmenkonzepts selbst (Filipović 2009) und insbesondere der libertären Eigentumstheorie, nach deren Muster Weber auch die Informationsfreiheit(en) zu modellieren versucht. Auf mittlerer Ebene lässt sich beispielsweise fragen, wie plausibel die angenommene Analogie zwischen Eigentumsrecht und Informationsfreiheiten und, damit zusammenhängend, Webers Deutung des Verhältnisses zwischen negativen und positiven Rechten ist.

Gewisse Zweifel an dieser Analogie werden schon in Webers eigenen Ausführungen deutlich. So räumt er in einer Fußnote unvermittelt eine nach eigenem Bekenntnis »wichtige Ausnahme« (Weber 2009: 21, Fußnote 31) von der sonst durchgängig absolut formulierten[17] These ein, dass die negative Informationsfreiheit niemals durch positive Rechte eingeschränkt werden dürfe. Weber verteidigt dort ein Anspruchsrecht auf staatlich finanzierte Schulbildung (nicht hingegen auf universitäre Bildung) mit der Überlegung, dass »sie Bedingung zur Wahrnehmung der eigenen negativen Rechte sei und ebenso Voraussetzung für die Erkenntnis der Schranken dieser Rechte« (ebd.). Dieses Zugeständnis legt verschiedene Anschlussfragen von übergreifender Bedeutung nahe.[18] Für den vorliegenden Kontext

17 Gegenüber seiner Monographie *Das Recht auf Informationszugang* aus dem Jahr 2005 hat Weber seine Position anscheinend stark zugespitzt. Dort findet sich noch folgende Überlegung: »Meinungsfreiheit setzt bspw. voraus, dass Faktoren wie die ökonomischen Möglichkeiten eines Meinungsverbreiters den Meinungsaustausch zwischen Personen oder Gruppen von Personen nicht völlig asymmetrisch gestalten. Überhaupt könnte sich als größte Gefahr für die freie Meinungsäußerung nicht die Einschränkung durch staatliche Institutionen erweisen, sondern durch die Durchsetzung von Eigentumsrechten an Informationen […]. Auch hier wird deutlich, dass bei der Betonung der individuellen Rechte, wie es libertärer Position geschieht, selbst der Minimalstaat aufgefordert ist, redistributiv zu wirken. Nicht etwa, um im Libertarismus gleichsam verteufelte positive Rechte herzustellen, sondern um klassische negative Freiheiten zu wahren. Diese Rechte können dadurch gewahrt werden, dass bei großen Ungleichheiten redistributive Maßnahmen ergriffen werden, um die sozialen Grundgüter Information bzw. Informationszugang allgemein zugänglich zu machen.« (Weber 2005: 221f.)

18 Dazu gehört auch die wichtige, wenngleich im vorliegenden Kontext weniger relevante Frage: Müsste mit dem von Weber angeführten Argument neben dem *Recht* auf freie Schulbildung nicht auch eine allgemeine Schul*pflicht* begründet werden? Schließlich scheint es mit

relevant ist die Frage, ob strukturell analoge Begründungen nicht noch weitere Einschränkungen des Abwehrrechts auf Informationsfreiheit rechtfertigen könnten. Gibt es neben der »elementaren Bildung« (ebd.) nicht noch weitere »Bedingung[en] zur Wahrnehmung der eigenen negativen Rechte […] und für die Erkenntnis der Schranken dieser Rechte«? Ist es nicht gerade im Interesse der Autonomie von Bürger:innen und Marktteilnehmer:innen nötig, beispielsweise Mindestbedingungen für die Transparenz politischer oder kommerzieller Kommunikation fest- und durchzusetzen, z.B. Regeln gegen unlautere Verträge, irreführende Marketingstrategien oder manipulative *dark patterns* in elektronischen Entscheidungssystemen (Susser/Rössler/Nissenbaum 2019)?

Der volle Sinn solcher Regulierungen erschließt sich freilich erst dann, wenn Kommunikation überhaupt als eine gemeinschaftliche, durch interne Normen regulierte Praxis interpretiert wird, in der Redebeiträge auf entsprechende Geltungsansprüche (z.B. der Wahrheit, Aufrichtigkeit etc.) bezogen sind. Diese Dimension gerät gar nicht erst in den Blick, wenn man Rede auf eine bestimmte Form der privaten Verfügung über Informationseigentum versteht. Der resultierende Kommunikationsbegriff ist schlichtweg blind gegenüber Differenzen wie der zwischen wahrer und falscher, aufrichtiger oder absichtlich irreführender Rede. Insoweit erscheint dann auch die Gleichbehandlung aller Redeäußerungen durch eine »unlimitierte Meinungsfreiheit« (Weber 2007) nur konsequent. Zugleich ist aber kaum zu sehen, wie ohne jede Regulierung irreführender Rede und bewusster Täuschung, Verleumdung oder Diskriminierung auch nur der Kernbereich liberaler Privatautonomie geschützt werden könnte.[19]

Interessanterweise stellt Weber selbst fest, dass die Gültigkeit privatrechtlicher Verträge (z.B. über die Verwendung von Informationen) »an die Erfüllung einiger wichtiger Bedingungen geknüpft« ist, darunter »auch an das, was im Englischen als *informed consent* bezeichnet wird – beide Seiten müssen über die jeweils gültigen *terms of trade* informiert sein« (Weber 2009: 25). Die naheliegende Frage, ob sich im Ausgang von dieser Bedingung nicht doch zwangsläufig ein Bedarf an gewissen Regulierungen der (vertraglichen) Rede ergibt, vermeidet er jedoch mit dem Hinweis: »Es soll hier allerdings nicht Gegenstand der Untersuchung sein, ob dies [die hinreichende Informiertheit der Vertragspartner:innen, MHW] immer der Fall ist

dem normativen Individualismus der liberalen Tradition kaum vereinbar, die Fähigkeit von Bürger:innen zur Wahrnehmung eigener und zur Respektierung fremder Abwehrrechte von früheren Entscheidungen anderer (zeitweise erziehungsberechtigter) Personen abhängig zu machen.

19 Anders als Weber nahelegt, verteidigt übrigens Susan J. Brison in ihrem Aufsatz *The Autonomy Defense of Free Speech* (Brison 1998) keineswegs die Auffassung »dass das Recht auf freie Meinungsäußerung keiner Grenzsetzung unterworfen sein sollte« (Weber 2007: 35). Ganz im Gegenteil plädiert sie explizit für die Legitimität von Regulierungen zum Schutz vor Diskriminierung oder Hassrede.

– es geht um grundsätzliche Erwägungen« (Weber 2009: 25). Tatsächlich gehört zu den ›grundsätzlichen Erwägungen‹ zwar wohl nicht die *empirische* Frage, inwieweit faktisch mit hinreichender Informiertheit von Vertragsparteien gerechnet werden kann, wohl aber die Frage, wie sich die Bedingungen eines *informed consent* überhaupt sichern ließe, wenn die vollständig deregulierte Redefreiheit nicht einmal die nachträgliche Ahndung von Missbräuchen zulassen würde (Weber 2007: 38). Gesteht man allerdings die Notwendigkeit gewisser Regulierungen der Rede im Kontext von Privatverträgen, also zur Sicherung der Privatautonomie, ein, wird es schwer zu sehen, warum nicht auch beispielsweise im Kontext der politischen Selbstbestimmung bestimmte Mindestbedingungen (etwa eine Minimalversorgung mit adäquaten Informationen) gegeben sein müssen, damit beispielsweise demokratische Wahlen (analog dem *informed consent*) als legitim gelten können (Brison 1998: 330f.).

Im Hinblick auf die grundlegende Analogie zwischen Informationsfreiheit und Eigentumsrecht als solche benennt wiederum Weber selbst den möglichen Einwand, wonach

> »Eigentum an Informationen letztlich keinen Sinn mache, denn Eigentum in seiner ursprünglichen Orientierung an materiellen Dingen bedeute einfach nur die alleinige Verfügungsgewalt über jene materiellen Dinge. Eigentum an diesen ist exklusiv, da materielle Dinge zu einer Zeit nur an einem Ort sein könnten und ihr Ge- oder Verbrauch engen Restriktionen unterliege. Für Informationen, so wird nun argumentiert, gelte dies jedoch nicht. Denn man könne Informationen an andere weitergeben und sie gleichzeitig selbst doch behalten – es entstünden dadurch weder Verluste noch Nachteile.« (Weber 2009: 25)

Auf diesen Einwand reagiert Weber mit dem an sich berechtigten Hinweis, dass dies »beileibe nicht für alle Informationen« (Weber 2009: 25) gelte. Offen bleibt damit allerdings, warum Informationen *generell* – auch diejenigen, die zu teilen *keine* Verluste oder Nachteile mit sich bringt – nach dem Muster des Privateigentums verstanden werden sollten. Eben dies nimmt Weber aber an, indem er *alle* Ansprüche auf Informationszugang der Kategorie der Ansprüche auf »Umverteilung von Gütern« zuordnet, die *prima facie* illegitim sind, »[d]a Umverteilung […] immer bedeutet, auf der einen Seite Menschen etwas zu nehmen, um es auf der anderen Seite anderen Menschen zu geben«, und damit in vorrangig geschützte Abwehrrechte eingreift (Weber 2009: 29).

Dass die Übertragung des libertären Modells vorpolitischer Eigentumsrechte (das schon in Bezug auf materielle Güter schwerwiegenden Einwänden ausgesetzt ist; einführend Werner 2012) auf den gesamten Bereich der Rede Probleme mit sich bringt, lässt sich auch an der ökonomischen Replik Eric Ch. Meyers (2009) ablesen. Nicht nur stößt die Durchsetzbarkeit von Eigentumsrechten auf die im digitalen

Kontext leicht kopierbaren Informationen auf ganz andere Herausforderungen als die des Eigentums an physischen Objekten. Vor allem aber deutet die von Meyer aufgezeigte »Mehrrelationalität« von Informationen (Meyer 2009: 52) darauf hin, dass es kaum möglich sein dürfte, ein plausibles Modell des Erwerbs von Informationseigentum zu entwickeln, das etwa der Locke'schen Theorie der ursprünglichen Inbesitznahme nachgebildet wäre:

> »Die Zuweisung von Eigentumsrechten an Informationen ist […] nicht einfach, da es nicht immer klar ist, wer eigentlich ein Eigentum an diesen Informationen für sich reklamieren darf. Im Fall der Bankdaten könnte dieses der Kontoinhaber, jedoch genauso gut auch die Bank sein, die diese Kontonummer schließlich an den Kontoinhaber vergeben hat.« (Meyer 2009: 50)

Außer der von Meyer betonten Mehrrelationalität vieler Informationen bestehen weitere Herausforderungen für den Versuch, vorpolitische Grenzen von Informationsverfügungsrechten ausfindig zu machen. Eine vorpolitische Theorie des Informationseigentums könnte jedenfalls kaum die Grenzen des Urheberrechts ›naturrechtlich‹ erklären (warum sollte es genau 70 Jahre über den Tod des Urhebers hinausreichen?) und vermutlich auch nicht so leicht den Umstand, dass ein Patent auch die Freiheiten derjenigen beschränkt, die nach dem Patenteintrag nachweislich unabhängig zur derselben Idee gelangt sind wie der *patent holder*. Faktisch ist die Institutionalisierung von Eigentums- und Verwertungsrechten im Hinblick auf geistiges Eigentum zweifellos durch Nutzen- und Gemeinwohlüberlegungen mitbestimmt.

Am überzeugendsten wirken Webers Überlegungen zum ersten Aspekt der Informationsfreiheit, nämlich in Bezug auf den Schutz privater Daten. Auch macht er zu Recht darauf aufmerksam, dass der Besitz von Informationen bereits als solcher einen Gebrauchs- oder Tauschwert für den:die Inhaber:in haben kann. Dieser Wert ist aber weder für das basale Recht auf Äußerungsfreiheit noch für den gesamten Wert der freien Rede ein hinreichendes Fundament. Äußerungsfreiheit schützt als Teil der allgemeinen Handlungsfreiheit neben symbolischen Äußerungen, durch die Informationen mitgeteilt werden, ebenso andere, z.B. ästhetisch-expressive, Handlungen, die für mögliche Rezipient:innen gar keinen Sinn (oder einen ganz anderen als für die:den Handelnden) haben mögen.

Der Wert der Beteiligung an sprachlichen Kommunikationsprozessen ist erst recht mehr als eine Funktion des vorgängigen Werts von Privatinformationen. Das liegt einmal daran, dass »Informationen […] häufig […] erst in Verbindung mit anderen Informationen […] an Wert gewinnen« (Meyer 2009: 47). In noch größerem Umfang liegt es jedoch daran, dass sowohl der soziale Wert der verschiedenen Kommunikationspraxen für die Gemeinschaft als auch der individuelle Wert der eigenen Kommunikationsteilnahme maßgeblich an den Praxen selbst haftet; entweder am intrinsischen Wert dieser Praxen und/oder an den durch sie beförderten gemein-

schaftlichen oder privaten – aber eben auch im letzten Fall vielfach nur kooperativ erreichbaren – Zielen.

Sicher werden Gespräche, wissenschaftliche Konferenzen, Briefwechsel oder digitale Chats auch unternommen, um ›Informationen auszutauschen‹, weil diese Informationen für die vorgängigen Privatzwecke der Beteiligten unmittelbar oder mittelbar, und sei es auch nur wegen ihres Tauschwerts in weiteren Transaktionen, wertvoll sind. Sie sind aber wesentlich auch sozialintegrative und kooperative Veranstaltungen.[20] Diese Veranstaltungen sind auf vielfältige Weise – als teilautonome Funktionselemente oder als kritische Steuerungsinstanzen – mit nicht-sprachlichen sozialen Praxen und Institutionen verwoben, die ihrerseits der Verfolgung gemeinschaftlicher oder privater Ziele dienen. Insofern besteht der Wert der freien Rede offenbar auch in der Möglichkeit der *Partizipation*, das heißt in den Chancen auf die *Zulassung zu*, die aktive *Mitwirkung an* und (je nach dem Grad des selbstorganisierten Charakters der jeweiligen Praxen) auch die in geteilter Verantwortung unternommene *Mitgestaltung von* gemeinschaftlichen Unternehmungen und gemeinschaftlich veranstalteten strategischen Interaktionen (für den Bereich der allgemeinen politischen Öffentlichkeit vgl. Habermas 1990: 332f.).

Dieses zentrale Element des Werts der freien Rede bleibt verdeckt, wenn man Rede nicht auf die kommunikative Beziehung zwischen Kommunikationspartner:innen bezieht, sondern daran nur die freie Verfügung über das Informationseigentum der Sprechenden betrachtet. Entsprechend thematisiert Webers Konzeption der Informationsfreiheit ausschließlich Rechte der ersten und der dritten Generation – (starke) Abwehrrechte gegen erzwungene Eingriffe in Informationseigentum einerseits und Rechte auf Informationszugang, die nach dem Muster von Anspruchsrechten auf die staatliche Umverteilung privater Ressourcen verstanden werden (und laut Weber als Eingriff in die Abwehrrechte grundsätzlich problematisch sind). Rechte der zweiten Generation – Partizipationsrechte, die auf gesellschaftliche Mitwirkung und Mitgestaltung zielen – treten hingegen gar nicht in den Blick.

8. Folgerungen

Schon vor dem digitalen Wandel gab es Gründe für die Einschätzung, dass im Rahmen mediatisierter Kommunikation neben staatlicher Zensur und Propaganda

20 Diese Veranstaltungen behalten sogar dort, wo offen strategisch kommuniziert wird (etwa zwischen Erpresser und Opfer) noch einen (wenn auch, in Habermas' Worten, ›parasitären‹) Bezug auf das *telos* der Verständigung (Habermas 1981: Bd. 1; zum Problem offen strategischer Sprechakte vgl. Werner 2003). Diese Annahme ist für die folgenden Ausführungen allerdings nicht notwendig.

auch der Einfluss nicht-staatlicher Akteure problematisch sein kann (Baker 1995), was gegen eine *laissez faire*-Konzeption der Kommunikationssphäre als unregulierter *markteplace of ideas* spricht (Brison 1998). Anders als zu Zeiten größerer Netzeuphorie oft angenommen (vgl. kritisch Baker 2006: 97ff.) verliert diese Einschätzung durch die Digitalisierung nicht an Plausibilität, sondern erhält mit dem Aufstieg der Intermediäre sogar neue Nahrung. Neben den faszinierenden neuen Möglichkeiten globaler und interaktiver Kommunikation bietet Digitalisierung nämlich auch neue Mittel des individualisierten Monitorings von Kommunikationsprozessen, der algorithmischen Priorisierung von Inhalten, des auf große Mengen individueller (Vergleichs-)Daten gestützten gezielten *targeting* von Nutzer:innen, der formalen Vorstrukturierung möglicher Redebeiträge durch Designvorgaben, der durch technische *restrictions* und *affordances* beeinflussten Inkubation von semiöffentlichen Kommunikationsgemeinschaften und der (Beeinflussung der) ›maschinellen‹ Inhaltsproduktion durch generative KI.

Aus diesen Möglichkeiten resultiert ein erhebliches Ungleichgewicht des kommunikativen Einflusses weniger Digitalunternehmen einerseits und privater Nutzer:innen andererseits, das deren normative Gleichstellung (etwa durch die Anerkennung von Priorisierungsalgorithmen als vor Eingriffen geschützte Instanzen freier Privatrede) unplausibel erscheinen lässt. Noch offenkundiger ist der Umstand, dass durch die Verschiebung des kommunikativen Flaschenhalses von den Äußerungschancen zu den Rezeptionschancen eine Regulierung, die allein auf Äußerungsfreiheit fokussiert ist, unbefriedigend bleiben muss. Nun wäre all dies zwar bedauerlich, aber doch nachrangig, wenn sich zeigen sollte, dass sich für eine strikt libertär-abwehrrechtliche Konzeption kommunikativer Grundrechte zwingende Gründe anführen lassen. Die exemplarische Diskussion der von Karsten Weber entwickelten Konzeption libertärer Informationsfreiheitsrechte (Weber 2009) und »unlimitierter Meinungsfreiheit« (Weber 2007) in Abschnitt 7 kann diese Einschätzung aber jedenfalls nicht stützen.

Was folgt nun aus diesen Überlegungen für philosophische Bemühungen, ein plausibles Konzept der Redefreiheit zu entwickeln, das dem digitalen Wandel Rechnung trägt?

Wie in der Zwischenbemerkung (6) angedeutet (und wie es grundsätzlich der Fall ist, wenn man in Bezug auf ethische Einzelnormen keinen platonischen Realismus vertritt), bieten sich verschiedene Möglichkeiten: Entweder könnte man das Konzept der Redefreiheit wesentlich als Abwehrrecht verstehen, das jedoch im Kontext mediatisierter Kommunikation durch weitere kommunikative Grundrechte ergänzt werden muss: Grundrechte auf faire Kommunikationschancen oder gegebenenfalls auch Rechte auf faire Möglichkeiten der Mitgestaltung von Kommunikationsumgebungen. Oder man könnte das Konzept der Redefreiheit selbst so begreifen, dass es als solches bereits partizipative Aspekte einschließt (Kenyon 2021).

Auch wenn man die erste Option wählt, ist das Verhältnis zwischen Redefreiheit und kommunikativen Partizipationsrechten vergleichbar mit dem Verhältnis zwischen dem Recht auf körperliche Bewegungsfreiheit (das vor Gefangennahme u.ä. schützt) einerseits und dem Recht auf Mobilität (das die Existenz und den fairen Zugang zu realen Möglichkeiten der Ortsbewegung mitmeint) andererseits: In dem Maße, in dem unsere Chancen auf für uns bedeutsame körperliche Ortsbewegungen von sozialer Unterstützung, von technischen Hilfsmitteln und/oder von einer für unsere Mobilitätsbedürfnisse dienliche Gestaltung der Infrastruktur abhängig werden, verschwimmt die Grenze zwischen den abwehr- und den anspruchsrechtlichen Aspekten. Eine Tür ohne für uns zugänglichen und betätigbaren Griff kann uns faktisch ebenso behindern wie eine verriegelte. Im Hinblick auf die resultierenden Behinderungen spielt es dabei keine Rolle, ob die Abhängigkeit von Mobilitätsbedingungen (einem niedrigen Türgriff, einem elektrischen Türöffner, einem Rollstuhl oder einem Automobil) durch eine individuelle Eigenheit unseres Bewegungsapparats mitbedingt ist, oder durch soziotechnische Entwicklungen verursacht, die unsere Partizipation am normalen gesellschaftlichen Leben von jenen Bedingungen abhängig machen – beispielsweise, weil Arztpraxen und Einkaufsmöglichkeiten durch Veränderungen der Stadtentwicklung (die selbst durch die Durchsetzung des motorisierten Individualverkehrs bedingt sein mögen) zu Fuß kaum noch zu erreichen sind. Die resultierende Verschiebung des kommunikativen Flaschenhalses von den Äußerungs- zu den Rezeptionschancen lässt im digitalen Kommunikationsraum ähnliche neue Abhängigkeiten entstehen, da unser (von der klassischen Redefreiheit geschützter) Einfluss auf die Kontrolle unserer Äußerungen beschränkt bleibt, während die Chancen auf Gehörtwerden zunehmend von dem Design digitaler Infrastrukturen abhängig wird, auf deren Gestaltung wir kaum Einfluss nehmen können.

Für eine Theorie freier Rede, die faire Rezeptionschancen mitbedenken möchte, ergeben sich allerdings ernste Herausforderungen.

Erstens stellen sich grundsätzliche Fragen wie die folgenden: Ist es überhaupt möglich, allgemein akzeptable Kriterien für die faire Partizipation an Kommunikationspraxen zu definieren? Wer wäre gegebenenfalls befugt, solche Kriterien verbindlich fest- und durchzusetzen? Wird nicht jede Ermächtigung, Eingriffe in Infrastrukturen freier Rede oder kommunikationsökonomische Zusammenhänge mit anspruchs-, oder partizipationsrechtlichen Argumenten zu rechtfertigen, von autokratischen Regierungen oder tyrannischen Behörden missbraucht werden? Solchen Bedenken lässt sich zunächst entgegenhalten, dass der Kommunikationsraum auch aktuell ein regulierter Raum ist, dass auch eine auf Abwehrrechte fokussierte Regulierung Konzentrationen kommunikativer Macht Vorschub leisten kann und dass auch im Rahmen dieser Regulierung Kontroversen und Konflikte auftreten, die beispielsweise das Verhältnis zwischen ökonomischen Freiheiten, Redefreiheiten und dem Schutz der Privatsphäre betreffen.

Zu vermuten ist trotzdem, dass durch die stärkere Berücksichtigung von Teilhabe- und Mitwirkungsrechten neue Komplikationen in die Diskussionen über Redefreiheit(en) einziehen. Verstärkt wird diese Befürchtung durch die folgende Überlegung: Wie schon erwähnt sind Kommunikationspraxen mit anderen gesellschaftlichen Praxen verwoben. Im Zusammenhang mit der funktionalen Ausdifferenzierung moderner Gesellschaften haben sich auch die sozialen Kommunikationspraxen ausdifferenziert. Ihre Unterschiede in Zielsetzung und institutioneller Struktur färben auch die im jeweiligen Kontext anerkannten kommunikationsbezogenen Normen und Konventionen. Entsprechend den Zielen des Wissenschafts-, Rechts- oder Marktsystems gelten für den wissenschaftlichen Diskurs, das Gerichtsverfahren oder die kommerzielle Kommunikation abgewandelte Erwartungen etwa bezüglich der Darstellung oder Bekräftigung von Sachverhalten oder der Prüfung von Hypothesen.

Vom jeweiligen Redekontext hängt auch ab, in welchen Hinsichten und in welchem Umfang regulative Kommunikationsideale wie die kontrafaktisch unterstellten Voraussetzungen rationaler Diskurse jeweils abgeschwächt werden dürfen. Davon sind auch die Fairness von Redechancen betreffende Ideale wie die der unbeschränkten Zugänglichkeit bzw. Offenheit für Teilnehmer:innen und Kommunikationsbeiträge nicht ausgenommen. Das bedeutet offenbar, dass auch philosophische Überlegungen zur Fairness der Verteilung von Kommunikationschancen nur kontextsensitiv und nur in der interdisziplinären Zusammenarbeit mit einer kritischen Institutionentheorie sinnvoll sind. Das gilt zum einen, weil Aussagen zur Fairness der Verteilung von Kommunikationschancen den funktionalen Eigensinn der jeweils relevanten Praxiskontexte berücksichtigen müssen und zum anderen, weil auch Prognosen über die vermutlichen Konsequenzen habitueller oder institutioneller Veränderungen ein Verständnis der vielgestaltigen Kommunikationslandschaft voraussetzen.

Freilich könnten sich Philosoph:innen damit begnügen, nach Regeln für ›die‹ Öffentlichkeit zu suchen, verstanden als thematisch offener und für alle Teilnehmer:innen gleichermaßen zugänglicher zentraler Marktplatz, dessen Betriebsamkeit gar nicht auf eine spezifische Funktion verpflichtet ist. Das erscheint jedoch aus mehreren Gründen problematisch (Bhagwat 2019).

Erstens ergeben sich die Redefreiheit betreffende Fragen und Konflikte oft gerade innerhalb von Institutionen oder in teilöffentlichen Kontexten (man denke an die Thematisierung politischer Themen innerhalb des Bildungssystems).

Zweitens existieren zahlreiche Übergangsfelder und Schleusen zwischen spezialisierten, klar funktionsbezogenen und/oder zugangsbeschränkten Redekontexten und ›der‹ Öffentlichkeit, von der Medizin und Wissenschaftskommunikation über die mehr oder weniger strategische ›Öffentlichkeitsarbeit‹ von Behörden, Parteien, Unternehmen und Verbänden – und auch hier treten nicht selten Konflikte auf.

Drittens sind zumal durch die sozialen Medien neue Formen teilöffentlicher Kommunikation entstanden, die auch die traditionellen Grenzziehungen zwischen privatem und öffentlichem Raum in Frage stellen (Habermas 2021).

Viertens ist ›der‹ Rahmendiskurs der ›allgemeinen‹ Öffentlichkeit, der traditionell professionalisiert über Zeitungen, Radio- und Fernsehsender hergestellt, inzwischen aber zunehmend auf interaktiven Medienplattformen geführt wird, zwar thematisch weitgehend offen. Auch er erfüllt jedoch zwar vielfältige, aber als solche doch klar benennbare Funktionen – etwa der Krisenanzeige, als Medium des Ausdrucks von Interessen und der spontanen Selbstorganisation, der Verständigung über Deutungsperspektiven und Identitätskonzeptionen, als kultur- oder institutionenkritischer Metadiskurs und Medium lebensweltlicher Rationalisierung. Habermas'sche Motive aufgreifend ließe sich sagen, dass er gerade um dieser *übergreifenden* Kritik- und Selbstverständigungsfunktionen willen nicht durch die funktionale Logik gesellschaftlicher Teilsysteme dominiert sein darf.

Der vierte Punkt weist zugleich darauf hin, dass die funktionale Ausdifferenzierung sozialer Kommunikationspraxen für die kommunikationsethische Reflexion nicht nur als Komplikation zu Buche schlägt, sondern auch wesentliche Kriterien für die normative Orientierung liefert.

Das gilt etwa auch für Überlegungen zur kommunikativen Fairness. So folgt aus der Tatsache, dass der globale wissenschaftliche Fachdiskurs in den funktionalen Kontext der Erkenntnisgewinnung eingebettet ist, dass auch die Verteilung von Rede- und Rezeptionschancen im Sinne dieses Ziels effizient sein muss. Die Teilnehmenden (denen man die Orientierung am gemeinsamen Erkenntnisziel normativ zumuten darf) können es daher nicht als unfair betrachten, wenn beispielsweise eigene Beiträge seltener zitiert werden, weil sie weniger innovativ sind. Dysfunktional und zumindest potentiell[21] auch unfair ist es hingegen, wenn Wissenschaftler:innen rein strategische Zitierkartelle bilden oder Studienergebnisse unnötig häppchenweise publizieren um ihren Beiträgen größere Verbreitung zu sichern. Unfair ist auch, wenn die Fähigkeit von Wissenschaftler:innen, in besonders sichtbaren wissenschaftlichen Journals zu publizieren, durch nicht erkenntnisbezogene Faktoren wie vermeidbare Sprachbarrieren, *gender biases* oder die begrenzte Fähigkeit von Wissenschaftsinstitutionen in ärmeren Ländern, Druckkosten- oder OpenAccess-Zuschüsse zu leisten, begrenzt wird. In ähnlicher Weise lässt sich argumentieren, dass die in Abschnitt 3 skizzierte Strategie von Online-Handelsplattformen, die Informationen von Konkurrenzunternehmen unter einer Masse von Partnerseiten zu begraben, unfair ist, weil sie die Angebotstransparenz vermindern, die wiederum eine Bedingung für das Funktionieren eines effizienten

21 Falls solche Publikationspraktiken klar der Regelfall wären, blieben sie dysfunktional; es ließe sich aber bestreiten, dass einzelne Wissenschaftler:innen *unfair* handeln würden, wenn sie sich daran beteiligten.

Marktsystems darstellt. Solche Beispiele deuten zugleich darauf hin, dass Fragen kommunikativer Fairness manchmal durchaus klar zu beantworten sind.

Dass Mängel im Hinblick auf die kommunikative Fairness zu verzeichnen sind, ist freilich nur ein möglicher, aber keineswegs schon ein hinreichender Grund für regulatorische Eingriffe, zumal in der Regel ganz unterschiedliche Maßnahmen der Abhilfe möglich sind. Das führt zu einer letzten Überlegung: Die relative Aufwertung der anspruchs- und partizipationsrechtlichen Aspekte kommunikativer Grundrechte legt es nahe, gesellschaftliche Bemühungen um eine vernünftige und faire Gestaltung digitaler Kommunikationsstrukturen nicht lediglich als Anlass reaktiver gesetzlicher Regulierungen, sondern stärker auch als *Projekt aktiver gesellschaftlicher und technologiepolitischer Gestaltung* zu verstehen. Der digitale Kommunikationswandel bietet durchaus Möglichkeiten und gut funktionierende Vorbilder zur Realisierung von öffentlich-rechtlichen oder *community*-basierten Medien und Plattformen, die eine problematische Dominanz von Regierungen oder mächtigen Privatinteressen sowie manipulative Kommunikationsformen zumindest erschweren. Für die Weiterentwicklung solcher Modelle kann die Philosophie nur allgemeine Orientierungen, konditionale Empfehlungen und laienhafte Anstöße geben. Selbst dem kritischen Diskurs auf dem öffentlichen Marktplatz verpflichtet, kann ihr die Gestaltung dieses Marktplatzes aber nicht völlig gleichgültig sein.

Literatur

Alm, N. et al. (Hg.) (2022): Die digitale Transformation der Medien. Leitmedien im Wandel, Wiesbaden: Springer Fachmedien.

Amazon (2023): PartnerNet Website. [https://partnernet.amazon.de/help/node/topic/GK5TZZ4AWML2QSLA] (Zugriff: 02.05.2023).

American Psychological Association (2023): Health Advisory on Social Media Use in Adolescence. [https://www.apa.org/topics/social-media-internet/health-advisory-adolescent-social-media-use.pdf] (Zugriff: 10.05.2024).

Asscher, L.F. (2002): Communicatiegrondrechten. Een onderzoek naar de constitutionele bescherming van het recht op vrijheid van meningsuiting en het communicatiegeheim in de informatiesamenleving, Amsterdam: Cramwinckel.

Averbeck-Lietz, S. (Hg.) (2017): Kommunikationswissenschaft im internationalen Vergleich, Wiesbaden: Springer Fachmedien.

Bacher, V. (2022): Regulierungsbedarf beim Einsatz von Algorithmen in sozialen Medien zum Schutz der Meinungsfreiheit, in: *ELSA Austria Law Review*, 7(1), 64.

Baker, C.E. (1995): Advertising and a Democratic Press, Princeton: Princeton University Press.

Baker, C.E. (2006): Media Concentration and Democracy. Why Ownership Matters, Cambridge (NY): Cambridge University Press.

Balkin, J.M. (2004): Digital Speech and Democratic Culture. A Theory of Freedom of Expression for the Information Society, in: *New York University Law Review* 79(1), 1–55.

Balkin, J.M. (2009): The Future of Free Expression in a Digital Age, in: *Pepperdine Law Review*, 36(2), 427–444.

Balkin, J.M. (2018): Free Speech in the Algorithmic Society. Big Data, Private Governance, and New School Speech Regulation, in: *U.C. Davis Law Review*, 51, 1149–1210.

Barendt, E. (2. Aufl. 2007): Freedom of Speech, Oxford: Oxford University Press.

Barnes, R.; Zakrzewski, C. (2023): Supreme Court rules for Google, Twitter on terror-related content, in: *Washington Post*, 18.05.2023.

Bastos, M.; Farkas, J. (2019): »Donald Trump Is My President!«. The Internet Research Agency Propaganda Machine, in: *Social Media + Society*, 5(3). 205630511986546.

Beck, K. (2021): Kommunikationsfreiheit, Wiesbaden: Springer Fachmedien.

Beck, K.; Schweiger, W. (2001): Attention please! Online-Kommunikation und Aufmerksamkeit, München: Fischer.

Bennett, W.L. (2012): The Personalization of Politics. Political Identity, Social Media, and Changing Patterns of Participation, in: *The ANNALS of the American Academy of Political and Social Science*, 644(1), 20–39.

Bennett, W.L.; Livingston, S. (2018): The Disinformation Order. Disruptive Communication and the Decline of Democratic Institutions, in: *European Journal of Communication*, 33(2), 122–139.

Bhagwat, A. (2019): Free Speech Categories in the Digital Age, in: Brison, S.J.; Gelber, K. (Hg.), Free Speech in the Digital Age, Oxford/New York: Oxford University Press, 88–103.

Brison, S.J. (1998): The Autonomy Defense of Free Speech, in: *Ethics*, 108(2), 312–339.

Bromell, D. (2022): Regulating Free Speech in a Digital Age. Hate, Harm and the Limits of Censorship, Cham: Springer Nature.

Büchner, G. (2016): Der Hessische Landbote [1834], Stuttgart: Reclam.

Bunker, M.D. (2012): Originalism 2.0 Meets the First Amendment. The »New Originalism«, Interpretive Methodology, and Freedom of Expression, in: *Communication Law and Policy*, 17(4), 329–354.

Chang, C.-C.; Lin, T.-H. (2020): Autocracy Login. Internet Censorship and Civil Society in the Digital Age, in: *Democratization*, 27(5), 874–895.

Cho, J. et al. (2020): Do Search Algorithms Endanger Democracy? An Experimental Investigation of Algorithm Effects on Political Polarization, in: *Journal of Broadcasting and Electronic Media*, 64(2), 150–172.

Dawson, A.; Innes, M. (2019): How Russia's Internet Research Agency Built its Disinformation Campaign, in: *The Political Quarterly*, 90(2), 245–256.

De Gregorio, G. (2022): Digital Constitutionalism in Europe. Reframing Rights and Powers in the Algorithmic Society, Cambridge (NY): Cambridge University Press.

Dechêne, A. et al. (2010): The Truth About the Truth. A Meta-Analytic Review of the Truth Effect, in: *Personality and Social Psychology Review*, 14(2), 238–257.

Deibert, R.J.; Villeneuve, N. (2004): Firewalls and Power. An Overview of Global State Censorship of the Internet, in: Klang, M.; Murray, A. (Hg.), Human Rights in the Digital Age, London: Routledge-Cavendish, 111–124.

Europäische Kommission (2021): Vorschlag für eine Verordnung des Europäischen Parlaments und des Rates über die Transparenz und das Targeting politischer Werbung. [https://eur-lex.europa.eu/legal-content/DE/TXT/HTML/?uri=CELEX:52021PC0731] (Zugriff: 23.05.2023).

Feldman, S.M. (2017): The History, Philosophy, and Law of Free Expression in the United States. Implications for the Digital Age, in: Price, M.; Stremlau, N. (Hg.), Speech and Society in Turbulent Times, Cambridge (NY): Cambridge University Press, 192–212.

Filipović, A. (2009): Die Informationsfreiheit und der Zusammenhang von Abwehr- und Anspruchsrechten. Koreferat zu Karsten Weber, in: Aufderheide, D.; Dabrowski, M. (Hg.), Internetökonomie und Ethik. Wirtschaftsethische und moralökonomische Perspektiven des Internets, Berlin: Duncker & Humblot, 35–43.

Franck, G. (1998): Ökonomie der Aufmerksamkeit. Ein Entwurf, München: Hanser.

Freud, S. (1999): Massenpsychologie und Ich-Analyse [1921], in: Ders. (Hg.), Gesammelte Werke, Bd. XIII, Frankfurt a.M.: S. Fischer, 71–161.

Frischlich, L. (2022): H@te Online. Die Bedeutung digitaler Kommunikation für Hass und Hetze, in: Weitzel, G.; Mündges, S. (Hg.), Hate Speech. Definitionen, Ausprägungen, Lösungen, Wiesbaden: Springer Fachmedien, 99–131.

Gabriel, L. (2023): Die Macht digitaler Plattformen. Möglichkeiten und Konsequenzen der Personalisierung, Wiesbaden: Springer Fachmedien.

Geise, S. et al. (2021a): Wie normativ ist die Kommunikationswissenschaft?, in: *Publizistik*, 66(1), 89–120.

Geise, S. et al. (2021b): The Normativity of Communication Research. A Content Analysis of Normative Claims in Peer-Reviewed Journal Articles (1970–2014), in: *Mass Communication and Society*, 25(4), 528–553.

Goldhaber, M.H. (1997): The attention economy and the Net, in: *First Monday*. [https://firstmonday.org/ojs/index.php/fm/article/view/519] (Zugriff: 12.05.2023).

Graber, M.A. (1991): Transforming Free Speech. The Ambiguous Legacy of Civil Libertarianism, Berkeley: University of California Press.

Greenawalt, K. (1989): Free Speech Justifications, in: *Columbia Law Review*, 89(1), 119–155.

Greenslade, R. (2014): PRs outnumber journalists in the US by a ratio of 4.6 to 1, in: *The Guardian*, 13.05.2023.

Habermas, J. (1981): Theorie des kommunikativen Handelns, Frankfurt a.M.: Suhrkamp.

Habermas, J. (1990): Strukturwandel der Öffentlichkeit. Untersuchungen zu einer Kategorie der bürgerlichen Gesellschaft. Mit einem Vorwort zur Neuauflage 1990, Frankfurt a.M.: Suhrkamp.

Habermas, J. (2021): Überlegungen und Hypothesen zu einem erneuten Strukturwandel der politischen Öffentlichkeit, in: Seeliger, M.; Sevignani, S. (Hg.), Ein neuer Strukturwandel der Öffentlichkeit?, Baden-Baden: Nomos, 470–500.

Hahn, K.; Langenohl, A. (Hg.) (2017): Kritische Öffentlichkeiten – Öffentlichkeiten in der Kritik, Wiesbaden: Springer Fachmedien.

Han, R. (2018): Contesting Cyberspace in China. Online Expression and Authoritarian Resilience, New York: Columbia University Press.

Hendricks, V.F; Vestergaard, M. (2019): Reality Lost. Markets of Attention, Misinformation and Manipulation, Cham: Springer.

Hoffmann-Riem, W. (2002): Medienregulierung als objektiv-rechtlicher Grundrechtsauftrag, in: *Medien & Kommunikationswissenschaft*, 50(2), 175–194.

Hoffmann-Riem, W. (2020): Digitale Disruption und Transformation. Herausforderungen für Recht und Rechtswissenschaft, in: Eifert, M. (Hg.), Digitale Disruption und Recht, Baden-Baden: Nomos, 143–195.

Jaursch, J. (2022): Politische Werbung. Die Zukunft des Microtargeting in der EU, in: netzpolitik.org. [https://netzpolitik.org/2022/politische-werbung-die-zukunft-des-microtargeting-in-der-eu/] (Zugriff: 23.05.2023).

Jongepier, F.; Klenk, M. (2022): The Philosophy of Online Manipulation, New York: Routledge.

Kenyon, A.T. (2021): Democracy of Expression. Positive Free Speech and Law, Cambridge (NY): Cambridge University Press.

Krempl, U.S. (2023): Missing Link. 5 Jahre DSGVO – »Die gezielte Panikmache hat sich gelegt«, in: *heise online*. [https://www.heise.de/hintergrund/Missing-Link-5-Jahre-DSGVO-Die-gezielte-Panikmache-hat-sich-gelegt-9059939.html] (Zugriff: 10.05.2024).

Kubicek, H. (1996): Informationelle Grundversorgung in der Informationsgesellschaft, Bremen: Telecommunications Research Group, University of Bremen.

Kubin, E.; von Sikorski, C. (2021): The role of (social) media in political polarization. A systematic review, in: *Annals of the International Communication Association*, 45(3), 188–206.

Livni, E.; Kessler, S.; Mattu, R. (2023): Who Is Liable for A.I. Creations?, in: *The New York Times*, 03.06.2023.

MacKinnon, R. (2012): Consent of the Networked. The Worldwide Struggle For Internet Freedom, New York: Hachette UK.

Meiklejohn, A. (1948): Free Speech and Its Relation to Self-Government, New York: Harper Brothers.

Meyer, E.C. (2009): Informationsfreiheit – eine ökonomische Analyse. Koreferat zu Karsten Weber, in: Aufderheide, D.; Dabrowski, M. (Hg.), Internetökonomie und Ethik. Wirtschaftsethische und moralökonomische Perspektiven des Internets, Berlin: Duncker & Humblot, 45–55.

Mill, J.S. (1977): On Liberty [1859], in: Ders. (Hg.) Collected Works, Bd. 18, Toronto: University of Toronto Press, 213–310.

Milmo, D.; Hern, A. (2023): Elections in UK and US at risk from AI-driven disinformation, say experts, in: *The Guardian*, 20.05.2023.

Peters, J. (2018): How the Law Protects Hate Speech on Social Media, in: *Columbia Journalism Review*. [https://www.cjr.org/analysis/gab-hate-speech.php] (Zugriff: 31.12,.2022).

Peukert, A. (2022): Das Netzwerkdurchsetzungsgesetz. Entwicklung, Auswirkungen, Zukunft, in: Spiecker gen. Döhmann, I.; Westland, M.; Campos, R. (Hg.), Demokratie und Öffentlichkeit im 21. Jahrhundert – zur Macht des Digitalen, Baden-Baden: Nomos, 229–248.

Ragland, C.P.; Heidt, S. (Hg.) (2001): What is Philosophy?, New Haven/London: Yale University Press.

Redish, M.H. (2021): Commercial Speech as Free Expression. The Case for First Amendment Protection, Cambridge (NY): Cambridge University Press.

Reicher, S. (2022): From Crisis to Opportunity. New Crowd Psychology and Public Order Policing Principles, in: Madensen, T.D.; Knutsson, J. (Hg.), Preventing Crowd Violence, Boulder: Lynne Rienner Publishers.

Riehm, K.E. et al. (2019): Associations Between Time Spent Using Social Media and Internalizing and Externalizing Problems Among US Youth, in: *JAMA Psychiatry*, 76(12), 1266–1273.

Riehm, U.; Krings, B.-J. (2006): Abschied vom »Internet für alle«? Der »blinde Fleck« in der Diskussion zur digitalen Spaltung, in: *M&K Medien & Kommunikationswissenschaft*, 54(1), 75–94.

Roose, K. (2019): The Making of a YouTube Radical, in: *The New York Times*, CLXVIII(58,353).

Roose, K. et al. (2020): Rabbit Hole, in: *The New York Times*, Podcast vom 17.05.20 [https://www.nytimes.com/2020/04/17/podcasts/the-daily/rabbit-hole.html] (Zugriff: 04.06.2024).

Rosen, J. (2011): Interpreting The Constitution In The Digital Era, in: *NPR*. [https://www.npr.org/2011/11/30/142714568/interpreting-the-constitution-in-the-digital-era] (21.05.2023).

Rosen, J. (2012): The Deciders. The Future of Privacy and Free Speech in the Age of Facebook and Google, in: *Fordham Law Review*, 80(4), 1525–1538.

Rössler, B. (2017): Autonomie. Ein Versuch über das gelungene Leben, Berlin: Suhrkamp.

Russ-Mohl, S. (2022): Aufmerksamkeitsökonomie | Journalistikon. [https://journalistikon.de/aufmerksamkeitsoekonomie/] (Zugriff: 12.05.2023).

Saunders, K.W. (2003): Saving Our Children from the First Amendment, New York: New York University Press.

Scanlon, T. (1972): A Theory of Freedom of Expression, in: *Philosophy & Public Affairs*, 1(2), 204–226.

Schauer, F.F. (1982): Free Speech. A Philosophical Enquiry, Cambridge (NY): Cambridge University Press.

Schauer, F. (1983): Free Speech and the Argument from Democracy, in: *Nomos*, 25, 241–256.

Schneider, M. (2018): There are now more than 6 PR pros for every journalist, in: *MuckRack*. [https://muckrack.com/blog/2018/09/06/there-are-now-more-than-6-pr-pros-for-every-journalist] (Zugriff: 13.05.2023).

Schönhagen, P. (2004): Soziale Kommunikation im Internet. Zur Theorie und Systematik computervermittelter Kommunikation vor dem Hintergrund der Kommunikationsgeschichte, Bern/New York: Peter Lang.

Schweiger, W.; Beck, K. (Hg.) (2019): Handbuch Online-Kommunikation, Wiesbaden: Springer Fachmedien.

Sell, S. (2017): Kommunikationsfreiheit, Wiesbaden: Springer Fachmedien.

Smolla, R.A. (2020): Competing Conceptions of Free Speech, in: Ders. (Hg.), Confessions of a Free Speech Lawyer, Ithaca/London: Cornell University Press, 61–74.

Spindler, G. (2022): Funktion und Verantwortung von Plattformen als Informations-Intermediäre, in: Spiecker gen. Döhmann, I.; Westland, M.; Campos, R. (Hg.), Demokratie und Öffentlichkeit im 21. Jahrhundert – zur Macht des Digitalen, Baden-Baden: Nomos, 75–126.

StatCounter (2023): Search Engine Market Share Worldwide, in: StatCounter Global Stats. [https://gs.statcounter.com/search-engine-market-share] (Zugriff: 13.05.2023).

Stephens, B. (2020): Donald Trump Is Our National Catastrophe, in: *The New York Times*, 06.06.2020.

Sunstein, C.R. (1993): Democracy and the Problem of Free Speech, New York: Free Press.

Susser, D.; Rössler, B.; Nissenbaum, H.F. (2019): Online Manipulation. Hidden Influences in a Digital World, in: *Georgetown Law Technology Review*, 4(1), 1–45.

Sutter, T.; Mehler, A. (Hg.) (2010): Medienwandel als Wandel von Interaktionsformen, Wiesbaden: VS, Verlag für Sozialwissenschaften.

The Social Dilemma (2023), in: Wikipedia. [https://en.wikipedia.org/w/index.php?title=The_Social_Dilemma&oldid=1152378558] (Zugriff: 24.05.2023).

The U.S. Surgeon General (2023): Social Media and Youth Mental Health. [https://www.hhs.gov/sites/default/files/sg-youth-mental-health-social-media-advisory.pdf] (Zugriff: 04.06.2024).

Tushnet, M.V.; Chen, A.; Blocher, J. (2017): Free Speech Beyond Words. The Surprising Reach of the First Amendment, New York: New York University Press.

van Mill, D. (2021): Freedom of Speech, in: Zalta, E.N. (Hg.), The Stanford Encyclopedia of Philosophy. [https://plato.stanford.edu/archives/spr2021/entries/freedom-speech/] (Zugriff: 01.05.2023).

Vermeulen, J. (2022): To Nudge or Not to Nudge. News Recommendation as a Tool to Achieve Online Media Pluralism, in: *Digital Journalism*, 10(10), 1671–1690.

Vosoughi, S.; Roy, D.; Aral, S. (2018): The spread of true and false news online, in: *Science*, 359(6380), 1146–1151.

Weber, K. (2005): Das Recht auf Informationszugang. Begründungsmuster der politischen Philosophie für informationelle Grundversorgung und Eingriffsfreiheit, Berlin: Frank & Timme.

Weber, K. (2007): Plädoyer für unlimitierte Meinungsfreiheit als Grundlage einer europäischen Medienethik, in: *Zeitschrift für Kommunikationsökologie und Medienethik*, 9(1), 35–39.

Weber, K. (2009): Die Informationsfreiheit und der Zusammenhang von Abwehr- und Anspruchsrechte, in: Aufderheide, D.; Dabrowski, M. (Hg.), Internetökonomie und Ethik. Wirtschaftsethische und moralökonomische Perspektiven des Internets, Berlin: Duncker & Humblot, 11–33.

Werner, M.H. (2003): Ist das Böse selbst-verständlich? Zur Diskussion über »einfache Imperative« – ein Versuch, mit Apel gegen Apel zu denken, in: Böhler, D.; Kettner, M.; Skirbekk, G. (Hg.), Reflexion und Verantwortung. Auseinandersetzungen mit Karl-Otto Apel, Frankfurt a.M.: Suhrkamp, 83–96.

Werner, M.H. (2012): Property Rights, in: Chadwick, R.F. (Hg.), Encyclopedia of Applied Ethics. Second Edition, San Diego: Academic Press, 624–631.

Whitney, H.M.; Simpson, R.M. (2019): Search Engines and Free Speech Coverage, in: Brison, S.J.; Gelber, K. (Hg.), Free Speech in the Digital Age, Oxford/New York: Oxford University Press, 33–51.

Wu, T. (2018): Is the First Amendment Obsolete?, in: *Michigan Law Review* 117(3), 547–581.

Zillich, A.F. et al. (2016): Werte und Normen als Sollensvorstellungen in der Kommunikationswissenschaft. Ein Operationalisierungsvorschlag, in: *Publizistik*, 61(4), 393–411.

Sucht oder Autonomie?
Neue ExpertInnen im Netz

Nicola Mößner

Abstract: *During the Covid-19 pandemic, a significant number of people has seemingly been lured in believing conspiracy theories. Many deliberately disregarded expert advices by virologists and physicians to reduce new infections. This turning away from traditional expert authorities exemplifies the »crisis of expertise« that has been discussed in the philosophy of science for some time, namely that many people seem to have lost their trust in the established authority of expert knowledge and are looking for epistemic alternatives, especially on the Internet and in particular on social media. In this article, this digital cultural trend will be analysed. Will people actually becoming more epistemically autonomous as a result of this new tend? Attention is drawn to the epistemic and moral vulnerability of people who opt for new media as epistemic alternatives instead of relying on traditional expert opinion. It will be shown that some important presuppositions about the Internet and, especially, social media tools as alternative ways to gathering information and find moral support in a group of likeminded people do not hold.*

Keywords: *algorithm; expert knowledge; epistemic individualism; social media; trust; moral vulnerability*

Ausnahmezeiten wie jene der Corona-Pandemie können tiefliegende Verunsicherungen über das richtige Verhalten sowohl auf individueller als auch auf gesellschaftlicher Ebene offenlegen. In der durch den neuartigen Virus verursachten globalen Krisensituation ging es immer wieder um Abwägungen im Spannungsfeld von wirtschaftlichen Erwägungen und solchen, die den allgemeinen Gesundheitsschutz betrafen. Viele Menschen empfanden die Vorgaben zum Verhalten während der Pandemie als verwirrend. Warum war es einerseits verboten, den Urlaub am heimischen Badesee oder Strand zu verbringen, andererseits aber erlaubt, Flugreisen anzutreten? Selten ist die Suche nach Orientierungswissen so überdeutlich zu Tage getreten wie in den zurückliegenden Jahren der Pandemie. Diese Zeit

lehrte uns so nicht nur etwas über moralische Dilemmata, sondern auch etwas über epistemische Schwierigkeiten, die im Zusammenhang mit der Pandemie auftraten. Wie kam es beispielsweise, dass eine nicht unerhebliche Anzahl von Menschen in den Bann von Verschwörungstheorien geriet, die man unter normalen Umständen wohl als irrationalen Nonsense abgetan hätte?[1] Die beschriebene Problemsituation kann als ein Beispiel dafür betrachtet werden, was im Kontext der Wissenschaftstheorie unter dem Stichwort »Experten-Krise« oder »Krise der Expertise« populär geworden ist. Diese Terminologie verweist auf die Beobachtung, dass in den letzten Jahren scheinbar mehr und mehr Leute das Vertrauen in klassische ExpertInnen[2] verloren haben. Die Skepsis bezüglich unserer epistemischen Abhängigkeit von anderen ist dabei keineswegs auf bestimmte Gebiete beschränkt: Misstrauen in wissenschaftliche Expertise bezüglich des Klimawandels gilt als das notorische Beispiel dieser epistemologischen Debatte im US-amerikanischen Raum,[3] während im europäischen Umfeld – auch schon vor der Corona-Pandemie – viele Zweifel hinsichtlich der wissenschaftlichen Tragfähigkeit von Experten-Empfehlungen in Bezug auf Impfkampagnen in Kindergärten und Schulen diskutiert wurden.[4] Ob man sich auf Expertenmeinungen verlassen will oder nicht, ist somit durchaus nicht allein eine Frage persönlicher Präferenzen. Individuelle Entscheidungen und das nachfolgende Verhalten der Menschen betreffen oft nachhaltig das gesellschaftliche Zusammenleben. Vor diesem Hintergrund haben PhilosophInnen darauf hingewiesen, dass die ExpertInnenkrise zu problematischen Entwicklungen in westlichen Demokratien führen könne, die wesentlich auf wohlinformierte Bürgerentscheidungen angewiesen sind (vgl. z.B. Anderson 2011; Kitcher 2011).

Zwei Fragen stehen dabei im Fokus der Debatte: zum einen Überlegungen, welche Gründe das Vertrauen in Expertenmeinungen untergraben, zum anderen die Frage, welche möglichen Auswege es aus dieser epistemischen Bredouille geben

1 Vgl. z.B. Krause 2021.

2 Der Begriff›Experte‹wird hier in dem von Alvin I. Goldman definierten Sinne verwendet: »[…] an expert (in the strong sense) in domain D is someone who possesses an extensive fund of knowledge […] and a set of skills or methods for apt and successful deployment of this knowledge to new questions in the domain« (Goldman 2011: 115).

3 Der Fotograf James Balog dokumentierte in seinem Dokumantarfilmprojekt »Chasing Ice« [https://chasingice.com/] (Zugriff: 26.04.2024) die dramatischen Entwicklungen, indem er über einen längeren Zeitraum hinweg die schmelzenden Gletscher in der Arktis im Bild festhielt. Seine Intention war dabei, den SkeptikerInnen des Klimawandels dessen Auswirkungen unmittelbar vor Augen zu führen.

4 Dies wurde vom European Centre for Disease Prevention and Control (ECDPC) insbesondere im Zusammenhang mit dem Ausbruch der Masern in Europa zwischen 2016 und 2019 dokumentiert, vgl. [https://ecdc.europa.eu/en/news-events/ecdc-insufficient-vaccination-coverage-eueea-fuels-continued-measles-circulation] (Zugriff: 23.08.2023). Michael Butter (Butter 2021) erklärt in einem Interview mit ZEIT-Online den Zusammenhang zwischen Verschwörungstheorien und Impfkampagnen gegen das Corona-Virus.

könnte. Der vorliegende Artikel soll einen Beitrag zur ersten Frage liefern: Welche Gründe können Menschen dazu motivieren, dem individualistischen Trend in ihren epistemischen Prozessen zu folgen? Zugestandenermaßen muss eine Vielzahl unterschiedlichster Aspekte zu dem beschriebenen Wandel des epistemischen Verhaltens angenommen werden (vgl. z.B. Gelfert 2011; Kitcher 2012; Oreskes/Conway 2012; Proctor/Schiebinger 2008). Im Rahmen der vorliegenden Untersuchung wird aus dieser Vielfalt lediglich ein Punkt herausgegriffen, nämlich die Annahme, dass Informations- und Kommunikationstechnologien (IuK-Technologien) den epistemischen Individualismus befeuern. Diese Technologien ermöglichen es den Menschen erst, sich unabhängig von klassischen ExpertInnen zu machen, indem Informationen dem Einzelnen (scheinbar) unmittelbar direkt zugänglich werden. Es zeigt sich aber auch, dass die Nutzung von IuK-Technologien ebenfalls zu neuen epistemischen Risiken für ihre AnwenderInnen geführt hat. Schon seit geraumer Zeit werden in diesem Zusammenhang die Gefahren des Eingeschlossenseins in »Filterblasen« (vgl. Pariser 2012), »Echokammern« oder »Informations-Kokons« (vgl. Sunstein 2006: 8ff.) diskutiert.

In der genaueren Analyse wird dabei deutlich, dass, obwohl die NutzerInnen sich von ExpertInnenmeinungen freizumachen meinen, die Verwendung von IuK-Technolgien sie in neue Abhängigkeiten führt. Besorgniserregend ist der Umstand, dass diese Abhängigkeiten gezielt von den Anbietern der verschiedenen Social-Media-Plattformen angestrebt werden, wie IT-ExpertInnen (vgl. z.B. Lanier 2018) nachgewiesen haben.

Während die NutzerInnen also in der virtuellen Welt nicht nur selbstständig nach Informationen suchen, sondern auch bestrebt sind, emotionalen Rückhalt und Bestärkung zu finden, welche sie bei klassischen ExpertInnen nicht mehr zu finden meinen, wird bei genauerer Betrachtung der Situation klar, dass es sich oftmals tatsächlich um bloß fingierte Akteure, nämlich um programmierte Bots handelt, mit welchen die NutzerInnen auf diesen Plattformen interagieren. AnwenderInnen geraten somit eventuell nicht bloß in eine epistemisch prekäre Situation, wie sie unter dem Stichwort der Filterblasen angesprochen wurde, sondern werden darüber hinaus in dem Sinne betrogen, als sie künstliche Akteure für menschliche Unterstützer ihrer Meinungen und Ansichten halten. Diese Thematik in ihren epistemologischen Dimensionen näher zu beleuchten ist Ziel des nachfolgenden Beitrags.

In einem ersten Schritt wird die Debatte rund um das Thema zur ExpertInnen-Krise genauer skizziert. Es wird herausgestellt, dass erst mit dem Aufkommen der IuK-Technologien den NutzerInnen eine Alternative zur Verfügung stand, die eine echte Abkehr von traditionellen Expertenmeinungen ermöglichte. In einem zweiten Schritt soll dann diese Nutzung von IuK-Technologien im Kontext von Wissensprojekten näher analysiert werden. Eine wichtige Frage in diesem Zusammenhang lautet: welche Art von epistemischem Individualismus wird durch dieses Vorgehen etabliert? Schließlich wird auf die epistemische und emotionale Abhängigkeit der

NutzerInnen von IuK-Technologien und sich daraus ergebender Problemstellungen eingegangen. Es wird sich zeigen, dass wesentliche Grundannahmen vieler NutzerInnen in diesem Zusammenhang nicht erfüllt werden, nämlich vor allem die Auffassung, man würde notwendig einen höheren Grad epistemischer Autonomie erzielen, wenn man sich auf IuK-Technologien statt auf klassische Expertenmeinungen verlässt.

Die Krise der Expertise

Sicherlich handelt es sich letztlich um eine empirische Frage, ob wir es tatsächlich mit einer allgemeinen Krise in Bezug auf Expertenmeinungen zu tun haben. Die Schwierigkeit, dies sicher zu entscheiden, lässt sich wiederum am Beispiel der Corona-Pandemie verdeutlichen: Einerseits wurden in dieser Zeit Verschwörungstheorien sehr populär,[5] gleichzeitig haben aber auch viele Menschen sehr genau darauf geachtet, was die unterschiedlichen ExpertInnen zu den Entwicklungen der Pandemielage zu sagen hatten.[6]

In diesem Artikel soll weder eine optimistische noch eine pessimistische Sicht hinsichtlich eines vermeintlichen Trends des Sich-Verlassens auf Expertenmeinungen verteidigt werden. Die Punkte, die im Folgenden herausgearbeitet werden, betreffen – unabhängig von einer solchen Trendanalyse – Personen, die sich von den klassischen ExpertInnen als RatgeberInnen abwenden.

Auch wird in der nachfolgenden Diskussion für keine spezielle Position innerhalb der philosophischen Debatte zum ExpertInnen-Problem argumentiert. Es soll lediglich hervorgehoben werden, welche Aspekte eine Rolle spielen können, wenn epistemische Subjekte sich von Expertenmeinungen abwenden und inwiefern dies von einem epistemologischen Gesichtspunkt aus sinnvoll erscheinen kann. Eine hilfreiche Zusammenfassung der Hauptargumente der Debatte findet sich bei Philip Kitcher.[7] Im Anschluss an deren Darstellung wird untersucht, welche Aspekte in dieser Auflistung fehlen.

Kitcher hebt drei Hauptgründe hervor, welche Personen dazu gebracht haben könnten, ihre epistemische Einstellung bezüglich Expertenmeinungen zu ändern

5 Dies implizierte oftmals, dass ›neue‹ ExpertInnen von diesen Personengruppen konsultiert wurden, welche diese Theorien verbreiteten. Allerdings muss festgehalten werden, dass diese vermeintlichen ExpertInnen nicht die Bedingungen erfüllen, welche Goldmans Begriffsdefinition erfordern, die letztlich die Relevanz wahrer Überzeugungen in den Vordergrund rückt.

6 Zum Problem der Expertenmeinungen während der Corona-Pandemie vgl. Hauswald/Schmechtig 2023.

7 Interessierte LeserInnen finden Details zu dieser Debatte z.B. in Kitcher 2011; Leuschner 2012; Nichols 2017; Oreskes 2019 sowie in den darin enthaltenen Literaturhinweise.

(vgl. Kitcher 2012: 212f.): Erstens scheinen einige ExpertInnen das Problem unterschätzt zu haben, das entsteht, wenn Hypothesen verbreitet werden, die im Vorwege nicht hinreichend genau überprüft wurden. Hier droht ein Reputationsverlust. Oftmals geht eine voreilige Verlautbarung von Ergebnissen darauf zurück, dass ExpertInnen dem sozialen Druck, der von MedienvertreterInnen ausgeübt wird, um Meinungsbilder einzuholen, wenig reflektiert nachgeben. Sollte sich im Nachhinein allerdings herausstellen, dass veröffentlichte Einschätzungen nicht den Fakten entsprechen, kann dies einen nachhaltig negativen Effekt auf das wissenschaftliche Ansehen sowie die Glaubwürdigkeit der InterviewpartnerInnen haben.[8]

Als zweiten Grund führt Kitcher an, dass wissenschaftliche Laien häufig nicht die Dynamiken und Eigenheiten wissenschaftlicher Kommunikationsprozesse verstünden. Forschungsergebnisse werden meist recht vorsichtig formuliert, z.B. wenn Ausdrucksweisen gewählt werden wie ›Wir haben zu zeigen versucht, dass…‹ oder ›Es ist normalerweise der Fall, dass…‹. Üblicherweise werden keine steilen Thesen formuliert. Schon Karl Popper hat auf die menschliche Fehlbarkeit in epistemischen Prozessen hingewiesen – ein Fakt, der innerhalb der wissenschaftlichen Gemeinschaft weithin Anerkennung findet (vgl. Popper 1987: 225ff.). Demgegenüber steht die Medienberichterstattung, die oft stark vereinfachte Schwarz-Weiß-Bilder der Dinge zeichnet. Da die Laien an diese Art der Informationsvermittlung gewöhnt sind, kann für sie folglich der Eindruck entstehen, dass die ForscherInnen mit ihrer vorsichtigen Ausdrucksweise nicht wirklich Ahnung von dem haben, was sie vermitteln wollen.

Schließlich weist Kitcher darauf hin, dass Werturteile in den Wissenschaften unvermeidbar seien (vgl. Kitcher 2012: 213).[9] Allerdings kann festgestellt werden, dass einige dieser Urteile schlicht falsch sind. Sie sind beispielsweise das Resultat narzisstischer Haltungen oder von persönlichem Profitstreben. Als solche können sie dann in Konflikt mit demokratischen Idealen einer wohlinformierten Entscheidungsfindung geraten, wie Naomi Oreskes und Erik M. Conway (vgl. Oreskes/Conway 2012) im Zusammenhang mit ihrer Analyse der Beschäftigung von WissenschaftlerInnen in der Tabakindustrie herausgearbeitet haben. Offensichtlich gibt es schwarze Schafe innerhalb der wissenschaftlichen Gemeinschaft, also Personen, die

8 Susanne Hahn (Hahn 2021) hebt hervor, dass in diesem Kontext das Phänomen des Bullshits, d.h. Äußerungen, die keinen Anspruch auf Wahrheit erheben, eine wesentliche Rolle spielen kann. Sie weist darauf hin, dass dieser Prozess insbesondere durch das konstante Streben nach öffentlicher Aufmerksamkeit zusätzlich angeheizt wird. WissenschaftlerInnen sind häufig ebenso wie JournalistInnen in einem Teufelskreis gefangen, der sie zur Produktion von Bullshit treibe. Vgl. Hahn 2021: 226.

9 Er folgt damit Richard Rudners These, dass die wissenschaftliche Praxis selbst notwendig wertgeladene Urteile erforderlich mache (vgl. Rudner 1953). Eine elaborierte Verteidigung dieser These findet sich bei Heather Douglas (Douglas 2009).

sich allein von ihren privaten Interessen leiten lassen, statt für die Wahrheit einzutreten. Wenn Laien jedoch von solchen Fällen erfahren, kann dies dazu führen, dass sie künftig wissenschaftlichen ExpertInnen im Allgemeinen misstrauen.[10] Insgesamt bietet Kitchers Analyse einen guten Überblick, wie üblicherweise innerhalb der Wissenschaftsphilosophie das Problem der Glaubwürdigkeitskrise wissenschaftlicher Expertise dargestellt wird. Allerdings muss festgehalten werden, dass in der bisherigen Diskussion wenigsten zwei Punkte fehlen: Einerseits muss es eine echte Alternative dafür geben, an Wissen zu gelangen, damit eine wissenschaftsskeptische Haltung der Laien auch tatsächlich dazu führt, dass sie sich von klassischen ExpertInnen als Informationsquelle abwenden. Diese Funktion erfüllen die schon genannten IuK-Technologien.

Andererseits erscheint die philosophische Analyse etwas unausgewogen. Der Fokus wird zumeist darauf gelegt, auf welche Art und Weise WissenschaftlerInnen ihr (Kommunikations-)Verhalten ändern sollten, um die besprochene Glaubwürdigkeitskrise wieder zu überwinden. Diese Sichtweise unterschlägt aber, dass Vertrauenssituationen immer mindestens zwei Parteien involvieren. PhilosophInnen diskutieren häufig Fälle, in denen sich WissenschaftlerInnen als unzuverlässig erweisen. Sie idealisieren aber dabei meist die andere Seite, indem sie voraussetzen, dass die Laien sich rational verhalten. Dass dies nicht notwendig der Fall sein muss, wird deutlich, wenn man die Rolle von Emotionen in Betracht zieht, die hier ebenfalls relevant sind.[11]

Im Folgenden soll gezeigt werden, dass die epistemische und die soziale Dimension im Kontext der Glaubwürdigkeitskrise der ExpertInnen eng verwoben sind. Diese problematische Verknüpfung wird anhand des Beispiels der #MeToo-Debatte erläutert, wobei der aufgezeigte Effekt weit über den genannten Kontext hinausreicht. Es wird sich zeigen, warum die sozialen Medien scheinbar so erfolgreich darin sind, die Lücke zu schließen, die die Abkehr von klassischen ExpertInnen hinterlässt.

Das Beispiel bezieht sich auf den Ausgangspunkt der #MeToo-Debatte im Jahr 2017. Zu diesem Zeitpunkt berichtete die New York Times über die Anklage des bekannten Medienproduzenten Harvey Weinstein wegen sexueller Belästigung.

---

10 Die Laien sind mit dem Problem konfrontiert, auf Basis welcher Belege sie wem trauen sollen. Goldman (Goldman 2011) hat sich explizit mit dieser Schwierigkeit auseinandergesetzt, wenn er diskutiert, auf Grund welcher Bedingungen ein Laie eine begründete Entscheidung zwischen zwei vermeintlichen Expertenmeinungen treffen könnte. Allerdings kann diese Ausgangslage noch wesentlich komplexer werden, wenn man sich vorstellt, dass ein Experte in einem Kontext durchaus vertrauenswürdig agiert, in einem anderen aber z.B. durch finanzielle Interessen zu einem Fehlverhalten motiviert wird.

11 Vincent F. Hendricks und Mads Vestergaard (Hendricks/Vestergaard 2018) führen einige der psychologischen Effekte an, die im Zusammenhang mit der Glaubwürdigkeitskrise in Expertenmeinungen eine Rolle spielen, vgl. Hendricks/Vestergaard: Kap. 5.

Nur zehn Tage nach dieser Veröffentlichung sammelten sich mehr und mehr Stimmen von Frauen, die Ähnliches erlebt hatten, unter dem Hashtag MeToo auf der Social-Media-Plattform Twitter.[12] Die Soziologin Eva Illouz fasst die Entwicklung folgendermaßen zusammen: »#MeToo ist die erste westliche Bewegung, die auf sozialen Medien beruht: Hier schildern Frauen ihre Erlebnisse unmittelbar, ohne dass eine lange Kette von Experten (Psychologen, Juristen, Journalisten) ihre Rede abschwächte oder verfälschte.« (Illouz 2018: 48) In dem Zitat wird deutlich, dass viele, die in der #MeToo-Bewegung aktiv sind, die sozialen Medien als ein Instrument des Empowerments betrachten. Gerade in diesem Kontext scheint die Nutzung von IuK-Technologien vorteilhaft, da sie es offenbar ermöglichen, sich unzensiert zu Wort zu melden. Klassische ExpertInnen, die als vermeintlich parteiisch wahrgenommen werden, können so umgangen werden. Insbesondere scheint so das Problem des sogenannten Silencing gelöst zu werden, d.h. die Schwierigkeit, dass die vorherrschenden paternalistischen Strukturen dafür sorgen, dass jene Personen in den relevanten Machtpositionen verbleiben, deren Fehlverhalten angeprangert werden soll, und die folglich eine öffentliche Kritik zu unterbinden suchen. Die sozialen Medien bieten den betroffenen Frauen in der #MeToo-Bewegung nun die Möglichkeit, Verbündete in ihrem Kampf um Gerechtigkeit zu finden.

Seit dem Aufkommen der Internettechnologie in den 1990er Jahren geht mit dieser auch das Versprechen der Demokratisierung des Wissens einher. Viele gehen nach wie vor davon aus, dass das Internet eine stetig wachsende Menge an unterschiedlichsten Informationen für immer mehr NutzerInnen verfügbar macht. Betont wird in diesem Kontext das Potential der IuK-Technologien für einen freien und unbeschränkten Zugang zu Informationen, von denen einige, so wird postuliert, zuvor nur bestimmten elitären ExpertInnengruppen zur Verfügung gestanden hätten. Darüber hinaus erlauben IuK-Technologien nicht allein den passiven Konsum bestehender Informationen, sondern binden die NutzerInnen durch die sozialen Medien vermehrt in die Produktion und Verbreitung von Informationen ein. Das epistemische Empowerment, das an den IuK-Technologien häufig hervorgehoben wird, beruht demnach auf der verbreiteten Annahme, dass Informationen nicht mehr ein von bestimmten Eliten unter Verschluss gehaltener Schatz, sondern für die Allgemeinheit zugänglich sind. Die These ist, dass NutzerInnen dadurch auch einen höheren Grad epistemischer Autonomie erlangen. Aber ist das tatsächlich der Fall?

Illouz' Zitat verdeutlicht, dass die Glaubwürdigkeitskrise der ExpertInnen nicht allein durch epistemische Erwägungen angeheizt wird, wie Kitcher es in seiner Analyse anspricht. Im Unterschied dazu zeigt sich im Beispielfall, dass oftmals auch der Wunsch nach moralischer Unterstützung wesentlich erscheint. Dass eine emotionale Komponente mit ins Spiel kommt, lässt sich auf den Umstand zurückführen,

12 Twitter wurde nach dem Erwerb der Plattform durch Elon Musk im Jahr 2022 in »X« umbenannt.

dass es sich letztlich um eine Vertrauenssituation handelt, die hier adressiert wird. Bernd Lahno hat in seiner Untersuchung des Vertrauensbegriffs klar herausgearbeitet, dass es sich bei diesem um einen mehrdimensionalen Begriff handelt (vgl. Lahno 2004: 38ff.). Der emotionale Faktor wird dann offenbar, wenn vom Vertrauenden angenommen wird, dass die Person, der Vertrauen geschenkt wird, dieselben Ziele und Werte teilt. Darüber hinaus reagieren viele Menschen auf verletztes Vertrauen nicht rational, sondern mit dem Gefühl der Enttäuschung.

Die Glaubwürdigkeitskrise der ExpertInnen betont damit nicht allein Zweifel an deren Rolle als neutrale InformationslieferantInnen, vielmehr wird auch mangelndes Einfühlungsvermögen und fehlende moralische Unterstützung moniert. Social-Media-Plattformen und die sich auf ihnen zusammenfindenden Communities scheinen diese Aufgabe in den Augen vieler NutzerInnen besser zu erfüllen. Es bleibt allerdings fraglich, ob diese Wahrnehmung gerechtfertigt ist, sprich, ob IuK-Technologien tatsächlich die angesprochenen Bedürfnisse besser befriedigen können.

Eine neue Form des epistemischen Individualismus?

Einige Personen scheinen anzunehmen, sie würden in epistemischer Hinsicht mehr Autonomie gewinnen, wenn sie sich von den klassischen ExpertInnen ab und den IuK-Technologien als Informationsquelle zuwenden. Diese Annahme wirft jedoch zwei Fragen auf: Zum einen, warum haben die Akteure nicht schon früher mehr epistemische Autonomie angestrebt? Zum anderen, mit welcher Art von epistemischer Autonomie und epistemischem Individualismus haben wir es hier konkret zu tun?

Natürlich haben die Menschen auch früher schon nach epistemischer Autonomie gestrebt. Die unterschiedlichsten Informationen sind bereits seit langem in Bibliotheken usw. zugänglich. Die beschriebene epistemische Haltung ist also kein neues Phänomen, wie ein Blick insbesondere auf die Zeit der Aufklärung zeigt, die als ein durchgängiger Versuch, sich von epistemischen Abhängigkeiten zu befreien, verstanden werden kann (vgl. z.B. Kant 1999). Nichtsdestotrotz kann aber ein wesentlich qualitativer Unterschied zu den Entwicklungen der jüngsten Dekade festgestellt werden, der sich vor allem auf die umfangreichen Neuerungen im Zusammenhang mit den IuK-Technologien zurückführen lässt. Nie war für viele Leute die Suche nach Informationen so einfach wie in der heutigen Zeit.

Bedeutet das Streben nach mehr Autonomie in den Wissensprojekten auch ein Wiederaufleben des epistemischen Individualismus? Das zugehörige Ideal des autonomen Wissenden beschreibt Elizabeth Fricker folgendermaßen: »The wholly autonomous knower will not accept any proposition unless she herself possesses the evidence establishing it. Thus she will not accept anything on the basis of another's

word for it, even when she has evidence of their trustworthiness on the topic in question.« (Fricker 2006: 225) Dieses Ideal wurde einst von John Locke verfochten, der sich überaus kritisch zum Wissen aus dem Zeugnis anderer Menschen äußerte: »For I think we may as rationally hope to see with other men's eyes, as to know by other men's understandings. So much as we ourselves consider and comprehend of truth and reason, so much we possess of real and true knowledge. The floating of other men's opinions in our brains, makes us not one jot the more knowing, though they happen to be true.« (Locke 1690: 84) Seine Position kann als testimonialer Nihilismus bezeichnet werden. Das Wort der anderen spielt für ihn nur insofern eine Rolle, als es den Rezipienten auf neue Informationen aufmerksam macht. Um jedoch zu Wissen zu gelangen, muss die relevante Neuigkeit stets selbst überprüft werden, indem die individuellen Erkenntnisquellen der Wahrnehmung und der Vernunft zu Rate gezogen werden.

Das gegenwärtige Streben nach epistemischer Autonomie fällt jedoch nicht mit dem radikalen lockeschen Ideal zusammen, das sich darüber hinaus in praktischer Hinsicht auch als vollkommen unerfüllbar erweist, wie Fricker hervorhebt (vgl. Fricker 2006: 227f.). Heutzutage verlassen sich die Menschen auch weiterhin auf das Wort der anderen. Der Unterschied besteht allein darin, dass sie selbst entscheiden wollen, wem sie ihr Vertrauen schenken und wem nicht. Sie wollen nicht länger den traditionellen Schemata folgen, dass der vermeintliche Expertenstatus einer Person – z.B. eines Lehrers oder einer Wissenschaftlerin – diese automatisch zu einer zuverlässigen Informationsquelle macht.

Fricker hebt in diesem Zusammenhang hervor, dass diese zwei epistemischen Phänomene – unsere testimoniale Abhängigkeit einerseits und der Wunsch nach epistemischer Autonomie andererseits – zunächst kontradiktorisch erscheinen. Und tatsächlich besteht hier ein gewisses Spannungsverhältnis, denn die Notwendigkeit, anderen als Informationsquelle zu vertrauen, impliziert das Risiko, zu falschen Überzeugungen zu gelangen. Schließlich könnte der Zeuge lügen oder auf Grund eigener Inkompetenz etwas Falsches vermitteln (vgl. Fricker 2006: 242).

Darüber hinaus ist das Wort der anderen oftmals relevant für die praktischen Ziele epistemischer Subjekte. In dieser Hinsicht betreffen die negativen Effekte eventuell nicht allein unsere epistemischen Bestrebungen, sondern auch unser praktisches Handeln. Die Rezipienten sind also doppelt verletzbar: zum einen gelangen sie eventuell zu falschen Überzeugungen. Zum anderen treffen sie, basierend auf diesen, unter Umständen falsche praktische Entscheidungen. Diese können wiederum negative Auswirkungen auf sie selbst, aber auch auf andere haben. Zum Beispiel schädigen sie sich vielleicht selbst, weil sie sich gegen eine notwendige medizinische Behandlung entscheiden. Ebenso können aber beispielsweise auch die eignen Kinder auf Grund falscher Überzeugungen der Eltern gesundheitliche Schäden davontragen, weil letztere sich fälschlicherweise gegen wichtige Impfungen (siehe das Beispiel der Masern-Epidemie) entschieden haben.

Ein in epistemischer Hinsicht überlegenes Subjekt, so scheint es, würde am besten fahren, wenn es sich vollständig aus der Abhängigkeit vom Wort der anderen befreien könnte. Allerdings, so führt Fricker aus, sind Menschen nicht in der Lage, diese Position einzunehmen. Jede Person ist letztlich auf die eine oder andere Weise in ihrer kognitiven Leistungsfähigkeit beschränkt. Eine epistemische Abhängigkeit von anderen ist unvermeidlich.

Diese Abhängigkeit ist Fricker zufolge aber kein Nachteil, denn sich auf andere zu verlassen und epistemisch autonom zu sein können kompatible epistemische Strategien sein. Entscheidend sei, dass das epistemische Subjekt eine sorgfältige Auswahl treffe, wem es sein Vertrauen schenken will.[13] Insbesondere gehe es hierbei natürlich um eine genaue Abwägung der vermeintlichen Glaubwürdigkeit des Sprechers, d.h. dessen Aufrichtigkeit und Kompetenz (vgl. Fricker 2006: 243). Sind diese Vorbedingungen erfüllt, dann kann man auch von den epistemischen Fähigkeiten anderer profitieren.

Zusammengefasst: Viele Personen, die sich von traditionellen ExpertInnen abwenden, substituieren die auftretende epistemische Lücke dadurch, dass sie meinen, die relevanten Informationen ebenso gut selbst im Internet – v.a. in den sozialen Medien – finden zu können. Zwar zeigt diese Praxis den Wunsch der RezipientInnen nach einer größeren epistemischen Autonomie auf, führt aber in keiner Weise zu einer grundsätzlichen Abkehr von der epistemischen Arbeitsteilung.

In den folgenden zwei Abschnitten wird nun ein genauerer Blick auf die Schwierigkeiten geworfen, die sich ergeben, wenn man sich in epistemischen Kontexten auf die sozialen Medien verlässt. Es wird sich zeigen, dass neben den von Fricker bereits angesprochenen epistemischen und praktischen Herausforderungen auch die Frage nach emotionaler Verbundenheit auftritt, die Online-Communities scheinbar anbieten.

Soziale Medien und die epistemische Verletzbarkeit der NutzerInnen

Typischerweise wird von traditionellen ExpertInnen nicht nur erwartet, Ratschläge auf individuelle Fragestellungen hin zu geben, sie spielen auch eine wichtige Rolle im öffentlichen Diskurs. Im Unterschied zu Gemeinschaften von Gleichgesinnten ist es nicht die Aufgabe der ExpertInnen, bequeme Geschichten zu erzählen, sondern die Wahrheit (oder zumindest das, was sie dafürhalten): beispielsweise Ratschläge bezüglich der Abstandsregeln während der Corona-Pandemie. Werden ihre

13 An dieser Stelle wird nicht näher darauf eingegangen, wie nach Fricker eine solche Auswahl sinnvoll getroffen werden kann und welche Schwierigkeiten mit der vorgeschlagenen Strategie verknüpft sind. Weitere Ausführungen zu diesen Punkten finden sich in Gelfert 2014: 110ff.; Mößner 2010: Kap. 3.2.3.2.

Mitteilungen durch die Nachrichtenströme der sozialen Medien ersetzt, verlieren die NutzerInnen dieser Plattformen ein wertvolles Korrektiv für ihre Überzeugungen, was in der Folge auch zu ernsthaften sozialen und politischen Problemen führen kann.

Eine der Hauptschwierigkeiten der Nachrichtenströme auf den Plattformen der sozialen Medien ist die Gefahr, in sogenannte Filterblasen eingeschlossen zu werden. Eli Pariser, der diese Begrifflichkeit eingeführt hat, beschreibt ihren Effekt als eine Art Informationsdiät (vgl. Pariser 2012: 14). Ein ähnliches Phänomen erläutert Cass R. Sunstein (vgl. Sunstein 2006) unter dem Terminus »echo chambers«.[14] Auch dieser Begriff bezieht sich auf Gemeinschaften, die ihren Mitgliedern wertvolle Informationen vorenthalten. Beide Phänomene führen letztlich zu einer oft stark verengten Perspektive auf bestimmte Fragestellungen – beispielsweise hinsichtlich der Herausforderungen der globalen Klimaerwärmung.

Filterblasen sind das Ergebnis von Algorithmen, die Informationen nach zuvor etablierten Profilen von Individuen filtern. Angezeigt wird also nur das, was vermeintlichen Interessen und Präferenzen des Individuums entspricht.[15] Echokammern sind dagegen das Resultat von psychologischen und sozialen Mechanismen in Gruppen. In solchen Kontexten werden oft die vorherrschenden Meinungen einfach wiederholt, wodurch diese über die Zeit hinweg dazu tendieren, immer extremer zu werden. Auch wenn einige Personen innerhalb der Gruppe die Mehrheitsmeinung eventuell nicht teilen, werden sich diese jedoch wahrscheinlich nicht offen gegen sie aussprechen, da der Wunsch nach Gruppenzugehörigkeit dominiert.

Aus epistemologischer Perspektive sind beide Phänomene – Filterblasen und Echokammern – in mehrfacher Hinsicht problematisch: Sie können dazu beitragen, dem epistemischen Subjekt wichtige Informationen vorzuenthalten, die notwendig wären, um zu Entscheidungen bezüglich konkreter Fragestellungen zu gelangen.[16]

---

14 Beide Begriffe und die postulierten negativen Effekte werden inzwischen kritisch betrachtet (vgl. Bruns 2019). Dir Kritik richtet sich vor allem gegen Vereinfachungen und übertriebene Verallgemeinerungen. Zugegebenermaßen werden sich nicht alle Online-Communities in Filterblasen verwandeln und nicht alle der letzteren Art haben ihren Ursprung in der virtuellen Welt. Ferner treffen nicht alle negativen Effekte, die Pariser und Sunstein anführen, auf alle Mitglieder solcher Gemeinschaften gleichermaßen zu. Ungeachtet dieser Vorbehalte verhelfen ihre Analysen doch zu einer Vorstellung, wie sich die genannten Technologien auswirken könnten.

15 Diese Filterprozesse, die für die individuelle Informationsversorgung in sozialen Medien typisch sind, wirken sich natürlich umso massiver aus, je weniger sich die Betroffenen dieser Mechanismen bewusst sind. Pariser (Pariser 2012) hebt diesen Aspekt explizit hervor. Eine philosophische Analyse dazu findet sich in Mößner/Kitcher 2017.

16 So wird angenommen, dass ein Effekt von Filterblasen darin besteht, dass Personen darin kaum mit abweichenden Meinungen konfrontiert werden, da die Filtermechanismen insbesondere Informationen zur Präsentation auswählen, die mit bisher schon vertretenen Ansichten übereinstimmen. Die Technik bedient hier das, was in der Psychologie als »Bestä-

Wie Fricker verdeutlicht hat, kann diese Informationseinschränkung das Individuum sowohl in epistemischer als auch in praktischer Hinsicht negativ betreffen. Der Ratschlag traditioneller ExpertInnen erscheint hier oft als die bessere Alternative, denn von diesen RatgeberInnen kann angenommen werden, dass sie den Leuten nicht nach dem Mund reden, sondern ihnen verdeutlichen, was tatsächlich der Fall ist.[17]

Allerdings geht es hier nicht nur um die epistemische Situation des Einzelnen. Sich von traditionellen ExpertInnen abzuwenden kann die Meinungs- und Willensbildung der Bevölkerung demokratischer Staaten insgesamt betreffen. In Demokratien müssen BürgerInnen über die relevanten Informationen verfügen, um (Wahl-)Entscheidungen bezüglich des Allgemeinwohls treffen zu können. Voraussetzung dafür ist, dass sie Meinungen austauschen und sich über die Interessen, Überzeugungen und Bedürfnisse ihrer MitbürgerInnen hinreichend informieren können. Empathie und Verständnis für fremde Lebensbedingungen sind in solchen Abwägungsprozessen wichtig. Sunstein schreibt: »As preconditions for a well-functioning democracy, these requirements – chance encounters and shared experiences – hold in any large country. They are especially important in a heterogeneous nation – one that faces an occasional danger of fragmentation.« (Sunstein 2018: 7)

Leider üben Filterblasen und Echokammern genau an diesem Punkt ihren schädlichen Einfluss aus. Diese Phänomene könnten unter Umständen den Austausch von Ideen in demokratischen Gesellschaften in einem solchen Ausmaß beeinträchtigen, dass – über die Zeit hinweg – die reale Gefahr entsteht, dass gesellschaftliche Spaltungen den demokratischen Prozess blockieren. ›Wir-gegen-die-anderen‹ und ähnliche populistische Rhetoriken könnten die Oberhand gewinnen, wenn immer mehr BürgerInnen immer weniger mit fremden Interessen, Überzeugungen und Bedürfnissen konfrontiert werden, sondern nur noch mit denen einer bestimmten (eigenen) Gruppe.

Aus diesem Grund betont Sunstein den wesentlichen Unterschied zwischen einem Bürger und einem Konsumenten. Für letztere mögen personalisierte Informationsangebote im Internet eine hilfreiche Angelegenheit sein, für erstere dagegen schaffen sie oftmals eine ganze Reihe von Problemen. In diesem Zusammenhang stellt Sunstein klar die epistemischen Pflichten der BürgerInnen heraus, wenn es um die epistemische Basis ihrer Entscheidungsfindung geht:

---

tigungsvorurteil« (confirmation bias) bekannt geworden ist, nämlich der Effekt, dass Menschen das erzählt wird, was sie gerne hören und wodurch sie in ihren bestehenden Meinungen bestärkt werden. Vgl. Hendricks/Vestergaard 2018: 126ff.

17 Impliziert wird dabei natürlich, dass ExpertInnen grundsätzlich bereit sind zu helfen. Das heißt, dass sie nicht darauf aus sind, ihre eigenen Ziele und Vorteile zu verwirklichen, sondern den Ratsuchenden bestmöglich zur Seite stehen wollen.

> »Citizens are not supposed merely to press their own self-interest narrowly conceived, nor are they to insulate themselves from the judgements of others. Even if they are concerned with the public good, they might make errors of fact or value – errors that can be reduced or corrected through the exchange of ideas. Insofar as people are acting in their capacity as citizens, their duty is to ›meet others‹ and ›consult,‹ sometimes through face-to-face discussions, and if not, through other routes, as, for example, by making sure to consider the views of those who think differently.« (Sunstein 2018: 51)

Wiederum spielen traditionelle ExpertInnen eine wichtige Rolle, da ihr Wissen und ihre diskursiven Fähigkeiten als Grundpfeiler eines öffentlichen Forums für das Teilen von Ideen und Erfahrungen gelten können. Natürlich können ExpertInnen sich irren. Doch können sie nichtsdestotrotz wichtige Informationen aus den jeweiligen Bereichen ihrer Expertise für die öffentliche Diskussion bereitstellen, z. B. wie man den Umweltschutz vorantreiben kann, ohne dadurch den Arbeitsmarkt zu gefährden, oder wenn es um die Frage geht, welche Auswirkungen wir bei einem weiteren Ausbau der Technologie des autonomen Fahrens zu erwarten haben.

Halten wir fest: Sowohl von einem epistemologischen als auch von einem politischen Standpunkt aus betrachtet steht viel auf dem Spiel, wenn BürgerInnen in Demokratien sich allzu gemütlich in ihren Online-Communities einrichten. Glücklicherweise wird hierüber inzwischen geforscht und kritisch nachgedacht.

Es gibt aber noch einen weiteren Grund, warum NutzerInnen sich in der digitalen Welt nicht blauäugig bewegen sollten. Im nächsten Abschnitt wird die schon angedeutete Schwierigkeit genauer expliziert. Sie hängt zusammen mit dem Wunsch nach emotionalem Rückhalt, den einige Personen nicht mehr bei traditionellen ExpertInnen finden können, dafür aber scheinbar in ihren Online-Communities erhalten. Sie suchen dann oft nach moralischer Unterstützung, nach dem, was man heute auch als Empowerment bezeichnet. Es stellt sich jedoch die Frage, wem die NutzerInnen hier ihr Vertrauen tatsächlich schenken, wenn sie meinen, auf Online-Plattformen diese Art von Unterstützung zu erhalten.

Der Betrug der Algorithmen

Im Folgenden wird die These kritisch diskutiert, dass man tatsächlich ein höheres Maß an epistemischer Autonomie gewinnen kann, wenn man sich auf IuK-Technologien stützt. Zweitens wird analysiert, auf wen oder, besser gesagt, auf was Menschen vertrauen, die sich in Online-Communities bewegen.

Die erste These besagt, dass Personen, die nach mehr epistemischer Autonomie streben, dies heutzutage besser können, weil ihnen ein neues Hilfsmittel – IuK-Technologien – enorme Mengen von Informationen verfügbar macht. Schnell

und bequem lassen sich diese Informationsmengen durchsuchen, indem man Suchmaschinen wie Google »Fragen stellt«. Darüber hinaus haben weitergehende technologische Entwicklungen dazu geführt, dass Online-Communities nun etwas anbieten können, was klassischen ExpertInnen abzugehen scheint, nämlich moralische und emotionale Unterstützung für ihre Mitglieder.

Unabhängig von der empirischen Frage, ob diese Angebote tatsächlich erfolgreich und/oder besser als frühere Alternativen sind, bildet die Annahme, dass NutzerInnen frei wählen können, auf wen sie sich bei ihren Online-Aktivitäten verlassen möchten, eine wesentliche Prämisse. Allerdings spricht einiges dafür, dass genau diese Prämisse falsch ist. Kann dies gezeigt werden, verliert die These vom Zugewinn an epistemischer Autonomie durch die Nutzung von IuK-Technologien eine entscheidende Stütze.

Dass hier tatsächlich die Crux liegt, beruht auf den technologischen Mechanismen, welche sich die sozialen Medien zunutze machen, und den psychologischen Annahmen, die dabei im Hintergrund eine Rolle spielen. Eine genauere Betrachtung zeigt, warum die Prämisse der freien Wahl falsch ist. Tatsächlich profitieren die IuK-Technologien der sozialen Medien von gewissen psychologischen Prädispositionen des menschlichen Gehirns. Wir sind geneigt, Handlungen zu wiederholen und Verhaltensmustern zu folgen, die durch unser Umfeld belohnt werden. Auf diesen Mechanismus setzt beispielsweise unser Ausbildungswesen. Lernaktivitäten werden durch positive Rückmeldungen angespornt.

Viele Menschen suchen nach Wegen, positive Unterstützung zu vervielfachen – insbesondere dann, wenn diese mit realen Belohnungen wie Zertifikaten, monetärer Entlohnung oder Reputationsgewinn verbunden ist. Manchmal mag es aber auch schon ausreichen, dass man schlicht gelobt wird. Und genau diesen Belohnungsmechanismus haben ProgrammiererInnen zum Bestandteil von Social-Media-Plattformen gemacht. Die Likes, Klicks und Kommentare, die NutzerInnen hier erhalten, funktionieren nach eben diesem Prinzip – also als eine Form von Online-Belohnung. In diesem Sinne beuten die IuK-Technologien die psychologischen Prädispositionen ihrer NutzerInnen aus. Letztere werden durch passende Anreize motiviert, mehr Zeit auf den Plattformen zu verbringen. Dieses Verhalten kann sich im Lauf der Zeit zu einer Sucht entwickeln.

Süchtige sind aber alles andere als autonom. Wenn NutzerInnen mehr oder wenig süchtig nach den Rückmeldungen auf den Plattformen sind, sind sie nicht mehr in der Lage, unabhängig und frei darüber zu entscheiden, welche Nachrichtenströme sie aufnehmen und auf welche Informationsquellen sie sich verlassen wollen. Epistemische Autonomie und ein Suchtverhalten im beschriebenen Sinne sind offensichtlich unvereinbar. Daher muss die These vom Zugewinn an epistemischer Autonomie durch Nutzung von IuK-Technologien in Form sozialer Medien, die nach den erläuterten Mechanismen arbeiten, als falsch zurückgewiesen werden.

Dass es diese Mechanismen gibt, ist keine philosophische Dystopie. Sie existieren tatsächlich in der beschriebenen Form, wie z.B. Jaron Lanier gezeigt hat. Selber ein Pionier der virtuellen Welt, weiß er glaubwürdig über die technologischen Entwicklungen zu berichten und gehört inzwischen zu ihren schärfsten Kritikern:[18] »How can you remain autonomous in a world where you are under constant surveillance and are constantly prodded by algorithms run by some of the richest corporations in history, which have no way of making money except by being paid to manipulate your behavior?« (Lanier 2018: 2)

Lanier bezweifelt also radikal, dass NutzerInnen der sozialen Medien autonom agieren können, denn die von ihnen verwendeten IuK-Technologien zielten unmittelbar darauf ab, die AnwenderInnen süchtig zu machen (vgl. Lanier 2018: 7ff.). Er erläutert, dass die Abhängigkeit zum Teil durch Implementierung zufälliger Rückmeldungen erzeugt wird, welche die NutzerInnen auf diesen Plattformen erhalten.[19] Die Likes, Klicks und Kommentare sind zum Teil künstlich geniert und führen dennoch dazu, dass NutzerInnen ihr Verhalten entsprechend anpassen, um mehr von diesen vermeintlichen Online-Belohnungen zu erhalten.

Der Computerspezialist Lanier ergänzt, dass diese von den Algorithmen reproduzierten Mechanismen längst nicht mehr so funktionieren wie vormals die Werbung für bestimmte Produkte. Vielmehr werden sie systematisch dazu verwendet, NutzerInnen psychologisch zu manipulieren. »The core process that allows social media to make money and that also does the damage to society is behavior modification.« (Lanier 2018: 10) Es wird wiederum offenbar, dass das Suchtverhalten bei der Nutzung von Online-Diensten, mit dem die Manipulation und Kontrolle von NutzerInnen einhergeht, mit der These vom Zugewinn an Autonomie nicht zusammenpasst.

Darüber hinaus sollte man sich klarmachen, auf wen oder was man sich eigentlich verlässt, wenn man sich in entsprechenden Online-Communities bewegt und nach Information und emotionaler oder moralischer Unterstützung sucht. Sind es (gleichgesinnte) Menschen, wie man meint?

Auch diese zweite Annahme in der These vom Zugewinn an Autonomie ist unzutreffend. Denn in vielen Fällen sind es gar keine realen menschlichen Wesen, die das Grundgerüst entsprechender Online-Communities bilden, sondern Algorithmen. Die NutzerInnen meinen also bloß, auf das Verständnis und Mitgefühl, Empathie

18 2014 erhielt Lanier den Friedenspreis des Deutschen Buchhandels für sein Werk Who Owns the Future? (Lanier 2014).

19 »The pioneers of the online exploitation of this intersection of math and the human brain were not the social media companies, but the creators of digital gambling machines like video poker, and then of online gambling sites. Occasionally, pioneers of the gambling world complain about how social media companies ripped off their ideas and made more money, but mostly they talk about how social media is helping them identify the easiest marks.« (Lanier 2018: 15f.)

und Interesse gleichgesinnter Menschen bauen zu können, während ihnen in Wirklichkeit Computerprogramme solches Einfühlungsvermögen bloß vorspielen.

Eine beunruhigende Konsequenz dieser Technologie ist dann, dass Filterblasen und Echokammern wiederum auf von Algorithmen generierten Meinungsäußerungen beruhen. Im Umkehrschluss heißt das auch, dass Polarisierungs- und Radikalisierungseffekte, wie sie von Pariser und Sunstein als Folge dieser Phänomene beschrieben werden, ihren Ausgangspunkt zum Teil in zufällig generierten Aussagen von Bots haben. Extremistische Haltungen, wie sie laut Sunstein in Chat-Gruppen entstehen können, basieren eventuell also auf bloß künstlich erzeugten Meinungen, denen kein wirkliches soziales oder politisches Programm zugrunde liegt, die vielmehr bloße Zufallsprodukte der Technologie sind, wie Lanier verdeutlicht. »Because the stimuli from the algorithms don't mean anything, because they genuinely are random, the brain [of the user] isn't adapting to anything real, but to a fiction.« (Lanier 2018: 15) NutzerInnen radikalisieren sich in ihren eigenen Meinungen, weil sie, ohne es zu wissen, von Algorithmen dazu gebracht werden – keine guten Aussichten für demokratische Gesellschaften!

Lanier führt aus: »Fake people are present in unknown but vast numbers and establish the ambiance. […] Invisible social vandalism ensues. Social pressure, which is so influential in human psychology and behavior, is synthesized.« Und er ergänzt: »Massive fake social activities turn out to influence real people. They indirectly create a genuine social reality, which means they make money. People are successfully manipulated by them.« (Lanier 2018: 36, 57) Er weist darauf hin, dass NutzerInnen oft nicht in der Lage sind, zwischen Meinungsäußerungen realer Personen und solchen der künstlichen Intelligenz zu differenzieren. Von Computerprogrammen gesetzte Likes und Kommentare erscheinen ununterscheidbar von solchen realer menschlicher NutzerInnen. Zudem meint Lanier, sei vielen AnwenderInnen nach wie vor nicht bewusst, dass diese Art von Manipulation auf den Plattformen der sozialen Medien erfolgt.

Auf computergenerierte positive Rückmeldungen reagiert das menschliche Gehirn nicht anders als auf die von menschlichen Kommunikationspartnern. Es ist daher möglich, dass NutzerInnen ihr Verhalten und ihre Meinungen künstlich generierten Thesen anzupassen versuchen, um im Belohnungssystem der Likes und Klicks weiter zu profitieren. Das läuft auf die dystopische Pointe hinaus, dass die psycho-sozialen Mechanismen, welche das Verhalten und die Meinungsbildung der NutzerInnen in Online-Communities steuern, letztlich von Technologien kontrolliert werden, die lediglich vorgeben, Menschen zu sein.

Die vermeintliche emotionale Unterstützung stiftet hier also häufig die sogenannte künstliche Intelligenz. Es ist davon auszugehen, dass dies nicht die Art von Rückhalt und Empathie ist, nach der die NutzerInnen ursprünglich gesucht haben.

In dieser Hinsicht kann man tatsächlich von einem Betrug der Algorithmen besprechen.[20]

Resümee

Die vorgehende Analyse begann mit der Beobachtung, dass eine nicht unerhebliche Zahl von Personen sich von ExpertInnen in konventionellem Verständnis und deren Ratschlägen zurückzieht, z.B. Personen, die während der Corona-Pandemie in das Universum der Verschwörungstheorien abdrifteten. An solchen Personengruppen wird deutlich, was gemeint ist, wenn von einer »Krise der Expertise« in der Wissenschaftsphilosophie die Rede ist. Es zeigte sich, dass neben drei Gründen, die üblicherweise in der Debatte genannt werden, warum Menschen nach mehr eigener epistemischer Autonomie streben, zwei zusätzliche Punkte angeführt werden müssen: Erstens die Verfügbarkeit technologischer Informationslieferanten, die IuK-Technologien, als Alternativen zu ExpertInnen; zweitens der Wunsch vieler NutzerInnen nach emotionaler und moralischer Unterstützung sowie ihre Wahrnehmung, dass klassische ExpertInnen diesen Wunsch nicht erfüllen.

Die Analyse machte deutlich, dass der gegenwärtige Trend zu mehr epistemischer Autonomie jedoch nicht mit einem radikalen epistemischen Individualismus à la Locke einhergeht, also nicht per se unsere Praxis der epistemischen Arbeitsteilung gefährdet. Allerdings wollen heute immer mehr epistemische Subjekte gerne selbst entscheiden, wem sie ihr Vertrauen schenken und wen sie um Rat fragen wollen und wen nicht. Ihr Anliegen wird unterstützt von der Überzeugung, dass das Internet mittlerweile alle Informationen zur Beantwortung ihrer Fragen bereithält. Außerdem erscheint es ihnen so, dass sie in den Online-Communities der sozialen Medien jenen emotionalen und moralischen Rückhalt finden, den sie vermissen. In diesem Sinne scheinen bestimmte IuK-Technologien sowohl die epistemischen als auch die sozialen Bedürfnisse ihrer NutzerInnen vollauf zu befriedigen.

Nimmt man die tatsächlichen Angebote, die im Web gemacht werden, jedoch genauer in den Blick, zeigt sich schnell, dass oftmals keiner der genannten Bedarfe wirklich gedeckt wird. Auf Grund der enormen Menge an Informationen im Netz wurden bereits früh Strategien der Personalisierung entwickelt, um das Angebot sinnvoll auf die NutzerInnen zuschneiden zu können. Diese Entwicklungen führten jedoch unter bestimmten Bedingungen zu Phänomenen, die als »Filterblasen« bekannt geworden sind und die letztlich die epistemischen Möglichkeiten

20 Die technologische Entwicklung hat hier sicherlich eine neue Stufe erreicht, auch wenn das Phänomen selbst nicht ganz neu ist. Viele Menschen zeigen bereits seit geraumer Zeit eine Tendenz, sich emotional auf Technologieprodukte einzulassen. Erinnert sei an dieser Stelle beispielhaft an die Tamagotchis, die virtuellen Haustiere der 1990er Jahre.

nicht ausweiten, sondern verengen. NutzerInnen erhalten einen Strom an bestätigenden Meinungen ihrer vorher geäußerten Überzeugungen. Irritierende oder falsifizierende Daten erreichen sie dagegen nur schwerlich.

Verstanden als soziale Konstrukte können die Filterblasen mit den »Echokammern«, die Sunstein diskutiert, verglichen werden. Meinungsbildung in solchen Kontexten wird immer von Vorurteilen behaftet und parteiisch sein. Darüber hinaus besteht eine Tendenz zur Radikalisierung, wie besonders Sunstein herausgearbeitet hat. Filterblasen und Echokammern schränken damit nicht nur die epistemischen Leistungen des Individuums ein, sondern stellen ferner auch eine Herausforderung dar für soziale und politische Aktivitäten, die auf solchen Informationsquellen aufgebaut werden. Sie reduzieren unter Umständen die Diversität der vertretenen Überzeugungen und unterbinden die Möglichkeit für Zufallsbegegnungen unter MitbürgerInnen mit abweichenden Meinungen, indem sie die Mitglieder der Online-Communities voneinander abschirmen. Letztlich bedeutet das ein wesentliches Defizit relevanter Informationen für die Meinungs- und Willensbildung in demokratischen Gesellschaften.

Darüber hinaus gibt es eine gewisse Tendenz zum Verfall der Diskussionskultur in geschlossenen Gemeinschaften gleichgesinnter Personen. Oftmals werden abweichende Meinungen nicht sachlich zur Kenntnis genommen und in Erwägung gezogen, sondern insbesondere bei politischen Themen entweder ignoriert oder zum Gegenstand verbaler Angriffe und Beleidigungen gemacht.[21] In diesem Sinne gerät eine weitere Stütze demokratischer Gesellschaften unter Druck, nämlich der argumentative Austausch von Meinungen – sprich, die »rationalen Diskurse«, von denen der demokratische Rechtsstaat zehrt (vgl. Habermas 1992).

Schließlich wurde in der obigen Analyse herausgearbeitet, dass viele dieser negativen Effekte tatsächlich das Resultat von Algorithmen sind. Computerprogramme erzeugen Stimuli bei den NutzerInnen, von denen bekannt ist, dass sie suchterzeugendes Potential besitzen. Somit erwies sich die zentrale optimistische Annahme, die Leute seien grundsätzlich frei bei der Wahl ihrer Informationsquellen im Netz, als unzutreffend. Sucht und Autonomie sind klarerweise einander entgegengesetzt.

Als falsch stellte sich auch die Annahme heraus, NutzerInnen könnten auf den Plattformen der sozialen Medien emotionalen und moralischen Rückhalt von (gleichgesinnten) KommunikationspartnerInnen gewinnen. Soweit auch hier bloße Simulations-Technologien am Werk sind (und dies nimmt mit dem Einsatz von KI zu), kann von echter Empathie und Unterstützung keine Rede sein.[22]

---

21 Hate speech und Shitstorms sind zu bekannten Phänomenen unseres Internetzeitalters geworden: vgl. z.B. Heinze 2016.

22 Für hilfreiche Kommentare zu einer früheren Fassung dieses Textes bedanke ich mich bei Philip Kitcher, Susanne Hahn und Matthias Kettner.

Literatur

Anderson, E. (2011): Democracy, Public Policy, and Lay Assessments of Scientific Testimony, in: *Episteme*, 8(2), 144–164.

Bruns, A. (2019): Are filter bubbles real?, Oxford: Polity.

Butter, M. (2021): Die Corona-Impfung ist ein Traum für Verschwörungstheoretiker, in: *Zeit Online*, 23.01.2021 [https://www.zeit.de/digital/internet/2021-01/michael-butter-verschwoerungstheorien-corona-impfung-soziale-medien-quer denken/] (Zugriff: 22.02.2022).

Douglas, H.E. (2009): Science, Policy, and the Value-free Ideal, Pittsburgh: University of Pittsburgh Press.

Fricker, E. (2006): Testimony and Epistemic Autonomy, in: Lackey, J.; Sosa, E. (Hg.): The Epistemology of Testimony, Oxford: Clarendon Press, 225–250.

Gelfert, A. (2011): Expertise, Argumentation, and the End of Inquiry, in: *Argumentation*, 25(3), 297–312.

Gelfert, A. (2014): A Critical Introduction to Testimony, London: Bloomsbury Publishing.

Goldman, A.I. (2011): Experts: Which Ones Should You Trust?, in: Goldman, A.I.; Whitcomb, D. (Hg.): Social Epistemology. Essential Readings, Oxford: Oxford University Press, 109–133.

Habermas, J. (1992): Faktizität und Geltung. Beiträge zur Diskurstheorie des Rechts und des demokratischen Rechtsstaats, Frankfurt a.M.: Suhrkamp.

Hahn, S. (2021): Bullshit in Science? On Epistemic Norms, Credibility and the Role of Science in Society, in: Michel, J.G. (Hg.): Making Scientific Discoveries. Interdisciplinary Reflections, Paderborn: Brill | mentis, 217–231.

Hauswald, R.; Schmechtig, P. (Hg.) (2023): Wissensproduktion und Wissenstransfer unter erschwerten Bedingungen. Der Einfluss der Corona-Krise auf die Erzeugung und Vermittlung von Wissen im öffentlichen Diskurs, Baden-Baden: Karl Alber.

Hendricks, V.F.; Vestergaard, M. (2018): Postfaktisch. Die neue Wirklichkeit in Zeiten von Bullshit, Fake News und Verschwörungstheorien, München: Karl Blessing.

Heinze, E. (2016): Hate Speech and Democratic Citizenship, Oxford: Oxford University Press.

Illouz, E. (2018): Es ist Krieg, in: *DIE ZEIT*, 2018/42, 48.

Kant, I. (1999): Was ist Aufklärung? Ausgewählte kleine Schriften, Hamburg: Felix Meiner.

Kitcher, P. (2011): Science in a Democratic Society, Amherst (NY): Prometheus Books.

Kitcher, P. (2012): Platons Rache. Undemokratische Nachricht von einem überhitzten Planeten, in: Hagner, M. (Hg.): Wissenschaft und Demokratie, Berlin: Suhrkamp, 189–214.

Krause, K. (2021): Ich liebe meine Mutter, aber ich verstehe sie nicht, in: *ZEITmagazin*, 2021/2, 16–25.

Lahno, B. (2004): Three Aspects of Interpersonal Trust, in: *Analyse & Kritik*, 26(1), 30–47.

Lanier, J. (2014): Who Owns the Future?, New York: Simon & Schuster.

Lanier, J. (2018): Ten Arguments for Deleting your Social Media Accounts right now, New York: Henry Holt and Company.

Leuschner, A. (2012): Die Glaubwürdigkeit der Wissenschaft. Eine wissenschafts- und erkenntnistheoretische Analyse am Beispiel der Klimaforschung, Bielefeld: transcript.

Locke, J. (1690): An Essay Concerning Human Understanding. Part I, London 1690, in: Projekt Gutenberg. [https://www.gutenberg.org/files/10615/10615-h/10615-h.htm] (Zugriff: 22.02.2022).

Mößner, N. (2010): Wissen aus dem Zeugnis anderer. Der Sonderfall medialer Berichterstattung, Paderborn: mentis.

Mößner, N.; Kitcher, P. (2017): Knowledge, Democracy, and the Internet, in: *Minerva*, 55(1), 1–24.

Nichols, T. (2017): The Death of Expertise. The Campaign against Established Knowledge and Why It Matters, New York: Oxford University Press.

O'Neil, C. (2016): Weapons of Math Destruction. How Big Data Increases Inequality and Threatens Democracy, London: Penguin Books.

Oreskes, N. (2019): Why Trust Science?, Princeton/Oxford: Princeton University Press.

Oreskes, N.; Conway, E.M. (2012): Merchants of Doubt. How a Handful of Scientists Obscured the Truth on Issues from Tobacco Smoke to Global Warming, New York: Bloomsbury Press.

Pariser, E. (2012): The Filter Bubble. What the Internet is Hiding from You, London: Penguin Books.

Popper, K.R. (2. Aufl. 1987): Duldsamkeit und intellektuelle Verantwortung, in: Ders. (Hg.): Auf der Suche nach einer besseren Welt. Vorträge und Aufsätze aus dreißig Jahren, München: Piper, 213–229.

Proctor, R.N.; Schiebinger, L. (Hg.) (2008): Agnotology. The Making and Unmaking of Ignorance, Stanford: Stanford University Press.

Rudner, R. (1953): The Scientist Qua Scientist Makes Value Judgments, in: *Philosophy of Science*, 20(1), 1–6.

Sunstein, C.R. (2006): Infotopia. How Many Minds Produce Knowledge, Oxford: Oxford University Press.

Sunstein, C.R. (2018): #Republic. Divided Democracy in the Age of Social Media, Princeton: Princeton University Press.

van Eimeren, B.; Simon, E.; Riedl, A. (2017): Medienvertrauen und Informationsverhalten von politischen Zweiflern und Entfremdeten, in: *Media Perspektiven*, 11(2017), 538–554.

Internetquellen

Chasing Ice: [https://chasingice.com/] (Zugriff: 26.04.2024).

European Centre for Disease Prevention and Control (ECDPC): [https://ecdc.europa.eu/en/news-events/ecdc-insufficient-vaccination-coverage-eueea-fuels-continued-measles-circulation] (Zugriff: 23.08.2023).

Friedenspreis des Deutschen Buchhandels“: [https://www.friedenspreis-des-deutschen-buchhandels.de/die-preistraeger/2010-2019/jaron-lanier] (Zugriff: 01.05.2024).

Roboter gegen Einsamkeit?

Zur Reproduktionsdynamik falscher und mangelnder Anerkennung durch »soziale« KI

Kerrin Artemis Jacobs

Abstract: *This analysis focuses on attempts to overcome loneliness with social AI and criticizes such attempts in the light of a theory of social recognition. The experience of loneliness significantly alters the experience of meaning, and often in ways that affect a person's relationship to self and others. The analysis reveals that interaction with social AI is often experienced as »successful« because it creates a strong illusion of interpersonal intersubjectivity. I argue that while social AI can promote human well-being and wellbeing in some areas of the lifeworld, it tends to be counterproductive in areas of distressing loneliness. The thesis states that the use of social AI to cope with the endemically increasing suffering from loneliness in technologically highly developed societies is even accelerated, as inadequate conditions of intersubjective recognition are themselves at the core of the clinical symptoms of painful loneliness. Empirical observation of the conditions of loneliness management, especially in Japan, shows that the use of »social robots« and other social AI does not necessarily lead to an overcoming of painful loneliness, but rather results in forms of digital loneliness.*

Keywords: *loneliness; isolation; AI companion; social recognition; social pathology*

1. Einleitung: X-Bots – harmlose Toys oder gefährliche Tools?

Dem Erleben von Einsamkeit ist eine signifikante Veränderung von Sinnzusammenhängen und Beziehungserleben inhärent, die das Verhältnis einer Person zu sich selbst, zu anderen Menschen und zu ihrer Umwelt grundlegend (zumeist negativ) beeinflusst. Um dies zu beschreiben kann man über eine rein subjektiv-phänomenologische Perspektive hinausgehen und Einsamkeit als den symptomatischen Ausdruck von veränderten Anerkennungsbeziehungen in hochdigitalisierten Gesellschaften verstehen: Wo Einsamkeit ist, hat sich ein Anerkennungsverhältnis verändert. Dieses Thema wird in Abschnitt 2 behandelt.

In Abschnitt 3 gehe ich der Frage nach, ob und wie digitale Strategien zum Problem und vielleicht auch zur Lösung der Einsamkeit beitragen. Zwar verändern Digitalisierungsprozesse unsere sozialen Beziehungen durchgreifend. Sie verstärken die kommunikative Vernetzung und begünstigen unter Umständen auch individuelles und kollektives *Empowerment*. Aber sie haben bisher nicht zur Senkung der Einsamkeitsquote in hochdigitalisierten Gesellschaften geführt. Diese steigt vielmehr an, und ein Teil des Anstiegs kann durch den digitalen Wandel unserer Beziehungsgestaltung erklärt werden. Deswegen konzentriere ich mich in dieser Untersuchung auf einen Spezialfall der digitalen Beziehungsgestaltung: auf die Beziehungen, die Menschen mit sogenannter sozialer KI, vor allem *X-Bots*[1] in Form von *Companion AI* (im Folgenden *CAI* genannt) eingehen. Meine Vermutung ist, dass Einsamkeit besonders dort endemische Ausmaße annimmt, wo ihre Eindämmung vornehmlich digital forciert wird, wie z.B. in Japan, wo Millionen in die Produktion von allen möglichen Formen von *X-Bots* investiert werden,[2] und sich *CAI*s längst als »Ersatz« bzw. als Erweiterungskomponente in und für allen möglichen Formen sozialer Beziehung im Alltag etabliert haben. In der Tat ist mein Beispiel der *CAI*-Mensch-Beziehung zur Darstellung der digitalen Reproduktion von Einsamkeit mit Blick auf die Kulturkreise, in denen Roboter Teil des Alltags sind und behandelt werden als

1 Mit dem Begriff »soziale X-Bots« (Jacobs 2023a; Jacobs 2024) kann man alle Arten von sozialer KI bezeichnen, d.h. digitale »Begleiter« (engl. *digital companion*), die in menschliche Beziehungsdynamiken (z.B. spezifische kommunikative Settings) integriert werden können oder darauf ausgelegt sind, sich diesen anzupassen, wobei es dabei erhebliche Unterschiede hinsichtlich der Spezifität der KI-Verkörperung, der Schnittstellenmöglichkeiten und der »autonomen« Systemkopplung des jeweiligen Gerätes gibt. Es muss hier auch nicht auf alle Unterschiede zwischen den verschiedenen Arten von Robotern wie *Winky*, *Aibo*, *MiRo*, *Paro*, *EmotiRob*, *Pepper*, *Dinsow*, *ElliQ*, *Atlas*, *Asimo*, *Harmony*, *LOVOT* oder Chatbots wie *ELIZA*, *Alexa*, Xiaolce, *Replika*, *Tess*, *Woebot* und *Wysa* usw. weiter eingegangen werden, denn meine These ist, dass sogar elaborierte X-Bots in der Regel nicht die Art von Anerkennung (re-)produzieren, die es tatsächlich bräuchte, um die Pathodynamik der Vereinsamung mit Hilfe einer Beziehung zu einem Roboter aufzuhalten.

2 Die Zahl der Geburten in Japan sank im Jahr 2016 das erste Mal seit 1899 unter eine Million, was in einem Land, das mit massiver Gerontifizierung zu kämpfen hat, zu einem hohen Anstieg der Produktion und Nutzung insbesondere von Pflege-KI geführt hat. Vereinsamung wird in Japan ein immer größeres Problem und Statistiken besagen, dass ungefähr 6.24 Millionen Japaner über 65 Jahre alt sind und 18.4 Millionen Erwachsene allein leben, mit steigender Tendenz: Prognostisch werden im Jahr 2040 ungefähr 40 % der Einwohner ohne Partner oder Familie leben. Das *Ministerium für Internationale Angelegenheiten und Kommunikation* hat ermittelt, dass die Zahl der Menschen über 65 Jahre auf 36.4 Millionen ansteigen kann. Das japanische *Wirtschafts- und Handelsministerium* prognostiziert, dass die Roboter-Industrie jährlich um $4 Billionen bis 2035 wachsen wird, was ungefähr dem 25-fachen ihres jetzigen Werts entspricht. Wir haben es also mit einer der umsatzstärksten Nischen in der Tech-Branche zu tun, für die die »Probleme« der wachsenden Einsamkeit und Überalterung ein sehr lukratives Geschäft verspricht.

wären sie menschlich, gar kein Sonderphänomen digitaler Beziehungsgestaltung, sondern steht für alle Formen, in denen wir uns mit sozialer KI im Alltags so verbinden, dass wir ihr meistens vertrauen (z.B. wenn wir *Alexas* Wissen nicht grundsätzlich in Frage stellen, wenn sie uns eine Frage beantwortet) oder wo wir sogar eine starke emotionale Bindung mit *X-Bots* eingehen: So wurde das Hologramm *Hatsune Miku*[3] geheiratet, Menschen werden intim mit Love-Dolls (z.B. mit *Harmony*[4]), manche KI-Partner (z.B. Roboter-Haustiere wie *Paro*[5]) können in offiziell autorisierten Abschiedszeremonien beerdigt werden (wie es z.B. in Japan möglich ist[6]), nicht wenige Menschen werden tagtäglich von *Pepper*[7] unterstützt (wie z.B. in Pflegeeinrichtungen in Japan, wo Assistenzrobotik im Alltag bereits viel verbreiteter ist als z.B. in Deutschland), Lernroboter wie *Winky*[8] werden als eine pädagogische Bereicherung für die eigenen Kindererziehung geschätzt, und manche Leute konsultieren

3 *Miku Hatsune* ist eine von der japanischen Mangaka und Illustrator Kei Garo im Auftrag von Crypton Future Media entworfene virtuelle Figur, die das Unternehmen Gatebox als Charakter für einen Companion Chatbot verwendet. Ein japanischer Mann namens Akihiko Kondo hat diese virtuelle Figur sogar offiziell geheiratet und dies ist ein gutes Beispiel dafür, dass man sich zukünftig verstärkt mit dem spezifischen Rechtsstatus (und entsprechend auch den Haftungsfragen) von X-Bots auseinander zu setzen hat. Siehe: [https://www.otaquest.com/hatsune-miku-gatebox-marriage/] (Zugriff: 25.11.2022).

4 *Harmony ist* ein Begleitroboter, der in Optik und Haptik realistisch menschlich wirken soll und von den Käufern an die eigenen Bedürfnisse angepasst werden kann, z.B. durch die Auswahlmöglichkeit von »Persona-Merkmalen«: [https://www.althumans.com/companion-robots/real-doll.html] (Zugriff: 24.11.2022).

5 *Paro* ist ein KI-Haustier, das einem kanadischen Sattelrobbenbaby ähnelt, und vor allem das Kindchenschema triggert. Dieser *X-Bot* wird seit 2003 in Pflegeeinrichtungen in Japan eingesetzt und reagiert auf taktile Reize und erkennt Temperatur, Körperhaltung sowie Lichtquellen und zielt im Design auf emotionales Bonding ab, indem Menschen z.B. das Gefühl haben, sich »kümmern« zu können.

6 Manche *CAIs* bedeuten ihren Besitzern so viel, dass sie auch eine angemessene Abschiedszeremonie (ähnlich einer Begräbniszeremonie) erhalten können, wie zum Beispiel bei dem von Leiya Arata gegründeten *Love Doll Funeral Services* in Osaka (Japan). Für den Preis von 800 Dollar können die Kunden ihre Roboter-Freunde auch »beerdigen« lassen. Geleitet werden die Trauerfeiern von dem Mönch Lay Kato.

7 *Pepper* ist in der Verkörperung ein klassischer »Semihumanoid« (hat also ein »Gesicht« und Arme) und wurde von *SoftBank Robotics* hergestellt und 2014 in Japan eingeführt. Die wichtigsten Zielfunktionen des Designs liegen in der Erkennung von Sprache, auch Emotionen, auf der Grundlage von Gesichts- und Stimmanalysen. *Pepper* wird z.B. an japanischen Flughäfen eingesetzt, um Reisenden mit Informationen weiterzuhelfen.

8 *Winky* ist ein Spielroboter des französischen Start-ups *Mainbot* und verfügt über Mikrofone, Sensoren, einen Lautsprecher, LEDs, einen rotierenden Kopf, rotierende Ohren, einen Bewegungs- und Entfernungsdetektor und ein Gyroskop, die es ihm ermöglichen, mit Kindern und der Umwelt zu interagieren.

Chat-Bots wie *Replika*[9] für therapeutische Zwecke oder ein romantisches Gespräch. Menschen können in *X-Bots* sogar eine Exklusivität finden, die sie hoch bewerten, z.B. dass sie mit einem *X-Bot* etwas »teilen« können, was negativ sanktioniert würde, wenn sie es mit anderen Menschen teilen wollten.[10] *CAIs* können wegen ihrer spezifischen Leistungen, vor allem aber wegen der ständigen Verfügbarkeit ihrer Dienste als vorteilhaft wahrgenommen werden. Zudem kann gerade die Künstlichkeit der *X-Bots*, ihre »Unheimlichkeit«, bisweilen zu ihrer Attraktivität beitragen, weil *CAIs*, den Spieltrieb, die Neugier, die Faszination für technischen Fortschritt etc. wecken, auch wenn sie vielleicht nicht gleich einem wirklichen Menschen vorgezogen werden (vgl. Jacobs 2023a: 56). Ich sehe in der steigenden Tendenz, *CAIs* für Reproduzenten von *adäquater* Anerkennung zu halten, das Problem, dass sich durch diese Tendenz bestimmte Verständnisse sozialer Anerkennung nachhaltig negativ verändern, – zwar nicht ausschließlich, aber eben *auch* in nicht-trivialer Weise negativ. Es geht mir also um diesen Wandel der Wahrnehmung dessen, was als adäquate soziale Anerkennungspraxis gelten kann, die m.E. ungeachtet aller Faszination für diese digitalen Lebensbegleiter in der Kritik sozialer KI thematisiert werden muss. Diese Kritik und ihre Schwierigkeiten behandle ich im vierten Abschnitt.

Was es so schwierig macht, die X-Bots von vorherein als »harmlose«, und u.U. als eher nützliche, denn als in irgendeiner Weise die menschlichen Anerkennungsverhältnisse erschütternde »Tools« zu beurteilen, wird deutlicher, wenn wir den *CAIs* als *relationale Artefakte* (Turkle et al. 2006) verstehen. Das Design dieser Maschinen ist darauf angelegt, dass Menschen eine Beziehung zu ihnen aufbauen sollen, d.h. wichtige Aspekte menschlicher Anerkennungsbeziehungen werden durch diese Maschinen simuliert.

Was gern vergessen wird zu erwähnen, ist, dass das *Beziehungsdesign* dieser digitalen Compagnons die pathogene Dynamik der Vereinsamung tendenziell verstärkt. Deshalb überlege ich in Abschnitt 4.1, welche alternativen KI-Szenarien es erlauben würden, Einsamkeit hinter sich zu lassen, was vornehmlich heißt, den paradoxen Effekt der *CAIs* zu vermeiden, der darin besteht, dass soziale Anerkennung

9 *Replika* ist ein generativer KI-Chatbot, der 2017 auf die Öffentlichkeit losgelassen und innerhalb eines Jahres von zwei Millionen Menschen genutzt wurde. Der Nutzer muss eine Reihe von Fragen beantworten, um ein Netzwerk zu erstellen, das als kontextueller Rahmen für die Entwicklung von »Freundschaften«, einschließlich romantischer und erotischer Beziehungen, dient, wobei die Funktion für erotische Gespräche von den Entwicklern im Jahr 2023 deaktiviert wurde.

10 Man denke hier an bestimmte Transgressionen, die im Allgemeinen moralisch nicht akzeptiert oder sogar verboten sind, weil sie Praktiken einschließen, die Menschen oder anderen fühlenden Wesen schaden. Die »Auslagerung« verpönter Praktiken in den Umgang mit Robotern mag eine Art Schadensbegrenzung darstellten, darf aber aus ethischer Sicht nicht darum schon als unproblematisch gelten.

zwar täuschend echt simuliert wird, aber Menschen, die an mangelnder sozialer Anerkennung in ihrer Einsamkeit leiden, trotzdem immer einsamer werden.

Einsamkeit wird in der digitalen Sphäre sozialer Anerkennungsbeziehungen mit *CAI*s reproduziert und nicht reduziert. Einsamkeit ist damit nicht nur als Einschränkung des individuellen Wohlergehens thematisierbar oder als potenzielle Bedrohung der individuellen Gesundheit verstehbar. Vielmehr kann sie als eine Ursache für eine Vielzahl sozialer Miseren in Betracht gezogen werden. Ich beschreibe sie daher in anerkennungstheoretischer Sicht als eine Form von Sozialpathologie und diskutiere sie im Abschnitt 5 als eine neue Form des Prekariats. Meine Schlussfolgerung ist, dass die sozialen Malaisen, die epidemische Einsamkeit mit sich bringt, das Ergebnis des Versuchs ist, ein nicht-triviales Leiden mit »künstlich-intelligenter Kognition« anstatt vorrangig mit humaner *Re*kognition zu lösen. Dieser Lösungsversuch dürfte dann als ethisch fehlgeleitet gelten, wenn die Praxis der Vermenschlichung von KI auf Kosten einer Entmenschlichung von Beziehungsformen geht.

2. Was ist (digitale) Einsamkeit?

> Wenn dich alles verlassen hat, kommt
> das Alleinsein.
> Wenn du alles verlassen hast, kommt die
> Einsamkeit.
> *(Alfred Polgar, 1873–1955)*

Mit Blick auf die unterschiedlichen Konzeptualisierungen von Einsamkeit hat Robert Weiss bemängelt, »dass sie nicht ausreichend auf den Status der Einsamkeit als reales Phänomen eingehen. […] Sie definieren sie über die Bedingungen, die sie theoretisch hervorrufen könnten. […] Tatsächlich sind dies nicht nur keine Beschreibungen, sondern auch keine Definitionen. Es sind Mini-Theorien. Indem sie die Identifizierung des Phänomens (»*das* ist Einsamkeit«) mit einer Erklärung für das Phänomen verbinden, schließen sie die entscheidende Forschungsfrage aus.« (Weiss 1987: 8; Übers. – KAJ). Um hier also nicht lediglich mit einer weiteren »Mini-Theorie« aufzuwarten, möchte ich meine Leseweise der Einsamkeit mit Blick auf verschiedene disziplinäre Perspektiven skizzieren, um verschiedene Konzeptualisierungen in mein Verständnis der Einsamkeit zu integrieren. Mit meiner anerkennungstheoretischen Verortung der Einsamkeit versuche ich das von Weiss bemängelte methodische Problem zu umgehen, indem ich eher eine heuristisch fruchtbare Rahmentheorie für Einsamkeitsforschung bereitstellen will und gerade nicht von einer prinzipiellen Abgeschlossenheit der Forschungsfrage ausgehe (vgl. Jacobs 2019b; Jacobs 2024).

2.1 Einsamkeit als Störung sozialer Anerkennungsverhältnisse

Zunächst einmal kann Einsamkeit als eine signifikante Veränderung von *bedeutungsvoller Bezogenheit* verstanden werden (vgl. Jacobs 2013: 2, 5ff.; Jacobs 2023a: 57; Jacobs 2024). Dies schließt positive Konnotationen des Begriffs zwar nicht grundsätzlich aus, allerdings sind es doch gerade die negativen Effekte von Vereinsamungsdynamiken, die vor dem Hintergrund des Bedürfnisses nach sozialer Anbindung und Anerkennung besonders hervorgehoben werden können. Zudem erachten wir Einsamkeit in der Regel als nicht sonderlich zuträglich für unser Wohlbefinden (Murphy/Kupshik 1992), was sich auch in unserem Verständnis psychosozialer Gesundheit niederschlägt, welches sich immer an Standards intersubjektiven Wohlergehens orientiert (Jacobs 2012: Kap.1, Kap.4; Jacobs 2017; Jacobs 2019; Jacobs 2020). Natürlich spielen individuellen Dispositionen und Resilienzfaktoren, z.B. bestimmte Persönlichkeitsmerkmale (vgl. Cacioppo et al. 2000; Buecker et al. 2020) und die persönliche Einstellung, eine große Rolle. Es kommt also darauf an, ob und wie die eigene Situation *als Einsamkeit* (positiv oder negativ) bewertet wird, aber allerspätestens dann, wenn Menschen in einer nicht-trivialen Weise an Vereinsamung leiden, besteht offensichtlich ein gewisser (sozialpolitischer, medizinischer, ethischer, etc.) Handlungsbedarf. Einsamkeit stellt damit eben nicht nur eine subjektiv erlebbare Grenzsituation dar, sondern ist als eine Sozialpathologie (Jacobs/Kettner 2017) zu begreifen. Die Exploration ihres soziopathogenen Potentials ist sicherlich auch eine empirische Fragestellung für globale Strategien der »Einsamkeitsprävention«. Erst in jüngster Zeit wurde dies vor dem Hintergrund aktueller epidemiologischer Daten erforscht. Es ist erwiesen, dass anhaltende Einsamkeitserfahrungen (Drageset 2004) neben anderen bekannten Faktoren wie schlechter Ernährung, Stress, Lärm oder niedrigem sozioökonomischen Status (Adler et al. 1994) erhebliche pathogene Auswirkungen auf die biopsychosoziale Gesamtsituation von Individuen hat (Orth-Gomer et al. 1993; Cacioppo et al. 2000; Wilson et al. 2007; Thurston/Kubzansky 2009), sodass eine wirksame Einsamkeitsprävention das Auftreten, die Manifestation und die Persistenz spezifischer Erkrankungen, z.B. von Depression (vgl. Vanhalst et al. 2012; Cacioppo et al. 2006) oder von *Kontaktmangelparanoia* (Janzarik 1973), in Gesellschaften signifikant verändern könnte. Darüber hinaus führen dysfunktionale Bewältigungsmechanismen der Einsamkeit zu noch schwerwiegenderen Folgen, wie z.B. zu Selbstmedikation durch Drogenmissbrauch oder zu Suizid(versuchen) (Meyer 2015; Nelson-Becker/Victor 2020).

Bemerkenswerterweise ist die bloße Quantität sozialer Beziehungen (eingeschlossen digitaler Interaktionen mit anderen) nicht das Entscheidende: Viele Menschen sind zwar einerseits optimal (sozial digital) vernetzt, fühlen sich aber dennoch sehr einsam, isoliert und sozial ausgeschlossen (Amichai-Hamburger/Ben-Artzi 2003; Jacobs/Uhle 2019). Einsamkeit ist ohne ihre soziale Einbettung gar nicht verstehbar, denn das individuelle Erleben von Einsamkeit kann seine

Bedeutung immer erst im Nachhall vergangener oder durch die Antizipation zukünftiger (Anerkennungs-)Beziehungen erhalten. Es ist zudem ein Gemeinplatz der theoretischen Einsamkeitsforschung, dass man allein sein kann, ohne sich einsam zu fühlen, und, umgekehrt, dass einem auch dann, oder vielmehr genau dann, wenn man in Gesellschaft anderer ist, die eigene Einsamkeit besonders stark bewusst wird (Zimmermann 2016[1784]; Maduschka 1933). Wenn es also die *Qualität der Beziehung* zu anderen ist, die sich verändert (Peplau/Perlman 1982) sollte der Zusammenhang zwischen »äußeren« Faktoren (wie etwa sozialer Ausgrenzung aus einer Gruppe) und den entsprechenden inneren Einstellungen dazu in einer Theorie der Einsamkeit berücksichtigt werden. Sozialphilosophische Theorien sozialer Anerkennung bieten sich dafür an, weil sie Typen von Anerkennungsbeziehungen systematisch beschreiben (Honneth 1994: 211), und eine sozialpsychologische Dimension der Einsamkeit (Brinkmann 1952) gut integrieren können. Das erlaubt die Beschreibung intersubjektiver, speziell interaffektiver Beziehungsdynamiken in den verschiedenen Sphären individueller, personaler und auch rechtlicher Beziehungsweisen: Wir konstituieren uns aktiv in Bezug auf unsere Umwelt und dies ermöglicht Integrität und persönliche Identität, d.h. als *relationale* Wesen stehen wir immer mit anderen Menschen und der Welt in einem Vermittlungsverhältnis. Diese Dynamik kennzeichnet Formen der *intersubjektiven* Bezogenheit (Fuchs/De Jaeger 2009; Schilbach et al. 2013), die alle möglichen Formen von *Interaktivität* einschließt und oftmals auch von einem »*Wir*-Gefühl« (Amodio/Frith 2006; Gallotti/Frith 2013) begleitet ist. In lebendigen, unmittelbaren – ganz maßgeblich auch *intuitionsbasierten* (Jacobs 2023b: 5) – Erfahrungen wird dieses »alltägliche Mitsein« (Heidegger 1967: § 27) in einer spezifischen Weise bedeutungsvoll für uns. Einsamkeit wird hier also eben nicht als ein »synthetisches Apriori« bzw. als ein monadischer Geisteszustand innerhalb der theoretischen Grenzen eines epistemischen Reduktionismus thematisiert (wie z.B. bei Mijuskovic 2012)), sondern gerade weil »unsere Mitmenschen der wichtigste Teil der Umwelt [sind]« (James 1884: 195), ist die Vorgängigkeit sozialer Anerkennung zu betonen (Honneth 2015: 46ff.; Honneth 2018: 17ff., 40ff.). Sie ist eine ontologische und konzeptuelle Vorbedingung dafür, überhaupt irgendeine »Mangelerfahrung« mit Blick auf bedeutungsvolles Beziehungserleben machen zu können. Dies wird verständlicher, wenn man sich die Grundlagen sozialer Kognition vergegenwärtigt: Menschliche Gehirne sind vermittelnde Organe (Papousek/Papousek 1992; Fuchs 2011; Colombetti 2014), die sich nur in Beziehung auf andere entsprechend »sozial« entwickeln, was ich an anderer Stelle bereits ausgeführt habe (Jacobs 2020; Jacobs 2022; Jacobs 2023b: 3, 59). Deswegen sind also *erstens* die (besonders interaffektiven) Erfahrungen vorausgegangener Interaktionen entscheidend für den persönlichen Erfahrungshorizont von Einsamkeit, und *zweitens* sind genau die Mini-Theorien der Einsamkeit, die die dem *cogito* vorgängigen Intersubjektivitätserfahrungen einfach ausblenden, nicht in der Lage, den vollen epistemischen Gehalt der Einsamkeitserfahrungen abzubilden. Ent-

sprechend berücksichtigen sie *drittens* auch nicht angemessen den Stellenwert von Einsamkeit aus entwicklungspsychologischer Sicht. Entwicklungspsychologisch werden Einsamkeitserfahrungen grundsätzlich als zum Person-Sein gehörige, sogar notwendige Lebenserfahrungen behandelt.[11] Einsamkeit kann nicht nur psychische Last, sondern zugleich auch ein tiefes menschliches Bedürfnis sein (Oppen 1967: 105), und begleitet Regenerierungs- und Individuierungsprozesse (Winnicott 1958). Im Kontrast hierzu ist sie als Störung sozialer Interaktionen in ihren Auswirkungen auf die seelische Entwicklung erforscht worden, weswegen Einsamkeit als potenzielle Ursache für psychische Erkrankungen und Traumata sehr ernst zu nehmen ist (Bowlby 1958; Winnicott 1964; Asher et al. 1984; Klein/Riviere 1992; Jerusalem et al. 2006). Ich möchte auch betonen, dass ich Einsamkeit nicht als *eigenständige Emotion* verstehe, obwohl sie gewiss eine starke affektive Dimension hat. Die »Gefühlskomponente« ist nur *ein* Aspekt des Einsamkeitserlebens. Wenn Einsamkeit Ausdruck einer aktiven Ausgrenzungserfahrung ist, geht mit ihr unter Umständen eben nicht nur ein gefühlsmäßiges Leiden, sondern eine Verletzung personaler Rechte einher (z.B. des Rechts, mit Respekt behandelt zu werden), d.h. sie tangiert das leidende Subjekt *als* Person. Einsamkeitserfahrungen sind immer ein Kompositum unterschiedlicher evaluativer Inhalte, d.h. im Einsamkeitserleben werden sowohl emotionale Episoden (z. B. Angst, Traurigkeit, Verlorenheit usw.) aber auch »einsamkeitstypische« Überzeugungen oder Wünsche miteinander in der konkreten Erfahrung verwoben.[12] Mit Blick auf die *Affektivität* können die emotionalen Reaktionen betont werden, die dann auftreten, wenn man überzeugt ist, unverstanden, sozial abgelehnt oder anderweitig in den Möglichkeiten emotionaler Intimität mit anderen eingeschränkt zu sein (Rook 1984: 1391). Dies geht normalerweise mit dem Wunsch einher, dieser Zustand möge sich ändern. Gemäß der neurobiologischen Forschung ist solches Leiden an Einsamkeit wortwörtlich zu nehmen, denn soziale Ausgrenzungserfahrungen gehen mit einer stärkeren Aktivierung des Anterioren Cingulären Cortex einher, der Region unseres Gehirns, in der auch körperliche Schmerzen verarbeitet werden (Price 2000; Eisenberger et al. 2003; Eisenberger/Liebermann 2004). Die sogenannte *Social Pain* Hypothese der Einsamkeit ist zudem auch anschlussfähig an evolutionsbiologische Theorien, die motivationale Effekte und adaptive Funktionen der Einsamkeit betonen, z.B. sich eben nicht noch weiter zurückzuziehen, sondern aktiv den Kontakt zu anderen zu suchen (Cacioppo/Patrick 2008: 202ff.).

11 Schon in der scholastischen Philosophie wurde formuliert, zum Menschsein gehöre auch eine radikale Einsamkeit, etwa von bei Duns Scotus (»ad personalitatem requiritur ultima solitudo«).

12 Das hat Erich Kästner (Kästner 1959) in seinem Gedicht *Apropos Einsamkeit* treffend zur Sprache gebracht.

Einsamkeitserleben resultiert aus einem Mangel an bzw. aus als mangelhaft erlebter sozialer Teilhabe, der sich negativ auf die Selbstbeziehung einer Person, d.h. auf ihr Selbstvertrauen, Selbstachtung, Selbstschätzung auswirken kann (Honneth 1994: 211). Sie ist mit gruppendynamischen Stigmatisierungsprozessen (Rotenberg/MacKie 1999) und oftmals auch mit Schamerfahrungen (Bohn 2008) verbunden. Offensichtlich leiden besonders vulnerable soziale Gruppen (z.B. Kranke, Alte, sozioökonomisch Benachteiligte, insbesondere Kinder: vgl. Asher/Parkhurst et al. 1990; Cassidy/Asher 1992; Lalayants/Price 2015) und junge Erwachsene (vgl. Thomas 2022) an Einsamkeit und werden sozial gemieden, *weil* sie einsam sind (Spitzer 2019: 71ff., bes. 86). Auffällig ist auch, dass Einsamkeit mit Prozessen mangelnder sozialer Anerkennung einhergeht, für die gerade eine »Invisibilisierung« des Leidens an Vereinsamung typisch ist. Das birgt auch eine nicht zu unterschätzende soziale Sprengkraft, etwa wenn Einsamkeit nicht nur ein stilles Leiden an mangelnder sozialer Teilhabe bleibt, sondern womöglich zur Quelle von kollektivem sozialem Neid, von Ressentiments und von Radikalisierungstendenzen wird, die sich unter Umständen gegen jene gesellschaftlichen Gruppen richten, die (vermeintlich) sozial stärker anerkannt sind bzw. denen deswegen mehr (soziales, ökonomisches, symbolisches) Kapital zu Verfügung steht, dessen Ermangelung sich die Einsamen schmerzlich bewusst werden. Die stetige Vergleichsmöglichkeit – das Leben der Anderen – ist meistens (in aller Beschönigung) voll digital durch *Social Media* zugänglich und scheint das Problem des im wahrsten Sinne des Wortes »*Sozial*-Neids« unter Umständen kausal mitzuverursachen.

2.2 Die Pathodynamik der Einsamkeit

Einsamkeit ist zu einem wichtigen Thema globaler Gesundheitspolitik und staatlichen Handelns geworden (WHO 2000; Mann et al. 2022), d.h. es sind in jüngster Zeit spezielle Ministerien[13] zur Entwicklung interdisziplinärer Interventionsstrategien[14] für die Reduktion sozialer Isolation (z.B. Findlay 2003) gegründet worden und

13 Seit Juni 2022 erarbeitet das Bundesministerium für Familie, Senioren, Frauen und Jugend eine Strategie gegen Einsamkeit. Im selben Jahr startete auch das Kompetenznetz Einsamkeit (KNE), welches das Ziel hat Informationen zu konkreten Angeboten für die Einsamkeitsprävention bereitzustellen: [https://www.bmfsfj.de/bmfsfj/themen/engagement-und-gesellschaft/strategie-gegen-einsamkeit-201642] (Zugriff: 21.11.2023).

14 Die Abteilung für demografischen Wandel und gesundes Altern der WHO hat die UN-Dekade *Healthy Ageing* (2021–2030) ausgerufen und soziale Isolation und Einsamkeit zu dringlichen Themen der Gesundheitsförderung erklärt, insbesondere für die ältere Bevölkerung, auch unter dem Vorzeichen digitaler Interventionen (wie Kompetenztraining, Community- und Selbsthilfegruppen sowie kognitive Verhaltenstherapie), die entwickelt werden, (1) um die soziale Isolation älterer Menschen zu verringern, (2) um den Zugang zu Informations- und Kommunikationstechnologien zu verbessern, und (3) um eine altersfreundlichere Gemein-

Einsamkeitsprävention ist ein Top-Ziel von Public Health »*built environment*«-Projekten (Srinivasan et al. 2003), weil es erwiesen ist, dass sie sich rasant auch in größeren Populationen ausbreitet (Cacioppo et al. 2009). Zu ihrer Sozialdynamik gehört, dass Einzelgänger oftmals keine Unterstützung erhalten, weil Einsamkeit – so der Neurowissenschaftler Manfred Spitzer (Spitzer 2019: 71ff.) – sozial (emotional) »ansteckend« ist. Daher verwundert es doch, dass all diese Erkenntnisse über die der Einsamkeit inhärente Pathodynamik der Selbstverstärkung bislang nicht dazu geführt haben, die kausale Rolle, die Einsamkeit in sozialethisch problematischen Prozessen wie der (kollektiven) Missachtung soziale Anerkennungspraxis haben könnten, ebenfalls in den Katalog der Miseren, die Einsamkeit bewirkt, mitaufzunehmen.[15]

Einsamkeit macht tatsächlich krank (Jacobs 2020): Wo durch Einsamkeit solche Lebensprobleme entstehen, die nicht-trivial leidvoll sind oder ein ersichtliches pathogenes Potenzial auch im medizinisch-klinischen Sinne mit sich bringen, haben wir guten Grund, die jeweiligen Theorien des Wohlergehens oder das geltende Verständnis von psychosozialer Gesundheit, oder beides, nicht etwa nur zur Bewertung und Einschätzung der leidvollen Beeinträchtigung heranzuziehen, sondern diese Theorien und Verständnisse ihrerseits ggf. zu kritisieren (Jacobs 2018). Denkbar wäre beispielsweise Kritik einer besonders ausgeprägten Arbeitskultur mit negativen Ausprägungen für die psychosoziale Gesundheit, und eine entsprechende Einschätzung des pathogenen Potenzials genau dieser gesellschaftlich akzeptierten Ideale, Normen und Werte für die Vereinsamung von Personen in bestimmten Gesellschaften und Organisationen.[16] Ein zeitgenössisches Beispiel für

---

schaft zu schaffen, was mit einem interventionistischen Anspruch einhergeht, der seit langem überhört worden zu sein scheint.

15 Weil sich die Pathodynamik der Einsamkeit tendenziell als ein chronischer Zustand, der »immer schlimmer« wird, manifestiert, kann man verschiedene Stadien der Vereinsamung unterscheiden. Im Endstadium der Einsamkeit (die ich als existenzielle *Verlassenheit* oder *Verlorenheit* beschrieben habe; Jacobs 2019b) sind Menschen dann auch gar nicht mehr dazu in der Lage, Erfahrungen von Nähe, affektiver Resonanz, existenzieller Sicherheit und Nähe zu anderen so erleben zu können, dass sie sich prosozial verhalten. Dieses Verlassensein z.B. aufgrund mangelnder sozialer Kontakte zeigt sich darin, dass Menschen nicht mehr in der Lage sind ein Gespräch zu führen, oder darin, dass sie, weil die negative Affektlage überwiegt, anderen Menschen das »soziale Glück« nicht mehr gönnen können und darüber so verbittern, dass sie von anderen Menschen als toxisch wahrgenommen werden (die dann noch stärker auf Distanz zu ihnen gehen). Unter Umständen gehen einsame Menschen sogar dazu über, andere Personen zu stalken, d.h. dezidiert antisoziale Verhaltensweisen zu zeigen, die strafrechtlich relevant werden können und ihre Ursache in einer tiefen Einsamkeit – einer Verlorenheit – haben.

16 Die funktionale Rolle der Einsamkeit für die Entstehung und Manifestation bestimmter Krankheiten (Cacioppo et al. 2006; Vanhalst et al. 2012) ist längst empirisch erwiesen. Sie erhöht bspw. die Vulnerabilität und das Risiko für eine körperliche und/oder psychische Er-

einen solchen chronifizierten Zustand von Einsamkeit ist der Fall der sogenannten *Hikikomoris* (jap. ひきこもり) in Japan. Bei manchen Menschen führt der soziale Erfolgsdruck dazu, dass sie sich völlig aus der Gesellschaft zurückziehen. Bestimmte Ideale, die Erfolgsziele vorgeben, die man »erreichen muss« (man denke z.B. an den »Perfektionismus« (*kodawari; jap.* こだわり) oder die kontinuierliche Verbesserung (*kaizen*; jap. 改善), die in Japan einen hohen Stellenwert haben), werden von einer wachsenden Zahl von (jüngeren) Menschen als etwas wahrgenommen, das man nicht erreichen kann. Dies führt dann eben nicht nur zur Tendenz, viel zu viel zu arbeiten, um zur Leistungsspitze zu gehören, (was unter anderem zu Fällen von *karō shi* (jap. 過労死), d.h. Tod durch Überarbeitung führen kann) oder besonders viel zu konsumieren, den sozialen Status z.B. durch den Besitz von Luxusgütern zumindest nach außen zu demonstrieren, sondern kann eben auch den sozialen Rückzug bedingen und in die Vereinsamung führen. Der japanische Psychologe Tamaki Saitō (Saitō 2013), der den Begriff der *Hikikomori* geprägt hat, spricht von einer »abnormen Vermeidung sozialer Kontakte«, was auch die Interpretation des Begriffs in der Übersetzung als »*Eingesperrt-Sein*« gut trifft.[17] Bei alleinstehenden *Hikikomori* endet dieser Zustand nicht selten im Phänomen des *kodoku-shi* (jap. 孤独死), d.h., führt zum »solitären« Tod (Takahiro et al. 2017).[18]

---

krankung, was auf die grundsätzlich *bidirektionale Beziehung* von Einsamkeit und (Psycho)Pathologie hinweist (z.B. Depression, Suchtverhalten und Paranoia bzw. *Kontaktmangelparanoid* (Janzarik 1973)).

17 Während Studien vor dem Jahr 2000 bestätigten, dass vor allem junge Menschen betroffen sind, war in jüngster Zeit eine deutliche Zunahme von *Hikikomori* bei Menschen mittleren Alters und älteren Menschen zu verzeichnen, wie eine Studie der japanischen Regierung aus dem Jahr 2000 zeigt, die von 500,000 Fällen von *Hikikomori* zwischen 16 und 39 Jahren ausgeht. Symptomatisch für diese selbstgewählte soziale Isolation ist, dass Personen mindestens sechs Monate lang weigern, das Haus zu verlassen und von Lieferdiensten oder den Angehörigen mit dem Nötigsten versorgt werden. Entscheidend ist, dass es zwar digitalen Kontakt zur Außenwelt gibt, aber so gut wie keine physischen Kontakte oder gelungene Kommunikation mit den »Versorgern«. Man steht zwar über soziale Medien in Verbindung mit anderen, allerdings wird die eigene Lebenssituation als so unbefriedigend und deprimierend empfunden (Cacioppo/Hawkley 2005), dass diese Form mit hoher sozialer Scham einhergeht, und es zu Stigmatisierungen auch der Verwandten dieser Personen kommen kann, denn viele *Hikikomoris* leben im Haushalt der Eltern, die sie mitversorgen, was dadurch verstärkt wird, dass Kinder in Japan generell länger bei ihren Eltern leben, z.B. auch noch während des Studiums, was sicherlich nicht nur einer bestimmten Traditionsverbundenheit in Bezug auf familiäre Bindungen geschuldet ist, sondern auch ökonomische Gründe hat.

18 Typischerweise bleibt der Leichnam sehr lange unentdeckt in der Wohnung und wird meistens nur deswegen gefunden, weil monatliche Rechnungen nicht bezahlt werden (Meyers 2015; Nelson-Becker/Victor 2020). Man geht von 30.000 einsamen Todesfällen pro Jahr aus, aber Unternehmen, die Wohnungen reinigen, wenn Fälle von *kodoku-shi* entdeckt werden, berichten, dass die Zahl zwei- oder dreimal so hoch sein könnte. Im Übrigen ist es einzigartig für Japan, dass dieses Phänomen künstlerisch reflektiert wird, z.B. von Miyu Kojima, die Mi-

Für eine Darstellung von Einsamkeit aus anerkennungstheoretischer Sicht sind also gerade auch die Dynamiken der sozialen Selbstisolation, der gefühlten Ausgrenzung und Ablehnung durch andere, wie auch das Gefühl »nicht mit anderen mithalten zu können« (was evtl. auch durch die omnipräsente Vergleichsmöglichkeit mit anderen Personen in den sozialen Medien getriggert wird) als Formen von Störungen der sozialen Anerkennung zu berücksichtigen.

Diese Erfahrungen sind wichtige Bezugskategorien für die ätiologische Erklärung der affektiven Dimension von Einsamkeit als eines *sozialen Schmerzes*. Wir *leiden* an Einsamkeit, weil wir wahrnehmen, dass uns etwas Wesentliches fehlt, nämlich die Verbundenheit mit anderen. Das setzt in der Regel auch voraus, dass wir normalerweise mit einem gewissen Maß bzw. mit konkreten Ausdrucksformen wechselseitiger sozialer Anerkennung rechnen (dürfen), etwa mit Empathie (einem Mindestmaß an Einfühlungsvermögen), mit Respekt (z.B. rechtskonformem Verhalten) und mit Solidarität (im Einklang mit Würde behandelt zu werden). Soziale *Anerkennung* ist, wie gesagt, der Kognition vorgängig (Honneth 2015: 46ff.; Honneth 2018: 17ff., 40ff.), d.h. die meisten unserer interaktiven Modi sind durch eine Form der interessierten Teilnahme oder sogar der dezidierten Anteilnahme (vor)strukturiert, weswegen Einsamkeit durch Formen *fehlender oder falscher sozialer Anerkennung* konzeptualisiert werden kann: Axel Honneth betont die »Anerkennungsvergessenheit« besonders im Zuge einer Theorie der Verdinglichung (Honneth 2015: 69) und spezifiziert damit Beziehungsformen, in denen »das Faktum vorgängiger Anerkennung verloren geht« (Honneth 2015: 70). Auf die Einsamkeit gemünzt, wird die sozialontologische Priorität des Anderen also »vergessen«, z.B., weil man so konzentriert auf eigene Motive und Zwecke ist, dass der Andere bisweilen gar nicht oder nicht »richtig« wahrgenommen wird. Anerkennungsstörungen wie z.B. aufgrund von Leugnung oder Abwehr basieren auf einer Form selektiver Interpretation sozialer Tatsachen – z.B., dass der Andere als »anders« (als nicht zugehörig) wahrgenommen wird und ihm daher sogar aktiv die Anerkennung verweigert wird. Die Verweigerung kann ein Ausmaß erreichen, dass Andere nur noch als Objekt der Verachtung betrachtet werden, vielleicht sogar nur noch »*mere enemies*« sind, wie Martin Sticker (Sticker 2023) es nennt, während Fälle falscher Anerkennung eher dadurch gekennzeichnet sind, dass jemand in einer Beziehung nicht um seiner selbst willen geschätzt wird, sondern lediglich als Mittel zum Zweck der Erreichung eigener Zielvorstellungen (»*mere mean to an end*«). Solche Instrumentalisierung muss nicht immer, kann aber moralisch problematisch werden, vor allem dann, wenn die Reziprozität und Angemessenheit mit der (vermeintlich) Anerkennung gezeigt wird, aus dem Gleichgewicht gerät (z.B. im Liebeswahn). Dies lässt sich nochmals unterscheiden von eben jener absichtlichen Zurückhaltung von Anerkennung, um

niaturszenen der Appartements in Tokyo nachbildet, die ihre Reinigungsfirma säubert nach einem endeckten Fall von *kodoku-shi*.

andere Menschen gezielt zu verletzten (so z.B. in Fällen von Mobbing) oder ganz bewusst so zu isolieren und auszugrenzen, dass sie nicht nur vereinsamen, sondern quasi den »sozialen Tod« sterben, weil ihnen durch die Isolation nicht nur die Ressource der sozialen Anerkennung selbst, sondern auch alle damit eng verbunden Formen sozialen Kapitals sowie die Möglichkeit, ein tragendes soziales Netzwerk aufzubauen, um neues soziales Kapital zu erwirtschaften, entzogen wird.

In diesem Kontext ist soziale Unsichtbarkeit (Honneth 2003: 10ff.) eine wichtige Bezugsgröße, um die sozialen Produktionsbedingungen von Einsamkeit als eine in den Dynamiken sozialer Desintegration und Marginalisierung fußenden Sozialpathologie zu begreifen. Zur Vereinsamung als eines Unsichbarwerdens mag es auch gehören, dass die sozialen Miseren, die mit ihr verbunden sind, in vielen Gesellschaften überhaupt nicht so thematisiert werden (können), dass eine signifikante Verbesserung der Lage dadurch eintreten könnte, dass man darüber »öffentlich« diskutiert. Zur Kritik am kollektiven Abwehrmechanismus *Invisibilisierung* sozialer Problematiken gehört nämlich auch, die Durchschlagskraft bestimmter Gegen-Narrative nicht zu unterschätzen, z.B. von Narrativen, die suggerieren, es handele sich bei der steigenden Einsamkeitsrate und den mit ihr korrelierenden Miseren lediglich um ein »gegebenes« Phänomen (das z.B. mit der Gentrifizierung von Gesellschaften einhergeht). Ein anderes narratives Stereotyp, mit dem z.B. in der Integrationsdebatte die angebliche Integrationsunwilligkeit bestimmter gesellschaftlicher Gruppen behauptet wird, besagt, die betreffenden Personen oder Kollektive seien an ihren Miseren selber schuld. Nehmen wir an, dass durchaus von der offiziellen Politik oder von zivilgesellschaftlichen Initiativen wahrgenommen wird, dass die Folgen und Quellen von Einsamkeit eigentlich eingedämmt werden müssten, ist es doch überraschend und daher erklärungsbedürftig, dass bestimmte ätiologische Erklärungsansätze nicht weiter berücksichtigt werden. Die Selektivität könnte ihre Erklärung darin finden, dass andernfalls genau jene gesellschaftlichen Dynamiken und Stereotype, die die soziale Exklusion begünstigen, kritisch beleuchtet werden müssten, z.B. Narzissmus als transpersonales Organisationsprinzip in bestimmten Arbeitsumwelten, überkommene Geschlechterrollen, Elitarismus, Fremdenhass u.a.m.

3. Einsamkeitsmanagement mittels *Companion AIs* in der Kritik

Rahel Jaeggi (Jaeggi 2009a; Jaeggi 2009b) argumentiert überzeugend, dass ein Anzeichen guter institutioneller (d.h. auch gesamtgesellschaftlicher) Praxis der Umfang ist, in dem fragwürdige Umstände kritisierbar bleiben können und dürfen. Auf Einsamkeit als Sozialpathologie angewandt, würde dies bedeuten, ihre *Gemachtheit* zu thematisieren statt sie als eine *Gegebenheit* hinzunehmen, und sie z.B. als Folge und Ausdruck einer digitalen (Re-)Produktionsdynamik von Anerkennungsstö-

rungen zu reflektieren. Gesellschaften in denen immer mehr, und zunehmend auch jüngere Menschen vereinsamen, müssten kritisierbar bleiben unter dem Verdacht, dass man das Prekariat der Einsamkeit systematisch verschleiert. Gesellschaften, aber auch Unternehmen und andere konkrete Organisationen, in denen das Gefühl der Vereinsamung quasi der voreingestellte Zustand ist, haben offenbar Schwierigkeiten damit, die Einsamkeitsmisere mit genau solchen Strategien in Verbindung zu bringen, die das neue Prekariat »ignorierbar« machen. Der Fall der *Hikkikomori* scheint mir prototypisch für eine schamhafte Vermeidung der Kritik fehlgeleiteter Anerkennungsdynamiken zu sein, die eben nicht durch Betonung anderweitiger Exzellenz- oder Effizienzstrategien in einem Kulturkreis einfach relativiert werden können.

Es ergibt sich also ein sozialethisches Legitimationsproblem. Der praktischen Frage, wie denn erfolgreiche sozialtechnische Interventionen aussehen könnten, ist die normativ schwierige Frage vorgelagert, *aus welchen Gründen* man überhaupt die Verursachung von Einsamkeit durch mangelnde und mangelhafte soziale Anerkennung kritisieren kann. Was, wenn überhaupt etwas, kann z.B. an Versuchen, drohende Vereinsamung mit »digitalen Begleitern« aufzuhalten, problematisch sein, und sogar dann, wenn die von Einsamkeit Betroffenen selber dies auf Anhieb vielleicht gar nicht problematisch finden und vielleicht die *CAI*s sogar als adäquate Reproduktionsmittel sozialer Anerkennung (für sich selbst) bewerten?

Im Licht von Rahel Jaeggis Unterscheidung von externen, internen und immanenten Formen der Kritik an Lebensformen erscheint *immanente* Kritik als diejenige Form von »kritischem Verhalten« (Jaeggi 2014: 258), das am tiefsten in die spezifischen Formen der Normativität eindringen kann, die bestimmten Lebensformen innewohnen. Wir könnten auf dem Wege immanenter Kritik versuchen, Personen, die mittels *CAI*s ihre Einsamkeit in den Griff zu bekommen möchten, davon zu überzeugen, dass ihr Interesse auf einer Fehleinschätzung der Pathodynamik von Einsamkeit beruht, eventuell sogar auf einem fehlgeleiteten Verständnis »digitaler Autonomie«, und dass, sollte sich so etwas wie Einsamkeitsmanagement mittels *CAI*s breit durchsetzen, womöglich das gängige Verständnis adäquater Anerkennung unterminiert und korrumpiert werden könnte. Skizzieren wir diese Kritik:

(1) Wir gehen von Werten und Normen aus, die konstitutiv für soziale Praktiken, d.h. nicht nur irgendwie gegeben, sondern auch begründete sind (Jaeggi 2009b: 266ff., bes. 286f.). Soziale Anerkennung ist solch ein Wert, schon weil gewisse Praxismodi sozialer Anerkennung Bedingungen der Möglichkeit des Einhaltens wichtiger sozialer Normen sind.

(2) Immanente Kritik operiert demaskierend: Wir machen eine defizitäre Verwirklichung von Normen in einer gegebenen Situation (z.B. unter Krisenbedingungen einer drohenden Einsamkeitsepidemie) sichtbar und betonen die Widersprüchlichkeiten, die durch die Wirksamkeit der Norm in der Lebensrealität erzeugt werden (Jaeggi 2009b: 286f.). Der größte Widerspruch besteht darin, dass Menschen zu

einer Maschine ein Verhältnis aufbauen und von künstlicher Intelligenz Zuspruch erhalten, ohne dass sich an ihrer prekären Lebenssituation (der faktischen Isolation, der monologischen Existenz, der Tatsache, dass da zwar »some-*thing*« ist, aber eben keine reale Person) etwas ändert.

Immanenten Kritik ist ein probates Mittel der Kritik des digitalen Einsamkeitsmanagements mit *X-Bot*s, weil sie sich an der *inneren Widersprüchlichkeit der Realität* orientiert (Jaeggi 2009b: 287), wobei es »im Charakter der Normen und in der Beschaffenheit der jeweiligen Praktiken und Institutionen liegende Gründe dafür [gibt], dass diese sich nicht widerspruchsfrei verwirklichen lassen« (ebd.). Das zeigt sich zum Beispiel darin, dass humanoide KI systemisch so implementiert bzw. institutionalisiert wird, dass viele Menschen kaum noch eine andere Wahl haben als sich mit ihr abzufinden (z.B. in Pflegeeinrichtungen), oder auch darin, dass immer mehr Roboter produziert werden, ohne dass ersichtlich wäre, dass es den Vereinsamten selbst dienlich ist (und nicht etwa nur denen taugt, die zu erbringende Leistungen an Roboter »outsourcen«, was zweifellos mit tatsächlichen Entlastungen z.B. für Menschen in Pflegeberufen einhergehen kann, letztlich aber nicht einfach als Nutzenertrag »verrechenbar« ist mit den potentiellen Verlusten an *menschlicher* Zuwendung, den andere Gruppen wie z.B. Patienten dadurch eventuell in Kauf nehmen müssen).

(3) Immanente Kritik zeichnet sich dadurch aus, dass sie auf eine *Transformation des Bestehenden* abzielt: eine widersprüchliche Situation (wir sind voll digital vernetzt und trotzdem einsam) soll in einen neuen Zustand überführt werden (den, in dem Menschen sich weniger einsam fühlen), und dafür ist entscheidend, dass dieser neue Zustand die überwundene Ausgangssituation in irgendeiner Form auch abbilden sollte, weil eine

> »nötig werdende Transformation[…] beides [beinhaltet]: Die defiziente Realität und die Normen selbst. Die Normen bleiben nicht unberührt von dem Umstand, dass sie in einer gegebenen Situation nicht realisiert worden sind. […] Immanente Kritik kritisiert also […] nicht nur eine defizitäre Realität anhand des Maßstabs der Norm, sondern verfährt auch umgekehrt.« (Jaeggi 2009b: 288)

Normativ-reflexiv sind wir mit unserer Kritik also dort, wo wir die schwierigen Verhältnisse der Einsamkeit so ausdeuten, dass wir die destabilisierenden Faktoren der Lebenswelt ernstnehmen. Wir wenden Kritik auf diese Weise auch zum Verbessern unserer eigenen Lebensform an. Das bedeute auch, dass wir nicht einfach bei Kapitalismuskritik am umsatzschweren Geschäft mit *CAI*s stehen bleiben, oder dass wir uns darüber mokieren, dass der Einsatz von *X-Bots* zwar in manchen Bereichen Entlastung bringen mag (z.B. größere Zeiteffizienz in der Pflege, Einsparung von Servicepersonal in Unternehmen), aber unter Inkaufnahme nachteiliger Folgen für vulnerable Personen. Wir müssen es auch nicht dabei belas-

sen, bestimmte Modi der Reproduktion von sozialer Anerkennung als inadäquate Verwirklichung von Werten oder Konventionen sozialer Anerkennungspraktiken zu verurteilen (z.B. die Suggestion, dass man sich mit *CAI*s Liebe, Geborgenheit und Verständnis einfach kaufen könne). Einsamkeit immanent kritisch als neues »Unbehagen an der Kultur« bzw. an unserer Lebensform zu begreifen, bedeutet, sie mit Blick auf das zu analysieren, was längst an bestehenden (kollektiven) Traumata, falschen Anerkennungsverhältnissen, Entfremdungserleben, spezifischen Immunisierungsstrategien einzelner Lebensformen präsent ist und durch die Einsamkeit selbst katalysiert wird. Im besten Falle entlarvt immanente Kritik, wie der Anschein von Legitimität von *CAI*s als »artificial *agents*« dadurch erzeugt wird, dass sie autorisiert und behandelt werden, als wären sie menschlich, und dass sie auf diesem Wege dann irgendwie doch als adäquate (Re-)Produzenten sozialer Anerkennung wahrgenommen und »benutzt« werden. Wo man also ausgerechnet in *CAI*s die Befreiung von Einsamkeit erkennen will, tritt das pathogene Potenzial einer digitalen Normalität voll zutage: einer im wahrsten Sinne des Wortes »kritischen« Situierung des atomistisierten Subjekts.

Halten wir also fest: Die sozialpathologische Dimension der Einsamkeit wird nicht nur in der (inter-)subjektiven Phänomenalität eines Leidens, sondern in der Implementierung spezifischer Leseweisen bestimmter Ideale oder Normen ansichtig, die auch hinsichtlich der Modi ihrer »angemessenen« Umsetzung kritisiert werden können. Wo genau über diese möglichen »Fehlverständnisse« (Honneth 2015: 230f.) nicht weiter debattiert wird, besteht die Gefahr einer simplen Reproduktion falscher Anerkennung unter digitalen Voraussetzungen. Wie schon erwähnt, betrifft das schon die Idee der sozialen Anerkennung selbst, d.h. unsere eingespielten und reflexiv einholbaren Vorstellungen davon, wie sie adäquat zum Ausdruck gebracht werden *sollte*. Wenn KI als »sozial« und als »Companion« propagiert und akzeptiert wird, bekommen *CAI*s einen Status in den in der Lebenswelt verteilten sozialen Anerkennungsverhältnissen zugeschrieben, den wir bisher eigentlich nicht für Maschinen vorgesehen haben. Was macht diese *CAI*s also so besonders und wie verändern CAIs womöglich unsere etablierten Angemessenheitsurteile über Anerkennung?

4. *Companion AIs* als Reproduzenten gestörter Anerkennungsverhältnisse

Gewiss gibt es auch einige positive und affirmative Lesarten von Einsamkeit, etwa ihre Lobpreisung als »Solitüde« (Tillich 1963: 17; Satorius 2006; Poschard 2007). Mir geht es im vorliegenden Beitrag betont um ihre Schattenseiten, insbesondere um problematische Veränderungen von Bezogenheit. Solche Veränderungen persönlicher Beziehungen im Zeitalter der Digitalisierung sind im Hinblick auf Ent-

fremdungsphänomene bereits skizziert worden, z.B. als Prozesse der Kommodifizierung (Illouz 2016: 73), als Folgen von Beschleunigung (Rosa 2019: 125ff., 141ff.), als soziale Atomisierung (Kucklick 2017: 105ff.; Taylor 1979). *CAI*s sind im wahrsten Sinne des Wortes *Kulturmaschinen*, d.h. Teil der digitalkulturellen Transformation, von denen man behaupten darf, dass sie die Singularisierung der Menschen (Reckwitz 2018: 225ff.) stark befördern. Der künstlich intelligente Gefährte ist der technische Andere, aber, anders als eine Kaffeemaschine oder ein Staubsauger, bisweilen *unheimlich* (Freud 1947[1919]: 232ff.) »menschelnd«. Wir reagieren und behandeln *X-Bot*s tatsächlich oftmals so, als wären sie menschlich, was natürlich durch das spezifische Design und ihre »gesellige« Funktionalität entscheidend mitbeeinflusst wird. Dass wir ein affektives Verhältnis zu diesen Maschinen aufbauen, wurde z.B. für die Chatbot-Nutzung empirisch nachgewiesen (Skjuve et al. 2021; Gao et al. 2018; Purington et al. 2017; Archer 2021), d.h. selbst dieser speziellen Form rein auditiv basierter KI werden solche reaktiven Einstellungen entgegengebracht, die man normalerweise nur gegenüber natürlichen Personen an den Tag legen würde. Es macht eigentlich gar keinen Sinn, auf Alexa wütend zu reagieren, sich bei *Hatsune Miku* zu entschuldigen, weil sie »so lange gewartet hat, bis wir nach Hause gekommen sind«, sich bei Pepper für die Serviceleistung zu bedanken, etc. aber wir tun es trotzdem. Es sind im Wesentlichen wir selbst, die »User«, die der Beziehung zu einem *CAI* eine besondere Bedeutung verleihen, weswegen es auch nahe zu liegen scheint, unser Einsamkeitserleben mit ihnen lindern zu wollen. Angesichts der verschiedenen funktionalen Rollen, die ein *CAI*-System im Leben eines Menschen spielen kann, möchte man meinen, das sei schon ausreichend um *X-Bot*s als legitime »(re-)produktive Quellen« sozialer Anerkennung ausweisen zu können. Genau dies stelle ich in Frage und erinnere daran, wie sehr sich Beziehungen von Menschen (immer noch) *grundlegend* von den diversen Beziehungen zu *CAI*s unterscheiden: Egal wie elaboriert das spezifische Funktionsdesign und die Verkörperungsformen der *CAI*s auch sein mögen, diesen Geräten fehlt im Grunde alles, was für soziale Anerkennungskompetenz, wie ich sie hier angesetzt habe, notwendig ist. Mindestens die folgenden fünf grundlegenden Unterschiede müssen betont werden (vgl. Jacobs 2023b):

Erster Unterschied: Während menschliches Beziehungserleben durch eine intersubjektive Dynamik wechselseitiger sozialer (Re-)Kognition geprägt ist, mangelt es *X-Bots* wie den *CAI*s an Inter*subjektivität*. Gewiss sind *CAI*s in der Regel zu vielfältigen Aktivitäten und so auch zu Interaktionen fähig. Und vielleicht ist ihnen sogar (technische) Selbstbezüglichkeit zuzuschreiben, insofern diese Maschinen auch Eigenzustände »registrieren« oder »protokollieren«, z.B. wann ein Software-Update notwendig wird oder ein Hardware-Schaden vorliegt. Es gibt hier allerdings kein Subjekt, das zu lebendiger Interaktion jenseits bloßer Aktion fähig wäre. Das ist übrigens einer der Gründe, weswegen Robotern eben (noch) kein Bewusstsein zu-

zuschreiben ist (so schon Dreyfus 1985), das die Voraussetzung wäre für einen intentionalen *Selbst*bezug (Böhme 2008).

An dieser Stelle könnte man eigentlich aufhören mit der Auflistung wesentlicher Mensch-Maschine-Unterschiede, denn die Abwesenheit von Intersubjektivität ist schon ein Ausschlusskriterium dafür, *CAIs* für gleichwertig befähigt zu menschlicher sozialer (Re-)Kognitionspraxis zu halten. *Nota bene*: Keineswegs ist damit ausgeschlossen, dass sie zu sozialen Anerkennungsbeziehungen irgendwie *beitragen können*, z.B. indem sie Menschen miteinander in Kontakt bringen können, quasi wie Smartphones.

Freilich sprechen einige AutorInnen auch vom »Bewusstsein« von KI-Systemen in mehr als nur metaphorischen Weise und behaupten starke Analogien zwischen menschlichem Bewusstsein und z.B. den *algorithmischen* Aktivitäten »sozialer KI« [19](Kurzweil 1993; Possati 2020 in Bezug auf Lacan 1978: 239). Possati z.B. geht von einem maschinellen Unbewussten aus und überträgt Funktionsweisen des psychischen Apparats auf KI-Systeme. Ich habe an anderer Stelle (Jacobs 2023a: 55) auf Nicolas Langlitz' Analyse »des Unbewussten als symbolische Maschine« (Langlitz 2005: 158ff.) verwiesen und finde sein Argument überzeugend, dass dieser Vergleich zwischen Mensch und Maschine, der z.B. die algorithmische Informationsverarbeitung in Analogie zum menschlichen Wiederholungszwang setzt, letztlich fehlgeleitet bleiben muss, weil eben dies auf ein »Jenseitiges« des Sinns verweist, der aber genau jener Sinn ist, um den es in der Psychoanalyse eigentlich geht (Langlitz 2005: 193ff.). Anders ausgedrückt: Die initiierten »Aktionen«, die auf einer binären Code-Reproduktion gründen, sind eben nicht ansatzweise in dem Maße »sinnhaft«, wie all das, was durch den Wiederholungszwang als menschliches Triebschicksal deutbar werden kann. Zudem können Menschen genau dieses Schicksal in der Durcharbeitung aufdecken. Das setzt eine Art der Freiheit in und durch symbolische »Verarbeitung« (als Deutung) voraus, die Maschinen nicht gegeben ist.

Zweiter Unterschied: Menschliche Beziehungen basieren auf inter-affektiver Praxis von empfindungsfähigen Wesen, die von Natur aus dialogisch ist und alle möglichen Formen des (z.B. symbolischen, physischen usw.) Austauschs vorsieht, die grundverschieden von sogenannter *emotiver* KI-Responsivität ist. Zwar können emotionale Ausdrücke maschinell erkannt werden und es gibt *X-Bots,* die Affektivität simulieren und auch definitiv beim menschlichen Gegenüber auslösen, aber mit »Gefühlsfähigkeit« hat das nichts zu tun. Selbst der affektkälteste Mensch hat noch ein Schmerzempfinden oder kann so etwas wie die Emotion der Emotionslosigkeit verspüren. Roboter sind zu Empfindungen nicht in der Lage. So beeindruckend also die jüngsten Entwicklungen des sogenannten *Affective Computing* (Picard 1997; Daily et al. 2017; Cambria et al. 2019), der *Sentimentanalyse* (Siegel/Melpomeni 2020) und

19 Zum Forschungsfeld sogenannter sozialer KI siehe: [https://socialai.nl/] (Zugriff: 28.05.2024).

der *Responsivität* (Asada 2015; McStay 2018) auch sein mögen, das, worauf es bei Anerkennung ankommt, ist doch grundsätzlich auf die Annahme gegründet, dass der responsive Andere auf geeignete Weise eine Fähigkeit zur *Sorge* (im Sinne emotionaler Anteilnahme und empathischer Einfühlung) aufweist. *X-Bots* sind (noch) nicht befähigt, auch wenn *CAI*s in einem technischen Sinne einen »emotiven Gehalt« transportieren oder tatsächlich »Versorgung« leisten, wie etwa in KI-gestützter Pflegeassistenz (Manzeschke 2022) und daher in bestimmten Anwendungsbereichen von instrumentellem Wert sind (Beck 2018: 773; zur Haftungsfrage von Care-Robotern s. Beck et al. 2023).

Dritter Unterschied: CAIs sind keine wie Organismen lebendigen Akteure, auch wenn sie nach ihren spezifischen Verkörperungsformen und algorithmischen Designs eine Art von Streben und sogar Spontaneität bzw. Unvorhersagbarkeit aufweisen können. Das ist jedoch nicht mit dem menschlichen trieb- und affektgebundenen Quasi-Code von Lust und Unlust zu vergleichen, und schon gar nicht mit dem, was die philosophische Psychologie seit alters her als *conatus* (Spinoza) und *impetus* thematisiert hat: lebendiges Selbsterhaltungsstreben. Es entbehrt nicht der Ironie, dass wir ein solches Streben (und vielleicht auch so etwas wie primordiale Affektivität) selbst Mikroorganismen zuschreiben können, aber nicht einem künstlichen intelligenten technischen Apparat, obwohl wir mit diesem eine Unterhaltung führen mögen. Da primordiale Affektivität eine der Vorbedingungen für Bewusstsein zu sein scheint,[20] sind *CAI*s auch im Kontext der enaktivistischen Affektforschung als nicht bewusstseinsfähig einzustufen: sie besitzen keine vitale Affektivität, nicht einmal in der Minimalform, die wir schon Mikroorganismen zuschreiben müssen. Auch können *X-Bots* nicht als frei oder unfrei gelten, denn zu Freiheit gehört wesentlich, dass man an ihr leiden kann, vornehmlich dann, wenn sie einem genommen wird. *CAI*s leiden nicht, wenn ihre Besitzer sie für zwei Wochen in den Schrank stellen, und sie erleben ihr gelegentliches »Scheitern« (z.B. nicht über die Türschwelle rollen zu können, um ein Ziel zu erreichen) nicht als irgendwie unangenehm, selbst wenn ihr Betriebssystem einen Fehler meldet und daraufhin die Sequenz ausgeführt wird, »mit Bedauern reagieren«.

Vierter Unterschied: *CAI*s sind nicht zur Autonomie in der Lage, die der von selbstreflexiven Wesen gleicht. Aus diesem Grund bereits müssen KI-Systeme a-moralisch betrachtet werden, auch wenn ihr Design die Ausrichtung an ethischen Standards (*Alignment*) vorsehen mag. Normalerweise (d.h. wenn nicht anders im Zieldesign beabsichtigt) lässt sich gewährleisten, dass *X-Bot*s bestimmte Nichtschädigungsnormen einhalten. Ausnahmen von der Regel könnten solche *X-Bots* sein, deren Design sie anfällig macht, für schädigende Aktivitäten instrumentalisiert zu werden. Denkbar ist der Fall, dass Malware-infizierte *X-Bots* »deviant« reagieren,

20 Siehe hierzu Colombetti (Colombetti 2014), die die Verbindung von Kognition mit Affektivität stark betont in ihrer Verkörperungstheorie des Geistes.

etwa aus dem Chatbot Alexa eine »*Malexa*« wird (Sharevski et al. 2021). *CAI*s sind also günstigenfalls ethisch kompatible Artefakte, aber eben keine moralischen oder ethischen Akteure, da sie Nichtschädigungsnormen *blind* befolgen, was dem Kerngehalt moralisch reflektierten Verhaltens grundsätzlich widerspricht: Dass Personen Regeln folgen, z.B. Moralregeln, setzt voraus, dass sie willentlich dagegen verstoßen können. Das ist im Zieldesign von KI-Systemen glücklicherweise nicht »programmatisch« vorgesehen (vgl. Floridi/Sanders 2004), wird allerdings in utopistischen KI-Narrativen (Bostrom 2014; Russel 2019; Christian 2020) als mögliche Gefahr ausgemalt.[21]

Fünfter Unterschied: Was heutige als KI bezeichnet wird, ist technisch gesehen größtenteils maschinelles Lernen, das Funktionsvarianten von Mustererkennung (*pattern recognition*) leistet. Mustererkennungsleistungen und soziale Anerkennung im Sinne der sozialethischen Erfahrungsweise von bedeutungsvoller Bezogenheit, die ich hier angesetzt habe, sind grundverschieden. Freilich ist es immer möglich, gewisse Maschinen, Apparate, Geräte, durch Definitionsmacht oder einfach durch überschießende Zuschreibungen von Intentionalität (Marchesi et al. 2019) zu »beseelen« und durch gezielte Analogiebildungen den Anschein zu erwecken, als wären *CAI*s Menschen ganz ähnlich.

Für derartige Anthropomorphisierungen spielt der Rekurs auf Kriterien oder spezifischer Lesarten von »Fähigkeiten« eine große Rolle, die dann als funktionale Äquivalente menschlicher Befähigungen gelten sollen. Wer Personsein durch eine kriteriale Definition mit Hilfe von Fähigkeiten meint begreifen zu können, hat dann vielleicht Grund genug, bestimmte *CAI*s als »Akteure« – wie natürliche Personen – zu sehen. Nach diesem ersten Schritt einer »Humanisierung« läge der zweite nicht weit, solche *CAI*s auch als hinreichend fähige Träger oder Geber sozialer Anerkennung auszuzeichnen. Solche Upgrading-Versuche erscheinen mir nicht nur unbefriedigend, weil hinsichtlich der breiten philosophischen Debatte über die Sinngrenzen des Akteur-Vokabulars oftmals auffällig uninformiert. Sie erscheinen mir im Grunde (immer noch) als eine Art Gretchenfrage sozialer KI. Da Selbstbewusstsein von Robotern (meines Wissens) noch nicht klar bewiesen ist, nehme ich hier eine pragmatische Sichtweise auf *CAI*s als »anerkennungsrelevante« Artefakte ein, die durchaus auch Pro-Argumente für *CAI*s als potenzielle Reproduzenten von Anerkennungserfahrungen einrechnet: Was man schlicht zu akzeptieren hat, ist, dass Objektbeziehungen zu Artefakten libidinöse Investitionen und emotionales Bindungsverhalten (Kernberg 1992) beinhalten können (Xie/Pentina 2022). Entwicklungspsychologisch gesehen könnten *CAI*s unter Umständen als Über-

21 Zudem lässt sich mit Rafael Capurro (Capurro 2006) anmerken, dass es schlicht falsch ist, moralische Verantwortlichkeit schon dort zuzuschreiben, wo Akteure etwas Gutes oder das Gegenteil davon bewirken können. Capurros Caveat gilt für Menschen und allemal für *CAI*s.

gangsobjekte fungieren,[22] wenn sie als *vorübergehender* (!) Platzhalter dienen, der dann allerdings auch zurückgelassen bzw. aufgegeben werden können müsste. Das Design vieler *CAI*s ist aber meistens darauf zugeschnitten, dass man möglichst viel Zeit mit ihnen verbringt. Das scheint eher ein einsames Verharren in der digitalen Echokammer der *CAI*-Beziehung (Jacobs 2023a: 53) zu begünstigen. Die Responsivität von *CAI*s, etwa der hochentwickelten *Sophia*,[23] könnten von Menschen subjektiv sogar als angemessen erfahren und positiv erlebt werden, insbesondere dann, wenn das Gerät als Mittel zur Erreichung eines menschlichen Gutes beiträgt (z.B. einem das Gefühl gibt, Aufmerksamkeit zu erhalten) oder wenn es Schaden von Menschen abwendet. Das erhöht die Argumentationslast dafür, *CAI*s grundsätzlich aus dem »anerkennungsrelevanten« Praxisbereich auszuschließen.

Halten wir fest: Es ist möglich, dass gerade einsame Menschen in der Beziehung zu einem *CAI* positive Erfahrungen machen können. Wo sich jemand von einem Roboter angemessen anerkannt (d.h. geliebt, respektiert, wahrhaftig gesehen, unterstützt, erwünscht usw.) erleben kann oder Vertrauen zu der und in die Maschine hat, hätten wir womöglich sogar gute Gründe, eine theoretische Erweiterung der Sphäre der Anerkennungsbeziehungen in Betracht zu ziehen. Wir haben es zwar nicht mit einem inter-affektiven Beziehungsmodus zu tun, aber die Einsatzmöglichkeiten von gewissen Arten von *X-Bots* sollten wir für die Zukunft des Einsamkeitsmanagements nicht ignorieren. Doch es bleibt ein Dilemma, das ich andernorts (Jacobs 2023a: 53) als Echokammer-Szenario bezeichnet habe: Man hat zwar einen positiven Effekt (das Oberflächenphänomen der Einsamkeit wird gemildert), dennoch bleibt dies problematisch, sofern der Grundzustand, der idealiter verändert werden sollte, reproduziert wird: Wir sind verbunden, aber trotzdem allein, und sogar dann, wenn die Einsamkeit punktuell verschwindet, besteht das strukturelle Einsamkeitsrisiko langfristig fort.

Mit Blick auf die beschriebenen fünf Unterschiede zwischen menschlichen und technisch gemachten Modi von Bezogenheit möchte ich betonen, dass sogar dann,

---

22 Der Psychoanalytiker Donald Winnicott (Winnicott 1953) bezeichnet als *Übergangsobjekte* die ersten Objekte, zu denen kleine Kinder emotionale Bindung aufbauen. Sie markieren eine Phase der psychosozialen Entwicklung, in der Kinder innere und äußere Realität stärker zu unterscheiden beginnen. Übergangsobjekte werden als Nicht-Ich und doch auch als zum eigenen Ich gehörig erlebt und dienen dem Übergang vom infantilen Allmachtsdenken zu objektiverem Denken in Beziehungen.

23 *Sophia* ist ein *X-Bot*, der von der in Hongkong ansässigen Firma *Hanson Robotics* entwickelt wurde und in vielerlei Hinsicht sehr außergewöhnlich ist, z.B. weil »sie« die erste Maschine war, die eine Staatsbürgerschaft erhielt und auch der erste Nicht-Mensch war, dem im Rahmen des *United Nations Development Programms* den Titel *Innovation Champion* verliehen wurde. *Sophia* wurde nicht nur in 25 Ländern vorgestellt und hat sich mit Staatsführern unterhalten (z.B. mit der ehemaligen Bundeskanzlerin Angela Merkel), sondern ist auch in Produktion gegangen, d.h. für 80.000 US-Dollar kann man »eine« *Sophia* kaufen.

wenn wir gewissen Arten von *CAI*s mit relevanten menschenähnlichen Fähigkeiten einen bestimmten »anerkennungsrelevanten« Status zuschreiben wollten, blieben spezifisch intersubjektive Fähigkeiten davon ausgeschlossen, die für die Bedeutsamkeit von interaffektiver Resonanz und Selbstreflexivität für Menschen in Anerkennungsbeziehungen unumgänglich sind. Ich bin überzeugt, dass interaffektive Resonanz in Anerkennungsbeziehungen dasjenige ist, was tief einsame Menschen brauchen, um sich aus ihrer Einsamkeit zu befreien. Trifft das zu, dann ist es kein Widerspruch, dass manche Menschen den Austausch mit *CAI*s so erfüllend erleben, dass sie sich momentan nicht mehr einsam fühlen.

4.1 Ethische Fragen

Wie wäre das vor dem Hintergrund normativer Standards für angemessene soziale Integration zu bewerten? Ich neige zu der Auffassung, dass selbst Szenarien, in denen Menschen *just fine* mit ihren digitalen Freunden sind, solche Standards nicht erfüllen. Hier stellt sich die schwierige Frage der ethischen Urteilsbildung über Praktiken, die den sozialen Schmerz der Einsamkeit mit Maschinen beheben wollen, die bestenfalls menschlich erscheinen. Man könnte fragen: Verdienen wir nicht mehr und besseres?

Zu fragen wäre m.E. auch, ob Beziehungen zu *CAI*s im Unterschied zu komplizierten zwischenmenschlichen Beziehungen nicht nur gleichsam komfortabler sind, sondern eine verdeckte Form der Missachtung sozialer Anerkennung darstellen, die sich wie auf einer abschüssigen Bahn auch auf zwischenmenschliche Handlungspraxis übertragen könnte.

Sicher spielen Gewöhnungseffekte hier eine Rolle, deren Ausmaß noch wenig erforscht ist. Wenn manche Menschen *X-Bots* als Objekte von totaler Kontrolle behandeln, leiden die *X-Bots* zwar nicht darunter. Aber die Eingewöhnung entsprechender Einstellungen könnte sich destruktiv auf die Gestaltung zwischenmenschlicher Beziehungsmuster auswirken. Dass diese »Übertragung« möglich ist, zeigt sich überall dort, wo man den Gewöhnungseffekt positiv zu nutzen versucht. Das geschieht z.B. in Settings, in denen man mit sozialer KI Angststörungen zu therapieren versucht, indem man die Menschen in virtueller Realität mit Objekten ihrer Angst (etwa virtuellen Spinnen) konfrontiert, oder in Settings für die Anwendung spezifischer Neurofeedback-Methoden, die AI zur neuronalen »Überschreibung« verwenden (Koizumi et al. 2016).

Wenn man überlegt, *wie* die Verbreitung von und Gewöhnung an *CAI*s zu einer allgemeinen Veränderung der gängigen Bewertungsstandards für die Wahrnehmung von sozialer Anerkennung beitragen würde, müsste mitbedacht werden, dass einige der neuen, aus Mensch-Maschine-Interaktionen emergierten Bewertungsstandards, z.B. ständige Verfügbarkeit und stetige Bestätigung, so wie man es von *CAI*s gewohnt ist, viele wertvolle Formen zwischenmenschlicher Beziehungen kor-

rumpieren würden. Mag sein, dass man diese unterschiedlichen Beziehungstypen auch trennen kann, es also nicht zwingend zu einer »Verwechslung« oder Verschiebung der Bewertungsmaßstäbe kommen muss. Gleichwohl sollten wir diese Janusköpfigkeit sozialer KI in der ethischen Bewertung nicht vergessen.

In die ethische Bewertung möchte ich an dieser Stelle nicht tiefer eintreten. Ich möchte nicht prinzipiell ausschließen, dass wir Praktiken der sozialen Anwendung von KI entwickeln könnten, die der Einsamkeitsbewältigung tatsächlich recht gut dienen würden. Mit dieser selbstkritischen Bemerkung möchte ich dem methodischen Anspruch immanenter Kritik Rechnung tragen, dass sie eine Transformation hin zum Besseren nicht ausschließt, sondern anstrebt.

4.2 Digitale Strategien zur Überwindung der Einsamkeit?

Trotz aller geäußerten Kritik ist praktisch-lebensweltlich nicht zu leugnen, dass KI-Begleiter eben doch als hilfreich von einsamen Menschen wahrgenommen werden. Man könnte betonen, dass in der Beziehung zu einem *CAI* doch immerhin paradigmatisch eine *digitale Autonomie* zum Ausdruck kommt, denn schließlich versuchen Akteure die eigene Situation zu verbessern, auch wenn es sich im eigentlichen Sinne um kein echtes Anerkennungsszenario handelt, bzw. die Anerkennung beschränkt bleibt auf die spezifischen Formen mit der sich eine Person allein in der Interaktion mit einem X-Bot soziale Anerkennung schenken kann.

Denken wir versuchsweise affirmativ: *CAI*s erweisen sich unter den persönlichkeitspsychologischen Gesichtspunkten der Selbstwirksamkeit und Psychohygiene als nützliche Tools, um das subjektive Einsamkeitserleben zu steuern und ihm gegenüber Autonomie zu gewinnen. Aus sozialpsychologischer Sicht muss dann allerdings auf den unausblendbaren Stellenwert von sozialer Interaktion hingewiesen werden. Die subjektive psychische Realität von *ego* ist eben nicht strikt von der praktischen Realität von *alter* zu trennen. Da »die Anderen« aber für den tief Einsamen außer Reichweite geraten sind, ist die »bedingungslose« Wertschätzung, die unkritische Bejahung, die stetige Verfügbarkeit, die *CAI*s versprechen, für sie attraktiv.[24]

Hebt man als ethisch relevanten Gesichtspunkt hervor, dass soziale KI tatsächlich *sozial* wirken, also Vergemeinschaftung bewirken soll, damit von einer Anerkennungspraxis die Rede sein kann (Jacobs 2023a: 62ff.), dann kann man fordern, *CAI*s so zu gestalten, dass sie diese *Übergangsfunktion* erfüllen, die es Menschen, die das brauchen, erleichtert, die Position des sozialen Ausschlusses zu verlassen und sich aktiv als ein Gegenüber zu positionieren. Wie man einen Gips für ein paar Wochen

24 Die Kritikerin wird zu bedenken geben, dass ein therapeutisches Potential dieser Mensch-Maschine-Interaktionsfiguren mit einem *geliebten Objekt* (Habermas 2020) allenfalls dann genutzt werden kann, wenn daneben der Sinn für das Bedeutungsvolle realer Bezogenheit zwischen Mitmenschen nicht ganz vergessen wird.

trägt, damit ein gebrochener Knochen heilen kann, könnte der Einsatz von *CAI*s vorübergehend für an Einsamkeit Leidende nützlich sein, sollte aber gleichsam mit einem Warnhinweis auf paradoxe Wirkungen bei langfristigem Gebrauch versehen sein.

Ich kann mir KI-Szenarien mit größerem Potenzial für die Einsamkeitsbewältigung als den Einsatz von *CAI*s vorstellen. Die aktuelle Einsamkeitsforschung ist auf das Potenzial von virtuell realen Erlebnisräumen (VR) aufmerksam geworden. Man kann sie nutzen, um einsame Menschen für die Beziehungsaufnahme zu Mitmenschen zu motivieren. Studien zeigen, dass auch ältere Menschen sich gut daran gewöhnen können, sich digital mit Freunden und Familienmitgliedern auszutauschen. *Avatar Mediated Conversation* (AMC) kann hierfür nützlich sein. Einiges deutet darauf hin, dass besonders introvertierte Menschen *AMC* in sozialen VRs als ein für sie adäquates Medium bewerten. Sie erlebten, dass sie sogar über sensible Themen, z.B. die eigene Einsamkeit, unter Verwendung eines Avatars besser kommunizieren konnten. Die Forschergruppe weist aber darauf hin, dass ein rein »technischer« Ansatz zur Prävention und Behandlung von Einsamkeit nicht ausreicht (Barbosa et al. 2021; Johnston 2022), weil Menschen letztlich Menschen brauchen.

Ein weiteres gutes Beispiel, dass es dafür gar nicht hyper-realistische humanoide Roboter mit einer auf emotionalem Computing basierter Software oder unbedingt Avatare bedarf, ist das vergleichsweise simple, auf Smartphones verfügbare Realitygame *Pokémon Go*. Hier wird mit genuin digitalkulturellen Mitteln eine Gemeinschaft von Mitspielern ins Leben gerufen und mit Straßenkarten ausgestattet für die Jagd auf virtuell reale Phantasiegeschöpfe (Pokémons). Auf chronische Einzelgänger hat das Spiel eine therapeutisch positive Nebenwirkung, sofern es sie motiviert, sich wieder »unter Menschen« zu bewegen (Kato et al. 2017a; Kato et al. 2017b).

In diesen Beispielen leisten soziale Anwendungen von KI mehr als nur einen minimalen Beitrag zur Bewältigung von Einsamkeit, denn sie bieten ein transformatives therapeutisches Potenzial zur Wiederherstellung zwischenmenschlichen Kontakts. Gewiss garantieren sie nicht die endgültige Überwindung pathogener Einsamkeit, bieten aber immerhin eine Synchronisierung von virtuellen digitalen und realen zwischenmenschlichen Beziehungen an. Natürlich müssen auch diese eine bestimmte Qualität aufweisen, sodass sich symptomatische Muster von verzerrten Anerkennungsbeziehungen nicht einfach wiederholen: dass man sich »vernetzt«, sich aber trotzdem sozial vernachlässigt und einsam fühlt. Mir kommt es an dieser Stelle drauf an, dass und wie diese empirisch erforschten Szenarien sich von anderen Szenarien unterscheiden, die an der Befindlichkeitslage der Einsamkeit nichts grundlegend ändern oder die Einsamkeit sogar noch verstärken.

Dass mangelhafte Formen digitalkulturellen Einsamkeitsmanagements unter Umständen sogar für Dritte zum Problem werden kann, wenn das Leiden an Vereinsamung in Fremdschädigung umschlägt, erläutere ich nun mit einem Ausblick

auf das »kritische« Potenzial digitaler Einsamkeit als einer neuen Form des Prekariats.

5. Die Masse der Einsamen: Das neue (digitale) Prekariat

Der sozialpathologische Anteil der Einsamkeit wird in einigen psychologischen und sozialwissenschaftlichen Erkundungen von Einsamkeit bereits mitgedacht, etwa als »Defekt in den sozialen Beziehungen« (Horney 1937: 87), und wird mit zunehmender (»großstädtischer«) Anonymisierung und Entfremdung in Verbindung gebracht (Lehmann 1967). Trotz *»digital connectedness«* (Schetsche/Lehman 2003; Nguyen/Gruber/Hargitti et al. 2021) wird Einsamkeit wie vor hundert Jahren als ein Zustand *innerer Heimatlosigkeit* beschrieben (Simmel 2013[1908]: 65ff.). Wurde sie einst mit Entfremdungserfahrungen und seelischen Störungen assoziiert, können wir sie heute, wenn wir Bindungsunfähigkeit oder mangelnde Identifikation mit den Werten einer kulturellen Wir-Gruppe hervorheben wollen (Fromm 1941: 20ff.) auch unter Rückgriff auf das Vokabular der Stoa als eine Störung der *Oikeiôsis* (οἰκείωσις), also der Selbst- und Weltaneignung von Personen, begreifen (Jacobs 2023b).[25]

In Anlehnung an das altbewährte marxistische Entfremdungsparadigma der Kritischen Theorie lässt Einsamkeit sich, aktualisiert mit Fromm (Fromm 2010: 59ff.), als eine stille Misere innerhalb der *pathologischen Normalität* einer (zu) rasant digitalisierten Kultur thematisieren. Die digitale Transformation von sozialer Anerkennung ist eine systemisch induzierte Veränderung der Lebenswelt, in der Systemimperative (»die digitale Agenda«), Habermasianisch gelesen (Habermas 1981: 489ff.), zu einer *Kolonialisierung der Lebenswelt* führen und damit wichtige Bezugsgrößen von Prozessen sozialer und personaler Integration destabilisieren können. Das dürfte insbesondere für kapitalistische Gesellschaften gelten, wobei Kapitalismus hier nicht exklusiv als rein ökonomisches System, sondern als institutionalisierte Gesellschaftsordnung (Fraser 2017: 141ff., bes. 152ff.) und als Lebensform (Jaeggi 2014: 67–134; Fromm 2010: 138ff.) verstanden werden kann.

25 Der Begriff ist fragmentarisch überliefert und seine Etymologie wurzelt in dem Wort *oikos* (οἶκος), dass das Wort für Haushalt, Haus oder Familie ist. »*Oik-*« verweist auf »das, was zu mir gehört«, während »*-eiôsis*« sich auf das *Vertrautwerden mit etwas oder dem, was zu mir gehört*, bezieht, was in der stoischen Tradition gleichbedeutend mit Selbsterkenntnis ist. In ähnlicher Weise bezieht sich der Begriff *oikeiôtes* auf die Wahrnehmung von etwas als das Eigene und spricht ganz allgemein ein *Gefühl der Zugehörigkeit* an (Sorabji 2018; Pembroke 1971: 114ff.). Der Begriff bezieht sich auch auf *das Gefühl, zu Hause zu sein* oder mit etwas vertraut zu werden (vgl. Jedan 2012), ist also etwas, was in (digitalen) Vereinsamungsprozessen grundsätzlich verloren gehen kann.

Hilft uns der Sozialpsychologe Erich Fromm, das Defizitäre und u.U. Pathogene nicht nur der Einsamkeit selbst, sondern auch des Versuchs ihrer digitalkulturellen Überwindung mittels sozialer KI zu erklären? Fromm würde argumentieren, dass dann, wenn Digitalität Teil unseres *Gesellschaftscharakters* (Fromm 1980: 81ff.) geworden ist, epidemische Einsamkeit mitverursacht wird durch Anpassung an eine krankmachende (»pathologische«), als solche aber gar nicht empfundene Normalität (Fromm 1980: 15). Als latent pathologische Normalität gilt Fromm eine Lebensform, in der gravierende Verkennungen dessen, was Menschen zur gesunden Lebensführung brauchen, endemisch und folglich unauffällig, eben »normal« ist (Fromm 2005). In dieser Perspektive können wir untersuchen, wieweit digitale Einsamkeit aus der Anpassung an Praxisnormen resultiert, denen gravierende Missverständnisse über »adäquate« Formen der Anerkennung eingeschrieben sind. Da Fromm eine kommunitaristisch-sozialistische Interpretation der Normen objektiven Wohlergehens vorschlägt, verwundert es nicht, dass die Vereinsamung für ihn Ausdruck einer »seelischen Isolierung« ist: Ausdruck einer fehlenden Beziehung zu bedeutsamen Werten, Symbolen und Verhaltensmustern einer Gemeinschaft, die gesundheitsrelevant sind (Fromm 2006[1941]: 22ff., 101ff., 185). Vereinsamung interpretiert Fromm zudem als eine »Flucht« bzw. eine regressive Tendenz hin zu ausschließlich negativer Freiheit (als einer bloßen Freiheit *von* hindernden äußeren Zwängen), die eine Selbstaneignung und Selbstverwirklichung (als eine positive Freiheit *zu* einem Projekt der Lebensführung) eher verhindert als ermöglicht (Fromm 2006[1941]: 186ff.). So gesehen, erscheint Einsamkeit als ein Zustand der Verarmung an bedeutungsvoll erlebter Bezogenheit, ein gefühlter Mangel an positiver Freiheit, die für Fromm seelische Integrität ausmacht. Fromm mutet uns zu, nicht im Zweifel und in der Hilflosigkeit der Einsamkeit zu erstarren, sondern aktiv ihre Bewältigung zu versuchen, auch und gerade wenn das womöglich bedeuten, Gesellschaftsveränderung anzustreben (ebd.).

So ergibt sich mit Fromm eine zweite, gewissermaßen dialektische Lesart von Einsamkeit: Wenn sich Einsame nämlich aufgrund mangelnder Identifikation mit den Standardnormen des sozialen Charakters »sinnentleert« fühlen, wäre dies unter Umständen als Ausdruck eines authentischen Leidens an dem zu deuten, was zwar soziale Anerkennung genannt wird, aber der individuellen Sinnerfahrung bedeutungsvoller Bezogenheit entgegensteht. Bleibt man bei dieser Sichtweise, wären einsame Menschen – eventuell besonders jene, die eine digitale »Standardlösung« ihres Leidens ablehnen – potenziell sogar eher systemkritische Akteure, da ihre Einsamkeit einer aktiven Weigerung gleichkommen könnte, sich der »Normalität« digitaler Anerkennungsszenarien anzupassen.

Hier ergibt sich ein Ausblick auf eine spannende Dichotomie zwischen Einsamkeit als Ausdruck eines Kampfs um Anerkennung (auch mit digitalen Mitteln) und der Frommschen Lesart, wonach Einsamkeit eher Ausdruck eines Ringens mit sich selbst ist. In dieser Erfahrung (hinter)fragen sich Menschen: mit was kann ich mich

wirklich identifizieren? Wer bin ich, und was ist notwendig dafür, dass ich mich selbst verwirklichen kann? Selbst wenn also die Beziehung mit Anderen in der Konsequenz die eigene Isolation bedeutet, kann dies zumindest eine sein, in die personale Autonomie und kritische Kompetenz einfließen.

Das ist der Gegensatz zum zuvor besprochenen Fall der Personen, die ganz bewusst Beziehungen zu *CAI*s aufbauen und dies als Selbstsorge begreifen. Im letzteren Fall soll mittels *CAI*s künstlich-technisch »reproduziert« werden, was schmerzlich vermisst wird, während die Einzelgänger sensu Fromm das genau ablehnen würden, zumindest dann, wenn sie digitalkulturelles Einsamkeitsmanagement skeptisch als systemkonforme Mainstream-Lösung bewerten. Ein aktiver Rückzug aus der Gesellschaft, auch um sich nicht *gemein zu machen*, wie Nietzsche gesagt haben würde, würde es zwar prinzipiell erlauben, die Einsamkeitserfahrung umzuwerten und statt als leidvolles Entfremdungserleben vielmehr als heroischer Selbstaneignung durch soziale Distanzierung zu verstehen, aber diese Umwertung erfordert eine seltene Virtuosität, abgesehen davon, dass es heutzutage zunehmend schwieriger ist, sich dem (digitalen) Anderen überhaupt grundsätzlich entziehen zu können.

Bleiben wir beim Unbehagen in der massenhaften Einsamkeit. In einer umsichtigen Debatte über digitale Einsamkeit als Störung sozialer Anerkennungsbeziehung wäre m.E. auch Dynamiken einer kollektiven, unter Umständen radikalen Re-Politisierung mehr Beachtung zu schenken. Was das Einsamkeits-Prekariat auszeichnet, ist der (zumeist als ausweglos und hoffnungslos wahrgenommene) Kampf um Anerkennung, z.B. als berechtigter Empfänger bestimmter Formen von (gesellschaftlicher) Unterstützung oder bestimmter Verpflichtungen und Träger bestimmter Rechte wahrgenommen zu werden. Neben den hohen monetären Kosten, die Einsamkeit und ihre Folgen verursacht (z.B. für das Gesundheitswesen), wird ein weitaus höherer Preis der sozialen Desintegration gezahlt.

Unzufriedenheit mit den realen Lebensbedingungen, gerade auch mit und innerhalb bestimmter digitaler Sphären in der Lebenswelt (der Einsamen), zeigt sich in einer anderen Form als das »kritische Potenzial« steigender Vereinsamung in Gesten der Auflehnung gegen ihren Status von *social outcasts*. Ein Beispiel gibt die offen zur Schau gestellte Frauenfeindlichkeit, mit der sich ein Teil der Incel-Community (*Involuntary Celibate*) identifiziert (Speckhard et al. 2020). Ungewollte Einsamkeit wird damit allerdings nicht bewältigt, sondern schlägt im Zuge reaktiver Projektionsbildung in Gewaltbereitschaft um: sie wird zu einem Katalysator für Identifikation mit dem Leiden an sozialer Isolation, schürt die Verbreitung bestimmter subkultureller Gegen-Narrative (z.B. über männliche und weibliche Identität in krude frauenverachtenden Formen) und befeuert Sozialneid hin zum Hass, amplifiziert durch die sozial-medialen Kommunikationsverhältnisse.[26] Dies

26 Vergleiche hierzu den Beitrag von Kai Denker im vorliegenden Band.

ist nur eine von vielen Formen der Reaktionsbildung auf die Frustration ungewollter Einsamkeit, die die soziale Dynamik der Ausgrenzung Anderer, des Fanatismus, der gruppenbezogenen Menschenfeindlichkeit und der Gewalt fördern kann, zu denen digitale Einsamkeit beitragen kann.

Wir sehen hier ein lautes Gegenstück zum stillen digitalen Leiden an Einsamkeit und ein neues ethisches Problem: In bestimmten digitalkulturellen Praktiken wird eine verzerrte Anerkennungspraxis als normal empfunden, hat aber destruktiv-aggressive Auswirkungen. Anerkennungsvergessene digitale Praktiken sind wohl nur deshalb möglich, weil vereinsamte Personen sich nicht miteinander solidarisieren wollen oder können, sondern gleichsam atomistisch im Schutz der in digitalkulturellen informellen Kommunikationsverhältnissen vorherrschenden Anonymität nur das eine gemeinsame Verbindenden wahrnehmen können: den eigenen sozialen Schmerz, der kraft sozialmedialer Plattformen ungehindert auf bestimmte Kollektive projiziert und externalisiert werden kann.[27]

Die unterschiedlichen Formen der Einsamkeit werden gegenwärtig immer sichtbarer, insbesondere die (digitalen) Nischen, in denen die Einsamkeit »gedeihen« kann, sei es als Ressentiment oder als Depression und Hoffnungslosigkeit über genau diesen Zustand, sich sozial benachteiligt zu fühlen bzw. es zu sein (Schmalenbach 1919; Wood 2020). Wenn Einsamkeit zu anomischen Zuständen führt, die intersubjektiv konsensfähig als schädlich oder belastend nicht nur für die direkt Betroffenen bewertet werden, sondern auch für andere gefährlich werden können, sollten Gesellschaften mit steigenden Vereinsamungsraten problematisieren, wir gut sozial angewandte KI, insbesondere *CAI*s, wirklich sind, um diesen sozial prekären Dynamiken zu begegnen. Ich meine: Bei allem Enthusiasmus über die Möglichkeiten, die mit der weiteren Humanisierung von KI einhergehen, sollten wir eine kritische (nicht unbedingt pessimistische) Perspektive einnehmen, wenn meine Vermutung triftig ist, dass der Aufschwung von *CAI*s einen Teil des Auftretens von spezifischen, nicht-trivial schädlichen Formen von Einsamkeit im Kulturprozess der Digitalisierung erklärt. Die Vorstellung, dass digitale Vernetzung und *Digital Companionship* das Prekariat der Einsamen automatisch zum Positiven

27 In solch gravierenden Formen der Anerkennungsvergessenheit wird der andere bisweilen wie ein rein technischer Anderer behandelt – wie ein Objekt, ein *CAI* – da er emotional genauso »unerreichbar« ist und als ähnlich (digital) »kontrollierbar« imaginiert wird; eben wie ein relationales Artefakt, dessen man sich einfach bemächtigen kann, was ich an anderer Stelle im Kontext narzisstischer Abwehr als »Einverleibung der Persona des anderen« beschrieben habe (Jacobs 2019a). In diesen dysfunktionalen Kompensationsversuchen der Einsamkeit sind illusorisch entgrenzte (typischerweise: grandiose) *Beziehungsideen* signifikant. Dies erklärt die Psycho(patho)dynamik verschiedenster Formen von Anerkennungsvergessenheit (z.B. des digitalen Stalkens, der Paranoia, die mit Kontaktmangel einhergehen kann; siehe dazu: Kretschmer 1918). In alledem wird deutlich, wie groß das Bedürfnis nach einer (und sei sie noch so destruktiv) irgendwie als »sinnvoll« erlebten Beziehung sein kann.

verändert, erscheint mir als ein mythisches Erfolgsnarrativ zwecks Verbreitung sozialer KI. In Zukunft werden die *X-Bots*, die von vielen Menschen behandelt werden, als wären sie Menschen, unsere Lebenswirklichkeit sicher noch viel stärker prägen. Dabei werden mit dem Ziel, massenhafter Einsamkeit primär mit Digitaltechnologie zu begegnen, genau jene Verpflichtungen verfehlt, die sich aus einer Ethik der sozialen Anerkennung ergeben.

6. Fazit

Ich habe einen kritischen Blick auf die Einsamkeit geworfen, die teilweise in digitalkulturellen Kommunikationspraktiken erst gedeihen, denen aber selbst mit avancierter Digitaltechnologie allein nicht beizukommen ist. *CAI*s und ähnliche *X-Bots* können zwischenmenschliche Anerkennung nicht so ersetzen oder ausbessern, dass es ethisch gut zu verantworten wäre, solche (Re-)Produktionsbedingungen sozialer Integration in Zukunft noch viel stärker an sozial angewandte KI auszulagern. Ich habe digitale Einsamkeit als einen Modus von Entfremdung beschrieben, der kausal mit der komplexen Dynamik gestörter Anerkennungsbeziehungen verbunden ist. Digitale Einsamkeit bezieht sich auf den Zustand, in dem man zwar digital mit allem Möglichen, auch anderen Menschen, verbunden ist und Nachrichten austauschen kann, und dennoch oft keine sinnvolle, qualitativ einsamkeitslösende Beziehung zu anderen Menschen (er)lebt. Das muss nicht ausnahmslos so sein. Andere Szenarien sind möglich. Wir sollten dann aber bedenken, ob sie langfristig zu Quellen von noch tiefgreifenderen Isolationserfahrungen werden.

Unter Rekurs auf Erich Fromms Sozialpsychologie habe ich an Einsamkeit ein »kritisches Potenzial« im doppelten Sinne aufgezeigt: sie ist nützlich für die Regeneration von Nonkonformismus und kritischer Reflektion, kann aber auch Katalysator für eine Reihe antisozialer Reaktionsbildungen sein, die vornehmlich auch digital (aus)gelebt werden.

Wir sollten den kritischen Stellenwert von massenhafter Einsamkeit nicht nur im Hinblick auf das subjektive empfundene Leiden der von Einsamkeit betroffenen Einzelnen einschätzen, sondern zukünftig weitaus mehr im Hinblick auf ihre selbst- *und fremdschädigenden* Auswirkungen (inklusive der dysfunktionalen Kompensationsstrategien für Einsamkeit). Diese Blickerweiterung könnte begünstigen, dass hochdigitalisierte Gesellschaften vorrangig in die Gestaltung von humanen, gedeihlichen Lebensverhältnissen entgegenkommenden Institutionen investieren, um ethisch und rechtlich gesicherte Einbettungsbedingungen für die weitere Institutionalisierung von humanoider KI zu schaffen. Andernfalls könnte die im Geist der Zeit liegende Tendenz, möglichst viel von den Prozessen, die an der Reproduktion der Lebenswelt beteiligt sind, an soziale KI auszulagern, einiges Unheil anrichten. Wo soziale KI als Antidot gegen massenhafte Einsamkeit geplant wird, wächst

gegen alle gute Absicht die Gefahr, dass das systembedingte Elend der Einsamkeit noch stärker privatisiert und invisibilisiert wird.

Literatur

Adler, N.E.; Boyce, T.; Chesney, M.A.; Cohen, S.; Folkman, S.; Kahn, R.L.; Syme, L.S. (1994): Socioeconomic Status and Health. The Challenge of the Gradient, in: *American Psychological Association*, 49(1), 15–24.

Amichai-Hamburger, Y.; Ben-Artzi, E. (2003): Loneliness and Internet use, in: *Computers in Human Behavior*, 19(1), 71–80.

Amodio, D.M.; Frith, C.D. (2006): Meeting of minds. The medial frontal cortex and social cognition, in: *Nature*, 7(4), 268–277.

Archer, M.S. (2021): Friendship Between Human Beings and AI Robots?, in: von Braun, J.V.; Archer, M.S.; Reichberg, G.M.; Sánchez Sorondo, M. (Hg.), Robotics, AI, and Humanity. Science, Ethics, and Policy. Cham: Springer, 177–189.

Asada, M. (2015): Development of artificial empathy, in: *Neuroscience Research*, 90, 41–50.

Asher, S.R.; Hymel, S.; Renshaw, P.D. (1984): Loneliness in Children, in: *Child Development*, 55(4), 1456–1464.

Asher, S.R.; Parkhurst, J.T.; Hymel, S.; Williams, G.A. (1990): Loneliness and peer relations in childhood, in: Asher, S.R.; Coie, J.D. (Hg.), Peer rejection in childhood, Cambridge: Cambridge University Press.

Barbosa, N.B., Waycott, J.; Maddox, A. (2021): When Technologies are Not Enough. The Challenges of Digital Interventions to Address Loneliness in Later Life, in: *Sociological Research Online*, 28(1), 150–170.

Beck, S. (2018): Zum Einsatz von Robotern im Palliativ- und Hospizbereich, in: *Medizinrecht*, 36, 773–778.

Beck, S.; Faber, M.; Gerndt, S. (2023): Rechtliche Aspekte des Einsatzes von KI und Robotik in Medizin und Pflege, in: *Ethik in der Medizin*, 35, 247–263.

Böhme, G. (2008): Ethik leiblicher Existenz. Über unseren moralischen Umgang mit der eigenen Natur, Frankfurt a.M.: Suhrkamp.

Bohn, C. (2008): Die soziale Dimension der Einsamkeit unter besonderer Berücksichtigung der Scham, Hamburg: Kovac Verlag.

Bostrom, N. (2014): Superintelligence. Paths, Dangers, Strategies, Oxford: Oxford University Press.

Bowlby, J. (1958): The nature of the child's tie to his mother, in: *International Journal of Psychoanalysis*, 39, 350–373.

Brinkmann, D. (1952): Der einsame Mensch und die Einsamkeit. Ein Beitrag zur Psychologie des Kontakts, in: *Psychologische Rundschau*, 3, 21–30.

Buecker, S.; Maes, M.; Jaap, J.A.; Dennissen, A.; Luhmann, M. (2020): Loneliness and the Big Five Personality Traits. A Meta-Analysis, in: *European Journal of Personality*, 34, 8–28.

Cacioppo, J.T.; Ernst, J.M.; Burleson, M.H.; McClintock, M.K.; Malarkey, W.B.; Hawkley, L.C.; Kowalewski, R.B.; Paulsen, A.; Hobson, J.A.; Hugdahl, K.; Spiegel, D.; Berntson, G.G. (2000): Lonely traits and concomitant physiological processes. The MacArthur social neuroscience studies, in: *International Journal of Psychophysiology*, 35(2-3), 143–54.

Cacioppo, J.T.; Hawkley, L. (2005): People Thinking About People. The Vicious Cycle of Being a Social Outcast in One's Own Mind, in: Williams, K.D.; Forgas, J.P.; von Hippel, W. (Hg.), The Social Outcast. Ostracism, Social Exclusion, Rejection, and Bullying, New York: Psychology Press, 91–108.

Cacioppo, J.T.; Hughes, M.E.; Waite, L.J.; Hawkley, L.C.; Thisted, R.A. (2006): Loneliness as a specific risk factor for depressive symptoms: cross-sectional and longitudinal analyses, in: *Psychology of Aging*, 21(1), 140–151.

Cacioppo, J.T.; Patrick, W. (2008): Loneliness. Human nature and the need for social connection, New York/London: Norton.

Cacioppo, J.T.; Fowler, J.H.; Christakis, N.A. (2009): Alone in the crowd. The structure and spread of loneliness in a large social network, in: *Journal of Personal and Social Psychology*, 97(6), 977–991.

Cambria, E.; Poria, S.; Hussain, A.; Liu, B. (2019): Computational Intelligence for Affective Computing and Sentiment Analysis [Guest Editorial], in: *IEEE Computational Intelligence Magazine*, 14(2), 16–17.

Capurro, R. (2006): Towards an ontological foundation of information ethics, in: *Ethics and Information Technology*, 8(4), 175–186.

Cassidy, J.; Asher, S.R. (1992): Loneliness and peer relations in young children, in: *Child Development*, 63(2), 350–365.

Christian, B. (2020): The Alignment Problem. Machine Learning and the Human Values, New York: Norton.

Colombetti, G. (2014): The Feeling Body. Affective Science meets the enactive mind, Cambridge: MIT Press.

Daily, S.B.; James, M.T.; Cherry, D.; Porter, J.J.III; Darnell, S.S.; Isaac, J.; Roy, T. (2017): Affective computing. Historical foundations, current applications, and future trends, in: Jeon, M. (Hg.), Emotions and affect in human factors and human-computer interaction, Amsterdam: Elsevier Academic Press, 213–231.

Drageset, J. (2004): The importance of activities of daily living and social contact for loneliness. A survey among residents in nursing homes, in: *Scandinavian Journal of Caring Science*, 18(1), 65–71.

Dreyfus, H.L. (1985): Was Computer nicht können. Die Grenzen künstlicher Intelligenz, Königsstein: Athenäum.

Eisenberger, N.I.; Lieberman, M.D.; Williams, K.D. (2003): Does rejection hurt? An FMRI study of social exclusion, in: *Science*, 302(5643), 290–292.

Eisenberger, N.I.; Lieberman, M.D. (2004): Why rejection hurts. A common neural alarm system for physical and social pain, in: *Trends in Cognitive Sciences*, 8(7), 294–300.

Findlay, R.A. (2003): Interventions to reduce social isolation amongst older people. Where is the evidence?, in: *Ageing & Society*, 23(5), 647–658.

Floridi, L.; Sanders, J.W. (2004): On the morality of artificial agents, in: *Minds and Machines*, 14, 349–379.

Fraser, N. (2017): Behind Marx's Hidden Abode. For an Expanded Concept of Capitalism, in: Deutscher, P.; Lafont, C. (Hg.), Critical theory in critical times. Transforming the global political and economic order, New York: Columbia University Press, 141–159.

Freud, S. (1947[1919]): Das Unheimliche, in: Ders., Gesammelte Werke, Bd. 12, Frankfurt a.M.: Fischer Verlag, 229–268.

Fromm, E. (2006[1941]): Die Furcht vor der Freiheit, München: DTB.

Fromm, E. (2005): Die Pathologie der Normalität. Zur Wissenschaft vom Menschen, Berlin: Ullstein.

Fromm, E. (10. Auflage 1980): Wege aus einer kranken Gesellschaft. Eine sozialpsychologische Untersuchung, Frankfurt am Main: Ullstein.

Fuchs, T. (2011): The brain – A mediating organ, in: *Journal of Consciousness Studies*, 18(78), 196–221.

Fuchs, T.; De Jaegher, H. (2009): Enactive Intersubjectivity. Participatory Sense-Making and Mutual Incorporation, in: *Phenomenology and Cognitive Science*, 8, 465–486.

Gallotti, M.L.; Frith C.D. (2013): Social cognition in the we-mode, in: *Trends in Cognitive Sciences*, 17(4), 160–165.

Gao, J.; Galley, M.; Li, L. (2018): Neural Approaches to Conversational AI, in: *arxiv*. [https://arxiv.org/abs/1809.08267v3] (Zugriff: 01.10.2023).

Habermas J. (1981): Theorie des kommunikativen Handelns. Bd. 2. Zur Kritik der funktionalistischen Vernunft, Frankfurt a.M.: Suhrkamp.

Habermas, T. (2020): Geliebte Objekte. Symbole und Instrumente der Identitätsbildung, Berlin: de Gruyter.

Heidegger, M. (11. Auflage 1967): Sein und Zeit, Tübingen: Niemeyer.

Honneth, A. (1994): Der Kampf um Anerkennung. Zur moralischen Grammatik sozialer Konflikte, Frankfurt a.M.: Suhrkamp.

Honneth, A. (2003): Unsichtbarkeit. Stationen einer Theorie der Intersubjektivität, Frankfurt a.M.: Suhrkamp.

Honneth, A. (2013): Das Recht der Freiheit. Grundriss einer demokratischen Sittlichkeit, Berlin: Suhrkamp.

Honneth, A. (2015): Verdinglichung. Eine anerkennungstheoretische Studie. Erweiterte Ausgabe, Berlin: Suhrkamp.

Honneth, A. (2018): Reification and Recognition, in: Jay, M. (Hg.), A New Look at an Old Idea, Oxford: Oxford University Press, 17–95.

Horney, K. (1937): The neurotic personality of our time, New York: Norton.

Illouz, E. (2016): Warum Liebe weh tut. Eine soziologische Erklärung, Berlin: Suhrkamp.

Jacobs, K.A. (2013): The depressive situation, in: *Frontiers of Theoretical Psychology*, 4(429), 1–11.

Jacobs, K.A.; Kettner, M. (2017): Zur Theorie »sozialer Pathologien« bei Freud, Fromm, Habermas und Honneth, in: Clemenz, M.; Zitko, H.; Büchsel, M.; Pflichthofer, D. (Hg.), *IMAGO. Interdisziplinäres Jahrbuch für Psychoanalyse und Ästhetik*, 4, 119–146.

Jacobs, K.A. (2018): Sozialdiagnostik und Lebensrat. Ärztliche Praxis als medizinischer Grenzgang, in: Wittwer, H. (Hg.), Was ist Medizin? Der Begriff der Medizin und seine ethischen Implikationen, Freiburg: Karl Alber, 283–304.

Jacobs, K.A. (2019a): Der zahnlose Vampir. Zur Pathologie der Ausbeutung, in: Brock, E.; Lerchner, T. (Hg.), Denken des Horrors, Horror des Denkens. Erschreckendes, Monströses und Unheimliches in Philosophie, Psychologie und Literatur, Würzburg: Königshausen & Neumann, 99–130.

Jacobs, K.A. (2019b): Einsamkeit aus psychologischer, philosophischer und spiritueller Perspektive. Vortrag und einwöchiger Workshop zur Einsamkeit mit gleichnamigem Titel für die 10. Sommerakademie für Integrative Medizin, Universität Witten/Herdecke.

Jacobs K.A. (2020): Einsamkeit macht krank. In der Welt allein mit anderen sein – philosophische Perspektive auf Einsamkeit, Vortrag gehalten am 18.03.2019 für die Konferenz *Einsamkeit*, Evangelische Akademie Hofgeismar (akademie-hofgeismar.de).

Jacobs K.A. (2022): Loneliness from an interdisciplinary perspective. What does it mean to be human? Vortrag gehalten am 14.11.2022 am Center of Artificial Intelligence, Human Nature and Neurosciences, Hokkaido University, Japan.

Jacobs K.A. (2023a): (Nothing) Human is Alien – AI companionship and Loneliness, in: Possati, L.M. (Hg.), Humanizing Artificial Intelligence. Psychoanalysis and the Problem of Control, Berlin/Boston: De Gruyter, 51–70.

Jacobs, K.A. (2023b): Changes of Intuition in Paranoid Personality Disorder, in: *Frontiers of Psychiatry*, Volume 14: 1307629 [Research Topic: Women in Psychiatry].

Jacobs K.A. (2024): Review: Digital Loneliness. Changes of social recognition through AI Companions, in: *Frontiers in Digital Health*, Volume 6: 1281037.

Jacobs, K.A.; Uhle, C. (2019): Vernetzt und doch allein. Podiumsdiskussion im Literaturforum im Brecht-Haus in der Veranstaltungsreihe *Netzdialoge! Philo-*

sophie des Digitalen. [https://lfbrecht.de/event/vernetzt-und-doch-allein/] (Zugriff: 28.05.2024).

Jaeggi, R. (2009a): Was ist eine (gute) Institution?, in: Forst, R.; Hartmann, M.; Jaeggi, R.; Saar, M. (Hg.), Sozialphilosophie und Kritik, Frankfurt a.M.: Suhrkamp, 528–545.

Jaeggi, R. (2009b): Was ist Ideologiekritik?, in: Jaeggi, R.; Wesche, T. (Hg.), Was ist Kritik?, Frankfurt a.M.: Suhrkamp, 266–298.

Jaeggi, R. (2014): Kritik von Lebensformen, Berlin: Suhrkamp.

James, W. (1884): What is an Emotion?, in: *Mind*, 9(34), 188–205.

Janzarik, W. (1973): Über das Kontaktmangelparanoid des höheren Alters und den Syndromcharakter schizophrenen Krankseins, in: *Der Nervenarzt*, 44(10), 515–526.

Jedan, C. (2012): Hierocles' Ethics – (I.) Ramelli Hierocles the Stoic. Elements of Ethics, Fragments, and Excerpts. Translated by David Konstan, in: *The Classical Review*, 62(2), 426–428.

Jerusalem, M.; Lohaus, A; Klein-Heßling, J. (Hg.) (2006): Gesundheitsförderung im Kindes- und Jugendalter, Göttingen: Hogrefe.

Johnston, C. (2022): Ethical Design and Use of Robotic Care of the Elderly, in: *Bioethical Inquiry*, 19(1), 11–14.

Kästner, E. (1959): Gedichte, Zürich: Atrium Verlag.

Kato, T.A.; Shinfuku, N.; Sartorius, N.; Kanba, S. (2017a): Loneliness and Single-Person Households. Issues of Kodoku-Shi and Hikikomori in Japan, in: Okkels, N.; Kristiansen, C.; Munk-Jorgensen, P. (Hg.), Mental Health and Illness in the City. Mental Health and Illness Worldwide, Singapore: Springer.

Kato, T.A.; Teo, A.R.; Tateno, M.; Watabe, M.; Kubo, H.; Kanba, S. (2017b): Can Pokémon GO rescue shut-ins (hikikomori) from their isolated world?, in: *Psychiatry and Clinical Neuroscience*, 71(1), 75–76.

Kernberg, O.F. (5. Auflage 1992): Objektbeziehungen und Praxis der Psychoanalyse, Stuttgart: Klett-Cotta.

Klein, M.; Riviere, J. (1992): Seelische Urkonflikte. Liebe, Hass und Schuldgefühl, Frankfurt a.M.: Fischer.

Koizumi, A.; Amano, K.; Cortese, A; Shibata, K; Yoshida, W.; Seymour, B.; Kawato, M.; Lau, H. (2016): Fear reduction without fear through reinforcement of neural activity that bypasses conscious exposure, in: *Nature Human Behaviour*, 1, 0006. [https://www.nature.com/articles/s41562-016-0006].

Kretschmer, E. (1918): Der sensitive Beziehungswahn. Ein Beitrag zur Paranoiafrage und zur psychiatrischen Charakterlehre, Berlin: Springer Verlag.

Kucklick, C. (2016): Die granulare Gesellschaft. Wie das Digitale unsere Wirklichkeit auflöst, München: Ullstein.

Kurzweil, R. (1993): KI. Das Zeitalter der Künstlichen Intelligenz, München: Hanser.

Lacan, J. (1978): Le Séminaire Livre II, tome 2. Le Moi dans la théorie de Freud et dans la technique de la psychanalyse (1954–1955), Paris: Seuil.

Lalayants, M.; Price, J.D. (2015): Loneliness and Depression or Depression-Related Factors Among Child Welfare-Involved Adolescent Females, in: *Child and Adolescent Social Work Journal*, 32(2), 167–176.

Langlitz, N. (2005): Die Zeit der Psychoanalyse. Lacan und das Problem der Sitzungsdauer, Frankfurt a.M.: Suhrkamp.

Lehmann, H. (1967): Die Einsamkeit des Menschen in der Großstadt, in: Bittner, W. (Hg.), Einsamkeit in medizinisch-psychologischer, theologischer und soziologischer Sicht, Stuttgart: Klett, 188–200.

Maduschka, L. (1932): Das Problem der Einsamkeit im 18. Jahrhundert, insbesondere bei J. G. Zimmermann, Weimar: Alexander Duncker.

Mann, F.; Wang, J.; Pearce, E.; Ma, R.; Schlief, M.; Lloyd-Evans, B.; Ikhtab, S.; Johnson, S. (2022): Loneliness and the onset of new mental health problems in the general population, in: *Social Psychiatry and Psychiatric Epidemiology*, 57, 2161–2178.

Manzeschke, A. (2022): Robots in care. On people, machines, and other helpful entities, in: Rubeis G; Hartmann, K.V.; Primc, N. (Hg.), Digitalisierung der Pflege. Interdisziplinäre Perspektiven auf digitale Transformationen in der pflegerischen Praxis, Göttingen: Vandenhoeck & Ruprecht, 201–210.

McStay, A. (2018): Emotional AI. The rise of empathic media, London: SAGE.

Meyers, C. (2015): The thankless task of cleaning up the aftermaths of lonely deaths in Japan, in: *Japanese Times*, 01.04.2025. [https://www.japantimes.co.jp/news/2015/04/01/national/social-issues/cleanup-crew-hand-spruce-japans-lonely-death-apartments] (Zugriff: 05.10.2023).

Mijuskovic, B.L. (2012): Loneliness in Philosophy, Psychology, and Literature, Bloomington: iUniverse.

Murphy, P.M.; Kupshik, G.A. (1992): Loneliness, Stress and Well-Being. A Helpers Guide, London: Routledge.

Nelson-Becker, H.; Victor, C. (2020): Dying alone and lonely dying. Media discourse and pandemic conditions, in: *Journal of Aging Studies*, 55. [doi: https://doi.org/10.1016/j.jaging.2020.100878].

Nguyen, M.H.; Gruber, J.; Marler, W.; Hunsaker, A.; Fuchs, J.; Hargittai, E. (2022): Staying connected while physically apart. Digital communication when face-to-face interactions are limited, in: *New Media & Society*, 24(9), 2046–2067.

Oppen, D. (1967): Einsamkeit als Last und Bedürfnis, in: Bittner, W. (Hg.), Einsamkeit in medizinisch-psychologischer, theologischer und soziologischer Sicht, Stuttgart: Klett, 104–110.

Orth-Gomer, K.; Rosengren, A; Wilhelmsen, L. (1993): Lack of Social Support and Incidence of Coronary Heart Disease in Middle-Aged Swedish Men, in: *Psychosomatic Medicine*, 55(1), 37–43.

Papoušek, H.; Papoušek, M. (1992): Beyond emotional bonding: The role of preverbal communication in mental growth and health, in: *Infant Mental Health Journal*, 13, 43–53.

Pembroke, S.G. (1971): Oikeiosis, in: Long, A.A. (Hg.), Problems in Stoicism, London: Athlone Press, 114–149.

Peplau, L.A.; Perlman, D. (1982): Perspectives on Loneliness, in: Peplau, L.A.; Perlman, D. (Hg.), Loneliness. A sourcebook of current theory, research, and therapy, New York: Wiley, 1–18.

Picard, R. (1997): Affective computing, Cambridge: MIT Press.

Poschardt, U. (2007): Einsamkeit. Die Entdeckung eines Lebensgefühls, München: Piper Verlag.

Possati, L.M. (2021): The Algorithmic Unconscious. How Psychoanalysis Helps in Understanding AI, London: Routledge.

Price, D.D. (2000): Psychological and neural mechanisms of the affective dimension of pain, in: *Science*, 288(5472), 1769–1772.

Purington, A.; Taft, J.G.; Sannon, S.; Bazarova, N.N.; Taylor, S.H. (2017): »Alexa is my new BFF«. Social Roles, User Satisfaction, and Personification of the Amazon Echo, in: Association for Computing Machinery (Hg.): Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems, New York, 2853–2859. [doi:/10.1145/3027063.3053246].

Reckwitz, A. (2018): Die Gesellschaft der Singularitäten. Zum Strukturwandel der Moderne, Berlin: Suhrkamp.

Rook, K.S. (1984): Promoting social bonding: Strategies for helping the lonely and socially isolated, in: *American Psychologist*, 39(12), 1389–1407.

Rosa, H. (2019): Beschleunigung als Entfremdung, Berlin: Suhrkamp.

Rotenberg, K.J.; MacKie, K. (1999): Stigmatization of social and intimacy loneliness, in: *Psychological Reports*, 84 (1), 147–148.

Russel, S.J. (2019): Human Compatible. Artificial Intelligence and the Problem of Control, New York: Viking Press.

Saito, T. (2013): Hikikomori. Adolescence without End, Minnesota: University of Minnesota Press.

Satorius, M. (2006): Die hohe Schule der Einsamkeit. Von der Kunst des Alleinseins, Gütersloh: Gütersloher Verlagshaus.

Schetsche, M.; Lehmann, K. (2003): Netzwerker-Perspektiven. Bausteine einer praktischen Soziologie des Internet, Regensburg: Roderer.

Schilbach, L.; Timmermans, B.; Reddy, V.; Costall, A.; Bente, G.; Schlicht, T.; Vogeley, K. (2013): Toward a second-person neuroscience, in: *Behavioral and Brain Sciences*, 36(4), 393–414.

Schmalenbach, H. (1919): Die Genealogie der Einsamkeit, in: *Logos. Internationale Zeitschrift für Philosophie der Kultur*, 8, Tübingen: Mohr, 62–96.

Sharevski, F.; Jachim, P.; Treebridge, P.; Li, A.; Babin, A.; Adadevoh, C. (2021): Meet Malexa, Alexa's malicious twin. Malware-induced misperception through intelligent voice assistants, in: *International Journal of Human-Computer Studies*, 149. ht tps://doi.org/10.1016/j.ijhcs.2021.102604].

Siegel, M.; Melpomeni, A. (2020): Sentiment-Analyse deutschsprachiger Meinungsäußerungen, Wiesbaden: Springer Vieweg.

Simmel, G. (2013[1908]): Soziologie. Untersuchungen über die Formen der Vergesellschaftung, Berlin: Suhrkamp.

Skjuve, M.; Følstad, A.; Fostervold, K.I.; Brandtzaeg, P.B. (2021): My chatbot companion. A Study of Human-Chatbot Relationships, in: *International Journal of Human-Computer-Studies*, 149, 1–14.

Sorabji, R. (2018): Animal minds and human morals. The origins of the western debate, Ithaca (NY): Cornell University Press.

Speckhard, A.; Ellenberg, M.; Morton, J.; Ash, A. (2020): Involuntary Celibates' Experiences of and Grievance over Sexual Exclusion and the Potential Threat of Violence among those Active in an Online Incel Forum, in: *Journal of Strategic Security*, 14(2), 89–121.

Spitzer, M. (2019): Einsamkeit, die unerkannte Krankheit: schmerzhaft, ansteckend, tödlich, München: Droemer Knaur.

Srinivasan, S.; O'Fallon, L.R.; Dearry, A. (2003): Creating healthy communities, healthy homes, healthy people. Initiating a research agenda on the built environment and public health, in: *American Journal of Public Health*, 93(9), 1446–1450.

Sticker, M. (2023): Poverty, Exploitation, Mere Things and Mere Means, in: *Ethical Theory and Moral Practice*, 26(2), 191–207.

Takahiro, A.K.; Shinfuku, N.; Sartorius, N.; Kanba, S. (2017): Loneliness and Single-Person Households. Issues of Kodoku-Shi and Hikikomori in Japan, in: Okkels, N.; Blanner Kristiansen, C.; Munk-Jørgensen, P. (Hg.): Mental Health and Illness in the City, Singapur: Springer, 205–219.

Taylor, C. (1979): Atomism, in: Kontos, A. (Hg.), Powers, Possessions and Freedom. Essays in Honour of C.B. Macpherson, Toronto: University of Toronto Press, 39–62.

Thomas, S. (2022): Einsamkeitserfahrungen junger Menschen – nicht nur in Zeiten der Pandemie, in: *Soziale Passagen*, 14, 97–112.

Thurston, R.C.; Kubzansky, L.D. (2009): Women, loneliness, and incident coronary heart disease, in: *Psychosomatic Medicine*, 71(8), 836–842.

Tillich, P. (1963): The Eternal Now, New York: Scribner.

Turkle, S.; Taggart, W.; Kidd, C.D.; Dasté, O. (2006): Relational artifacts with children and elder. The complexities of cypercompanionship, in: *Connection Science*, 18(4), 347–361.

Vanhalst, J.; Klimstra, T.A.; Luyckx, K.; Scholte, R.H.; Engels, R.C.; Goossens, L. (2012): The interplay of loneliness and depressive symptoms across adolescence.

Exploring the role of personality traits, in: *Journal of Youth and Adolescence*, 41(6), 776–787.

Weiss, R.S. (1987): Reflections on the present state of loneliness research, in: *Journal of Social Behavior & Personality*, 2(2), 1–16.

Wilson, R.S.; Krueger, K.R.; Arnold, S.E.; Schneider, J.A.; Kelly, J.F.; Barnes, L.L.; Tang, Y.; Bennett, D.A. (2007): Loneliness and risk of Alzheimer disease, in: *Archives of General Psychiatry*, 64(2), 234–240.

Winnicott, D.W. (1953): Transitional objects and transitional phenomena, in: *International Journal of Psychoanalysis*, 34, 89–97.

Winnicott, D.W. (1958): The capacity to be alone, in: *International Journal of Psychoanalysis*, 39(5), 416–420.

Winnicott, D.W. (1964): The Child, the Family, and the Outside World, Harmondsworth: Penguin.

Winnicott, D.W. (1971): The use of an object and relating through identificationism, in: Ders. (Hg.), Playing and reality, London: Penguin, 86–94.

Wood, M.M. (2020): Paths of Loneliness. The Individual Isolated in Modern Society, New York: Columbia University Press.

World Health Organization (2000): World Report on Ageing and Health. A Policy Framework for Healthy Aging.

Xie, T.; Pentina, I. (2022): Attachment Theory as a Framework to Understand Relationships with Social Chat-bots. A Case Study of Replika, in: Proceedings of the 55th Hawaii International Conference on System Sciences. [doi: http://dx.doi.org/10.24251/HICSS.2022.258].

Zimmermann, J.G. (2016[1784]): Über die Einsamkeit [Nachdruck der Ausgabe von 1784], Norderstedt: Hansebooks GmbH.

Hass, Wut und Zorn

Beobachtungen zum Imageboard 4chan/pol*

Kai Denker

Abstract: *The article focuses on the non-mainstream imageboard 4chan, and in particular its board 4chan/pol, which is notorious for politically incorrect communication. It analyzes how technical and social affordances such as volatility, anonymity, and transgressive »humor« facilitate the emergence and expression of aggressive emotional states, and how this media configuration is exploited for political purposes, primarily to promote right-wing extremist ideologies. The article examines the intentional staging of anger and its political exploitability. Two case analyses support the hypothesis that concepts of aggressive emotional states such as resentment, indignation, anger, and hate fail to address the political narratives on 4chan/pol, while rage, with its normative and political underpinnings, allows for an analysis of right-wing extremist narrative strategies. Conclusion: The term »hate speech« is inadequate for categorizing extremism prevention, regulation, and digitization research on online hostility. The new fascism on the Internet is not the result of unregulated Internet companies; rather, it is deliberately orchestrated in niches by exploiting specific affordances.*

Keywords: *4Chan/pol; group-focused misanthropy; emotional work; Red Pill(ing); right-wing extremism*

Hass hat sich im Internet breit gemacht und wurde von dort auch offline wirksam. Diese zum Gemeinplatz gewordene Diagnose speist sich nicht zuletzt aus Belästigungen und Bedrohungen prominenter Politiker*innen sowie von Amts- und Mandatsträger*innen (vgl. Riemenschneider/Lutz 2021: 374). Hatten Demmerling und Landweer (Demmerling/Landweer 2007: 299) noch eine Tabuisierung von direktem Hass diagnostiziert, haben mittlerweile die Demonstrationen gegen die Corona-Politik »Ausmaß und Schärfe der Feindseligkeit […] spürbar« und öffentlich sichtbar

* Dieser Beitrag entstand im Rahmen des BMBF-geförderten Verbundvorhabens »Meme, Ideen, Strategien rechtsextremistischer Internetkommunikation (MISRIK)«.

gemacht (Brockhaus 2020: 87f.). Zunehmend dringt daher ins Bewusstsein, dass von den ›Hass-Maschinen‹ im Internet eine Spur zu neuen Formen des Terrorismus führt (vgl. Wentz 2019: 3). Ziel dieses Beitrags ist es erstens, diese Spur nachzuzeichnen – genauer: die Verbindung von technischen Eigenschaften von Internet-Plattformen – hier am Beispiel des Imageboards 4chan – zu einer an faschistische Tatgemeinschaften (vgl. Reichardt 2014) erinnernden Lust an Hass, Wut, Zorn, … im Netz. Diese Lust drückt sich aus in Körpertechniken, in Gemeinschaftserleben und in einem kruden Humor. Zweitens soll aber die Rede vom Hass differenziert werden: Es ist nicht einfach Hass, der als bloß technisch begünstigter Affekt einer verrohten Kommunikationskultur auftritt. Vielmehr haben wir es mit einem strategisch befeuerten und orchestrierten Geschehen zu tun (Riemenschneider/Lutz 2021: 371), in dem Aggressionsaffekte eine politische – nämlich im Regelfall rechtsextreme[1] – *Umcodierung zu Zorn* erfahren. Kurz: Hass und Wut werden gesät und als politisch verwertbar geerntet – von »Hassunternehmern«, wie Eisenberg sie nennt (Eisenberg 2020: 656f.). Die Umcodierung reproduziert nicht einfach Hass, sondern sie produziert Zorn als einen dienstbar gemachten, moralisch markierten Affekt. Der rechte Zorn im Netz bedient dabei bekannte Narrative: ›der Westen‹ und ›die weiße Rasse‹ seien dekadent und im Niedergang begriffen. Sie müssten durch eine Art Wiedergeburt gerettet und erneuert werden. Man fühlt sich an Griffins Definition des Faschismus durch palingenetische Vorstellungen erinnert (Griffin 2020: 83). Im Netz wird diese Definition des Faschismus durch den Rechtsakzelerationismus komplementiert. Dabei handelt es sich um eine Gruppe von Strategien, die – teils gewaltsam, teils nur sekundierend – in einer perversen Dialektik auf den Systemumsturz zielt, ja, diesen beschleunigt herbeiführen will. Der rechte Zorn im Netz ist keine Spielerei, wie eine naive technik- und kulturpessimistische Position nahelegen könnte. Er zielt auch nicht nur darauf, Gegner*innen ›mundtot‹ zu machen (Riemenschneider/Lutz 2021: 373) und so Zustimmung einer schweigenden Mehrheit zu simulieren (Brockhaus 2020: 90). Es geht stets um *Gewalt*, die sich im Extremfall in ›Amokläufen‹ aktualisiert. ›Amokläufe‹ deshalb in Anführungszeichen, da Sicherheitsbehörden das Phänomen lange heruntergespielt und zu Taten einzelner Wirrköpfe erklärt haben. Dass es sich aber um eine neue Form von Terrorismus handelt, wird immer deutlicher (Eisenberg 2020: 657). Die bekannte, ›intellektualistische‹ Strategie der Metapolitik (Strick 2021: 80), also die Verschiebung von Werten innerhalb eines vorpolitischen Raums, ist dabei nicht verschwunden. Die ›Neue Rechte‹ hat aber dazugelernt: Der Rechtsakzelerationismus tritt auf als ›stochastischer Terrorismus‹, d.h. in Form von fluiden Netzen, die eher aus Diskursen denn aus persönlichen Treueverhältnissen innerhalb irgendwelcher Kampf-

1 Es soll nicht unterschlagen werden, dass es auch jenseits des rechten Randes ähnliche Strategien gibt, wenngleich in deutlich geringerem Umfang und ohne dieselbe Wirkung, wie Rothbart herausstellt (Rothbart 2021: 681).

gruppen bestehen. Gewalt wird nicht wie im ›klassischen‹ Terrorismus durch einen Führungskader angeordnet, sondern im Rauschen und Gemurmel des Netzes bloß ›wahrscheinlicher‹. Ziel bleibt der Umsturz ›des Systems‹ und die ultra-nationale[2] Wiedergeburt. Beides nimmt man nun gleich selbst in die Hand, anstatt es nur in pseudointellektuellen Zirkeln am Kamin zu besingen. Mussolinis Erklärung des Faschismus als Tat lässt grüßen (Griffin 2020: 101), und die Gleichung ›Faschismus = Gewalt‹ behält ihre Gültigkeit – allerdings unter nun *auch* digitalen Verhältnissen.

Hass und Zorn im Netz werden erst in den letzten Jahren auch in der Forschung mit Merkmalen des Faschismus charakterisiert (Fielitz/Marcks 2020; Griffin 2020: 174). Dies hängt mit der hohen Fluidität rechtsextremer Phänomene im Netz zusammen, die sich selten in klar abgrenzbaren und als solche erkennbaren ›Neonazi-Gruppen‹ zeigen. Diskursive Phänomene können sich nichtsdestoweniger in Gewalt aktualisieren. Vordergründig bescheidener ist die Charakterisierung als Hatespeech, also die »Verrohung der Sprache sowie eine ausgeprägte verbale Radikalität […], die in Form von Postings im Internet verbreitet wird«, was »eine Mobilisierung, Radikalisierung und Vernetzung innerhalb der rechtsextremen Szene« sichtbar macht (Horten/Gräber 2021: 92). Dennoch hat die Rede vom Hatespeech das Phänomen lange individualisiert und sich beispielsweise auf die bloße Erkennung von Hass-Postings verengt, ohne Strategien der Verbreitung und Umcodierung des Hasses zu diskutieren (Hine et al. 2017: 94). Auch die Verantwortung von Plattformbetreiber*innen wurde lange ausgeblendet (Gillespie 2018: 37) und auf deren Kooperation mit Strafverfolgungsbehörden vereinseitigt, kurz: Man hat den Hass im Netz begrifflich verengt und lediglich »verwaltet« (Ceffinato 2020: 561). Damit wurde der Hass im Netz in erster Linie ein Problem des Ehrschutzes und des öffentlichen Friedens, also letztlich des Strafrechts und somit der Strafverfolgung, nicht der zivilen Sicherheit oder der Kriminal- bzw. Extremismusprävention (Ceffinato 2020: 554).

Jenseits der pessimistischen Zeitdiagnostik und der verführerischen Metaphorik zeigt die kurze Einführung bereits die schwierige Begriffsverwendung an: Nicht nur in technikphilosophischer Hinsicht ist die Beschreibung der einschlägigen Kommunikationsplattformen nicht trivial, auch die Abgrenzung der in Frage kommenden Aggressionsaffekte – Hass, Wut (oder Ärger), Zorn, … – verkompliziert sich, je genauer man hinsieht: Erstens treten Aggressionsaffekte im Netz meist in einem humoristischen Gewand und gespickt mit Doppeldeutigkeiten auf, so dass ihre politische Dimension oft nicht erfasst wird und sie wohl gerade auch deshalb zur Rekrutierung ›von rechts‹ taugen (Doğru 2020: 31).[3] Zweitens ist der

2 Die vergleichende Faschismusforschung hat herausgearbeitet, dass Faschismus nicht auf eine Wiedergeburt klassischer Nationalstaaten zielt, sondern auf die Wiedergeburt einer imaginierten Ultra-Nation, was supranationale Allianzen gerade nicht ausschließt. Vgl. Griffin 2020: 82.

3 Auch für den historischen Faschismus bleibt (bösartiger) Humor noch immer untererforscht.

für die Philosophie der Gefühle wichtige Leib abwesend und tritt in einer rechten *Körper*technik nur indirekt auf. Drittens sind die soziale Funktion (Szanto 2020: 456) und ihre moralische Dimension – nicht zuletzt im Sinne eines gerechten Zorns (Coleman 2015: 111) – unübersehbar. Hinzu kommt: Anders als Hass, Wut und Ärger ist der Zorn ein »anerkannter philosophischer Gegenstand«, dies aber dank seiner moralischen, weniger wegen seiner aggressiven Dimension (Demmerling/Landweer 2007: 287f.). Die digitalen Aggressionsaffekte, wie sie uns beispielsweise auf 4chan begegnen, unter dem Begriff des Zorns zu verhandeln, ist also begründungsbedürftig.

Das 2003 von Christopher Poole begründete und 2015 an Hiroyuki Nishimura verkaufte Imageboard 4chan (s. Hine et al. 2017: 92), das in diesem Beitrag in den Blick genommen wird, gehört nicht zu den großen und bekannten Sozialen Medien, die meist im Zentrum kritischer Auseinandersetzungen mit der Netzkultur stehen. Die meisten Kritiker*innen der Netzkultur sind tatsächlich Zaungäste, und sie beschäftigen sich allzu oft nur mit Phänomenen und Plattformen, von denen sie lediglich im Feuilleton gelesen haben, um dann mit großen Pinselstrichen irgendeinen neuen digitalen Strukturwandel zu diagnostizieren. Man ist dann schon glücklich, wenn es um die Rolle von Algorithmen geht, auch wenn dieser technische Begriff oft eher miss- als gebraucht wird.[4] Immerhin ist zuzugestehen, dass Algorithmen mitunter bestimmte Inhalte bevorzugen und damit als oft unkontrollierte, technische Kuratoren funktionieren (Martin/Vukadinović 2020: 106; Boulianne/Lee 2022: 32). 4chan unterscheidet sich hier fundamental: Es gibt lediglich eine basale algorithmische Inhaltskuratierung, was den Fokus zurück auf Nutzungskulturen verschiebt. Wenig überraschend spielen auch im Fall von algorithmisch ›basalen‹ Plattformen Affordanzen eine Rolle und allein hierdurch werden bestimmte Inhalte – und sei es bloß der Form nach – bevorzugt. Die Neigung ›nach rechts‹, die so eine Bevorzugung der Form nach beispielsweise auf 4chan offenbar hervorbringt, ist erklärungsbedürftig, und zwar ohne einen *algorithmic bias*, wie er in der Digitalisierungsforschung in erster Linie anhand von avancierten algorithmischen Kuratierungssystemen – in der Regel aus dem Bereich des *machine learning* – diskutiert wird. Die These hier lautet, dass 4chan und dort das Board /pol (für ›Politically Incorrect‹, kurz: 4chan/pol) ein Beispiel für so ein basales System ist, auf dem bevorzugte Inhalte mit aggressiven Affekten verbunden sind und gezielt – d.h. zur Durchsetzung von Interessen (Riemenschneider/Lutz 2021: 371) – zu Zorn umcodiert werden können. Die mit aggressiven Affekten verbundenen Inhalte auf 4chan/pol sind entspre-

4 Die Diskussion zu den algorithmischen Eigenschaften digitaler Kommunikationsplattformen hat den technischen Begriff des Algorithmus verunklart, sogar um ausschließlich Verfahren des maschinellen Lernens zu beschreiben. Auch ohne maschinelles Lernen sind alle informationstechnischen Systeme zugleich algorithmisch – und als solche sind sie auch zu problematisieren.

chend oft mit gruppenbezogener Menschenfeindlichkeit verbunden, die hier offen ausgelebt werden kann. Das erinnert nicht nur der Form nach an zentrale Motive faschistischer Affektpolitiken und Gemeinschaftspraktiken gegen einen imaginierten Feind (Brockhaus 2020: 94f.). Voraussetzung für den Erfolg rechter Affektpolitiken online und auf 4chan im Besonderen ist, dass ihre Inhalte in der Regel nicht offen und nicht ohne Hintergrundwissen verstehbar sind. Mehr noch: Sie können weiterverbreitet werden, ohne dass sich die Verbreiter*innen dem bewusst sind – es reicht schon, dass sie den geteilten Inhalt witzig oder vielleicht empörend finden (Martin/Vukadinović 2020: 109f.).

4chan/pol – die dunkle Ecke des Internets

4chan hat sich zu einem so beliebten wie anstößigen Imageboard entwickelt (Coleman 2015: 41). Gegründet wurde 4chan zur Pflege japanischer Anime-Kultur, die noch immer einen wichtigen Teil seiner Inhalte bildet. Bekannt wurde es aber letztlich für seine Anonymität, d.h. dem Fehlen re-identifizierbarer Nutzerprofile (Coleman 2020: 148), für die aggressiven Versuche seiner Nutzer*innen, ihre eigene ›Netzkultur‹ – etwa mittels ›Trolling‹[5] – zu verbreiten, und natürlich für seine Rolle bei der Entstehung des Hacker-Kollektivs Anonymous, womit es schließlich zu einer ›dunklen Ecke‹ des Internets avancierte (Colley/Moore 2020: 4). Tatsächlich gilt 4chan als Quelle für erhebliche Teile der Netzkultur, etwa von bekannten Memen wie *lolcat* (Hine et al. 2017: 92; Coleman 2015: 44). Das gilt allerdings wohl eher für 4chan/b als für 4chan/pol (Bernstein et al. 2011: 56). 4chan ist in sogenannte Boards organisiert, die ähnlich wie Bulletin-Boards 4chan in thematisch getrennte Bereiche, wie z.B. /mu für Musik, aufteilen (Wagener 2017: 304). /b steht für ›random‹, also für praktisch jeden beliebigen Inhalt, und es hat zeitweise mehr Beiträge gezählt als alle anderen Boards auf 4chan zusammen (Herwig 2011). Entsprechend wurde /b medial zeitweise aufmerksamer verfolgt und auch öfter wissenschaftlich untersucht als /pol (Hine et al. 2017: 93).

Trotz der medialen Aufmerksamkeit auf 4chan diagnostizieren Hine et al. aber 2017 (Hine et al. 2017: 92) noch immer das Fehlen systematischer Untersuchungen. Tatsächlich finden sich schon deutlich früher insbesondere ethnologische Studien zu 4chan, jedoch vergleichsweise wenige quantitative Studien – Ausnahme: Bernstein et al. 2011 –, die etwa Nutzungsdynamiken oder die Verbreitung von Hatespeech vermessen. Unbefriedigend bleiben auch die wenigen diskursanalytischen Studien, die auf Nutzungskulturen und Sinngebungsprozesse abzielen (Herwig 2011). Dies dürfte zunächst im Wesentlichen darauf zurückzuführen sein, dass die

5 Unter ›trolling‹ versteht man Strategien vorsätzlicher, wiederholter und oft bösartiger Provokationen online, um etwaige Reaktionen der Opfer humoristisch auszuschlachten.

Ethnologie des Digitalen sich auf das Hacker-Kollektiv *Anonymous* konzentriert hat, das zwar 4chan entstammt, sich aber um das Jahr 2011 von 4chan löste (Colley/Moore 2020: 5). Qualitative Studien zu 4chan untersuchen insbesondere den Jargon, die Verwendung von Humor und Ironie sowie Sozialisierungsprozesse (Wagener 2017: 304). Beispielsweise können unerfahrene Nutzer*innen, die mit den sich schnell entwickelnden sprachlichen Codes noch nicht vertraut sind, mit ihren Postings meist nicht reüssieren und werden selbst zur Zielscheibe. Im besten Fall werden sie aufgefordert, erst einmal mehr mitzulesen (»lurk moar«), um den Jargon und die akzeptierten Verhaltensweisen zu erlernen. Derartige Sozialisierungsprozesse sind seitens der digitalen Ethnologie untersucht worden. Viele Fragen bleiben aber ungeklärt, etwa zu den Nutzer*innen, ihren Beziehungen untereinander und Verschiebungen ihrer politischen Positionsnahmen (Colley/Moore 2020: 5). Einige Aufmerksamkeit haben die 4chan-typischen Affordanzen Anonymität und Flüchtigkeit erfahren (Bernstein et al. 2011: 50), wenngleich plattformspezifische Effekte – gerade im Vergleich zu Untersuchungen an Mainstream-Plattformen wie Twitter und Facebook – noch immer unterkritisch thematisiert werden (Boulianne/Lee 2022: 31). Zu fragen wäre hier insbesondere, wie die Plattform selbst gegenüber ihren Nutzer*innen eine Agentialität entwickelt haben (Bucher/Helmond 2018: 249). Am Rande Aufmerksamkeit haben forschungspraktische Erwägungen erfahren, wozu einerseits Belastungssituationen für Forschende im Umgang mit den teils extremen Inhalten von 4chan ebenso gehören (Brockhaus 2020: 97) wie die Beobachtung, dass 4chan-Nutzer*innen sich gegen ihre Erforschung wehren, indem sie Forschungsansätze beispielsweise durch Verschiebungen in sprachlichen Codes zu unterlaufen versuchen (Colley/Moore 2020: 19).

Inkubator für rechte Bewegungen

Schon bei flüchtiger Beobachtung fallen auf dem praktisch unmoderierten 4chan/pol rechtsaffine Diskurse sowie ›Hatespeech‹ auf (Hine et al. 2017: 93; Colley/Moore 2020: 13).[6] /pol und 4chan im Allgemeinen sind für die Netzkultur und besonders für die Internet-Ästhetik ein Versuchslabor, etwa für Internet-Meme (Douglas 2014: 315), also für multimodale Text-Bild-Arrangements,[7] die in der Regel hochgradig repetitiv und intertextuell Versatzstücke u.a. der Popkultur zusammenführen und die eine eigene Ästhetik und Sprache aufweisen. Internet-Meme sind praktisch immer humorvoll lesbar (Jäger et al. 2021: 19) und können

6 Hine et al. haben in 12% der Postings auf /pol hassbezogene Signalwörter entdeckt – im Unterschied zu 2,2% der Postings auf anderen Plattformen. Konkret das N-Wort werde etwa 120-mal pro Stunde verwendet (Hine et al. 2017: 98).

7 Internet-Meme können auch als Videos, als Bild ohne Text oder auch als reiner Text auftreten.

auch emanzipativ etwa zur Bildung von Gruppenidentitäten von Minderheiten beitragen, sind aber zunehmend Gegenstand der Kritik geworden: Internet-Meme erzählen nicht nur Geschichten oder Witze, sondern wirken persuasiv und dienen der Mobilisierung – auch ›von rechts‹ (Hakoköngäs et al. 2020: 1). Sie zielten dabei, Hakoköngäs et al. weiter, auf Affekte, nicht auf rationale Argumente. Die hier implizierte Position, dass Emotionen der Rationalität gegenüberstehen, lässt sich nicht halten, wird aber meines Erachtens von Hakoköngäs et al. so auch nicht vertreten. Ich werde später darauf zurückkommen. Internet-Meme funktionieren in rhetorischer Hinsicht enthymetisch, d.h. sie sind in der Interpretation gleichzeitig unter- und überbestimmt: Die in ihnen vorkommenden Symbole und Verweise sind zitathafte Ausschnitte, die ähnlich einer Collage miteinander verbunden werden. Symbole und Verweise bringen ihre eigenen Entstehungs- und Bedeutungskontexte mit, die für die Interpretation des Internet-Mems gekannt werden müssen, wodurch sie das Mem überdeterminieren. Umgekehrt lassen Internet-Meme Leerstellen, durch die ihre Interpretation uneindeutig wird, was zugleich ermöglicht, dass arglose Nutzer*innen Meme verbreiten, die z.B. bei Kenntnis der verwendeten Symbolik eindeutig als rechtsextremistisch identifizierbar wären. Aber auch diesseits der Schwelle zum Rechtsextremismus können Meme in einer Interpretation als harmloser Witz, in der anderen als entwürdigender Angriff auf Personen oder Gruppen gelesen werden (Martin/Vukadinović 2020: 107). Die Verbreitung von Memen besteht dabei nicht nur in ihrer simplen Wiederholung im Sinne einer perfekten Digitalkopie, sondern im Remixing, d.h. in ihrer veränderten Wiederholung. Kurz: Internet-Meme weisen nicht bloß eine Verbreitungsdynamik, sondern insbesondere eine Entwicklungsdynamik auf, was 4chan, wie wir gleich sehen werden, zur idealen Plattform für memetische Kommunikationsstrategien macht (Herwig 2011; Wentz 2019: 8) und hier auch spezifische, memetische Ästhetiken hervorbringt, die sich aus den Nutzungsmöglichkeiten von 4chan ergeben (Douglas 2014: 334).

Free speech is non-negotiable

Die Nutzungskultur auf 4chan stellt stark auf freie Rede ab: »free speech is non-negotiable« (Coleman 2020: 147). Die Offenheit auch für extreme Themen steht im engen Zusammenhang mit einer Kultur der Anonymität (Coleman 2015: 41), in der freie Rede eine der wenigen durchgehend geteilten politischen Positionen darstellt (Colley/Moore 2020: 4) und in engem Zusammenhang zu einer Kultur der permanenten Überschreitung der Grenzen des Sagbaren steht (Herwig 2011). 4chan ist in enger Verbindung zur US-amerikanischen Kultur des *libertarianism* zu sehen, der sich auf die ersten beiden Verfassungszusätze konzentriert. Der erste Zusatzartikel zur US-Verfassung setzt der gesetzlichen Regulation der freien Rede

bekanntlich enge Grenzen (Eisenberg 2020: 655), so dass auch die Hatespeech-Gesetzgebung nicht weit entwickelt ist und sich eher auf Hasskriminalität bezieht (vgl. Eisenberg 2020: 644): »Der hohe Stellenwert der freien Rede schließt eine Intervention etwa in Fällen aus, in denen eine Person hetzerische, vorurteilsgeleitete Aussagen online gegen eine bestimmte Gruppe veröffentlicht und dabei auch Gewalt gegen Mitglieder dieser Gruppe gutheißt.« (Eisenberg 2020: 646) Rückblickend verwundert es nicht, dass auch eine ursprünglich eher links-liberale Plattform wie 4chan eine diskursive Radikalisierung ›nach rechts‹ durchgemacht hat und reaktionäre Positionen eines ›Kulturkampfes‹ gegen ›*Social Justice Warriors*‹ aufgenommen hat: Nicht nur wird linken Positionen unterstellt, sie seien gegen die freie Rede gerichtet, sondern insbesondere sind sie mit der oft gegen Minderheiten und Schwächere gerichteten ›Trollkultur‹ von 4chan unvereinbar (Colley & Moore 2020: 4). Die diskursive Entwicklung von 4chan entspricht aber nicht nur einfach der zunehmenden Polarisierung der US-amerikanischen politischen Kultur, sondern befeuert diese aktiv. 4chan gilt als Brutstätte für die US-amerikanische *alt-right*, einer eng mit der Internetkultur verknüpften rechtsextremen Bewegung, auf die die Öffentlichkeit wohl vor allem im Zusammenhang mit dem Wahlkampf Donald Trumps 2016 aufmerksam wurde (Elley 2021: 2). Der britische Faschismusforscher Roger Griffin spricht in diesem Zusammenhang gar von Cyberfaschismus (Griffin 2020: 175f.). Obgleich hiermit die politische und kulturelle Heterogenität der 4chan-Nutzer*innen, die vielleicht nicht alle, aber wohl doch mehrheitlich aus dem US-amerikanischen Raum stammen (Bernstein et al. 2011: 54; Hine et al. 2017: 94f.), sowie die Heterogenität der verschiedenen Boards auf 4chan unterschätzt zu werden droht (Colley/Moore 2020: 13), ist die inhaltliche und kommunikationsstrategische Passung zwischen 4chan/pol und *alt-right*, ja rechtsextremen Bewegungen überhaupt, nur schwer zu übersehen. 4chan/pol liefert unter dem Deckmantel transgressiven, memetischen Humors und unbeschränkter freier Rede gewissermaßen den weitgehend unmoderierten Rückzugsraum für rechtsextreme Positionen, Empörungs- und Zornkulturen, Verschwörungsnarrative und andere klassisch rechte Topoi wie eine Skepsis gegenüber akademischer Bildung und vermeintlich linken Institutionen (Elley 2021: 2f.). Zu nennen ist hier insbesondere die um 2017 auf 4chan entstandene und bald darauf auf das Imageboard 8chan ausgewichene QAnon-Verschwörungsbewegung, die im Zuge der Erstürmung des US-Kapitols am 6. Januar 2021 auch international größere Aufmerksamkeit erfuhr. In dieser sich pseudoreligiös präsentierenden, antisemitischen Bewegung, die sich längst über Imageboards hinaus auch auf Plattformen wie Telegram ausgebreitet hat, gelten die wirren Sentenzen (›Q drops‹) eines anonymen ›Q‹, der angeblich Zugriff auf streng geheime Regierungsunterlagen habe,[8] als Verkündung

8 ›Q‹ bezieht sich offenbar auf die ›Q clearance‹ des US-amerikanischen Energieministeriums, die ungefähr der deutschen Einstufung ›streng geheim‹ entspricht.

einer von Eliten unterdrückten Wahrheit (Hodge/Hallgrimsdottir 2020: 14). Die kommunikative Heterogenität dieser Szene, die zwischen pseudoreligiösem – pardon – Bullshit, memetischen Humor, libertären und – hier nur anzudeutenden – pseudointellektuellen Debatten schwankt, hat sich zu einer auf diskursive Massenbewegungen setzenden rechtsextremen Strategie entwickelt (Munn 2019: 151), die mit den rechtsakzelerationistischen Mordanschlägen der letzten Jahre zur Tat geworden ist (Colley/Moore 2020: 5; Doğru 2020: 16).

Anonymität und Flüchtigkeit

Die Eingangsthese des Beitrags, dass die Dynamiken von Hass, Wut und Zorn auf 4chan/pol ebenso sehr mit extremistischen Strategien wie mit den technischen Eigenschaften von 4chan zu erklären sind, erfordert, die Affordanzen der Plattform zu beschreiben. Die Darstellung konzentriert sich auch hier auf 4chan als einzelne Plattform und unterschlägt damit, dass 4chan Teil eines ganzen Ökosystems von Internet-Plattformen ist, die mit jeweils unterschiedlichen Affordanzen und entsprechenden Nutzungskulturen miteinander interagieren. Dies zeigt sich nicht zuletzt an auf 4chan geplanten ›raids‹ auf andere Plattformen – sei es konzertiertes ›Trolling‹ oder technisch nur wenig anspruchsvoll DDoS-Angriffe (Coleman 2015: 44; Hine et al. 2017: 92).

4chan ist ein Imageboard: Eine Nutzer*in erstellt einen neuen Beitrag (ein ›Posting‹), indem sie ein Bild hochlädt und ggf. mit einem Kommentar versieht. Das Posting ist nicht mit einem Nutzerprofil verknüpft, und bekannte Interaktionsformen wie ›share‹ und ›like‹ fehlen (Hine et al. 2017: 96). Es ist lediglich möglich, auf Postings mit weiteren Postings in Form von Kommentaren (mit und ohne Bild) zu reagieren. Diese Postings zusammen bilden einen Thread. Postings erhalten eine Identifikationsnummer (ID) mittels derer es möglich ist, innerhalb von Postings auf andere zu verweisen (Hine et al. 2017: 92). Auf 4chan/pol werden die Postings zusätzlich mit einer Landesfahne als Icon versehen, welche anhand der IP-Adresse der Nutzer*in eine grobe geographische Zuordnung erlaubt und offenbar mit den jeweils verhandelten Themen korreliert (Hine et al. 2017: 98). Postings und Kommentare sind zumindest auf /pol zudem mit einer Poster-ID versehen, die zwar nicht die langfristige Zuordnung zu Nutzer*innen erlaubt, es aber zumindest gestattet, die Kommentare derselben Nutzer*in einem Thread zuzuordnen (Hine et al. 2017: 93). Zudem wird mit einfachen, d.h. leicht zu umgehenden Mitteln verhindert, dass Nutzer*innen direkt auf eigene Postings oder Kommentare reagieren (Herwig 2011). Mittels eines ›*tripcode*‹ können sich Nutzer*innen in Grenzen wiedererkennbar machen und einen Nutzernamen wählen, was jedoch kaum genutzt wird (Bernstein et al. 2011: 53). Davon abgesehen hält 4chan keine Funktionen vor, die Anonymität der Nutzer*innen zu gewährleisten (Coleman 2015: 42f.). Die berühmte Anony-

mität von 4chan basiert also lediglich auf der Abwesenheit expliziter Nutzerprofile, während 4chan selbst die IP-Adressen der Nutzer*innen erfassen und verarbeiten kann. Dennoch ist eine Kultur der Anonymität entstanden, die von Anfang an auch eine gegen Autoritäten gerichtete Haltung begünstigt hat (Hannan 2018: 219). In der Literatur wird die Auswirkung von Anonymität kontrovers diskutiert: Sie könne zum einen die Kreativität des Austauschs erhöhen, da sie Hierarchien beseitigt und einen offenen Umgang fördert, ohne persönliche Bindungen zu erzeugen (Bernstein et al. 2011: 51), zum anderen impliziere dies das Fehlen sozialer Kontrollmechanismen (Herwig 2011), was enthemme und nicht zuletzt zu einer Atmosphäre aggressiven Humors und der Lust am Verbotenen führe (Martin/Vukadinović 2020: 107). Dem Vertrauensverlust durch anonyme Kommunikation begegnet 4chan mit einer insbesondere auf Bildern und Sprachstilen basierten Nutzungskultur, die es zwar nicht ermöglicht, einzelne Nutzer*innen wiederzuerkennen, aber dennoch auf Basis von Ausdrucksstilen eine Art virtuelle Gruppenidentität zu erzeugen, der gegenüber neue und uneingeweihte Nutzer*innen schnell auffallen (Bernstein et al. 2011: 51). Die US-amerikanische Ethnologin Gabriella Coleman macht ein Zusammenspiel der Funktion der Anonymität in Form reinen Wettbewerbs ohne soziales Kapital und dem Effekt der Affordanz der Anonymität, etwa im Fall memetischer Kommunikation, aus (Coleman 2015: 45f.): 4chan ist in dem Sinne eine radikal inhaltsbezogene Aufmerksamkeitsökonomie ohne Reputationserwerb par excellence (Dunn Cavelty/Jaeger 2015: 183).

Die Unmöglichkeit, eine langfristige Nutzeridentität und entsprechende Reputation aufzubauen, hängt eng mit einer zweiten Affordanz zusammen: der Flüchtigkeit. Anders als viele andere Internet-Plattformen speichert 4chan Threads nicht langfristig. Wird ein Board wie /pol aufgerufen, erscheinen die Postings in der Reihenfolge der letzten Interaktion mit ihnen. So wird ein neu erstellter Thread zunächst an den Anfang einsortiert, dort aber bald durch andere Threads verdrängt. Sobald mit einem Thread (durch Kommentar) interagiert wird, rutscht er wieder an die erste Stelle. Dies produziert zunächst einen Matthäus-Effekt: Erfolgreiche Threads erscheinen am Anfang, werden so als erstes gesehen und haben daher eine höhere Wahrscheinlichkeit, dass eine andere Nutzer*in mit ihnen interagiert. Threads, mit denen nicht oder wenig interagiert wird, rutschen schnell nach unten und haben hierdurch geringere Chancen. Threads, die ans Ende eines Boards gelangen, werden unwiederbringlich gelöscht. Zugleich wird die Lebenszeit von Threads weiter beschränkt, indem spätere Interaktionen den Thread nicht mehr an die erste Stelle des Boards springen lassen. Innerhalb von 4chan ist es also nicht möglich, sich auf ältere, da gelöschte Threads zu beziehen. Threads werden jedoch auf Webseiten wie archive.4plebs.org gespeichert und lassen sich so untersuchen.

Annäherungen an Hass, Wut und Zorn

Die Philosophie der Gefühle hat über die Alltagssprache hinaus Unterscheidungsmerkmale der Klassen affektiver Phänomene entwickelt. Die Differenzierung gegenüber der Alltagssprache erschwert die Übertragung philosophischer Unterscheidungen auf die explizite Verhandlung von Affekten auf einer Plattform wie 4chan/pol, auch wenn sie sich in analytischer Absicht als fruchtbar erweist. In diesem Abschnitt werden daher zunächst einige Unterscheidungen an Hass, Wut und Zorn nachgezeichnet, um die Beschreibung der Fallstudie des nächsten Abschnitts zu informieren.[9]

Zu bemerken ist zunächst der Widerfahrnischarakter der Gefühle: Sie stoßen uns eher zu, als dass sie uns verfügbar und für uns veränderbar wären (Demmerling/Landweer 2007: 9). Verfügbar sind demgegenüber Gefühlsdispositionen, die etwa durch Übung durch allmähliche Änderungen von (Wert)Haltungen änderbar sind (Demmerling/Landweer 2007: 25), etwa mit psychotherapeutischen Mitteln (Demmerling/Landweer 2007: 9). Es lassen sich daher Haltungen von Gefühlen unterscheiden (Demmerling/Landweer 2007: 5f.), die beide mit (rationalen und irrationalen) Gedanken zusammenhängen, ohne auf diese reduzibel zu sein (Demmerling/Landweer 2007: 33). Haltungen und Gefühle, in denen jene sich aktualisieren, hängen also mit (Un)Werturteilen zusammen, abermals, ohne auf diese reduzibel zu sein. Die Bewertung von Situationen gibt entsprechend eher den Anlass für das Gefühl, determiniert aber nicht den Gehalt des Gefühls. Dies steht der Beobachtung, dass Gefühle eine Intentionalität aufweisen, also auf ein Objekt gerichtet sein können, nicht entgegen: Anders als Stimmungen, die – von Übergangsphänomenen abgesehen – regelmäßig keinen Objektbezug aufweisen, haben Gefühle einen Gehalt, auf den sie sich richten (Demmerling/Landweer 2007: 292f.). In Frage steht also nicht in erster Linie, ob Gefühle einen intentionalen Gehalt aufweisen, sondern eher *wie* sie sich auf ihre Objekte beziehen. Während Affekte wie Ekel eine Abkehr vom Objekt aufweisen und nicht auf Gewalt ausgehen (Demmerling/Landweer 2007: 108), trifft dies für die in Frage stehenden

9 Was ich im Folgenden unterschlage, ist die Leibdimension von Gefühlen, die sich einer verbreiteten Meinung nach nicht von Körperempfindungen abtrennen lassen (vgl. dazu Demmerling und Landweer 2007: 27). Es ist zwar mehr als naheliegend, dass auch ›online‹ ausgelöste Gefühle körperlich empfunden werden, jedoch lässt sich dies im vorliegenden Material freilich nicht untersuchen. Die Affordanzen und Nutzungskulturen von 4chan/pol machen es auch schwierig, die soziale Funktion von Gefühlen – etwa bei der Herausbildung von sozialen Netzwerken – im Material zu berücksichtigen, so dass ich mich auf explizite ›Gefühlsnarrative‹ beschränken muss. Die von Engelen hervorgehobene Ordnungsfunktion für Gemeinschaften bei der Verhandlung von Normen lassen sich hingegen gut im empirischen Material beobachten (Engelen 2008: 43).

Aggressionsaffekte – Ärger, Wut, Zorn, Empörung, Hass, die in einem Steigerungsverhältnis zu stehen scheinen (Demmerling/Landweer 2007: 289) – nicht zu: Diese sind durch eine aggressive Hinwendung zum Objekt charakterisiert, etwa dem Objekt schaden zu wollen (Demmerling/Landweer 2007: 287): »Aggression in diesem Sinne gilt als eine radikal und bedrohlich desorganisierende Kraft, die unkontrollierbare Destruktivität freisetzen kann« (Demmerling/Landweer 2007: 290). Während Ärger, Wut und Zorn mitunter plötzlich hervorbrechen, bleibt Hass (womit er eher an eine Haltung erinnert) unter der Oberfläche, wo er regelrecht lauert und über lange Zeit bestehen bleiben kann, von wo aus er dazu beiträgt, soziale Gruppen zu organisieren, soziale Typisierungen anzuleiten und Feindschaften zu pflegen (Szanto 2020: 454). Hass bezieht sich entsprechend nicht auf einzelne Situationen, sondern auf gesamte Personen oder Gruppen, zu denen er eine sie völlig umgreifende Beziehung herstellt – anders als etwa Wut und Ärger, die unfokussiert sein können (Demmerling/Landweer 2007: 288, 308f.). Es lassen sich etwa gerichtete »Hassausbrüche« ausmachen, und der Hass gilt damit schon für Aristoteles als »maßlos und unheilbar« (Demmerling/Landweer 2007: 299). Hass erscheint irrational und dysfunktional, er macht unfrei, ist schädlich für die Hassende selbst, gleichwohl hat Hass Gründe (Demmerling/Landweer 2007: 295f.), die ihn an vergangene (möglicherweise imaginierte) Schadenserfahrung binden und so für die Produktion durch narrative Strukturen öffnen. Kurz: »Hass verlangt notwendigerweise ein personales oder personalisiertes Objekt und dieses Objekt wird in einem schwachen Sinne für etwas Schädliches verantwortlich gemacht.« (Demmerling/Landweer 2007: 308) Vielleicht ist hier auch ein Grund dafür zu suchen, dass die Philosophie unter den Aggressionsaffekten vor allem den Neid und den Zorn in den Blick genommen und damit gewissermaßen geadelt hat, während Analysen zum Ärger und zur Wut weitgehend fehlen (Demmerling/Landweer 2007: 287). Die eigentliche Ursache dürfte aber darin bestehen, dass der Zorn (wie die Empörung) nicht bloß aggressiv ist, sondern zu den klassischen Unrechtsaffekten zählt (Demmerling/Landweer 2007: 299), also moralisch überlagert werden kann. Zorn ist ein moralisches, maßvolles Gefühl, das in Reaktion auf ein wahrgenommenes Unrecht gegen den Zürnenden selbst, gegen einen Nahestehenden oder in Reaktion auf eine (tatsächliche oder vermeintliche) Ehrverletzung oder Herabsetzung durch eine bestimmte Handlung entsteht (Rothbart 2021: 682; Engelen 2008: 44, 52; Demmerling/Landweer 2007: 297, 308). Als solches muss Zorn also ein spezifisches, personales Objekt besitzen, »dem gezürnt wird«, indem es verantwortlich gemacht wird, ohne aber – wie beim maßlosen Hass – immer zugleich dessen Vernichtung zu begehren: Zorn scheint sich im Unterschied zum Hass auf das rechte Maß der Rache zu begrenzen und sich mit moralischen Erzählungen zu verbinden (Demmerling/Landweer 2007: 305, 308; Engelen 2008: 62f.; Szanto 2020: 456), die allerdings durch den Objektbezug gerade nicht auf »ein System« hinauslaufen, sondern wenigstens narrativ immer wieder zur Personalisierung

zwingen (Brockhaus 2020: 93). Wie wir sehen werden, bedeutet das keineswegs, dass in den moralischen Erzählungen des Zorns »das System« nicht vorkommen kann, sondern ›nur‹, dass »das System« in der Erzählung personalisiert werden muss.

Es ist wohl vor allem Peter Sloterdijk zu verdanken, im Rückgriff auf den *thymós* wieder an die politische Funktion des Stolzes und damit mittelt über die Kränkung an die des Zorns erinnert zu haben: Thymotische Energie stiftet Gemeinschaft und so ist Zorn auch nicht die Gemeinschaft und Freiheit zerstörende Kraft, als die er mitunter denunziert wurde (Sloterdijk 2006: 26; Engelen 2008: 64). Ungeachtet der Notwendigkeit seiner Begrenzung gegen etwaige Exzesse, und anders als auf Zerstörung ausgehender Hass gilt Zorn also immer wieder als legitime, politisch gar erforderliche Emotion, die soziale Ordnungs- und Korrekturfunktion besitzt (Demmerling/Landweer 2007: 308; Engelen 2008: 41, 50; Szanto 2020: 456). Kurz gesagt zürnt man nicht alleine, während der Hass offenbar durchaus eine bloße Privatangelegenheit sein kann, womit die entpolitisierende Funktion der Rede vom »Hass im Netz« offen hervortritt. Sloterdijk macht dagegen eine thymotische Spannung in politischen Gruppen – etwa im Fall ehrgeiziger Individuen – aus, deren Gefälle Aktionen in von selbstaffirmativen Kräften durchzogenen politischen Feldern, nicht zuletzt gesteuert durch rhetorische Affektlenkung, in Gang setzt (Sloterdijk 2006: 36f.), womit nicht zuletzt auch Normen in Gruppen zur Geltung gebracht werden (Demmerling/Landweer 2007: 302). Beim Zorn handelt es sich also um einen in der Sprache der Moral ausgedrückten sozialen Affekt, der um Narrative des Unrechts, der (Ehr)Verletzung und der (gezügelten) Rache oszilliert (Demmerling/Landweer 2007: 289).[10] Der Zorn bleibt damit zugleich eng auf die Figur der Held*in – und damit der Täter*in und des Opfers – bezogen (Sloterdijk 2006: 14).[11] Neben moralischen Narrativen finden wir im Zorn also auch Ermächtigungsnarrative, die ihn für propagandistische Sozialtechniken öffnen, die durch eine »Reinigung« auf die Ermächtigung des Individuums in einer neuen Gemeinschaft hinauslaufen (Aikin 2019: 431; Griffin 2020: 51, 76, 79f.; Hakoköngäs et al. 2020: 2).

---

10 Die Empörung kann eine analoge Funktion erfüllen, kann aber im Gegensatz zum Zorn abstrakt bleiben und so auf personale Objekte verzichten: »Ich kann mich zwar auch über jemanden empören, aber ebenso über Verhältnisse, die nicht eindeutig personell zurechenbar, wohl aber Menschen gemacht sind.« (Demmerling und Landweer 2007: 309). Wir werden im Folgenden sehen, dass die auf 4chan/pol beobachteten Narrative meist personale oder personalisierbare Objekte besitzen.

11 Folgt man Griffin, so ist es gerade der *homo heroicus*, in dem sich der neue Mensch des Faschismus ausdrückt (vgl. Griffin 2020: 85). Die Rolle der Heldenerzählungen für den gegenwärtig Dynamik gewinnenden Cyberfaschismus wird erst in der jüngeren Forschung anhand der Darstellung rechtsextremistischer Gewalttäter als religiöse Märtyrerfiguren deutlicher. (Vgl. Thorleifsson 2022) Die geschlechtergerechte Sprache darf nicht darüber hinwegtäuschen, dass auch Sloterdijk dies als entschieden männliche Angelegenheit ausweist.

Gegenüber der oft objektlosen Wut und der abstrakten Empörung können wir also intentionale aggressive Affekte identifizieren, die sich im Unterschied zum Ekel ihrem Objekt zuwenden: Hass und Zorn. Während der Hass eher als isolierend gilt und mit unbedingtem Vernichtungswillen ausgestattet ist, gilt der Zorn als gezügelt, moralisch überformt und in der neueren Literatur auch als gemeinschaftsstiftend sowie in korrektiver Absicht soziale Normen affirmierend.

Use your rage!

Anhand zweier kleiner Fallstudien soll nun die formulierte These untersucht werden, dass sich auf 4chan/pol Affektstrategien beobachten lassen, in denen Hass (als eine Disposition zu einem Affekt) durch empörende Erzählungen produziert und Wut (als eine Aktualisierung des Hasses) in Zorn umcodiert wird.

Gaskammer für den Tierquäler

Am 30.6.2016 um 17:12:22[12] fragt eine aus Großbritannien stammende Anon auf 4chan/pol im Posting #79293113[13] nach der Meinung der anderen zur Misshandlung von Tieren. Der Post enthält das Bild einer jungen Katze und einen Link auf eine BBC-Meldung vom Vortag über die Verurteilung eines britischen Mannes für die Misshandlung eines Tieres, das daraufhin eingeschläfert werden musste. Nach 96 Sekunden kommen die ersten Reaktionen (jeweils mit Landesangabe): »If it were up to me I'd execute people who do it« (USA), »Should be punishable by death« (USA), »Should be shot« (Australien), »I think people who abuse animals should be forced into slavery« (USA), »It shows lack of empathy, a lack of consciousness and that's a n***** trait« (Slowenien),[14] »Public flogging, 10 lashes« (Großbritannien) lauten die Antworten in den ersten vier Minuten seit der Eröffnung des Threads. Um 17:17:30 fordert ein Beitrag, der als einziger verknüpft mit einem Profilnamen – »Reichswehr« – versehen ist: »Gas chambers« (Kanada). Die folgenden Antworten – durchgehend wieder ohne Profilnamen – sind zustimmend: »Basically this. It's literally the most degenerate thing imaginable along with child abuse. Kill em all« (Großbritannien), »Came here to post exactly this« (USA),[15] »People that do this should be beaten to death in public« (USA). Es geht in diesem Tenor weiter,

12 Die Zeitstempel werden im Folgenden auf Basis des auf archive.4plebs.org archivierten Threads wiedergegeben, der keine Zeitzone anzeigt. Es kommt im Folgenden auf die Zeitabstände zwischen den Postings an, nicht auf die absoluten Zeitangaben.

13 Der Thread zu #79293113 ist unter [https://archive.4plebs.org/pol/thread/79293113/] (Zugriff: 25.07.2023) abrufbar. Es ist unsicher, ob die Archivversion vollständig ist.

14 Das N-Wort wird im Posting ausgeschrieben.

15 Dieser Beitrag bezieht sich auf den zitierten Beitrag aus Slowenien.

wobei rassistische Beiträge sich häufen, z.B. mit abermaliger Forderung nach der Gaskammer: »The ability to show empathy, care and train a being of lower existence is an implicit part of white and western identity. / That being said if it was up to me they'd go straight to a gas chamber as they've proved themselves to be lower than even the animals we've trained« (USA), und aus Belgien wird gepostet: »Look Hitler was pro animal rights, what do you think we think about it? / Degenerate and should be punishable!«. Ab 17:23:30 postet ein Anon (USA) wiederholt historische Fotografien aus der NS-Zeit: Hitler mit Schäferhündin »Blondi«, Hitler mit anderem Hund, Hitler mit Rehkitzen, Hitler in einer Gruppe von Offizieren ein Löwenjunges streichelnd, abermals Hitler mit Schäferhündin »Blondi«, ein gezeichneter Göring vor einer Menge den rechten Vorderlauf hebender Labortiere.[16] Die vorgenannten Bilder wurden von ein und demselben Anon (ID: yATJXOfe) innerhalb von 124 Sekunden gepostet. Wer sich auch hinter der ID yATJXOfe verborgen haben mag, es ist nicht die einzige User*in, die auf historische Bilder zurückgreift: Um 17:38:24 postet ID vJAZ2fmx die historische Fotografie eines SS-Offiziers, der zwei Katzen streichelt mit dem Kommentar: »[…] Anyway, pic related. Animal welfare is a real thing, not some Left wing project […]« (USA). Dieselbe User*in postet um 18:13:46 noch das Bild eines Wehrmachtsoffiziers, der offenbar eine Katze füttert. Die Diskussion lockt – wie üblich – Trolle an und keineswegs besteht 4chan/pol ausschließlich aus Katzenfans (»Hey you fucking retard, keep your cats inside«, Kanada; »My cat kills parrots and native birds, you mad you greeny faggot?«, Australien; »Cats are for faggots«, Portugal). Das letzte Posting des Threads findet sich um 18:51:17 (»Haha good«, Australien).

An dem hier vorgestellten Thread lassen sich zunächst die genannten Affordanzen von 4chan bzw. 4chan/pol beobachten: In der Regel wird als Anonymous gepostet, wobei IDs die User*innen in Grenzen wiedererkennbar machen. Ausnahme bildet hier das Posting durch »Reichswehr«, ohne dass die Verwendung eines Profilnamens erkennbaren Einfluss auf den Threadverlauf gehabt hätte. Die Herkunft der User*innen ist – die Korrektheit der Landesangaben unterstellt – in erster Linie in englischsprachigen ›westlichen‹ Ländern zu finden. Die Zeitstempel geben einen Eindruck von der Geschwindigkeit des Austauschs: Es wird schnell und eher emotional reagiert. Abwägende Argumente finden sich nicht und wären vermutlich auch nicht erfolgreich, d.h. würden eher keine weiteren Reaktionen nach sich ziehen. Auffällig ist zudem, dass einzelne User*innen – hier unter der ID yATJXOfe – Sammlungen mit einschlägigen Bildern vorrätig zu halten scheinen oder wissen, wo

16 Es handelt sich um eine Karikatur aus der den Nationalsozialismus unterstützenden Satire-Zeitschrift Kladderadatsch vom 3.9.1933 zum Verbot der Vivisektion. Online abrufbar unter [https://digi.ub.uni-heidelberg.de/diglit/kla1933/0569/image,info] (Zugriff: 30.07.2023). Nicht nur mit Blick auf das 1933 eingeführte Schächtungsverbot ist »Der Jude als Tierquäler« ein auch im Nationalsozialismus bedienter Topos.

sie diese schnell finden.[17] Auch die beschriebene Eskalation der Diskussion ist deutlich zu beobachten. Mit Händen sind die überkochenden Emotionen zu greifen, die hier kaum durch den 4chan-typischen Humor überformt sind: Es finden sich nur wenige ›lustige‹ Meme und auch das Trolling scheint sich zunächst in Grenzen zu halten. Was sich nicht in Grenzen zu halten scheint, ist das Rache- und Vernichtungsbedürfnis: Forderungen nach der Gaskammer, der Versklavung, der öffentlichen Auspeitschung erscheinen Außenstehenden maßlos und von unbedingtem Vernichtungswillen geprägt, aber grenzüberschreitende Sprache ist auf 4chan/pol der Normalfall. Kurz: So schwer es fallen mag, wir können aus dieser Beobachtung heraus nicht ausschließen, dass wir es mit Zorn zu tun haben. Es ist auch nicht bloß eine Stimmung, und es handelt sich ebenso nicht einfach um eine abstrakte Empörung über »die Verhältnisse«, auch wenn sich die Affekte mit einer Empörung im Namen unschuldiger Tiere verbinden. Die Anons wenden sich auch nicht in Ekel ab, sondern wir beobachten eine Hinwendung: Die aggressiven Affekte richten sich auf personale oder personalisierbare Objekte. Was hier aber auffällt: Es geht in der Diskussion allenfalls am Rande um den in der BBC-Meldung genannten Täter, einem 48jährigen, offenbar weißen Briten aus Sussex. Das personale Objekt des Zorns fällt hier aber nur scheinbar aus. Stattdessen treten die entgrenzten Gewaltvorstellungen generalisiert und rassistisch hervor: Die BBC-Meldung liefert den Anlass, Objekt sind trotz des Täters andere, über die bereits ein Unwerturteil getroffen wurde: »Mudslimes[18] are already killing people's dogs in Europe« (USA). Unübersehbar ist die Verhandlung der Normverletzung, nach der Tiere nicht zum Spaß gequält werden dürfen, unter rassistischen Vorzeichen. Offenbar handelt es sich bei den ausgedrückten Affekten zusammen mit Fantasien der Vergeltung sowie der Reaffirmation von Normen um Zorn. Die BBC-Meldung liefert den Anlass, bestehende, rassistische und hasserfüllte Unwerturteile in Form einer Normverhandlung zu reaktualisieren. Und auch wenn die Objekte des Zorns unter rassistischen Vorzeichen verhandelt werden, wirkt er hier gemeinschaftsstiftend: Es scheint eine rassistische Gruppe zu zürnen, die scheinbar klare moralische Normen vertritt. Es fällt auf: Offenbar gibt es eine (kleine) Zahl von User*innen, die den Thread zum Anlass nehmen, die moralischen Narrative zu vereinnahmen: Der Misshandlung einer Katze wird die Tierliebe der Nationalsozialisten entgegengesetzt, und zwar ›bewiesen‹ in Form historischer Bildaufnahmen. Eine direkte Wirksamkeit ist nicht zu beobachten, es findet aber auch keine erkennbare Distanzierung statt. Die Verflechtung mit dem Narrativ des nationalsozialistischen Tierschutzes mit rassistisch codiertem Zorn scheint nicht zu irritieren.

17 Für Bevorratung spricht, dass solche Bilder immer wieder auf 4chan/pol gepostet werden sowie dass rechte »Anleitungen« zu memetischer Kommunikation die Anlage entsprechender Sammlungen anraten.

18 Ein nicht nur auf 4chan/pol gebräuchlicher, abwertender Ausdruck für Muslime.

Surviving the Red Pill mit dem Self-Improvement General

Thematisch auf 4chan/pol wurde im Kontext der »iron pill« in den letzten Jahren der »Self-Improvement General« – ein oft mit /SIG/ abgekürztes Mem, das um Selbstoptimierung kreist (Abb. 1) und dabei nicht nur auf die persönliche Besserung, sondern auch darauf zielt, sich der extrem rechten Bewegung besser andienen zu können (Elley 2021: 2). Das Mem kann – was für Meme ungewöhnlich ist – über mehrere Jahre praktisch unverändert[19] beobachtet werden: Zwischen Ende 2017 und Anfang 2023 wird das Bild-Mem immer wieder gepostet, während etwa 70% der Bilder auf 4chan nur einmal gepostet werden (Hine et al. 2017: 97).[20] Bildsprachlich ist das Mem komplex: Es ist zunächst ein klassisches Vorher-Nachher-Mem, das in zwei Bildteilen die Situation »Before /SIG/« und nach »After /SIG/« darstellt. Beide stilistisch typischen und daher unauffälligen Bildteile sind strukturanalog: Für jedes Element des einen Bildteils gibt es eine Entsprechung im anderen Bildteil, an dem die Vorher/Nachher-Differenz markiert werden kann. In beiden Bildteilen finden wir ein Fenster, ein Regal, Wanddekoration, eine Fahne an der Wand, einen Schreibtisch mit Computer sowie eine männlich gelesene Person, die am Computer sitzt. Zeigt der Vorher-Teil ein unaufgeräumtes, vermülltes Zimmer mit herumliegender Wäsche, einem offenen Pizzakarton und einem überlaufenden Mülleimer, ist dieses im Nachher-Teil aufgeräumt und sauber. Im Vorher-Teil zeigt das Fenster Gebäude, was auf eine städtische Lebensweise hindeutet. Im Nachher-Teil handelt es sich um eine Landschaftsaufnahme. Das Regal im Vorher-Teil ist unordentlich und enthält u.a. typische Gegenstände aus der Nerd-Kultur. Im Nachher-Teil ist das Regal mit Büchern,[21] Hanteln und Gegenständen für die Gartenarbeit gefüllt. Neben dem

19 Elley stellt das Mem anders dar: Bei Elley fehlt das »I ♥ ISRAEL« Bildelement (Elley 2021: 3). Die Darstellung von Elley konnte auf 4chan/pol nicht gefunden werden. Auch die von Elley angegebene Quelle zeigt das Mem mit dem genannten Bildelement.

20 Hine et al. weisen allerdings darauf hin, dass einige Bilder immer wieder gepostet werden. Dazu gehört etwa »Pepe, der Frosch«, von dessen zahlreichen Varianten Hine et al. eine im Beobachtungszeitraum 838-mal beobachten konnten. Auch die historischen Bildaufnahmen aus der Zeit des Nationalsozialismus lassen sich immer wieder beobachten. Vgl. Hine et al. 2017: 97.

21 Die Auflösung des Bild-Mems auf 4chan/pol ist hoch genug, um die bewusst eingefügten Autorennamen der gezeichneten Buchrücken zu erkennen. Neben Namen des klassischen Philosophie-Kanons (neben Aristoteles, Rousseau, Spinoza, Kant, Homer, Hume, Aurel, Hobbes, Locke, Platon, Hegel, Calvin und Mill finden sich auch in rechten Kreisen besonders wohlgelittene Namen wie Spengler, Heidegger, Nietzsche und mit Edmund Burke der Vater des Konservativismus). Hinzu kommen Schriftsteller wie Orwell und Huxley. Schließlich finden sich Autoren, die der (extremen) Rechten zuzuordnen sind und sich auch auf entsprechenden Leselisten finden: der rechtsextreme US-amerikanische Kulturphilosoph Ulick Varange (ein Pseudonym von Franic Parker Yockey), der belgische Rexist und Offizier der Waffen-SS Léon Degrelle, der paläolibertäre Vertreter des Anarchokapitalismus Hans-Hermann Hoppe

Regal stehen nun Sportschuhe und auf der anderen Seite lehnt eine Axt. Auch die Wanddekoration wurde verändert, z.B. wurde das Bild des britischen Evolutionsbiologen Richard Dawkins (hier wohl als Vertreter des *new atheism* angesprochen) durch ein christliches Kreuz ersetzt und statt dem Bild eines bekannten Internettrolls hängt ein Maschinengewehr und ein Boxsack. Die Flagge der Republik Kekistan – einer auf 4chan/pol gerne genannten fiktiven politischen Entität – wurde durch die Flagge einer lorbeerumkränzten Lebensrune ersetzt. Fanden sich zuvor auf dem Schreibtisch eine Bong (eine Wasserpfeife), ein Energy-Drink und eine Tüte Kartoffelchips, sehen wir nun ein Glas Wasser. Der ansonsten unveränderte Computer zeigte zuvor einen »I♥ISRAEL« -Aufkleber sowie das Logo der misogynen und antifeministischen Online-Bewegung »Men Going Their Own Way (MGTOW)«. Beide sind nun durch das Logo der dem Rechtsextremismus zugeordneten Identitären Bewegung ersetzt. Die abgebildete Person wurde zuvor als deutlich übergewichtig, mit Hut und längeren Haaren in einem schwarzen T-Shirt dargestellt. Nun ist sie schlank, besitzt einen Undercut-Haarschnitt[22] und ist mit einem hellen Poloshirt bekleidet.

Die Umcodierung von Topoi »von rechts« im SIG-Mem ist unübersehbar: Stadt wird durch Land, Chaos durch Ordnung, Übergewicht durch Sportlichkeit, Atheismus durch Christentum, Trollkultur durch rechte »Kontrakultur« ersetzt. Die Sportlichkeit wird mit Waffen verbunden und verweist auf den (nun durch philosophische Lektüre gebildeten) *homo heroicus*, indem typisch für faschistische Propaganda an Männlichkeitsbilder appelliert wird, was sich vor allem an weiße und heterosexuelle Männer richtet (Jäger et al. 2021: 10) und was de facto immer die Abwertung des Weiblichen und Queeren einschließt (Jäger et al. 2021: 7). Das SIG-Mem richtet sich unübersehbar an einen selbstbestimmten »neuen Mann« (Griffin 2020: 88), der aus der Misere von Feminismus, städtischem Leben, Konsum und Trollkultur herausfindet. Gerade der Körper, dessen Verfassung sich noch im Zustand der Wohnung ausdrückt, wird zum Schauplatz rechter Kommunikationsstrategien: Selbstverbesserung wird politisch und zwar nicht nur im einzelnen Körper, sondern vermittelt über die Verweise auf politische Bewegungen auch kollektiv (Elley 2021: 7f.). Das SIG-Mem ist damit politisch, antinihilistisch und ermächtigend – allerdings zugleich entschieden extrem rechts. Durch die Selbstverbesserung von Seele, Körper und Geist soll sich der neue Mann nicht nur selbst retten, sondern sich in den Dienst eines Rassenkrieges stellen können: »/SIG/ is

und schließlich der Gründer der *British Union of Fascists* Oswald Mosley. Der Name »Paine« ließ sich nicht eindeutig zuordnen. Möglicherweise ist der Gründervater der USA Thomas Paine gemeint.

22 Obzwar der Undercut nicht spezifisch für die rechte Szene ist, ist diese auch in den 1930er und 40er Jahren beliebte Frisur dort oft anzutreffen.

political, because politics begins with you. Change yourself, and you will change your community and in time, the nation will change.«[23]

Abbildung. »Self-Improvement General«-Mem von 4chan/pol

Quelle: [https://archive.4plebs.org/pol/thread/223955517/#223958154] (Zugriff: 30.07.2023)

23 Vgl. unter [https://shoebat.com/2019/09/25/the-one-thread-that-keeps-getting-shut-down-on-4chan-that-tells-you-what-they-dont-want-you-to-know/] (Zugriff: 30.07.2023). Der Autor Andrew Bieszad berichtet auf der vom offenbar evangelikalen Islamkritiker Walid Shoebat betriebenen Webseite shoebat.com 2019 davon, dass SIG-Postings auf 4chan/pol nun innerhalb weniger Minuten gelöscht würden und führt dies darauf zurück, dass »sie« Gedankenkontrolle ausüben wollten. Ein derartiges Zensieren von SIG-Postings lässt sich nicht nachvollziehen.

Das SIG-Mem gehört zu der Diskursformation, die unter dem Titelwort »iron pill« gefasst werden kann. *iron pill* verweist auf das »red pill«-Mem, das auf die berühmte Szene des Films Matrix zurückgeht, in der Morpheus Neo zwei Pillen zur Wahl anbietet: Die blaue Pille lässt Neo in der Matrix und alles für einen bösen Traum halten, während die rote Pille Neo aus der Matrix aufwachen und ihn die Welt sehen lässt, wie sie ist. Die *red pill* markiert also eine Aufklärungs- und Ermächtigungserzählung (Aikin 2019: 429; Wentz 2019: 4). Aus der Sicht der *iron pill* zeigt die *red pill* eine degenerierte, derangierte und perverse Welt, die von einer globalen Elite eigennützig manipuliert wird – ohne aber etwas gegen diese Zustände und die von ihnen ausgelösten Frustrationen zu unternehmen (Aikin 2019: 431; Elley 2021: 2). Hier tritt die Idee der *iron pill* auf den Plan: Durch Lebenshilfe sollen gleichzeitig die moderne Dekadenz im Namen einer Tradition und ungesunde Angewohnheiten der 4chan-User*innen sowie ihre Frustrationen – wenigstens im eigenen männlichen Körper – überwunden werden (Elley 2021: 3ff., 9; Strick 2021: 235f.). Die *iron pill* verweist hier auf die Vorstellung einer (kreativen) Zerstörung der alten dekadenten Welt und ihre gereinigte Wiedergeburt im Sinne einer faschistischen Palingenese (Griffin 2020: 181, 200). Die Wiedergeburt wird zugleich durch den neuen Menschen ausgelöst (Griffin 2020: 90).

Entscheidend für das Verständnis der »iron pill« und des SIG-Mems ist neben den palingenetischen Narrativen die Gefühlsarbeit: Zielloser Hass und Wut müssen produziert und abgeschöpft werden. Es geht nicht einfach nur darum, negative Gefühle zu induzieren und die Induktion dann positiv umzuwerten (Strick 2021: 81), sondern sie sollen zu einer Ressource zum Nutzen der rechten Bewegung werden (Elley 2021: 9). Diese Gefühlsarbeit wird explizit verhandelt: Am 26.4.2022 antwortet jemand auf »god i hate jews so fucking much it is unreal« (USA) mit »channel that rage into self-improvement, anon. make yourself smarter, stronger, and more aware than they are« (USA).[24] Und in einem Thread mit dem Titel »How to survive redpilling on /pol/« liefert ein Anon (USA) am 15.11.2017 elf Ratschläge, die sich hier ausführlich wiederzugeben lohnen:

> »4. At this point, you will fall into despair. You will feel alone. You will feel like everyone around you are just idiots going through life without a single rational thought. *And then, you will become angry. You will know what true rage is. The strength and magnitude of this rage will surprise you.*
>
> *Now, this rage is the crucial fork in the road.* It is at this point that either /pol/ will make you a god or an ostracized hateful racist neo-nazi. *You can use this anger and become destructive (to yourself and others) or you can use it to become something better.* It is up to you.

24 [https://archive.4plebs.org/pol/thread/374692288/#q374699755] (Zugriff: 30.07.2023).

[…]

7. You will lift weights. Slowly at first. It matters not how much you bench. It only matters that you are there and that you keep going there. And it matters that you are there for the right reason. You are not there to impress girls. You are not there to make yourself look sexy. You are not there to become a model. *You are there to improve yourself. Period. You are doing this for yourself and no one else.* When you get home, eat a healthy meal. Under no circumstances are you allowed fast food or processed foods. *Spend 30 minutes on /pol/ or some other right-wing site, especially Libertarian sites.* Read a redpilling book. No TV. No Netflix. No video games. No porn. Try to go to bed early.

8. Repeat this for 3 months. It will be the hardest thing you've ever done in your life. But don't give up. Above all, do not think. Thinking only leads to laziness and excuses. Every time you start thinking about making excuses not to go to the gym, *GET ANGRY!!! And this is where you use that RAGE I talked about earlier. THIS IS WHERE THAT RAGE WILL HELP YOU INSTEAD OF DESTROYING YOU.* If you find yourself lacking the motivation, summon that anger that /pol/ has instilled in you. Bring it to the brim. *USE THE RAGE. LET IT CONSUME YOU. Then, just get up and go to the gym and burn out the rage by lifting.* As you lift, think of all the redpills you've swallowed and think about how mad it makes you. Think about how hypocritical and how evil this world truly is. *Think about how when shit hits the fan, there are only 2 things that are going to save you; your physical strength and your guns.* Think about how the government and people in power force you to be less than what you're capable of becoming. How they want you to conform to social norms simply for the sake of political correctness. Think of how unfair things in this world truly are. And finally, know that no one on this god damn planet can help you EXCEPT yourself.«[25]

Uns begegnet auf 4chan/pol also eine explizite Gefühlsarbeit in enger Verbindung mit Selbstverbesserungsdiskursen, die politisch nach rechts zeigen. Die Gefühlsarbeit codiert aggressive Affekte, wo sie noch unbestimmt oder unscharf sind. Ist etwa ein vager Judenhass vorhanden, wird vorgeschlagen, diesen produktiv zu nutzen und, ohne das Objekt des Hasses zu ändern, ihn mit einer Zwecksetzung auszustatten, nämlich um »more aware than they are« zu werden. Die Sprache, die an die ältere *Anonymous*-Selbsthilfekultur auf 4chan anschließt, ist dabei streckenweise völlig untypisch, nämlich eindeutig, arm an ironischer Brechung und sogar wertschätzend (Elley 2021: 4f.). Dennoch ist, wie Elley weiter ausführt, die SIG-Selbsthilfekultur von aggressiven Affekten nicht abzulösen. Dies lässt sich auch im zitierten

25 [https://archive.4plebs.org/pol/thread/149588819/#q149588980] (Zugriff: 30.07.2023), Hervorhebung von mir – KD.

Auszug nachzeichnen: »*rage*«, was eher Wut (*anger*) als Zorn (*wrath*) bedeutet, ist die affektive Quelle, die frappant an den thymos erinnert. Er entstammt einer weitgehend abstrakten Empörung, ist eine positive, letztlich politische Kraft, er muss entfesselt und genutzt werden, um ihn schließlich gleichzeitig in die eigene Verbesserung im Fitnessstudie wie gegen »sie«, die Konformität und politische Korrektheit erzwingen wollen, zu verwenden. Was verlangt wird, ist die Freisetzung der Energie zur Tat (nicht zur Reflexion: »Above all, do not think«), aber einer Tat, in der dann zürnend gedacht werden muss: »As you lift, think of all the redpills you've swallowed and think about how mad it makes you […] Think about how the government and people in power force you to be less that what you're capable of becoming.« Es geht also nicht um freies Denken oder Denken überhaupt, sondern zürnend sollen bestimmte Narrative ›bedacht‹ werden – seien sie bloß libertär oder aber antisemitisch oder rassistisch. In jedem Fall handelt es sich auch hier um Ermächtigungsnarrative, in denen man vom Opfer zur Held*in wird, die zumindest zur Selbsthilfe mit körperlicher Stärker und Waffen in der Lage ist. Entsprechend imaginiert man unter der »iron pill« auch die Rebellion gegen einen dekadenten und doch totalitären Staat, der einerseits an Huxleys *Schöne Neue Welt* erinnert (Elley 2021: 8), andererseits schlicht der Unterdrückung der »white people« dient, gegen die ein »white thymos« in einer impliziten Tatgemeinschaft freigesetzt werden muss (Ganesh 2020: 7). Auch hier sind also die Zutaten des Zorns versammelt: Es wird ein aggressiver Affekt verhandelt, der sich an einer Kränkung (nämlich der red pill, die ihr Versprechen der Befreiung nicht eingelöst hat) entzündet, gezürnt wird nicht einem abstrakten System, sondern der Regierung, den Eliten, den Juden, und schließlich werden gesundheitliche Normen verhandelt, die nicht im Dienst des Aussehens oder des sexuellen Begehrens stehen dürfen, sondern die als Selbstoptimierungstechnik zugleich politischen Zwecken zu dienen haben.

Schluss

Hass gilt als ein blinder, umfassender, auf völlige Zerstörung zielender aggressiver Affekt, der von dem begrenzten, gelenkten und moralisch überformten Zorn klar unterschieden werden kann. Hat der Zorn in den letzten Jahren in der Diskussion wieder eine politische Aufwertung erfahren, die im Namen des *thymos* auf antike Zornvorstellungen zurückgreift, gilt der Hass als mit den Mitteln des Rechts zu bekämpfende, mindestens aber zu verwaltende »elementare Gewalt« (so der Titel von André Glucksmann Studie zum *Hass*: Glucksmann 2005). Entsprechend haben sich die Initiativen, den Hass im Netz zu bekämpfen, unüberschaubar vermehrt und »hate speech« als in der deutschsprachigen Diskussion etabliert. Hier erscheint Hass fast immer als eine ungelenkte, anonyme Kraft, die insbesondere Folge unzureichend regulierter digitaler Kommunikationsplattformen ist, deren algorithmi-

sche Affordanzen Hass und politischen Extremismus begünstigen. Affektstrategien – also zweckgerichtete Produktion und Verwertung von Affekten – durch politische Bewegungen auch jenseits komplexer algorithmischer Systeme kommen in der Diskussion nicht vor, sind aber in den dunklen Ecken des Internets eine Realität, die mit der simplen Rede vom Hass nicht erfasst werden kann. Von so einem Bild individuell verfasster Hasspostings kann in den beiden Fallbeispielen kaum die Rede sein. Selbst wenn diese hier und da auftreten: Sie sind abgeschöpfte affektive Energie, die im Rahmen einer rechtsgerichteten Strategie mit Narrativen verknüpft und so zu einem rechten Zorn umgedeutet werden. Dies alles geschieht ohne Eingriff avancierter algorithmischer Systeme im Rahmen einer simplen Aufmerksamkeitsökonomie, die den starken Affekt einer sich selbst verstärkenden Schleife belohnt und jenseits dessen, was durch die simple rechtliche Verwaltung und Moderation von Hatespeech greifbar wäre. Die Verrohung der Sprache, die im Netz zu finden ist, mag lästig, für vom Hass Betroffene mehr als nur störend und verletzend, sondern regelrecht bedrohlich sein. Extrem rechte Strategien spielen aber auf anderen Schauplätzen und bedienen sich Mechanismen, die die bisherige Diskussion zu digitalen Verständigungsverhältnissen kaum in den Blick genommen hat. Der neue Faschismus im Netz ist nicht Ergebnis unregulierter Internetkonzerne, sondern wird in Nischen bewusst orchestriert und von dort wirksam. Wenn wir diesem Phänomen wirksam begegnen wollen, müssen wir es als Ausdruck des Zorns ernst nehmen. Das bedeutet nicht, die wirren Sorgen vor einem Untergang des Abendlands durch eine angebliche liberale oder multikulturelle Dekadenz zu adeln, sondern anzuerkennen, dass wir es überhaupt mit narrativen Mustern im Sinne des Zorns zu tun haben, die unter der Verengung auf Hass nicht verhandelt werden können.

Literatur

Aikin, S.F. (2019): Deep Disagreement, the Dark Enlightenment, and the Rhetoric of the Red Pill, in: *Journal of Applied Philosophy*, 36(3), 420-435.

Bernstein, M.; Monroy-Hernández, A.; Harry, D.; André, P.; Panovich, K; Vargas, G. (2011): 4chan and /b/. An Analysis of Anonymity and Ephemerality in a Large Online Community, in: *Proceedings of the International AAAI Conference on Web and Social Media*, 5(1), 50-57.

Boulianne, S.; Lee, S. (2022): Conspiracy Beliefs, Misinformation, Social Media Platforms, and Protest Participation, in: *Media and Communication*, 10(4), 30-41. [https://doi.org/10.17645/mac.v10i4.5667].

Brockhaus, G. (2020): Emotionale Dilemmata im Umgang mit Hasspolitik, in: *Freie Assoziation*, 23(1+2), 84-104.

Bucher, T.; Helmond, A. (2018): The Affordances of Social Media Platforms, in: Burgess, J.; Marwick, A.; Poell, T. (Hg.), The SAGE Handbook of Social Media, Los Angeles u.a.: SAGE Publications Ltd., 233-252.

Ceffinato, T. (2020): Zur Regulierung des Internet durch Strafrecht bei Hass und Hetze auf Onlineplattformen, in: *Zeitschrift für die gesamte Strafrechtswissenschaft*, 132(3), 544-563.

Coleman, E.G. (2015): Hacker, hoaxer, whistleblower, spy. The many faces of anonymous, London/New York: Verso Books.

Coleman, E.G. (2020): Logics and Legacy of Anonymous, in: Hunsinger J.; Allen M.M.; Klastrup, L. (Hg.), Second International Handbook of Internet Research, Dordrecht: Springer Netherlands, 145-166.

Colley, T.; Moore, M. (2020): The challenges of studying 4chan and the Alt-Right. ›Come on in the water's fine‹, in: *New Media & Society*, 24(1), 5-30. [https://doi.org/10.1177/1461444820948803].

Demmerling, C.; Landweer, H. (2007): Philosophie der Gefühle. Von Achtung bis Zorn, Stuttgart: J.B. Metzler.

Doğru, B.I. (2020): »For the lulz, mein Fuehrer«. Humor als strategisches Element der Enthemmung in der »Neuen« Rechten, in: *Freie Assoziation*, 23(1+2), 15-34.

Douglas, N. (2014): It's Supposed to Look Like Shit. The Internet Ugly Aesthetic, in: *Journal of Visual Culture*, 13(3), 314-339.

Dunn Cavelty, M.; Jaeger, M.D. (2015): (In)visible Ghosts in the Machine and the Powers that Bind. The Relational Securitization of Anonymous, in: *International Political Sociology*, 9(2), 176-194. [https://doi.org/10.1111/ips.12090].

Eisenberg, A.K. (2020): Den Hass kriminalisieren: Rechtsgrundlagen und Vollzugsdefizite bei der strafrechtlichen Verfolgung von Hassdelikten in den Vereinigten Staaten, in: *Zeitschrift für die gesamte Strafrechtswissenschaft*, 132(3), 644-665.

Elley, B. (2021): »The rebirth of the West begins with you!« – Self-improvement as radicalisation on 4chan, in: *Humanities and Social Sciences Communications*, 8(1), 1–10. [https://doi.org/10.1057/s41599-021-00732-x].

Engelen, E.-M. (2008): Eine kurze Geschichte von Zorn und Scham, in: *Archiv für Begriffsgeschichte*, 50(1), 41-73.

Fielitz, M.; Marcks, H. (2020): Digitaler Faschismus. Die sozialen Medien als Motor des Rechtsextremismus, Berlin: Dudenverlag.

Ganesh, B. (2020): Weaponizing white thymos. flows of rage in the online audiences of the alt-right, in: *Cultural Studies*, 34(6), 892-924.

Gillespie, T. (2018): Custodians of the Internet. Platforms, content moderation, and the hidden decisions that shape social media, New Haven (CT): Yale University Press.

Glucksmann, A. (2005): Hass. Die Rückkehr einer elementaren Gewalt, München/Wien: Nagel & Kimche.

Griffin, R. (2020): Faschismus. Eine Einführung in die vergleichende Faschismusforschung, Stuttgart: ibidem-Verlag.

Hakoköngäs, E.; Halmesvaara, O.; Sakki, I. (2020): Persuasion Through Bitter Humor. Multimodal Discourse Analysis of Rhetoric in Internet Memes of Two Far-Right Groups in Finland, in: *Social Media + Society*, 6(2), 1–11. [https://doi.org/10.1177/2056305120921575].

Hannan, J. (2018): Trolling ourselves to death? Social media and post-truth politics, in: *European Journal of Communication*, 33(2), 214-226.

Herwig, J. (2011): Fluktuierende Kollektive, lebendiges Archiv. semiologische Praktiken im Imageboard 4chan, in: Ehardt, C.; Pillgrab, D.; Alge, B. (Hg.), Inszenierung von »Weiblichkeit«. Zur Konstruktion von Körperbildern in der Kunst, Wien: Löcker.

Hine, G.E.; Onaolapo, J.; De Cristofaro, E.; Kourtellis, N.; Leontiadis, I.; Samaras, R.; Stringhini, G.; Blackburn, J. (2017): Kek, Cucks, and God Emperor Trump. A Measurement Study of 4chan's Politically Incorrect Forum and Its Effects on the Web, in: Ruths, D. (Hg.), Proceedings of the Twelfth International AAAI Conference on Web and Social Media (ICWSM 2017), Cambridge (MA): AAAI Press, 92-101.

Hodge, E.; Hallgrimsdottir, H. (2020): Networks of Hate. The Alt-right, »Troll Culture«, and the Cultural Geography of Social Movement Spaces Online, in: *Journal of Borderlands Studies*, 35(4), 563-580.

Horten, B.; Gräber, M. (2021): »Hatespeech« – Der Hass im Netz. Kriminologischer Beitrag, in: *Forensische Psychiatrie, Psychologie, Kriminologie*, 15(1), 91-94. [https://doi.org/10.1007/s11757-020-00644-7].

Jäger, L.; Kracher, V.; Manemann, T. (2021): Fashwave. Rechtsextremer Hass in Retro-Optik, in: de:hate report Nr. 2, Berlin. [https://www.amadeu-antonio-stiftung.de/wp-content/uploads/2021/06/de.hate_Report02_Fashwave-1.pdf] (Zugriff: 13.04.2024).

Martin, S.D.; Vukadinović, V.S. (2020): Humor ist, wenn man trotzdem lacht. Anmerkungen zu Rechtsextremismus, Ressentiment und Lustgewinn im Internet, in: *Freie Assoziation*, 23(1+2), 106-111.

Munn, L. (2019): The High/Low Toggle. Alt-Right Code Switching on 8chan, in: *Navigationen – Zeitschrift für Medien- und Kulturwissenschaften*, 19(2), 149-160.

Reichardt, S. (2014): Faschistische Tatgemeinschaften. Anmerkungen zu einer praxeologischen Analyse, in: Schlemmer, T.; Woller, H. (Hg.), Der Faschismus in Europa. Wege der Forschung, München: Oldenbourg, 73-88.

Riemenschneider, S.; Lutz, M. (2021): #HateSpeech – Shitstorms als Kampfmittel organisierter Strukturen. Zugleich eine Anmerkung zum zivilrechtlichen Schutz des allgemeinen Persönlichkeitsrechts im Fall Künast, in: *Datenschutz und Datensicherheit – DuD*, 45(6), 371-374. [https://doi.org/10.1007/s11623-021-1453-y].

Rothbart, D. (2021): Righteous rage as political power, in: *Peace and Conflict: Journal of Peace Psychology*, 27(4), 681-684. [https://doi.org/10.1037/pac0000544].

Sloterdijk, P. (2006): Zorn und Zeit. Politisch-psychologischer Versuch, Berlin: Suhrkamp.

Strick, S. (2021): Rechte Gefühle. Affekte und Strategien des digitalen Faschismus, Bielefeld: transcript Verlag.

Szanto, T. (2020): In hate we trust. The collectivization and habitualization of hatred, in: *Phenomenology and the Cognitive Sciences*, 19(3), 453-480. [https://doi.org/10.1007/s11097-018-9604-9].

Thorleifsson, C. (2022): From cyberfascism to terrorism. On 4chan/pol/ culture and the transnational production of memetic violence, in: *Nations and Nationalism*, 28(1), 286-301. [https://doi.org/10.1111/nana.12780].

Wagener, A. (2017): Lauren Mayberry vs. 4chan's online misogyny. A Critical Discourse Analysis Perspective, in: *Lodz Papers in Pragmatics*, 13(2), 303–325. [https://doi.org/10.1515/lpp-2017-0015].

Wentz, D. (2019): Krieg der Trolle. Digitale Reproduzierbarkeit und ›Memetic Warfare‹, in: *Navigationen – Zeitschrift für Medien- und Kulturwissenschaften*, 19(2) [Themenheft: Neue Rechte und Universität], 135–148. [https://doi.org/10.25969/mediarep/13811].

III
Vernunftverhältnisse

Die Philosophie des Digitalen

Zur Struktur, Signatur und Phänomenologie des Digitalen

Gabriele Gramelsberger

Abstract: *The digital has become an integral part of everyday life. It is everywhere and all-encompassing. The relevance of the digital is a challenge for philosophy. How can we understand the digital as an all-encompassing phenomenon? What are the ontological, phenomenological and epistemological aspects of digitality? What is machinic rationality and how does it change our lifeworld? The contribution aims at a systematic approach to the digital. It deals with the structure and signature of the digital.*

Keywords: *digitality; semiotics; digital objects; digital signature*

1. Allgegenwart des Digitalen

Die Allgegenwärtigkeit des Digitalen nimmt nicht nur tiefgreifende Veränderungen am Lebensweltlichen vor, sondern lässt sich für die Philosophie aufgrund seines schieren Umfangs nicht mehr marginalisieren. Vor diesem Hintergrund formiert sich aktuell eine Philosophie des Digitalen respektive der Digitalität, die weniger die technischen Errungenschaften der Digitalisierung, als vielmehr die lebensweltlichen Effekte der Digitalität in den Blick nimmt. Nach Jahren der technischen Euphorie, so formuliert dies Felix Stadler in seiner *Kultur der Digitalität*, nachdem »die Faszination für die Technologie abgeflaut ist [… wird deutlich, dass] Kultur und Gesellschaft in einem umfassenden Sinne durch Digitalität geprägt [sind]« (Stadler 2016: 20). Erst diese Ernüchterung ermöglicht die nötige Distanz, um über die »Verfestigung des Digitalen, die Präsenz der Digitalität jenseits der digitalen Medien, [… die] der Kultur der Digitalität ihre Dominanz« verleiht, nachzudenken (ibid.). Doch wie soll die Philosophie über das Digitale nachdenken? Bislang ist sie mit einem heterogenen Feld und disparaten Diskursen zu Teilaspekten der Digitalität konfrontiert, die von Bereichsethiken (Künstliche Intelligenz, Robotik, Daten, etc.) bis zu Disziplinen orientierten Philosophien (Medienphilosophie, Computerphilosophie, Simulations- und Modellierungsdebatte, etc.) reichen. Die Frage,

die sich daher stellt, ist die nach einem systematischen Ansatz einer Philosophie des Digitalen respektive der Digitalität, die ganz im Sinne der philosophischen Tradition »tradierte Fragestellungen […] durch die Digitalisierung revitalisiert und neu motiviert« (Noller 2022: 9). Eine solche Revitalisierung resultiert aus der lebensweltlichen Wirkmächtigkeit des Digitalen in Form seiner spezifischen Raum-Zeit-Logik (Ubipräsenz), seines spezifischen Verhältnisses von Subjektivität und Objektivität (Interobjektivität) sowie den neuen Bedingungen des Individuums durch die digitale Vernetzung (Transsubjektivität). Es liegt auf der Hand, dass sich durch diese neuen Verhältnisse, wie dies Jörg Noller in seinem Buch *Digitalität. Zur Philosophie der digitalen Lebenswelt* (Noller 2022) ausführlich diskutiert, die traditionellen Kategorien maßgeblich modifizieren.

Auch Yuk Hui hat mit seiner Untersuchung *On the Existence of Digital Objects* (Hui 2016) auf die ontologischen Transformationen des Digitalen aufmerksam gemacht. Die doppelte Verfasstheit des Digitalen, »objects that take shape on a screen or hide in the back end of a computer program, composed of data and metadata regulated by structures or schemas. […] They exist both on the screen, where we can interact with them, and in the back end, or inside the computer program« (Hui 2016: 1f.), sowie dessen Folgen sind hier von Interesse. Durch die immer dichter werdenden Datenstrukturen des »semantic web« erlangen die digitalen Entitäten nahezu einen Objektstatus und verweben sich zunehmend mit Subjektdaten zu einer eigenartigen Interobjektivität. »Digital humans«, die mit künstlicher Intelligenz und individuellen Personendaten ausgestatteten Avatare unserer selbst, werden zukünftig der Inbegriff interobjektiver-transsubjektiver Objekte sein.

Digitalität als Begriff für die lebensweltlichen Effekte des Digitalen – in Abgrenzung zur Digitalisierung als Begriff der technischen Effekte des Digitalen – legt den Fokus zumeist auf ontologische und ethische Fragestellungen.[1] Im Unterschied dazu stellt der hier skizzierte und in dem Buch *Philosophie des Digitalen zur Einführung* (Gramelsberger 2023) näher ausgeführte Ansatz eine andere Systematik respektive Forschungsprogramm zum Digitalen vor. Gefragt wird zum einen nach den Bedingungen der Möglichkeit des Digitalen, also nach seiner maschinenlogischen Struktur. Insbesondere die Ersetzungsverhältnisse, die im Prozess der Digitalisierung auftreten, sind hier von Interesse. Gefragt wird aber zum anderen auch nach den Selbstverständlichkeiten des Digitalen im lebensweltlichen Umfeld und damit nach seiner maschinenrationalen Signatur. Durch das Wechselverhältnis von Struktur und Signatur, also von Ersetzungsverhältnissen und Selbstverständlichkeiten, las-

1 Das CAIS Projekt *Philosophische Digitalisierungsforschung* (2019–2022), aus dem der vorliegende Band hervorging, und seit 2021 die *Arbeitsgemeinschaft Digitalitätsforschung* [https://digitale-philosophie.de/] der Deutschen Gesellschaft für Philosophie widmen sich grundlegenden Fragen zum Digitalen.

sen sich vor allem die epistemischen und phänomenologischen Aspekte des Digitalen systematisch erfassen und reflektieren.

2. Struktur des Digitalen

2.1 Grundlage des Digitalen und der Digitalisierung

Digitalisierung wird üblicherweise mit Binarisierung verwechselt, also der Beschreibung eines Zustandes in zwei Vorkommnissen: on/off, 1/0, alt/neu, etc. Doch Digitalisierung meint weitaus mehr und leitet sich vom englischen Begriff für Ziffer (*digit*) ab. Zu Beginn des 20. Jahrhunderts im Kontext der Elektrifizierung der Sprache in der Telefonie eingeführt, meint Digitalisierung die Diskretisierung und Quantisierung eines kontinuierlichen Sprachsignals. Das durch technische Vorgänge der Diskretisierung generierte Digitalsignal besteht aus diskreten Zahlwerten, die mit einem Binärcode dargestellt werden können, aber auch mit Dezimalzahlen oder anderen Einheiten. Erst die Kodierung der quantisierten Zustände mit Binärzahlen erzeugt die Grundlage digitaler Zustände im Computer, die wir heute gemeinhin als Digitalisierung bezeichnen.

Mögen diese technischen Details als Spitzfindigkeiten erscheinen, so offenbaren sie bei genauerer Betrachtung ein grundlegendes Ersetzungsverhältnis, das die Struktur des Digitalen maßgeblich prägt. Denn um die Quantisierung eines diskreten Signals in Bits (*binary digits*) zu ermöglichen, ändert sich seine Logik von einer statistischen in eine stochastische. Gemeint ist damit die in den 1940er-Jahren von Claude Shannon entwickelte *Mathematical Theory of Communication* (Shannon 1948), die als Informationsgehalt eines Signals die Informationsentropie (H) einführt. Dazu legt Shannon die Annahme zugrunde, dass ein Signal eine Folge diskreter Symbole einer bestimmten Sprache transportiert und dass diese Folge nicht zufällig ist, sondern eine gewisse statistische Struktur aufweist. »In general, they form sentences and have the statistical structure of, say, English. The letter E occurs more frequently than Q, the sequence TH more frequently than XP, etc.« (Shannon 1948: 385) Das entscheidende Ersetzungsverhältnis, das Shannon nun einführt, ist, die statistische Häufigkeit eines Zeichens in einer Sprache durch ein stochastisches Modell zu ersetzen. Dazu greift er auf die Untersuchung von Andrei Markov zurück, der zu Beginn des 20. Jahrhunderts anhand von Buchstabensequenzen in der russischen Literatur seine Wahrscheinlichkeitsrechnung (Markov-Ketten) entwickelt hatte (vgl. von Hilgers/Velminski 2007). In diesen Untersuchungen hat nicht nur das Digitale seinen Ursprung, sondern auch die Digital Humanities.[2]

2 Bereits vor Andrei Markov hatte Friedrich W. Kaeding 1898 das *Häufigkeitswörterbuch der deutschen Sprache mit Wort- und Silbenzählungen* herausgegeben. Dessen Korpus umfasste über 10

Der Grund, warum Shannon hier auf Markov zurückgreift, ist folgender: »We may consider a discrete source, therefore, to be represented by a stochastic process. Conversely, any stochastic process which produces a discrete sequence of symbols chosen from a finite set may be considered a discrete source.« (Shannon 1948: 385) Mit diesen stochastischen Prozessen wird nicht die gegebene Häufigkeit, sondern die mögliche Wahrscheinlichkeit des zukünftigen Eintretens einer Zeichenfolge angebbar. Je komplexer die Approximation durch ein stochastisches Modell (Markow-Kette) ist, desto mehr nähert sie sich der tatsächlichen Wahrscheinlichkeit des Vorkommens eines Symbols in einer natürlichen Sprache an; je niedriger die Komplexität, desto artifizieller ist die Häufigkeitsverteilung bis hin zur Gleichverteilung der Symbole. Erst auf dieser Basis kann Shannon den Informationsgehalt H einer Nachricht bestimmen. Statistische Regelmäßigkeit (Redundanz) bedeutet eine hohe Entropie. Basic English mit etwa 850 Wörtern weist dementsprechend eine hohe Entropie auf im Unterschied zu James Joyces Buch *Finnegans Wake*. Da die Entropie einer Symbolfolge (Nachricht) wiederum die notwendigen Übertragungskapazitäten (Bits pro Sekunde) bestimmt, erfordert eine verlustfreie Datenübertragung eine möglichst hohe Redundanz. Shannon definiert nun, dass eine Nachricht bestehend aus zwei gleichwahrscheinlich Zeichen 1 Bit zur Übertragung benötigt. Ob ein Bit mit den Ziffern 0/1 oder mit den logischen Werten falsch/wahr dargestellt wird, hängt von der gewählten Konvention des Alphabets und der Verwendung ab. Die materiale Verbindung zur Schaltungsalgebra eines Digitalcomputers, die Shannon bereits 1937 formuliert hatte, ergibt sich dann daraus, dass »a relay or a flip-flop circuit can store one bit of information« (Shannon 1948: 379). Aus diesen Flip-Flop-Schaltungen lassen sich die logischen Grundschaltungen, Speicher und Zähler eines Computers aufbauen und so durch logische Verschaltung Energie symbolisch urbar machen.

2.2 Grundlage der Digitalität

Die Reduktion der Sprache auf Bits sowie die Zusammensetzung einfacher Schaltelemente zu komplexen Schaltungen charakterisiert das Digitale in Form des Digitalcomputers als cartesische Maschine par excellence (vgl. Weizenbaum 1978). Doch Digitalität benötigt kybernetische Maschinen, um das Digitale in all seiner Komplexität zur Anwendung zu bringen. Die entscheidende Frage aus Perspektive einer Philosophie des Digitalen ist daher, wie sich diese Transformation von der cartesischen zur kybernetischen Maschine vollzieht, welche Ersetzungen, aber auch Erweiterungen dafür nötig sind.

Millionen Worte und 60 Millionen Buchstaben, die von über tausend freiwilligen Helfern auf Karteikarten zusammengetragen wurden. Kaedings Akribie diente der Verbesserung der Stenographie. Vgl. Bernhart 2015.

Diese Transformation ist eng mit den frühen Programmen der Entwicklung Künstlicher Intelligenz (KI) verknüpft und der Idee, auch Maschinen »Verhalten« zuzusprechen. Es ist die Grundidee der Kybernetik, Verhalten als Verallgemeinerung zwischen Menschen, Maschinen, Lebewesen und Systemen jeglicher Art zu verstehen. Dies ist nur möglich, wenn regelungsbasiertes Steuern zum zentralen Organisationsprinzip maschinellen, aber auch menschlichen Verhaltens erklärt wird, und wenn absichtsvolles Handeln als reglungstechnische Rückkopplung interpretiert wird: »Purposeful active behavior may be subdivided into two classes: ›feed-back‹ (or ›teleological‹) and ›non-feed-back‹ (or ›non-teleological‹). […] The feed-back is then negative, that is, the signals from the goal are used to restrict outputs which would otherwise go beyond the goal. All purposeful behavior may be considered to require negative feed-back.« (Rosenblueth et al. 1943: 19f.) Basierend auf diesen Ersetzungsverhältnissen werden Lebewesen wie Maschinen als Rückkopplungsmaschinen beschreibbar und Feedbacksignale als Informationen eines Systems über seinen Zustand interpretierbar. Versteht man dann noch Rückkopplung als Selbststeuerung hat man den Paradigmenwechsel von von-außen-gesteuerten zu sich-selbst-steuernden und -regelnden Maschinen vollzogen. Damit ist die Synthese des Digitalen in Form von Schaltalgebra, Informationsübertragung sowie reglungstechnischem, rückgekoppeltem Verschalten in zunehmend komplexere Datenprozessierungseinheiten vollendet. Diese Synthese ist die Grundlage, dass die Digitalität ihre lebensweltlichen Einflüsse entfalten kann.

Erst auf dieser konzeptuellen wie materialen Synthese des Digitalen kann nach der »Intelligenz« der Maschinen gefragt werden und zwar als Urbarmachung der maschinenlogischen Teleologie des »absichtsvollen« Handelns und der Selbststeuerung. Es ist bezeichnend, dass bereits zu Beginn der KI-Forschung »Denken« als Rechnen und Ableiten durch »Denken« als Problemlösungsverhalten abgelöst wurde. Nicht die Programmierung einer Maschine zur Lösung logischer Probleme, sondern die Beobachtung des Verhaltens »of a test person solving a logical problem, […] which leads to a psychological theory of human problem solving« (Newell/Simon 1961: 109), stand im Mittelpunkt. Dadurch wurde der introspektive Vorgang des Lösens logischer oder mathematischer Probleme durch beobachtbares, regelbasiertes Verhalten bei der Lösung solcher Probleme substituiert. Letzteres lässt sich dann problemlos als (behavioristische) Lernstrategie in Form negativer und positiver Verstärkung verstehen. Schon Alan Turing hatte 1950 in seinem berühmten Artikel *Computing machinery and intelligence* festgestellt:

> »The machine has to be so constructed that events which shortly preceded the occurrence of a punishment signal are unlikely to be repeated, whereas a reward signal increased the probability of repetition of the events which led up to it. These definitions do not presuppose any feelings on the part of the machine, I have done

> some experiments with one such child machine, and succeeded in teaching it a few things. […] Now the learning process may be regarded as a search for a form of behavior which will satisfy the teacher (or some other criterion).« (Turing 1950: 457ff.)

Es sind diese Ersetzungen »X als U« – Intelligenz »als« Problemlösen, Problemlösen »als« Problemlösungsverhalten, Problemlösungsverhalten »als« bestärkendes Verhalten, bestärkendes Verhalten »als« Lernen, etc. – die von der klassischen KI zum heute so erfolgreichen maschinellen Lernen mit Künstlichen Neuronalen Netzen (KNNs) führen (vgl. Gramelsberger 2022).

3. Signatur des Digitalen

Maschinelles Lernen (ML) hat in rasant kurzer Zeit die Signatur des Digitalen umgestaltet, wie kaum eine digitale Entwicklung zuvor. Der Begriff »Signatur« meint die charakteristische Erscheinungsweise des Digitalen im Sinne einer bestimmten Weise des Zeichen-Setzens (sīgnāre) und Anzeigens (sīgnificāre). Das Digitale ist ontologisch betrachtet rein semiotisch verfasst, noch dazu formal-semiotisch, und dennoch schreiben wir ihm Objekthaftigkeit, Akteurialität, Lebensweltlichkeit, Inhaltlichkeit, Intentionalität (Dennett 1971; Dennett 1984) und vieles mehr zu. Wie kann dies sein? Um die spezifischen Erscheinungsweisen des Digitalen aufzuklären, bietet es sich als philosophisches Forschungsprogramm an, die Selbstverständlichkeiten der Signatur des Digitalen offenzulegen. Schon 1963 forderte Hans Blumenberg in dem Artikel *Lebenswelt und Technisierung unter Aspekten der Phänomenologie* die Offenlegung der Selbstverständlichkeiten des Technischen. Dabei berief sich Blumenberg auf Edmund Husserls Konzept der Lebenswelt als Universum vorgegebener Selbstverständlichkeiten, wobei

> »das Selbstverständliche […] der Gegenbegriff zu jener ›Selbstverständigung‹ [ist], die für Husserl die eigentliche Aufgabe einer phänomenologischen Philosophie zu sein hat. […] Von dieser Art ist die Lebenswelt – unabhängig davon, ob sie jetzt als Vorwelt oder als Mitwelt betrachtet wird – als der zu jeder Zeit unerschöpfliche Vorrat des fraglos Vorhandenen, Vertrauten und gerade in diesem Vertrautsein Unbekannten. Alles, was in der Lebenswelt wirklich ist, spielt in das Leben hinein, wird genutzt und verbraucht, gesucht und geflohen, aber es bleibt in seiner *Kontingenz* verdeckt, d.h. nicht als auch-anderssein-könnend empfunden.« (Blumenberg 1963: 26)

Mittlerweile gehört das Digitale zum unerschöpflichen Vorrat des fraglos Vorhandenen und damit zur Lebenswelt. Philosophisches Forschen heißt dann, die Fraglosigkeit in Frage zu stellen, die Selbstverständlichkeiten offenzulegen. Dieses, an

Blumenberg und Husserl angelegte Forschungsprogramm führt zwangsläufig zu einer Phänomenologie des Digitalen und wird an drei Aspekten des Digitalen näher ausgeführt.

3.1 Notationale Ontik des Digitalen

Der erste Aspekt nimmt auf die notationale Ontik des Digitalen Bezug. Mit Nelson Goodmans Symboltheorie lässt sich das Digitale semiotisch (analog zu Notationen) durch »syntaktische und semantische Disjunktivität und Differenzierung« beschreiben (Goodman 1968: 150f.). Im Prozess der Digitalisierung wird ein kontinuierliches (Goodman: syntaktisch dichtes) Zeichensystem in ein diskretes (Goodman: syntaktisch disjunktes und differenziertes) Zeichensystem transformiert. Im Unterschied zu Notationen, die eindeutig (Goodman: semantisch disjunkt und differenziert) auf etwas Bezug nehmen (bspw. Tanzschritte in Tanznotationen), nimmt das Digitale aufgrund seiner formalen Struktur erst einmal nicht Bezug auf etwas. Die Inskriptionen der digitalen Zeichenträger (bits) formal-operativ verwendeter Zeichensysteme sind vakant und besitzen keine extrasymbolischen Erfüllungsgegenstände. Das Digitale verdankt sich ja gerade der »Kalkülisierbarkeit und Mechanisierbarkeit« von Zeichen, und beide sind zusammengenommen die »zwei Seiten jener Münze […], die wir Formalisierbarkeit« (Krämer 1988: 129) nennen. Digitale Welten sind Ansammlungen von Bits analog zur realen Welt, die aus Molekülansammlungen besteht. Digitalisierung lässt sich daher so verstehen, dass formal korrekte Welten von Bitansammlungen ohne Bedeutung generiert werden, in die wir Bedeutung hineinlesen.

Nichtsdestotrotz weist das Digitale eine (formale) notationale Ontik auf, die rein syntaktisch die Vielzahl digitaler Objekte und damit die phänomenologische Fülle der digitalen Wirklichkeit generiert. Die These ist, dass die vakanten (formal-operativen) Zeichenträger einer Sättigung bedürfen, die jedoch wiederum nur formal-operativ sein kann. In anderen Worten: Ein Programmierer oder eine Programmiererin muss angeben, was ein formal-operativer Maschinenbefehl bedeutet, also wie dieser dargestellt werden soll. Die Vakanz der Zeichenträger hat verschiedene Folgen: Zum einen lassen sich dieselben Bits unterschiedlich darstellen, d.h. ob derselbe Maschinencode ein Schriftzeichen, ein Zahlzeichen, ein Bildpixel oder einen Ton notiert, lässt sich in einem Computerprogramm frei regeln. Dies bedeutet aber auch andererseits, dass Datensammlungen ohne dazugehörige Programme nur Abfolgen von 0 und 1 ohne Bedeutungszuweisungen sind. Zweitens wird durch die Vakanz der formalen Zeichenträger Bedeutung durch formale Korrektheit ersetzt. Solange eine Zeichenfolge formal-korrekt gebildet ist, ist sie zulässig. Auch wenn im Objekt-Englisch »weder ein ›ktn‹ noch ein ›k‹ irgendeinen Erfüllungsgegenstand« hat und damit semantisch betrachtet sinnlos ist (Goodman 1968: 141), so kann ›ktn‹ im Formalen dennoch formal-korrekt sein, ab-

hängig von den Regeln der Zeichenbildung in formalen Systemen. Formal lässt sich allenfalls die Korrektheit prüfen, aber nicht die semantische Bedeutung. Schließlich hat die Korrektheit formaler Zeichenfolgen den Effekt zur Folge, dass sich eine Zeichenfolge durch eine andere »salva veritate« ersetzen lässt, wie dies schon Gottfried Wilhelm Leibniz 1693 in *Kalkül der Lage* formuliert hatte. Leibniz hatte im Kalkül der Lage den anschaulichen Ähnlichkeitsbegriff durch einen formalen Identitätsbegriff ersetzt und dieses Ersetzungsverhältnis ist derart grundlegend, dass ohne dieses keine formalen Welten generierbar wären. Formal lässt sich ein X durch ein Y ersetzen, wenn dies wahrheitserhaltend korrekt ist, und ein Y für ein Z, und ein Z für ein U. Auf diese Weise lassen sich Ketten an Ersetzungen generieren, die, neben expliziten Setzungen in Programmen, die notationale Ontik des Digitalen wesentlich bestimmen. Dies führt zur Frage, wie lange diese Setzungen und Ersetzungen noch von Programmierern und Programmiererinnen (Menschen) gemacht werden und wann dies KI-Algorithmen übernehmen werden. Noch viel tiefgreifender ist jedoch die Frage, wie lange notationale Zuschreibungen dann noch an unserem menschlichen Verständnis von Begriffsbedeutungen orientiert sein werden. Für Maschinen ist diese Darstellungsebene nicht nötig.

Vor diesem semiotischen Hintergrund wird auch die zunehmende Dichte digitaler Objekte verständlich, wie sie Hui beschrieben hat. Das Stichwort ist hier Interoperabilität, also die Verknüpfbarkeit von Daten mit Daten und noch mehr Daten zu kontextuell dichten Gebilden – eben digitalen Objekte. Mit Hilfe von Ontologien lassen sich diese Verknüpfungen auch semantisch verdichten und führen zu dem globalen Wissensgraphen des semantischen Webs. Diese Ontologien, wie sie aktuell zahlreich im Digitalen entstehen, sind nichts anderes als die konkrete Umsetzung von Platons Definitionsmethode der Dihairesis, um zur semantischen Bedeutung von Wörtern (Begriffen) zu gelangen. Ein schönes Beispiel gibt Platon im *Kritias*, wenn er zu einer Klassifikation von Früchten durch die Unterscheidung aller möglichen Arten von Früchten gelangt. In ganz ähnlicher Weise werden digitale Objekte zunehmend komplexere, interoperable (interobjektive) Wissensobjekte, die durch die Digitalisierung und Datafizierung zudem indexikale Realitätshaltigkeit beanspruchen. Doch diese indexikale Realitätshaltigkeit ist mit Vorsicht zu genießen, denn »not only do the objects have identities, but their components or predicates also have identities and are thus subject to control and manipulation« (Hui 2016: 68).

3.2 Temporale Latenz des Digitalen

Der zweite Aspekt des Digitalen ist phänomenologisch entscheidend und generiert sich aus der unfassbaren Schnelligkeit der Datenprozessierung heutiger Computer. Elektronische Computer konnten von Beginn an schneller rechnen als Menschen und aktuelle Supercomputer führen etwa 30 Billiarden Operationen in der Sekun-

de aus. Diese enorme Performanz sorgt dafür, dass das Digitale weitgehend latent bleibt und nur dann für uns zugänglich ist, wenn uns Nutzeroberflächen Zugang gewähren. Hinzukommt, dass das, was prozessiert wird, ikonoklastische Messdaten sind. Messdaten, die die »erfahrbaren spezifisch sinnlichen Qualitäten (Füllen) […] ex datis« rekonstruieren (Husserl 1996[1935]: 36ff.), werden heutzutage von nahezu jedem »smart object« erhoben. Ein Smartphone, beispielsweise, ist mittlerweile weniger ein Telefon als vielmehr eine Umgebungsmessstation. Winzige, millimetergroße Sensoren gehen permanent auf Tuchfühlung mit uns und vermessen unser Verhalten und den Zustand der Umgebung. Die Daten werden in Bruchteilen von Sekunden weitergeleitet, verarbeitet, interpretiert und in Form von adaptiven Feedbacks an uns rückgemeldet. Was uns Menschen als Echtzeit erscheint, entspricht Wochen oder Monaten für die Maschinen-Maschinen-Kommunikation. Diese Hyperfluidität der Daten umgibt uns mittlerweile wie eine zweite, zumeist latente »Natur«. Doch dies wirft die grundlegende Frage nach den phänomenologischen Effekten dieser temporalen Latenz auf. Wenn sich schon ab 16 Bildern ein Bewegtbildeffekt einstellt, welche phänomenologischen Effekte haben dann erst die enormen Datenmengen und Rechengeschwindigkeiten in ihrer Hyperfluidität für uns zur Folge (vgl. Gramelsberger 2016; Gramelsberger 2023)?

Eines ist aber gewiss: Durch die Hyperfluidität und Ikonoklastizität entzieht sich uns das Digitale immer stärker. Die Automatisierung durch KI tut dann ihr Übriges zur phänomenologischen Abkopplung von uns Menschen. Auf diese Weise mündet das Digitale zunehmend in eine unterschwellige Parallelwelt der Maschinen, die nur noch mit ebenso unterschwelligen Entscheidungsalgorithmen kontrollierbar und zugänglich bleibt. »Unterschwellig« meint jenseits unserer Wahrnehmungsmöglichkeiten. Diese Unterschwelligkeit prägt nicht nur das Digitale in seiner Erscheinungsweise, sondern befördert auch den Datenkapitalismus, ohne dass wir uns der Ausbeutung überhaupt bewusst werden können: zu schnell werden zu viele Daten von uns, unserer Umgebung, unseren Aktivitäten generiert, weitergeleitet und analysiert, denn der »Überwachungskapitalismus beansprucht einseitig menschliche Erfahrung als Rohstoff zur Umwandlung in Verhaltensdaten« (Zuboff 2018: 22).

3.3 Panoptikumseffekt des Digitalen

Dies hat den dritten Aspekt des Digitalen zur Folge, der aus der eklatanten Asymmetrie von Menschen und digitalen Objekten respektive von Nutzern und digitalen Plattformen resultiert. Die Rede ist vom Panoptikum des Digitalen, das durch wenige, kommerzielle Technologieplattformen orchestriert wird. Während sich uns das Digitale zunehmend entzieht und nur noch selektiv durch den »space of anticipation« der verhaltensbasierten Empfehlungsalgorithmen zugänglich ist (Thrift 2004: 175), offeriert sich das Digitale den Analysealgorithmen der Technologieplattformen

in großer Luzidität und Transparenz. Michel Foucaults Begriff des Panoptikums aus seiner Studie zu Gefängnissen, *Überwachen und Strafen* (Foucault 1993[1975]), trifft als Metapher für heutige Gesellschaftsformationen zu, aber mehr noch auf das Digitale in seiner aktuellen Form als Plattformökonomie (Huszár 2022). Doch neben ökonomischen Effekten sind es vor allem soziale Effekte, die es wert wären, philosophisch erforschet zu werden. Denn mit Foucault gesprochen, verstärkt sich die Isolierung des Subjekts in der Bubble oder der Echochamber durch die nicht-diskursiven Praktiken digitaler Technologieplattformen, aus welchen diese ihre Marktmacht wie gesellschaftlichen Machtansprüche als »soziale Medien« generieren. Diese Isolierung wird umso wirkmächtiger, je akteurialer und affektiver die Algorithmen werden.

Dieser Panoptikumseffekt zeigt sich in den aktuellen Geschäftsmodellen des Digitalen, die mögliche Handlungen von Individuen auf Basis statistischer Auswertungen vorhersagen. Nicht nur wird uns entsprechend auf unserem Verhalten angepasste Werbung präsentiert, Suchresultate optimiert oder vorsorglich Bestellungen vorbereitet, sondern eine statistisch falsch veranlasste Bewertung kann lebensweltlich extreme Auswirkungen für Individuen haben, wie dies Louise Amoore in *The Politics of Possibility* (Amoore 2013) ausführlich beschrieben hat. In dieser statistischen Analyse und daraus abgeleiteten Vorhersage liegt die Handlungsmächtigkeit der Algorithmen begründet, wie auch ihr lebensweltliches Potenzial.

4. Phänomenologisch-anthropologische Grundprobleme des Digitalen – ein kurzes Fazit

Mit dem skizzierten Forschungsprogramm lassen sich die Grundprobleme des Digitalen analysieren und aufzeigen. Da uns das Digitale mittlerweile wie eine zweite »Natur« umgibt und durch seine Unterschwelligkeit zumeist verborgen, aber auch unverständlich entgegentritt, dennoch von uns technisch kreiert wurde, konstituiert das Digitale einen eigenartigen Seinsmodus wie Erscheinungsweisen. Die zunehmende Abkopplung des Digitalen ruft aber vor allem phänomenologisch-anthropologische Grundprobleme einer Parallelwelt der Maschinen und Algorithmen auf. Nicht nur werden die technologischen Sphären zunehmend environmental, unterschwellig und unzugänglich (Weiser 1991; Hörl 2011), mit der Entwicklung der künstlichen Intelligenz und der Verfahren des maschinellen Lernens wird das Digitale auch zunehmend kognitiv unabhängig von uns. Die Frage, wie die Philosophie damit umgehen kann und soll, ist offen. Ohne auf technologische Singularitätsthesen (Chalmers 2010) oder transhumanistische Theorien (Bostrom 2005) eingehen zu wollen, stellt sich insbesondere aus technikphilosophischer wie phänomenologischer Perspektive die Grundproblematik des Auseinanderdriftens menschlicher und technologischer Sphären. Denn trotz zunehmender Abkopplungstendenz, sind wir Menschen immer enger mit dem environmental Digitalen verwoben.

Diese Verwebung und ihre Konsequenzen für das Individuum gilt es in den kommenden Jahren näher zu untersuchen und bezüglich seiner anthropologischen Konsequenzen zu hinterfragen (vgl. Gramelsberger 2024).

Literatur

Amoore, L. (2018): The Politics of Possibility. Risk and Security Beyond Probability, Durham (NC): Duke University Press.

Bernhart, T. (2015): Von Aalschwanzspekulanten bis Abendrotlicht. Buchstäbliche Materialität und Pathos im Häufigkeitswörterbuch der deutschen Sprache von Friedrich Wilhelm Kaeding, in: Klausnitzer, R.; Spoerhase, C.; Werle, D. (Hg.), Ethos und Pathos der Geisteswissenschaften, Berlin/Boston: De Gruyter, 165–190.

Blumenberg, H. (1981): Lebenswelt und Technisierung unter Aspekten der Phänomenologie, in: Ders., Wirklichkeiten, in denen wir leben. Aufsätze und eine Rede, Stuttgart: Reclam, 9–58.

Bostrom, N. (2005): In Defense of Posthuman Dignity, in: *Bioethics*, 19(3), 202–214.

Chalmers, D.J. (2010): The Singularity. A Philosophical Analysis, in: *Journal of Consciousness Studies*, 17(9-10), 7–65.

Dennett, D.C. (1971): Intentional Systems, in: *The Journal of Philosophy*, 68, 87–106.

Dennett, D.C. (1984): The intentional stance, Cambridge (MA): MIT Press.

Foucault, M. (1993[1975]): Überwachen und Strafen. Die Geburt des Gefängnisses, Frankfurt a.M.: Suhrkamp.

Goodman, N. (1968): Sprachen der Kunst. Entwurf einer Symboltheorie, Frankfurt a.M.: Suhrkamp.

Gramelsberger, G. (2016): Es schleimt, es lebt, es denkt. Eine Rheologie des Medialen, in: *Zeitschrift für Medien- und Kulturforschung*, 7(2), 155–167.

Gramelsberger, G. (2023): Philosophie des Digitalen zur Einführung, Hamburg: Junius.

Gramelsberger, G. (2024): Phänomenologisch-anthropologische Grundprobleme des Digitalen, in: Krämer, S.; Noller, J. (Hg.), Was ist digitale Philosophie? Phänomene, Formen und Methode, Paderborn: Brill | mentis, 31–47.

von Hilgers, P.; Velminski, W. (Hg.) (2007): Andrej A. Markov. Berechenbare Künste, Zürich/Berlin: Diaphanes.

Hörl, E. (2011): Die technologische Bedingung, Berlin: Suhrkamp.

Hui, Y. (2016): On the Existence of Digital Objects, Minneapolis: University of Minnesota Press.

Husserl, E. (1996[1935]): Die Krisis der europäischen Wissenschaften und die transzendentale Phänomenologie, Hamburg: Meiner.

Huszár, S.A. (2022): Die normalisierende Macht der Digitalisierung in der Arbeitswelt. Eine Analyse nach Michel Foucaults Machttheorie, in: *Junior Management Science*, 7(4), 932–944.

Krämer, S. (1988): Symbolische Maschinen. Die Geschichte der Formalisierung in historischem Abriß, Darmstadt: Wissenschaftliche Buchgesellschaft.

Leibniz, G.W. (1996): Kalkül der Lage, in: Ders., Philosophische Werke, hg. Buchenau, A.; Cassirer, E., 1. Bd., Hamburg: Meiner, 69–76.

Newell, A.; Simon, H. (1961): GPS, a Program That Simulates Human Thought, in: Billing, H. (Hg.), Lernende Automaten, München: Oldenbourg, 109–124.

Noller, J. (2022): Digitalität. Zur Philosophie der digitalen Lebenswelt, Basel: Schwabe Verlag.

Rosenblueth, A.; Wiener, N.; Bigelow, J. (1943): Behavior, purpose and teleology, in: *Philosophy of Science*, 10, 18–24.

Shannon, C. (1948): A Mathematical Theory of Communication, in: *Bell System Technical Journal*, 27, 379–423, 623–656.

Stadler, F. (2016): Kultur des Digitalen, Berlin: Suhrkamp.

Thrift, N. (2004): Remembering the technological unconscious by foregrounding knowledges of position, in: *Environment and Planning D: Society and Space*, 22(1), 175–190.

Turing, A. (1950): Computing machinery and intelligence, in: *Mind*, 49, 433–460.

Weiser, M. (1991): The computer for the 21st century, in: *Scientific American*, 265(3), 94–104.

Weizenbaum, J. (1978): Die Macht der Computer und die Ohnmacht der Vernunft, Frankfurt a.M.: Suhrkamp.

Zuboff, S. (2018): Das Zeitalter des Überwachungskapitalismus, Frankfurt a.M./New York: Campus.

Die Nicht-Vernunft der Chatbots

Was macht auf Large Language Models beruhende Künstliche Intelligenz philosophisch interessant?

Sybille Krämer

Abstract: *Humans and machines are constitutively different; but at the same time, technology is a genuine dimension of human existence. What does this ambivalence mean for the interpretation of contemporary Large-Language Models-based artificial intelligence? An anthropomorphizing interpretation should be avoided, as – this is the thesis – artificial intelligence is becoming a cultural technique that forms a dimension of digital literacy. The arguments are developed with reference to a phenomenon that is based on the written character of training data in the scale of entire collective memories. From the perspective of tokenization, the fundamental difference between human language interpretation and algorithmic token statistics becomes obvious: Humans and machines realize completely different ways of using language. What does this mean for issues such as epistemic blackboxing or the ethical trustworthiness of chatbot-generated texts? One thing is clear: artificial intelligence systems belong to the genre of ›non-reason‹ (*Nicht-Vernunft*). The exercise of reason and unreasonableness remains the prerogative of their human designers and users.*

Keywords: *anthropomorphism; chatbots; digital literacy; tokenization; Large Language Models (LLM); distributional semantics; cultural technique of flattening; Digital Humanities; trust*

1. Die konstitutive Andersartigkeit des Technischen

Seit es einen Diskurs über Künstliche Intelligenz gibt, artikuliert sich ein drängendes, immer wieder auftauchendes Ansinnen, das, was Maschinen können, in eine anthropomorphisierende Perspektive einzurücken. ›Anthropomorphisierend‹ heißt dabei: etwas, das aus menschlichen Zusammenhängen – seien diese biologisch oder kulturell – vertraut ist, auf die Technik als ein anzustrebendes Leistungsmuster zu projizieren. Damit werden Mensch und Maschine auf eine Vergleichsebene gerückt, die beide in ein Wettrennen einspannt, bei dem es natur-

gemäß zwei Optionen gibt: die Maschine wird – jetzt schon oder auf die Länge der Zeit – die Menschen übertreffen, ihn gar ersetzen, was für viele, wenn auch nicht für alle, apokalyptisch als Kontrollverlust und Entmachtungsszenario imaginiert wird. Oder es wird eine Domäne unersetzbarer Einzigartigkeit des Menschen identifiziert, in die einzudringen oder diese gar zu erobern allem, was maschinenhaft ist, grundständig verwehrt bleibt.

Schon der Turingtest (Turing 1950) – zumindest so, wie er von vielen verstanden wurde – evozierte die Idee einer auszutestenden Ununterscheidbarkeit von Mensch und Maschine. Wie umgekehrt die kritische Diskussion künstlicher Intelligenz gerne die situierte Verkörperung in einer Lebensform (Dreyfus 1972) oder die Beschränkung auf rein syntaktische Operationen (Searle 1980; Searle 1999) als unüberschreitbare Demarkationslinie gezogen hat.

Doch es gibt auch einen dritten Weg jenseits der Ersetzbarkeit respektive Unersetzbarkeit des Menschen durch technische Apparate. Er besteht in einem veränderten Ausgangspunkt: Mensch und Maschine als kategorial verschiedenartig anzusehen. Es ist daher ein Kategorienfehler, das, was die Maschine leistet, am Vorbild menschlicher Befähigung zu qualifizieren (Crockett 2024). Was immer die Maschine bewerkstelligt, auch dann, wenn das Ergebnis dem Menschenwerk ähnelt, macht sie auf eine ganz andersartige Weise. Allerdings: das ist nur die eine Seite. Denn die Pointe dieses Gedankens konstitutiver Andersartigkeit besteht darin, dass diese Grunddifferenz zugleich die Bedingung der Möglichkeit ist, dass Mensch und Technik in ein enges Interaktionsverhältnis treten können. Ein Verhältnis, das anthropologisch so tief verwurzelt ist und zugleich so weit reicht, dass die Dimension technischer Operativität genuin zur ›Natur‹ des Menschen gehört und eine allen seinen Praktiken inhärente Dimension verkörpert. Die Diversität von Mensch und Technik grundiert und eröffnet erst deren produktives Zusammenspiel.

Wir sind immer zugleich auch ein anderer. Das ist der methodische Ausgangspunkt der folgenden Reflexionen über die zeitgenössischen Chatbot Technologien Künstlicher Intelligenz. Also über eine Technik, die erstaunlicher- oder auch irritierenderweise Texte zu produzieren vermag, die häufig – wenn auch nicht immer – von Texten, die Menschen erzeugten, nicht mehr zu unterscheiden ist.

2. Worum es geht

Eine Familie von Algorithmen bzw. Sprachmodellen macht gegenwärtig Furore (Durt et al. 2023), die als game-changer oder iPhone Moment charakterisiert werden und – da die Chatbots GPT3/4 auch öffentlich verfügbar sind – von Millionen von Nutzer:innen gebraucht, sowie in fast allen gesellschaftlichen Bereichen von Bildungsinstitutionen, über den Journalismus bis zu Stammtischrunden mit Ausdauer kommentiert werden. Firmen (Microsoft, Google sowie die chinesischen

Baidu und Alibaba) arbeiten mit Hochdruck an diesen Systemen; auch daran, sie mit Suchmaschinen und Office-Lösungen zu verknüpfen. Kurzum: eine steilere Aufmerksamkeitskurve als jene, seit ChatGPT3 2022 von Open AI veröffentlicht wurde, hat Künstliche Intelligenz noch kaum erlebt; und dies trotz einer bis in die 50er Jahre des 20. Jahrhunderts zurückreichenden Vergangenheit, die reich war an – gerne mit saisonalen Metaphern bezeichneten – Auf- und Abschwüngen.

Was an dieser neuesten Entwicklung ist philosophisch aufschlussreich? Und was ist überhaupt das Sujet, ein irgendwie bemerkenswerter Kern dieser Entwicklung?

Es sind verschiedene Termini, die dabei kursieren: Im weiteren Sinne wird von Synthetischen Medien und Generativer Künstlicher Intelligenz gesprochen. Damit ist gemeint, dass Künstliche Intelligenz eingesetzt wird um Texte, Bilder, Animationen oder Quellcode zu erzeugen, die in dieser Form und Konstellation gerade nicht als Digitalobjekte im Netz existierten; daher vom System nicht einfach übernommen bzw. adaptiert, vielmehr synthetisch erzeugt werden. Keine Kopien und Plagiate werden produziert, vielmehr Originale, allerdings indem empirisch vorliegende Datenschnipsel oder Datenteilstücke – die Token – synthetisch kombiniert werden. Hinzu kommt ein entscheidendes weiteres Attribut: die Instruktionen und Anfragen – ›prompts‹ genannt – sind in Alltagssprache verfasst, setzen keine Programmierkenntnisse voraus und sind daher – wird das System öffentlich zur Verfügung gestellt – auch von allen einsetzbar, welche die entsprechende App bzw. den Link herunterzuladen und die eigene Emailadresse anzugeben bereit sind.

In einer eingeschränkteren, nur auf die Erstellung von *Texten* fokussierten Sicht geht es um Künstliche Neuronale Netze, die mit hunderttausenden von Textkorpora, Billionen von Worten trainiert werden. Indem als Grundlage das digitalisierte kulturelle Gedächtnis bevorzugt englischsprachiger Kulturen, riesige Büchersammlungen, sowie Webpages und Blogs dienen, entstehen sukzessive Large Language Models (LLMs), die in ihren intern aufgebauten Parametern weder für die Ingenieure, noch für ihre Nutzer außerhalb des je produzierten Outputs transparent sind.

Die Familie der Large Language Models hat viele zeitgenössische Anwendungen selbst auf dem Gebiet der Bildproduktion: ChatBot GPT in seinen verschiedenen Versionen ist nur eine davon. Dass Large Language Models nicht als Modelle für das menschliche Denken zu interpretieren sind, sei hier festgehalten (Mahowald et al. 2023). – Das also ist in grober Skizze das Panorama, vor dem sich unsere Überlegungen entwickeln.

Auch wenn philosophische Reflexionen – zumindest im deutschsprachigen Diskurs – oftmals orientiert sind darauf, mit Künstlicher Intelligenz verbundene technische Entwicklungen kritisch zu kommentieren, ist unsere Absicht eine andere: Unser Ziel ist diese Entwicklung zu *verstehen, zu begreifen,* vielleicht auch: darüber *aufzuklären*. Das allerdings ist nicht möglich, ohne zugleich einzusehen, um welches Verhältnis von Kontinuität *und* Bruch, Tradition *und* Disruption innerhalb der Evo-

lution unserer Medien es dabei geht. Denn auch wenn häufig ein revolutionäres Vokabular geboten erscheint, um den faszinierenden Leistungssprung gegenwärtiger Chatbots zu charakterisieren, hat diese computergenerierte produktive Kraft Vorläufer, an denen bereits zutage tritt, was sich gegenwärtig – wie durch einen Brennspiegel – verstärkt radikalisiert.

3. Was ist neu an der gegenwärtigen Chatbot Technologie?

Medien verändern sich in Schüben; ihr Einsickern in die Alltagspraktiken wiederum hängt von vielfältigen sozialen, also nicht-medialen Bedingungen ab und erstreckt sich meist über lange Zeiträume, um schließlich in ganz unterschiedlichen Medienkulturen zu resultieren. Es gibt also eine Kluft zwischen Medieninnovation und ihrer kulturtechnischen Diffusion. Doch das Beispiel der zeitgenössischen Chatbots ist von markant anderer Dynamik: die Medieninnovation und ihr Gebrauch – dessen Internationalisierung beflügelt wird durch das Übersetzungspotenzial der Maschinen, welche viele Muttersprachen akzeptieren – erfolgt nahezu in ›Echtzeit‹.

Dieses Schrumpfen des Zeitintervalls zwischen Innovation und massenweiser Nutzung verweist auf einen elementaren Tatbestand: die Komponenten, deren Zusammenwirken die Effekte der Synthetischen Medien hervorbringen – so einschneidend neuartig das alles auch erscheint – verfügen über eine meist gut sondierte Vorgeschichte. Allen voran geht es um die Technik Künstlicher Neuronaler Netze, die ein Thema ist schon seit den 40er Jahren des letzten Jahrhunderts (McCulloch/Pitts 1943). Selbstlernende und -optimierende Neuronale Netze zeigen dann seit ca. 2009 (Bengio 2009) überraschende Erfolge in Form selbstadaptiver Lernalgorithmen unter dem Stichwort des ›Deep Learning‹. Allerdings ist das ›Selbstlernen‹ eher Metapher, wenn nicht gar Mythos: Nicht nur werden die Künstlichen Neuronalen Netze trainiert anhand von Daten (Texten, Bildern, Videos), die Menschen mit kollektiver Intelligenz erzeugt, oft auch etikettiert und ausgezeichnet haben; sondern klar ist auch – dem höflich-rücksichtsvollen Kommunikationsgestus gegenwärtiger ChatGPTs unschwer ablesbar – dass Tausende von Clickworkern dafür sorgen, die ins Netz gestellten Chatbots nicht binnen kürzester Zeit zu Rassisten und Hetzern mutieren zu lassen – wie in der Vergangenheit allzu oft geschehen. Gleichwohl: Dass zeitgenössische Lernalgorithmen durch Training interne Modelle bilden, die meist im weiteren Gebrauch sich optimieren, ist eine bereits gut eingeführte Technik.

Auch der Umstand einer dialogisch anmutenden Interaktion mit der Maschine ist nicht neu. Joseph Weizenbaums Computerprogramm ELIZA (Weizenbaum 1966) ließ in den 70er Jahren des 20. Jahrhunderts die Wogen einer kritischen Diskussion von Künstlicher Intelligenz hochschlagen (Weizenbaum 2000[1972]): Denn dieses Programm wurde von einigen Nutzern als ein ›menschlicher‹ Kommunikati-

onspartner empfunden, der mit seinem ›Einfühlungsvermögen‹ psychotherapeutisches Resonanzvermögen nicht nur zeigte, sondern dieses gar übertreffe. Und das obwohl ELIZA lediglich auf geschickte Weise die jeweils zuletzt eingesetzten Worte in Fragen umformulierte – ohne nur irgendetwas wie ein Sinnverstehen dazu zu benötigen.

Das, was in der zeitgenössischen Chatbot Technologie und den ihr zugrundeliegenden Lernalgorithmen tatsächlich neu ist, realisiert sich vor allem auf der Ebene der *Datenarchitekturen* und des *Datenumgangs* und bezieht sich auf zwei Komponenten: (i) auf die exorbitant große Datengrundlage, die als Trainingsreservoir dient und diese seit ca. 2017 (ii) kombiniert mit einer ›Transformer-Technologie‹, welche sich auf ›aufmerksamkeitsgesteuerte‹ Lernverfahren des Systems bezieht – übrigens ein durch und durch fehlgeleiteter psychologisierender Ausdruck!

Zu (i): Die jede Vorstellungskraft sprengende Größe ist schnell charakterisiert. Die internen LLMs sind das Resultat von ca. 570 Gigabyte Texteingabe. Die Parameterzahl, welche die Modelle ausbilden und Voraussetzung ihres Inputanalysepotenzials sowie ihrer synthetischen Leistungen sind, umfassen 175 Milliarden, in anderen Ansätzen sind es schon weit mehr. Für das menschliche Vorstellungsvermögen – von unseren operativen oder gar kontrollierenden Fähigkeiten erst gar nicht zu sprechen – geht es um Größen angesiedelt jenseits aller Vorstellungskraft. Quantität schlägt definitiv um in bemerkenswerten Qualitätszuwachs. Dass das Quantitative hier eine neue Bedeutung erhält – je mehr Trainingsdaten, umso besser die Leistung frei nach dem Motto: ›mehr bringt mehr‹ –, wird von vielen vermerkt.

Zu (ii): Genau auf der Schwelle dieses Umgangs mit unvorstellbar großen Datenmengen, ist die Transformer-Technologie entscheidend, welche Rechenschritte eines Systems zu reduzieren hilft (Vaswani et al. 2017). Das zu entwickelnde Modell erschließt die ›Bedeutung‹ von Worten aus den Kontexten, in denen diese jeweils positioniert sind. Das für sich ist nicht neu und aus der Linguistik bekannt als Distributionelle Semantik, aber auch schon philosophisch sondiert mit der Idee, dass die Bedeutung eines Wortes der Inbegriff seiner Verwendungsweisen und d.h. auch seiner Wortkontexte und Wortnachbarschaften ist. Bei riesigen Datenkorpora steigt allerdings die Anzahl der zu den Wortkontextermittlungen nötigen Rechenschritte unermesslich. Transformer-Technologie nun ist ein Mechanismus, der mit dem operiert, was psychologisch als ›Aufmerksamkeit‹ bekannt ist und – seiner anthropomorphen Hülle entkleidet – datentechnisch lediglich Folgendes meint: Wenn das Wort ›Bank‹ zu disambiguieren ist im Sinne von ›Ruhebank‹ oder ›Finanzinstitut‹, dann heißt ›mit Aufmerksamkeit‹ zu lernen, dass das System nicht *alle*, sondern nur bestimmte Worte, die in der Vorgeschichte auftauchen, dann zur Spezifizierung dieses Wortes überhaupt in Betracht zieht. Also: Worte wie ›Park, Bäume, Spazierengehen etc.‹ bildet das eine und ›Gebäude, Stadt, Kunden, Kreditvergabe, Vorstand…‹ das andere Wortfeld, auf welche die Aufmerksamkeit zu richten ist, um den Verwendungskontext von ›Bank‹ zu ermitteln. Solche Worte mit Aufmerksamkeits-

index werden im internen Modell mit hohen Zahlen belegt bzw. dargestellt, denn das Modell arbeitet immer – daher das Wort ›Transformer‹: ›Umwandler‹ – mit der Übertragung von Texten in Vektorräume, letztlich in *Zahlendarstellungen,* mit denen gerechnet wird, um das plausibelste nächste Wort vorzuschlagen. Was statistisch dann ermittelt wird, ist die Verteilung von Elementen in den Punkte-Populationen in Vektorräumen, in welche die Daten zu transformieren sind. Deshalb spielt das Räumliche, also Nähen und Distanzen zwischen den Worten – letztlich zwischen Datenpunkten – eine so grundlegende Rolle. Dabei ist klar, dass die Plausibilität dessen, was der Chatbot produziert, sich an der Eloquenz menschlicher Kommunikation ausrichtet – und nicht an menschlicher Intelligenz, Denkkraft oder gar ›Wahrheit‹. Elena Esposito (Esposito 2022) schlägt daher vor, dass Künstliche Intelligenz als apparative Künstliche Kommunikation zu verstehen sei.

Das ist natürlich alles eine allzu grobschlächtige Beschreibung. Doch geht es nur darum deutlich zu machen, dass die Neuartigkeit der zeitgenössischen Chatbots in der Dimension des operativen Datenumgangs liegt, verbunden mit sich beständig steigernder Rechenkraft der Hardware.

Doch wir richten nun unseren Fokus auf eine Dimension, die vielleicht so selbstverständlich ist, dass sie in Kommentaren kaum reflektiert wird: Und das ist der grundständige *Schriftcharakter* der Ein- und Ausgaben. Wobei ›Schriftcharakter‹ auch die prinzipielle Übertragbarkeit akustischer Ereignisse in Notation einschließt.

4. Operativität der Token

Es sind nicht einfach Worte, die das Operationsfeld der LLMs bilden, sondern Token. Token sind bedeutungslose kurze Buchstabenzusammenstellungen, wobei auch der Leerraum vor oder hinter einem Buchstaben zum Token zählen kann. Diese ›Leerräume‹ sind ein typisches Phänomen der Schriftbildlichkeit (Giertler/Köppel 2012), nicht unserer Lautlichkeit: die selbstverständlichen Pausen im Sprechen – zumeist solche zum Atemholen – sind der diskreten Zwischenräumlichkeit im Schriftbild, also den regulären Leerstellen und Lücken im Weißraum der Texte, gerade nicht kongruent.

Doch Digitalität bedarf der unterscheidenden Disjunktivität, also der Diskretheit und eindeutigen Differenzierbarkeit zwischen den Zeichen. Zu digitalisieren heißt, etwas, das relativ kontinuierlich und in dieser Hinsicht ›analog‹ ist, in diskrete Elemente zu zerteilen, die codiert und arbiträr miteinander kombinierbar sind (Gramelsberger 2023: 90). Wobei das, was als kontinuierlich und das, was als diskret gilt, relativ sind. Wird ein Drucktext digitalisiert, so nimmt er den Platz des Analogen ein und diese Rollenzuweisung des funktionell ›analogen Parts‹ erfolgt auch bei der Sequenzialisierung bereits digitalisierter Texte in eine Abfolge von Token. Wir

vernachlässigen an dieser Stelle, dass diese Token dann wiederum als Zahlen dargestellt und bearbeitet werden, sondern halten nur fest: Ein- und Ausgabe erfolgen in natürlicher Sprache, doch die maschinelle Verarbeitungsbasis sind Tokenstrukturen der Schrift. Die ›Tokenisierung‹ – ein in diesem Zusammenhang neu geprägtes Wort – bildet die Voraussetzung der generativen Künstliche Intelligenz.

Doch warum ist das bemerkenswert? Besteht unser Sprechen nicht ebenfalls darin, Phoneme – gemäß der Linguistik die kleinsten bedeutungslosen, aber Bedeutungen spezifizierenden Sprechelemente – aufeinanderfolgend hervorzubringen im akustischen Fluss der Rede? Doch eben dies täuscht: Wir können durch die Aneinanderreihung von Phonemen keinen natürlichsprachlichen Ausdruck erzeugen, so wie Buchstabenfolgen einen Text ergeben (Lüdke 1969). Vielmehr erweist sich das Phonem als ein linguistisches Konstrukt und Abkömmling des Graphems: entstanden aus der Projektion der Buchstabenstruktur auf die mündliche Sprache (Stetter 1997). Derrida hatte recht mit seiner zuerst einmal irritierenden Annahme, dass die Schrift der Sprache vorausginge (Derrida 1988). Denn erst die alphabetische Transkription des Sprechens löst die kommunikative Gesamtäußerung mit ihrem typischen Zusammenspiel von Prosodie, Gestik, Mimik, Verbalität und Deixis auf, isoliert und präpariert den verbalen Strang durch Transkription und erzeugt das Sprachliche als ein solitäres Objekt, das als eigenständige Entität – eben als ›die Sprache‹ – erst beobachtbar wird.

Halten wir fest: Die Bedingung der Möglichkeit, dass Künstliche Intelligenz in der zeitgenössisch generativen Form wirksam werden kann, ist die im Schriftcharakter der Datenkorpora gründende ›Tokenisierung‹.

Diese Token-Perspektive verkörperte eine Dimension der Schriftsprachlichkeit, zu der Menschen gewöhnlich keinen Zugang haben. Eine verschriftete Sprache zeigt Regelmäßigkeiten in der Häufigkeit, der Verteilung und Zusammenstellung der Buchstaben. Andrei Markov (Markov 1912) berechnete früh schon stochastisch Buchstabensequenzen in der russischen Literatur. Doch die Mathematik dieser Regelmäßigkeiten spielt für die Kulturtechnik alphabetischer Literalität, also in der Perspektive menschlicher Praktiken, keine Rolle – jedenfalls außerhalb der Kryptologie, die Geheimschriften zu decodieren hat. Denn selbstverständlich kann die Aufdeckung der mathematischen Buchstabenrelationen der entscheidende Schlüssel sein in der Entzifferung einer Geheimschrift. Doch gewöhnlich bleiben diese mathematisch identifizierbaren und analysierbaren Muster unterhalb der Oberfläche dessen, was beobachtbar und den Sprechenden, Schreibenden und Lesenden bewusst ist. Sinnhaftes Sprachverständnis und berechenbare Sprachstatistik scheinen einmal mehr den konstitutiven Unterschied von Mensch und Technik aufzurufen.

Und doch ist diese definitive Unterscheidbarkeit nun zu relativieren. Die Differenz zwischen human verstehbarem Sinn und maschinell berechenbarer Statistik als zwei unterschiedlichen Erschließungsformen von Sprache, zeigt nur die ›halbe

Wahrheit‹. Denn es gibt einen Punkt, wo beide sich berühren und auch dieses Zusammenspiel gründet wiederum im Schriftcharakter der Sprachdaten. Wenn – wie Wittgenstein, aber auch pragmatische Sprachtheorien es nahe legen – die Bedeutung eines Wortes sein Gebrauch ist (Durt et al. 2023), dann kommt die Verteilung von Worten in den Schriftbildern von Texten ins Spiel, denn diese Distribution ist signifikant für Wortbedeutungen. Worte sind charakterisierbar durch diejenigen Worte, die sie begleiten. Wie nah beisammen oder wie entfernt voneinander Worte in Textkorpora vorkommen, kann in ihren räumlichen Entfernungen jeweils gemessen und analysiert werden. Tritt ›Bank‹ in räumlicher Nachbarschaft zu ›Park‹ und ›Spazierengehen‹ etc. oder eher zu ›Kreditkonditionen‹ und ›Bilanzen‹ auf?

In einem viel radikaleren Sinne als Philosophen, Sprachwissenschaftler und Informatiker dies vermutet hätten, können aus diesen auf Textoberflächen in Erscheinung tretenden Wortrelationen – allerdings nur, sofern das Korpus der Sprachdaten genügend groß ist – maschinell Schlüsse gezogen, also Informationen mit dem Status eines ›Weltwissens‹ abgeleitet werden. Und dies, obwohl diese Informationen nicht explizit in den Trainingsdaten vorgelegen haben und ein LLM basiertes System – gewöhnlich, aber das wird sich ändern – auch keinen Zugang zur Außenwelt hat. Die Systeme können implizite Strukturen in Textkorpora erkennen und explizit machen – unter Umständen diese auch herbei phantasieren – kraft ihres synthetischen Vermögens. Wir kommen auf die halluzinatorischen Kräfte der LLMs noch zurück.

Die Linguistik der Distributionellen Semantik hat diesen Blickwinkel der Kontextabhängigkeit von Wortbedeutungen bereits entfaltet und damit eine Bedeutungstheorie vorgeschlagen, die unabhängig ist von einer denotativen, externen Weltreferenz (Rieger 1991). Worte haben eine ähnliche Bedeutung, sofern sie in ähnlichen Wortumfeldern vorkommen (Firth 1957). Und auch dabei ist klar, dass die Distributionelle Semantik Sprache in ihrer schriftsprachlichen Version zur Grundlage machen muss: Schriften sind räumliche Anordnungssysteme (Ehlich 2012). Deshalb erst kann mit der Schrift die Idee eines Sprach*raumes* entstehen, in dem Verteilungen, Positionierungen, Entfernungen etc. analysierbar werden.

Halten wir fest: Die embryonale Digitalität (Krämer 2022) des alphanumerischen Zeichenraumes, welcher Alphabete ebenso umfasst wie das dezimale Positionssystem, bildet eine wesentliche Grundlage gegenwärtiger Chatbot-Techniken und markiert eine notwendige Bedingung der zeitgenössischen synthetischen Künstlichen Intelligenz. Es ist zugleich der Faden, den die gegenwärtige Technik mit der Tradition und Historie der Kulturtechniken der Literalität verbindet. Neu ist dabei, dass das tradierte Nadelöhr zur Maschineninstruktion, das darin besteht, Programmiersprachen, also formale Schriften einsetzen zu müssen, bei Interaktionen mit den LLM basierten Chabots entfällt. Durch die Möglichkeit natürlichsprachlichen In- und Outputs kann diese Technologie den Status einer Kulturtechnik annehmen, die nicht nur in Expertenkreisen, sondern im

Alltag angekommen ist und überdies durch jede bestätigte Anwendung sich optimiert. Gerade die Effizienz, mit der die zeitgenössische Künstliche Intelligenz als ›Interaktionspartner‹ auf der Operationsbasis von Token agiert zeigt, dass jedweder Anthropomorphismus fehl am Platze ist: Ein Chatbot versteht nicht, was er kommuniziert. Allerdings sollten wir nicht vergessen, dass auch Menschen sich miteinander verständigen können, ohne sich – im emphatischen Sinne – verstehen zu müssen.

5. Der Computer – eine Oberflächentechnologie

Wir wollen uns der grundständigen Differenz zwischen Menschen und Computer noch einmal in anderer Perspektive annähern. Der Terminus ›Computer‹ sei hier eine Chiffre für alle möglichen Varianten digitalisierter Algorithmensysteme inklusive entsprechender Hardwareapparaturen. Die unvorstellbar großen Datenreservoire, welche Computer auf Muster hin durchsuchen und aus denen auch neue Muster synthetisiert werden können, verkörpert eine – um es mit einem Wortmonster auszudrücken – ›Oberflächenbezugnahmetechnologie‹. Denn es gibt einen Zusammenhang zwischen ›ein Muster sein‹ und ›Oberflächen inskribiert zu sein‹. Selbst so komplexe Figuren wie komponierte Musik können in Teilaspekten ihrer Performanz als Schriftmuster auf den Notenblättern der Partituren dargestellt werden. Überdies können Muster den Status von Spuren annehmen und in der Forensik der Spur übertrifft das Analysevermögen von Computern jeden Menschen. Um eine etwas schiefe optische Analogie zu bemühen: Computer sind Teleskope und Mikroskope in Datenuniversen. ›Schief‹, weil bei diesen Metaphern der synthetische, generative Aspekt entfällt.

Doch unser Fokus hier ist ein anderer: Zwar scheint wieder der Anlass gegeben, die Unterschiedlichkeit von Menschen und Maschine zu betonen: Die Maschine vermag die umfangreichen, inskribierten und illustrierten Oberflächen sozialer Gedächtnisse zu durchmustern, während Menschen bei den nur wenigen Texten, die sie lesend zu erschließen in ihrem Leben überhaupt in der Lage sind, auf die Hermeneutik einer Tiefeninterpretation angewiesen sind. Und doch wäre es falsch der ›Oberflächlichkeit‹ computergenerierter Verfahren die ›Tiefe‹ hermeneutischer Interpretation seitens der Menschen entgegenzuhalten (Krämer 2023b). Wir sind zwar sozialisiert mit dem Narrativ einer Rhetorik, das tiefgründiges Denken erwünscht und fruchtbar, auf die Oberflächen orientiert zu bleiben dagegen diskriminiert, wenn nicht gar als ›oberflächlich‹ tabuisiert. Doch menschliche Erkenntnisarbeit – und nicht nur diese – ist undenkbar ohne den Einsatz von Bildern, Schriften, Diagrammen, Graphen oder Karten, mithin inskribierter und illustrierter artifizieller Flächen (Krämer 2016; Krämer 2022b). Alle Wissenschaften, viele Künste, komplexe Architektur und Technik sowie die Verwaltung großer Organisationen

sind nicht machbar ohne eine ›Kulturtechnik der Verflachung‹. In der Ubiquität des Smartphones kulminiert diese Entwicklung. Die Projektion in die Zweidimensionalität bleibt kein schlichtes Reduktions- oder Abstraktionsverfahren, sondern verkörpert eine kulturstiftende und epistemische Produktivkraft (Krämer 2016).

Dieser Einsatz der artifiziellen Flächigkeit ist die ›Ebene‹, auf der sich Mensch und Maschine treffen. Eine Ebene, die den Nährboden operativer Kontrollierbarkeit zugleich bereitstellt, wie sie ihn auch wieder zunichtemachen kann.

Alles was körperlich ist, inklusive unsere leiblich-räumliche Situierung, ist durch drei senkrecht aufeinander stehende Achsen charakterisierbar: oben/unten, rechts/links, vorne/hinten. Notgedrungen bleibt im dreidimensionalen Raum erst einmal für menschliche Wesen verborgen und außer Kontrolle, was sich hinter ihnen befindet: Die Erfindung inskribierter und illustrierter Flächen mit ihren zwei Anordnungsrichtungen: oben/unten und rechts/links amputiert genau diese dritte, unübersichtliche, die verborgene Tiefendimension und schafft oder suggeriert einen Raum perfekter Übersicht und Kontrolle, erst recht wenn dieser – daher der Siegeszug des Papiers – überaus mobil ist (Krämer 2022a). Nicht zufällig hat das erste niedergeschriebene Computerprogramm von Ada Lovelace 1843 die Form einer Tabelle (Lovelace 1843; Krämer 2015): In horizontalen Zeilen einerseits und vertikalen Kolumnen andererseits entblättern sich die verschiedenen Maschinenzustände im Akt des Berechnens einer bestimmten Zahlenfolge (in diesem Falle der Bernoulli Zahlen). Mit dieser tabellarischen Instruktion wurde die ›analytical engine‹, die Charles Babbage als ersten Universalcomputer auf Papier entwarf – im Prinzip – von Ada Lovelace in eine konkrete Maschine für spezielle Zwecke verwandelt.

Der Computer ist also nicht nur eine forensische, er ist eine diagrammatische Maschine (Mackenzie 2017), die von zweidimensionalen Inskriptionen zehrt, auch wenn diese wiederum in die Linearität von Bitfolgen zu verwandeln sind. Die Turingmaschine, die Alan Turing als mathematische Präzisierung des Algorithmenbegriffs entwickelte und mit der er auf grundständige Weise ausbuchstabiert, was es heißt einem Algorithmus zu folgen, hat die Form einer Tabelle bzw. Tafel, in welche die vier grundlegenden Rechenschritte einer Maschine, nämlich die Beschriftung eines Arbeitsfeldes, das Nach-rechts-gehen, das Nach-links-gehen, sowie das Stoppen in zweidimensionaler Anordnung notiert ist.

Die Stationen computeraffiner Diagrammatizität sind hier nicht abzugehen. Doch in dem Augenblick, in dem die Datenkonvolute inskribierter Flächen zu Massendaten werden, die als Trainingsgrundlage selbstadaptierender Lernalgorithmen fungieren, ändert sich etwas grundsätzlich. Ursprünglich ist der epistemische Gebrauch artifizieller Flächigkeit auch eine Strategie der Schaffung von Übersicht, Transparenz und Kontrolle. Doch genau dies ist angesichts der gesammelten Mannigfaltigkeit kollektiver digitaler Textproduktion im Umfang kultureller Ge-

dächtnisse nicht mehr zu realisieren. Jorge Louis Borges ›Unendliche Bibliothek‹ (Borges 2013) liefert dazu ein passendes Bild.

Und so kommen mit den Large Language Models die Dunkelkammern unübersichtlicher Tiefenregionen zurück. Und nicht erst mit diesen. Für die vernetzten Nutzer:innen, die vor den Bildschirmen Texte und Bilder anschauen und bearbeiten, ist immer schon klar, dass die rhizomartige Verknüpfung und Interaktion zwischen den Computern, Algorithmen und Protokollen, die sich im Hintergrund der Bildschirme vollzieht, ihnen ebenso entzogen bleiben, wie die ins Unendliche sich vervielfältigenden Navigationsrouten, welche die Web-links eröffnen und die doch immer nur in den ersten Passagen individuell aktiviert werden können: Dass so unüberschaubar viele Wissenszugänge anzusteuern sind, wird untrüglich zum Anzeichen für all das, was zugleich *nicht* begangen, liegen gelassen wird und also gewusst werden kann. Doch nun gilt dieser Entzug von Wissen, einfach weil keine Übersicht und Kontrolle der intern entwickelten Modelle mit ihren in die Millionen gehenden Parameter überhaupt möglich ist, auch für die Ingenieure selbst.

Davon zeugt ein interessantes Phänomen: die Unvorhersehbarkeit eruptiver Leistungssteigerung im Zuge des Trainings der LLMs. Wenn Systeme mit einer fortschreitenden Erweiterung der Massendatenumfänge trainiert werden, zeigen sie über lange Phasen keinen mathematisch über dem Zufall liegenden Zugewinn hinsichtlich der anzutrainierenden Fähigkeiten. Doch dann zeigt sich wie in plötzlicher Eruption, ein steiler Zuwachs in der Ausübung der anvisierten Fähigkeit: die Lernkurve steigt steil (Wei et al. 2022) empor. Der Zeitpunkt der Emergenz dieser signifikant neuen Qualität auf Grundlage der Steigerung des Umfangs der Trainingsdaten, bleibt für die Beteiligten unvorhersehbar.

Diese Unvorhersehbarkeit gilt auch für jene Faktoren, die nicht einfach als Leistungssteigerung zu verbuchen sind, sondern umgekehrt die Risiken zeitgenössischer Chatbots markieren. In der technischen Literatur zu den LLMs werden viele solcher Risikobereiche sondiert. Wei et al. (Wei et al. 2022) heben drei Risiken hervor: Wahrhaftigkeit, Vorurteilsbehaftetheit, Toxikalität. Wobei unter ›Toxikalität‹ verstanden wird, dass die Reaktionen eines Chatbots auf Prompts gerade nicht hilfreich, nicht frei von Beleidigungen und nicht geprägt durch Respekt ausfallen. Für alle drei Bereichen hat sich gezeigt, dass mit Steigerung des Umfangs der Trainingsdaten, auch das Vorkommen dieser unerwünschten Eigenheiten von Chatbots zunehmen kann. Zugleich wird mit Hochdruck daran gearbeitet, wie dieses so unerwartete wie kommerziell desaströse Verhalten von Chatbots zu verändern ist. Überdies werden die Diagnosen dieser Risiken begleitet von Entwürfen für detaillierte Umgangsregeln für die Chatbot Nutzung (s. Antoniak et al. 2023).[1]

1 Antoniak et al. schreiben: »While the output of LLMs often looks very convincing, we recommend that you ask yourself the following questions before trusting it[:] Is this an appropriate use of an LLM, given the limitations of LLMs and the risks of my intended application? [/]

Dies macht klar, dass das Blackboxing, das zu einer Dimension der großen Sprachmodelle geworden ist, keineswegs identifiziert werden darf mit dem generellen Verlust an Kontrolle gegenüber den Systemen Künstlicher Intelligenz. Nur erstreckt die Kontrollierbarkeit sich auf den Output, nicht auf die interne Modellbildung. Gerade der Umstand, dass die gegenwärtigen Chatbots nicht mehr umkippen in Maschinen der Hetze und Hassrede, legt davon Zeugnis ab, in welchem Maße Regulierbarkeit und Kontrolle möglich sind. Kontrolle ist keine technische, sondern eine politische Frage.

Halten wir fest: Die ins Vielfache angewachsene Zugänglichkeit des kollektiven Wissens – von den Suchmaschinen bis hin zu den Chatbots – wird erkauft durch einen Zuwachs des Nichtwissens, durch das Blackboxing bezüglich interner Funktionsbereiche der Technologie. Und dies ist kein Betriebsunfall, der demnächst zu bereinigen wäre, sondern ist strukturell. Die Proportionalität zwischen Wissenszuwachs und Nichtwissen bildet nur das Echo jener Proportionalität, bei der die Erhöhung der Datenvolumina als Trainingsgrundlage, tatsächlich zu einer Erhöhung der Leistungen führt.

6. Künstliche Intelligenz wird zur Kulturtechnik

Unabhängig der Auf- und Abschwünge Künstlicher Intelligenz (Floridi 2020) galt doch zumeist, dass der Werkzeugkasten ihrer Verfahren für solche Domänen eingesetzt wurde, die als Expertenwissen zählten. Zwar änderte sich das, als mit Spamfiltern, Gesichts- und Spracherkennung, Smartphonefotografie, Marketingprofiling etc. auch die technischen Zurüstungen des digitalen Alltags durchdrungen werden von Künstlicher Intelligenz – doch diese bleiben zumeist ›stumme‹ Hintergrundverfahren, weder als spezifische KI-Technologie zugänglich noch überhaupt den Nutzern bewusst. Doch nun vollzieht sich darin eine signifikante Wende: Indem Chatbots der jüngsten Generation öffentlich nutzbar sind und das mit natürlichsprachlichen Eingaben, avanciert Künstliche Intelligenz zu einer sich im Alltag verankernden Kulturtechnik. Sie wird zu einer Facette der ›Kulturtechnik digitaler Literalität‹.

Is this an appropriate use of an LLM, given my own vulnerabilities or the vulnerabilities of people using the LLM? [/] Am I ok with my prompts being stored and shared with others? Is there any private information(medical history, finances) in my prompts? [/] Have I checked the accuracy of the output? Does the output contain information that I didn't ask for? [/] Am I asking the kind of questions where giving credit would be important, and if so, am I be able to identify the authors of the model's output so that I can credit them? [/] Does the output contain any opinions or advice, and if so, am I ok with my own opinions being influenced on this topic? [/] Do I have enough distance from the LLM, or am I interacting with the LLM as if it were a person (or encouraging others to interact with the LLM as if it were a person)?«. (Antoniak et al. 2023)

Was bedeutet ›Kulturtechnik‹? Um die Jahrtausendwende fand sich in Berlin eine Gruppe von acht Wissenschaftler:innen zusammen,[2] die das Helmholtz-Zentrum für Kulturtechnik an der Humboldt-Universität sowie einen neuartigen interdisziplinären Forschungsschwerpunkt ›Bild, Schrift, Zahl‹ begründeten (Krämer/Bredekamp 2003). Das Ziel war eine Umorientierung und Neujustierung im gängigen Kulturkonzept: Kultur sollte nicht länger primär als geistiges Phänomen begriffen werden, ausgerichtet an den Gipfelpunkten abendländischer Kunst und Bildung. Vielmehr galt es die konstituierende und konstruktive Rolle der Materialität, Medialität und Technizität alltäglicher Praktiken als kulturstiftende Dimensionen und zivilisatorische Partizipationsmöglichkeit zu begreifen und zu rekonstruieren.

Eingerückt in diesen Horizont befinden wir uns in einem Umbruch, bei dem die alphanumerische Literalität übergeht in die digitale Literalität und dieser Wandel schließt sukzessive Verfahren Künstlicher Intelligenz als genuine Facetten zeitgenössischer digitaler Kulturtechniken mit ein.

Dass Künstliche Intelligenz den Status einer Kulturtechnik anzunehmen beginnt, zeigt sich untrüglich darin, dass Institutionen wie Universitäten und Schulen darüber reflektieren und debattieren müssen, wie der Umgang mit dieser Technologie in Bildung und Ausbildung zu organisieren ist, wie das Prüfwesen sich verändern wird und welche Kontrollmöglichkeiten überhaupt zu erschließen und praktisch angewendet werden können. Fragen über Fragen, deren Antworten schwierig zu finden sein werden.

Auf einen weiteren Aspekt sei noch aufmerksam gemacht. Die Geisteswissenschaften haben ein Selbstbild gepflegt, das Interpretation und Hermeneutik gerne als Königsweg und Alleinstellungsmerkmal (ver)klärte (Krämer 2018). Doch solche Sicht ist problematisch, wenn nicht gar: falsch. Denn auch die Geisteswissenschaften haben von Anbeginn – denken wir an historische Datierungen, an Seitenzahlen, Konkordanzen, Werkkataloge – nicht nur mit Zahlen und Daten, sondern ›buchstäblich‹ auch mit Objekten und Materialien zu tun, welche gefunden, gesammelt, geordnet, ausgezeichnet, annotiert, verglichen, archiviert etc. werden müssen. Ohne dieses ›Handwerk des Geistes‹ wäre geisteswissenschaftliche Interpretation nicht möglich.

Es ist keine Frage, dass sich dieses Handwerk der Gelehrtenarbeit unter digitalen Bedingungen ändert. Die meisten Geisteswissenschaftler:innen sind – um es im Jargon zu sagen – ›analog unterwegs‹; sie setzen also gerade nicht die datengetriebenen Verfahren der Digital Humanities ein. Gleichwohl ist akademische Arbeit ohne das ›Abc‹ digitaler Grundoperationen kaum mehr praktizierbar. Das Lesen

2 Horst Bredekamp, Kunstgeschichte; Jochen Brüning, Mathematik; Wolfgang Coy, Informatik; Friedrich Kittler, Kulturwissenschaften; Sybille Krämer, Philosophie; Thomas Macho; Bernd Mahr, Informatik; Horst Wenzel, Mediävistik.

und Schreiben am Bildschirm, Kommunikation über Emails, gemeinsame Arbeit an Files, Nutzung von Suchmaschinen, digitalisierter Editionen und Quellen bestimmen den Alltag nahezu aller geisteswissenschaftlichen Arbeit.

Nun ist zu erwarten, dass diese digitale Literalität sich unter dem Einfluss jüngster Chatbots verändern wird. So wie heute Suchanfragen im Netz eine unerlässliche Dimension geisteswissenschaftlicher Arbeit sind, so werden in Zukunft Assistenzarbeiten von ChatGPT als da sind Literaturlisten zusammenstellen, Kurzinfos über Buchinhalte/Aufsätze geben, abstracts schreiben etc., viele Bereiche geisteswissenschaftlicher Arbeit unterstützen. Dass es um ›unterstützen‹ und nicht um ›ersetzen‹ geht, ist bedeutsam. Denn das Damoklesschwert über der akademischen Nutzung GPT produzierter Texte besteht in deren Fiktionalität – oft als deren Bereitschaft zu halluzinieren beschönigt. Ohne Überprüfung von Wahrheit und Faktizität, ist nachhaltige Chatbot-Aussagekraft nicht zu haben. Allerdings wird an der Sollbruchstelle ›Wahrheitsbezug‹ intensiv gearbeitet. Nur: gilt solche Unsicherheit bezüglich der Wahrheit eines Textes nicht letztlich für *jeden* Text? Doch die menschliche Kultur hat zur Bewältigung dieses Problems einen ›epistemischen Service‹ hervorgebracht. Was dieser mit dem Problemfeld der Chatbot Halluzinationen zu tun hat, sei im letzten Schritt des Essays sondiert.

7. Das Problem des Vertrauens

In 95% dessen, was wir wissen, verlassen wir uns auf Worte, Schriften und Bilder anderer. Vielleicht hören wir das nicht gerne, vielleicht ist uns das gar nicht bewusst: Dass wir unsere Überzeugungen eigenhändig auch rechtfertigen könn(t)en – wie es die Philosophie vorgibt bzw. erwartet – gilt nur für allzu wenige Wissensbereiche (Krämer 2023a). Zwar machte es historisch Sinn, dass die europäische Aufklärung in ihrem Versuch Erkenntnis von der kirchlichen, aber auch politischen Nabelschnur zu lösen, das Individuum stärkte; und also die Einzelnen ermutigte sich ihres eigenen Verstandes und ihrer Urteilskraft zu bedienen. Doch die Kollektivität unserer Intelligenz ist ebenso unumgänglich wie die Sozialität von Erkenntnis und Wissen. In vielen Hinsichten müssen wir dem, was andere sagen, schreiben oder zeigen, vertrauen. Erst dieses Vertrauen in Personen wie auch in Institutionen bereitet den Boden, aus dem ein Wissen durch die Worte anderer sich entwickeln kann.

Als Menschen neigen wir dazu denen zu vertrauen, die uns am ähnlichsten sind. Nicht zufällig waren es zu Beginn der Neuzeit gerade die ›Gentlemen‹, welche angesichts der Nicht-Reproduzierbarkeit naturwissenschaftlicher Experimente eingesetzt wurden, um deren Befunde zu bezeugen und zu verbürgen. Wir sind epistemisch abhängige Wesen (Krämer 2017): Das Vertrauen in andere ist also keineswegs nur praktisch, sondern auch epistemologisch von Belang. Dem Erkennen, geht das Anerkennen voraus.

Doch wie verhält es sich mit der Vertrauenswürdigkeit im Umgang mit Chat GPT erzeugten Texten? Hier zeigt sich ein Problem, das keineswegs auf Künstliche Intelligenz beschränkt ist, sondern immer schon maschinelle Textproduktionen und die Interaktion mit Technik z.B. in der Robotik begleitete. Es ist die Ambivalenz von Misstrauen einerseits und Übervertrauen andererseits in das, was eine Maschine, was Algorithmen tun.

Übervertrauen ist das Phänomen, wenn Menschen einem technischen System jenseits seiner aktual vorhandenen Kapazitäten trauen. Verstärkt wird Übervertrauen durch den psychologisch naheliegenden Mechanismus aus gelungenen vergangenen Erfahrungen auf die Zukunft zu schließen. Und dies, obwohl bezüglich vergangener ›guter‹ Erfahrungen keineswegs klar ist, ob deren Gelingen in der Richtigkeit und Angemessenheit eines Tuns wurzelte, oder einfach nur ›Glück‹ war. In der Robotik, insbesondere beim Einsatz sozialer Roboter, ist dieses Problem gründlich untersucht (Bahner 2008).

Die Eleganz, Natürlichkeit und Plausibilität, mit der die Large Language Models basierten Kommunikationsassistenten auf Eingaben reagieren, kann nicht nur einer anthropomorphisierenden Deutung der Chatbots als ›verstehenden Entitäten‹ Vorschub leisten, sondern kann auch massiv Phänomene wie eben das Übervertrauen befördern. Doch spätestens hier wird klar, mit welcher Skepsis und Distanz den Chatbots dann zu begegnen ist, wenn ihre Aussagen für ›bare Münze‹ gelten, sobald sie also als Instrumente akademischer oder pädagogischer Arbeit eingesetzt werden. Chatbots sind keine Suchmaschinen und keine Internetportale enzyklopädischer Information. Es ist unabdingbar die Variabilität der Informationsquellen des Internets zu nutzen, sobald Wahrheit im Spiele ist.

›Wahrheit‹ als ein Kriterium akademischer Arbeit zu etablieren, gehörte zum Telos der Europäischen Aufklärung. Und auch wenn die neuzeitliche Aufklärung sich selbst durch ihre kolonialen Verstrickungen ein Stück weit desavouierte, bleibt die Überprüfung des Wahrheitsbezuges ein durch sie auf die Agenda gesetztes epistemisches Potenzial, das auch ein Kernstück der Digitalen Aufklärung zu bleiben hat. Künstliche Intelligenz kann weder denken und erst recht nicht vernünftig oder unvernünftig sein. Ihre Systeme gehören dem Genre der ›Nicht-Vernunft‹ an. Vernunft und Unvernunft zu praktizieren bleibt ein Vorrecht ihrer menschlichen Konstrukteure und Nutzer.

Literatur

Antoniak, M. et al. (2023): Using Large Language Models With Care. How to be mindful of current risks when using chatbots and writing assistants, in: *AI2 Newsletter*, 2023(7). [https://blog.allenai.org/using-large-language-models-with-care-eeb17b0aed27] (Zugriff: 08.07.2023).

Bahner, J.E. (2008): Übersteigertes Vertrauen in Automation: Der Einfluss von Fehlererfahrungen auf Complacency und Automation Bias (TU Berlin Doctoral Thesis). [https://depositonce.tu-berlin.de/items/287a4685-ca1d-4972-94c5-beba68a10612] (Zugriff: 08.07.2023).

Bengio, Y. (2009): Learning Deep Architectures for AI, in: *Foundations and Trends in Machine Learning*, 2(1), 1–127.

Borges, J.L. (2013): Die unendliche Bibliothek, Frankfurt a.M.: Fischer.

Crockett, J. (2024): How to Raise Your Artificial Intelligence. A Conversation with Alison Gopnik and Melanie Mitchell May 31, 2024. [https://lareviewofbooks.org/article/how-to-raise-your-artificial-intelligence-a-conversation-with-alison-gopnik-and-melanie-mitchell/] (Zugriff: 08.06.2024).

Dreyfus, H. (1972): What Computers Can't Do, New York: MIT Press.

Durt, C. et al. (2023): Against AI Understanding and Sentience. Large Language Models, Meaning, and the Patterns of Human Language Use. [http://philsci-archive.pitt.edu/21983/ preprint].

Derrida, J. (1988): Signature Event Context, in: Ders. (Hg.), Limited Inc, Evanston: Northwestern University Press, 1–23.

Ehlich, K. (2012): Schrifträume, in: Krämer, S. et al. (Hg.), Schriftbildlichkeit, Berlin: Akademie Verlag, 39–60.

Esposito, E. (2022): Artificial Communication. How algorithms produce social intelligence, Cambridge (MA): MIT Press.

Firth, J.R. (1957): A synopsis of linguistic theory 1930–1955, in: *Studies in Linguistic Analysis*, 1957, 1–32 [Reprinted in: F.R. Palmer (Hg.) (1968), Selected Papers of J.R. Firth 1952–1959, London: Longman].

Floridi, L. (2020): AI and its New Winter. From Myths to Realities, in: *Philosophy & Technology*, 2020(33), 1–3.

Giertler, M.; Köppel, R. (Hg.) (2012): Von Lettern und Lücken. Zur Ordnung der Schrift im Bleisatz, München: Fink.

Gramelsberger, G. (2023): Philosophie des Digitalen zur Einführung, Hamburg: Junius.

Krämer, S. (2015): Wieso gilt Ada Lovelace als die »erste Programmiererin« und was bedeutet überhaupt »programmieren«?, in: Dies. (Hg.), Ada Lovelace. Die Pionierin der Computertechnik und ihre Nachfolgerinnen, München: Fink, 75–90.

Krämer, S. (2016): Figuration, Anschauung, Erkenntnis. Grundlinien einer Diagrammatologie, Berlin: Suhrkamp.

Krämer, S. (2017): Epistemic Dependence and Trust. On witnessing in the first, second, and third-person perspectives, in: Krämer, S.; Weigel, S. (Hg.), Testimony/Bearing Witness. Epistemology, Ethics, History, and Culture, London: Rowman & Littlefield, 247–259.

Krämer, S. (2018): Der »Stachel des Digitalen« – ein Anreiz zur Selbstreflexion in den Geisteswissenschaften? Ein philosophischer Kommentar zu den Digital Huma-

nities in neun Thesen, in: *Digital Classics Online,* 4(1). [https://doi.org/10.11588/dco.2018.0].

Krämer, S. (2021): Reflections on »operative iconicity« and »artificial flatness«, in: Wengrow, D. (Hg.), Image, Thought, and the Making of Social Worlds, Freiburger Studien zur Archäologie & Visuellen Kultur Bd. 3, Heidelberg: Propylaeum, 252–272.

Krämer, S. (2022a): Kulturgeschichte der Digitalisierung. Über die embryonale Digitalität der Alphanumerik, in: *APuZ: Aus Politik und Zeitgeschichte*, 72(2022), 10–17.

Krämer, S. (2022b): The Artificiality of the Human Mind: A Reflection on Natural and Artificial Intelligence, in: Nagl-Docekal, H.; Zacharasiewicz, W. (Hg.), Artificial Intelligence and Humane Enhancement, Berlin/New York: de Gruyter, 17–32.

Krämer, S. (2023a): Bearing Witness as Truth Practic: The Twofold – Discursive and Existential – Character of Telling Truth in Testimony, in: Jones, S.; Woods, R. (Hg.), The Palgrave Handbook of Testimony and Culture, Cham: Palgrave Macmillan, 23–38.

Krämer, S. (2023b): Should we really »hermeneutise« the Digital Humanities? A plea for the epistemic productivity of a »cultural technique of flattening« in the Humanities, in: *Journal Cultural Analytics*, 7(4). [https://doi.org/10.11588/dco.2018.0].

Krämer, S.; Bredekamp, H. (Hg.) (2003): Bild Schrift Zahl, Kulturtechnik Bd. 1, München: Fink.

Lovelace, A.A. (1843): Notes by A.A.L. (Augusta Ada Lovelace), Taylor's Scientific Memoirs, London, Vol. iii, wieder gedruckt in: Morrison P.; Morrison E. (Hg.) (1961), Charles Babbage and His Calculating Enginges. Selected writings by Charles Babbages and Others, New York: Dover Publications; dt. Version: Grundriß der von Charles Babbage erfundenen Analytical Engine, aus dem Französischen übersetzt und kommentiert von Ada Augusta Lovelace, in: Dotzler, B. (Hg.) (1996): Babbages Rechen-Automate. Ausgewählte Schriften, Computerkultur Bd. 6, Wien/New York: Springer, 666–731.

Lüdke, H. (1969): Die Alphabetschrift und das Problem der Lautsegmentierung, in: *Phonetik*, 20(2-4), 147–176.

Mackenzie, A. (2017): Machine Learners. Archaeology of Data Practice, Cambridge (MA): MIT Press.

Mahowald, K. et al. (2023): Dissociating Language and Thought in Large Language Models. A Cognitive Perspective. [https://doi.org/10.48550/arXiv.2301.06627].

Markov, A.A. (1912): Wahrscheinlichkeitsrechnung, Leipzig: Teubner.

McCulloch, W.; Pitts, W. (1943): A logical calculus of the ideas immanent in nervous activity, in: *Bulletin of Mathematical Biophysics*, 5, 115–133.

Rieger, B.B. (1991): On Distributed Representations in Word Semantics (Report). ICSI Berkeley 12–1991. CiteSeerX 10.1.1.37.7976.

Searle, J.R. (1980): Minds, Brains, and Programs, in: *The Behavioral and Brain Sciences*, 3(3), 337–356.

Searle, J.R. (1999): Chinese Room Argument, in: Wilson R.A.; Keil, F. (Hg.), The MIT Encyclopedia of the Cognitive Sciences, Cambridge (MA): MIT Press, 115f.

Stetter, C. (1997): Schrift und Sprache, Frankfurt a.M.: Suhrkamp.

Turing, A. (1950): Computing Machinery and Intelligence, in: *Mind*, 59(236), 433–460.

Vaswani, A. et al. (2017): Attention Is All You Need. [https://arxiv.org/abs/1706.03762].

Weizenbaum, J. (1966): ELIZA – A Computer Program For the Study of Natural Language Communication Between Man and Machine, in: *Communications of the ACM*, 9(1), 36–45.

Weizenbaum, J. (2000[1972]): Die Macht der Computer und die Ohnmacht der Vernunft, Frankfurt a.M.: Suhrkamp.

Wei, J. et al. (Hg.) (2022): Advances in Neural Information Processing Systems. [https://openreview.net/forum?id=_VjQlMeSB_J] (Zugriff: 08.07.2023).

Der zwanglose Zwang des besseren Tweets

Über kommunikative Rationalität in Sozialen Medien

Matthias Kettner

Abstract: *Two case studies of discussion projects in social media demonstrate that these projects are unsuccessful in meeting expectations of success based on norms of communicative rationality that should be followed in discourse-friendly communication communities where purposes of discussion and argumentation can be realized. This finding motivates the question of how we can explain systematic restrictions on communicatively rational communication by reference to specific features of digital-cultural communication practices in social media. Five explanatory hypotheses are developed: (1) The normality of structural non-commitment; (2) structural uncooperativeness; (3) the marginalization of indirect speech; (4) the structural decoupling of retweet decisions from reflection on the value of tweets for purposes of discussion and argumentative discourse. I propose a research design for articulating the space of reasons in which participants make their decisions. This rationalistic approach could serve to complement prevailing approaches of predicting retweet decisions probabilistically. The rationalist approach is then complemented by psychodynamic considerations. It is claimed (5) that the concept of primary and secondary process is particularly relevant for research perspectives on communication in social media because some properties describable in terms of media technology correspond functionally to the psychologically describable properties of primary-process thinking.*

Keywords: *community of communication; argumentative discourse; space of reasons; retweet decisions; primary process*

1. Zur Realität der Sozialen Medien

> Ohne umfassende Kommunikation kann keine Gemeinschaft oder Gesellschaft entstehen und (fort)bestehen.[1]

Die Geschichte der Erfindung immer neuer Organisationsformen eines umfassenden, für den gesellschaftlichen Zusammenhalt notwendigen Austauschs von Mitteilbarem ist die Entwicklungsgeschichte der Kommunikationsmedien. Was in der Betrachtung dieser Geschichte in welchem fachdisziplinären, begrifflichen oder theoretischen Rahmen als Austausch, als Mitteilbares, als Entwicklung, als Kommunikation gefasst wird, variiert enorm, und damit auch der seit den 1960er Jahren immer stärker beachtete Medienbegriff selbst. Dürfen Zeichen, Sprache, Schrift noch relativ unstrittig als die kulturgeschichtlich uranfänglichen »Mittler von Kommunikation« (Bösch 2016) gelten, gehen die Besonderungen, Einteilungen und Abteilungen des Medienbegriffs inzwischen so heterogen auseinander, dass niemand seriöse Aussichten auf eine integrative »Supertheorie« bieten kann. Die Medialität der Medien bleibt ein nahezu magisches Thema.

Nicht unwesentlich dazu beigetragen hat Marshall McLuhans Engführung des technischen und des gesellschaftlichen Moments in seinem einflussreichen Vorschlag, Medien (pragmatistisch) als Mittel der Ausdehnung (»extensions«) menschlicher sensorischer, körperlicher und geistiger Vermögen, und ihre Botschaft (konsequentialistisch) zu begreifen als Gesamtheit der Veränderung der Maßstäbe, des Tempos und Schemas, die sie in das Zusammenleben der Menschen bringen,[2] – ein Vorschlag, der sich nicht nur für das Verständnis der Massenmedien seiner Zeit (Fernsehen, Radio und Zeitungswesen) als sehr aufschlussreich erweisen sollte, sondern auch für die Technikphilosophie. Es wäre ganz im Sinne McLuhans, die Basistechnologie der Verarbeitung digitalisierter Information in

1 Schönhagen/Meißner 2022: 22.

2 »[T]he ›message‹ of any medium or technology is the change of scale or pace or pattern that it introduces into human affairs.« (McLuhan 1994: 7) Botschaft sensu McLuhan ist nicht nur der Sinngehalt eines medial mitgeteilten Items, sondern »the personal and social consequences of any medium – that is, of any extension of ourselves – [which] result from the new scale that is introduced into our affairs by each extension of ourselves, or by any new technology«, »because it is the medium that shapes and controls the scale and form of human association and action«. »The use of any kind of medium or extension of man alters the patterns of interdependence among people, as it alters the ratios among our senses.« (McLuhan 1994: 8f., 90) Kritisieren kann man an diesem begrifflichen Zuschnitt (»as extensions«), dass er womöglich zu weit ist, da McLuhan auch Straßen, Zahlen, Kleidung, Wohnen als Medien behandelt. Ein Vorteil aber ist sicher, dass Ausdehnung weder Verbesserung noch gar Fortschritt impliziert, die Bewertung von Medienveränderungen also offenlässt.

vernetzten Computern, das Internet, als das neue Medium anzusprechen, dessen Botschaft die »digitale Transformation« unserer Lebensverhältnisse ist.[3]

Die folgenden Überlegungen richten sich auf einen kleinen, aber stark transformativen Ausschnitt dieses Mediums, nämlich auf diejenigen Kommunikationsmedien, in denen der Austausch von Mitteilbarem in Formen organisiert ist, die wir seit dreißig Jahren, und inzwischen wie selbstverständlich, als die *Sozialen Medien* ansprechen. Die soziologische Digitalisierungsforschung fasst sie zusammen als »Angebote auf Grundlage digital vernetzter Technologien, die es Menschen ermöglichen, Informationen aller Art zugänglich zu machen und davon ausgehend soziale Beziehungen zu knüpfen und/oder zu pflegen« (Schmidt/Taddicken 2022: 5). Soziologisch gängige Klassifikationen enthalten: Soziale Netzwerkplattformen, Diskussionsplattformen, Kreativplattformen (für Videoformate z.B. YouTube, TikTok, für Photoformate z.B. Flickr, für Audioformate z.B. Soundcloud), Personal-Publishing-Dienste (für Blogging z.B. Wordpress, für Microblogging z.B. Twitter), Instant-Messaging Dienste für synchrone Kommunikation (z.B. WhatsApp, Telegram, Videokonferenzsysteme wie Zoom), Wikis (z.B. Wikipedia) in denen die Netizens das Abonnieren, Annotieren, Erstellen, Modifizieren, Vernetzen, Veröffentlichen, Weiterleiten von medial mitteilbaren Sinngehalten praktizieren, für unterschiedlichste Zwecke und in unterschiedlichsten Sozialgebilden (z.B. als amorphe Massen bzw. Crowds, Schwärme, Mobs, Szenen, als organisierte Bewegungen, volatile kulturelle Wir-Gruppen bzw. Communities) (vgl. Schmidt/Taddiken 2022: 3–34). Begriffsgeschichtlich interessant sind gewisse Verschiebungen im Vokabular (Aichner et al. 2021): Während anfängliche Definitionen die Interaktivität betonen, geht in wissenschaftlichen *und* populären Definitionen nach 2010 die Terminologie von »Nutzern«, die »Inhalte generieren«, semantisch in Führung. Hierin schlägt sich vermutlich die massenhafte Verbreitung von Smartphones nach 2007 nieder und die damit einhergehende globale Mobilisierung des Internet in Form von Apps, sowie die wachsende Bedeutsamkeit von Anonymität.[4]

Nun zur Fragestellung. Ich möchte im Folgenden einige Thesen begründen, die empirisch-normativ hybride Forschung zur Bewertung der relativen Rationalität

3 Zur digitalen Kommunikation und Kommunikationsgeschichte gut fokussierte Beiträge in Schwarzenegger et al. 2022. Synthetisch zum Begriff digitaler Medien s. Bateman 2021. Für wissenschaftlich informierten Alarmismus angesichts nicht intendierter disruptiver Folgen kollektiver digitalkultureller Kommunikationspraktiken s. Bak-Coleman et al. 2021. Ein Lagebild zum digitalen Wandel in Deutschland zeichnet jährlich der D21-Digital-Index (https://initiatived21.de/).

4 Für eine detaillierte Chronologie siehe [https://en.wikipedia.org/wiki/Timeline_of_social_media]. Demographisch differenzierte Information zur aktuellen Social Media Nutzung in den USA ist auf [https://www.pewresearch.org/internet/fact-sheet/social-media/] zu finden, weltweite Übersichten gibt [https://www.statista.com/topics/1164/social-networks/].

von informeller Kommunikation in Sozialen Medien anleiten könnten.[5] Die Thesen enthalten beschreibende, erklärende und bewertende Aspekte:

Empirische Erfahrungen mit zwei großangelegten Versuchen, Soziale Medien so zu gestalten, dass die informelle Kommunikation in diesen Medien die beteiligten Netizens zu Diskussion und argumentativem Diskurs anregt, zeigen, wie ich in Abschnitt 4 darlege, überraschend schlechte Ergebnisse, wenn man den Erfolgserwartungen gewisse Anforderungen kommunikativer Rationalität zugrunde legt, die in diskursfähigen Kommunikationsgemeinschaften für Zwecke von Diskussion und Argumentation zum Tragen kommen sollten. Diese skeptische Einschätzung begründet die Fragestellung: *Wie können wir Einschränkungen kommunikativer Rationalität erfassen und wieweit können wir ihr Zustandekommen aus Besonderheiten der digitalkulturellen Kommunikationspraktiken in Sozialen Medien erklären, statt aus persönlichen Rationalitätsdefiziten der beteiligten Netizens?* Ich entwickele vier Erklärungshypothesen: Was die Realisierung kommunikativer Rationalität medienspezifisch einschränkt, ist (1) die Normalität struktureller Unverbindlichkeit; (2) strukturelle Unkooperativität; (3) die Marginalisierung von indirekter Rede; (4) die strukturelle Entkopplung von Retweet-Entscheidungen, Mitteilungen weiterzuverbreiten, von der Reflexion auf den Wert der Tweets im Licht der Zwecke von Diskussion und argumentativem Diskurs. Ich entwerfe ein Forschungsdesign, das die vorherrschenden Ansätze, Retweet-Entscheidungen mit Hilfe probabilistischer Modelle vorherzusagen, ergänzen könnte durch die Artikulation des Raums der Gründe, in dem die Beteiligten ihre Entscheidungen treffen. Diesen rationalistischen Ansatz relativiere und ergänze ich zum Schluss durch die Integration von psychodynamischen Überlegungen. Die These ist: (5) Das Konzept von Primär- und Sekundärprozess ist für Forschungsperspektiven auf Kommunikation in Sozialen Medien besonders relevant, weil einige medientechnisch beschreibbaren Eigenschaften den psychologisch beschreibbaren Eigenschaften von primärprozesshaftem Denken entgegenkommen.

Zuvor aber nehme ich in den Abschnitt 2 und 3 einige Rekonzeptualisierungen vor, die die unübersichtliche Realität der internetbasierten Kommunikationsmedien für die Zwecke meiner Fragestellung hilfreich vereinfachen.

5 Mit informeller i.U. zu formeller Kommunikation meine ich, dass sie aus eigenem Antrieb erfolgt, ungezwungen, ohne Formalität und Feierlichkeit, frei von amtlich oder beruflich autorisierten Forderungen (von vorgesetzten Stellen, Behörden, Regierungen u.ä.), weitgehend frei auch von Moderation, Kuratierung oder Redaktion durch autorisierte Dritte. Gewiss ist diese Unterscheidung nicht binär, hat aber klare Fälle: Wenn Olaf Scholz auf TikTok seine Aktentasche einem unbestimmten Publikum präsentiert, wie im April 2024 geschehen, ist die Medienkommunikation informell; wenn das Finanzamt mich per Email zur Zahlung auffordert, formell.

2. Mediengestalten und Netizens

Denken wir im Rahmen der von Berner Lee (2006) angeregten *Webscience*[6] die informationstechnische Seite des Mediums Internet als einen mittels formaler Sprachen und Protokolle konstruierten Raum der Datenverarbeitung, dann können wir alle Formen organisierter Kommunikation, die diesen Raum als technische Infrastruktur nutzen, eigensinnige Nutzungspraktiken ausbilden und (zumindest auf gewisse Zeit) verstetigen, als *Mediengestalten digitalkultureller Kommunikation* zusammenfassen. Wenn Mediengestalten sich unterscheiden, dann in den konstitutiven Eigenschaften der jeweiligen Kommunikationspraktiken, die wir im, und immer begrenzt nur vom, Spielraum der von uns konstruierten technischen Infrastruktur kultivieren.

Der Kommunikationsbegriff ist bekanntlich so weit, man könnte auch sagen: unbestimmt, dass man ihn auf den Fluss digital codierter Informationen zwischen Computern ebenso anwenden kann wie auf Mitteilungen zwischen Personen.[7] Grenzen wir den Kommunikationsbegriff ein auf *Mitteilungen von Äußerungen, die Menschen als von Menschen sinnvoll Gemeinte auffassen können*, und abstrahieren von der weiteren Extension des Begriffs, mit der er z.B. in der Informatik verwendet wird, dann lässt sich eine soziale Eigenschaft aller Medienteilnehmer, die im Sinne der Fragestellung interessieren, gut kennzeichnen: Wenn wir, durch Mediengestalten digitalkultureller Kommunikation verbunden, Mitteilbares austauschen, sind wir dadurch *Netizens*, d.h. an der Kommunikationsgemeinschaft aller, die solche Mediengestalten nutzen können, Beteiligte.

Netizen bin ich dadurch und genau soweit, wie ich mich an Mediengestalten, die durch die Infosphäre des Internet infrastrukturiert sind, *beteilige*. Um nicht gleich voraussetzungsvoll von »Autorschaft« zu sprechen,[8] erscheint es mir sinnvoll, die

6 Siehe [https://webscience.org] zum weltweiten *Web Science Trust Network*. Das Desiderat wird so beschrieben: »A science of the Web arguably comprises the study of the networks that underpin the Web (thus Network Science, Internet Science), the vast quantities of information/transactions that are generated by this global artefact (thus Data Science and Big Data) as well the various ways in which both humans and artificial constructs produce, consume and react to the data (thus ML, AI, Psychology, Sociology) and the larger scale impact and management of such systems (thus Law, Ethics and Philosophy). Not to forget perspectives on education, health, politics, innovation/business et al we see how broad, interdisciplinary and universal the Web Science perspective can be.« (Ebd.)

7 Kritisch zur zu Missverständnissen einladenden Unschärfe des Informationsbegriffs: Janich 2006.

8 Dass Habermas (2022) seine Denkfigur der Autor-Adressat-Beziehung, die im Kontext der Rechtfertigung demokratischer Rechtsetzung erhellend ist, nahezu umstandslos auch auf mediale Öffentlichkeiten, die spezifisch durch digitalkulturelle Medienpraktiken konstituiert sind, anwendet (sie »*ermächtigen* alle potentiellen Nutzer prinzipiell zu selbständigen und gleichberechtigten Autoren«, Habermas 2022: 44), halte ich für eine der wenigen Schwä-

geringfügigste Beteiligung dieser Art von der aktiven Seite her darin zu sehen, *sich zu äußern*, und von der passiven Seite darin, etwas *als* Äußerungen anderer qua Netizens *auf sich wirken zu lassen*. Beide Seiten sind verschränkt, denn wer wollte sich in einem Medium äußern ohne zumindest zu unterstellen, dass das Medium die Voraussetzung der Zugänglichkeit erfüllt, dass andere die Äußerung als eine Äußerung, also als etwas möglicherweise sinnerfüllt Gemeintes, wahrnehmen können.

Ich handele in der, mir durch meine Teilnahme objektiv, d.h. ob ich will oder nicht, zukommenden Eigenschaft des Netizens, gleichviel ob ich nun eine eingegangene SMS lese oder eine neu schreibe und sende, auf Youtube ein Video als anstößig melde, anschaue oder selber hochlade, einen Tweet retweete oder es unterlasse, einen neuen Hashtag in Umlauf zu bringen versuche, das Photo eines handgemalten Bildes in meiner WhatsApp-Familiengruppe poste oder lösche, den Link einer Website für späteren Wiederaufruf in meinen Browser kopiere, Hyperlinks in meinen Blog hineinsetze oder wieder herausnehme, mich zum Follower mache oder einen Follower sperre, auf der Website von Change.org eine Petition digital unterschreibe oder lieber weiterklicke, Preisvergleiche auf Check24 durchführe usw., gleich in welchen sonstigen sozialen Rollen-, Mitglieds- oder Teilnehmereigenschaften aus allen sonstigen Praktiken, die mein Leben ausmachen, ich dies tue: als Familienvater, als Hochschullehrer, als Freund, als *Citoyen*, als *Bourgeois* usw. Der minimalistische Begriff von Äußerungen hat zudem den Vorteil, die semiotische Modalität des intramedial Äußerbaren (visuell, auditiv, textuell) offenzulassen und auch sonstige Kategorien zunächst auf Distanz zu halten, mit denen wir von Menschen hervorgebrachte Äußerungen rubrizieren (z.B. Redegenres, Kunstgenres, Textsorten usw.).

Keine Mediengestalt, ob digitalkulturell oder »alt«, kommt für sich alleine aus, sondern bildet mit allen übrigen, die noch nicht passé sind, etwas, was die Medienforschung früher ein System nannte und heute, unter dem Eindruck der schnell mutierenden und offenbar endlos emulationsfähigen Mediengestalten im Hypermedium Internet, oft schon Umwelten (Caliandro 2017) oder Ökologie nennt (Giesecke 2002; Scolari 2012; Bayer et al. 2020). Für mein Argument genügt der sparsame Begriff einer Kommunikationsmedienumgebung. Sie beinhaltet die Mediengestalten und deren medieneigene Praktiken, auf die sich einlassen muss, wer im Rahmen dieser Praktiken Äußerungen entweder auf sich wirken lassen oder sich für andere äußern will, oder beides. Nichts weniger kann man durch Betätigung in Kommunikationsmedienumgebungen tun. *Wie* man es kann und wie nicht, und was alles man *durch* solche Betätigung darüber hinaus erreichen kann und was nicht, wird durch die Mediengestalt(en) einerseits freigestellt, anderseits begrenzt. Sie ermächtigen und beschränken die Handlungsmächtigkeit (*Agency*) sämtlicher Beteiligter. Wie ich im Medium der Umgangssprache zwar jemanden die Tür zu öffnen bitten,

---

chen seiner hoch synthetischen und interessanten Überlegungen zum neuen Strukturwandel der Öffentlichkeit im Kulturprozess der Digitalisierung.

aber nicht die Tür öffnen kann, so kann ich z.B. im Medium X/Twitter zwar Äußerungen von Querdenkern lesen, auch auf sie antworten (allerdings ohne einen »Wunsch nach Gewalt« zu äußern),[9] könnte jedoch nicht mit ihnen telefonieren, wenn ich wollte, dass wir miteinander ins Gespräch kämen. Zeitungsleser dürfen Leserbriefe schreiben und zur Veröffentlichung anbieten, aber, selbst als Abonnenten, keine Traktate; angemeldete Twitterer haben Platz für 260, zahlende Mitglieder für 4000 Zeichen Text, usw.

Wer sich auf eine Umgebung einlässt, versucht mit ihr zurechtzukommen, sich in ihr zu orientieren und sein eigenes Verhalten zielführend an sie anzupassen.[10] Dito für Medienumgebungen. Digitalkulturelle stellen einerseits wegen der nötigen Technikbeherrschung vergleichsweise hohe Anforderungen, obwohl die Betreiber alles tun, um das Einlassen kinderleicht und schier unwiderstehlich zu machen. Spezifische Anforderungen des Zurechtkommens machen eine Umgebung, wie ich sagen möchte, zugleich zu einer Rationalitätsumgebung: eine Umgebung, in der alle, die in ihr unterwegs sind, neben allem sonstigen *irgendwie* immer auch ihre intelligenten Fähigkeiten und geistigen Kräfte einbringen müssen, entsprechend den jeweiligen Anforderungen mehr oder weniger. Mit Seitenblick auf die umgangssprachlich vermittelte Kommunikation gesagt: Man muss die Sprache kompetent beherrschen, wenn man wirklich nutzen will, was alles sich mit Worten erfolgreich tun lässt.

Wenn wir reden, werden wir nicht dadurch zu *bloßen* Sprechern und Hörern, sondern bleiben die konkreten Individuen, die sich nun im Medium der Rede äußern; wenn wir digitalkulturell kommunizieren, werden wir nicht dadurch zu *bloßen* Netizens, sondern bleiben die konkreten Individuen, die sich nun gerade in digitalkulturellen Mediengestalten äußern.

Dieser Punkt ist nicht so banal, wie er vielleicht erscheint, denn in Verlängerung folgt daraus, dass die konkreten Individuen im ganzen Umfang ihrer Subjek-

9 Das verbietet die aktuelle Version der Netiquette von X, siehe [https://help.twitter.com/de/rules-and-policies/x-rules].

10 Im kommunikationswissenschaftlichen Vokabular greift man neuerdings auf einen Schlüsselbegriff der ökologischen Psychologie von James Gibson zurück: Affordanzen. *Affordances* sind mit den Werten, Vorstellungen, Fähigkeiten und Absichten der Akteure zusammenhängende Wahrnehmungen von Handlungsmöglichkeiten, die Objekte in der Handlungsumgebung der Akteure den Akteuren geben. Begriffliche Klärungen führen zur folgenden, auf Soziale Medien zugeschnittenen Definition: »die wahrgenommenen tatsächlichen oder vorgestellten Eigenschaften Sozialer Medien, die sich aus der Beziehung zwischen technologischen, sozialen und kontextuellen Eigenschaften ergeben und die spezifische Nutzungen der Plattformen ermöglichen und einschränken« (Ronzhyn et al. 2022: 3178, Übers. von mir – MK).

tivität engagiert sind (Perzeption, Kognition, Motivation, Volition, Affektivität),[11] wenn sie sich zu medial Beteiligten machen und als solche eben nur so »dünn«, wie eine bestimmte Mediengestalt es zulässt, füreinander in Erscheinung treten, z.B. als Netizens. Ihre Aktivitäten sind daher nicht nur soziologisch zu erforschen, sondern unter sämtlichen Perspektiven, die die Wissenschaften vom Menschen anbieten. Im Hinblick auf Soziale Medien interessiert mich besonders, wie psychologische, insbesondere psychodynamische Perspektiven in die philosophische Digitalisierungsforschung integriert werden können. Psychologische Arbeiten im Zusammenhang mit sozialen Medien fallen bis jetzt vorwiegend in die Persönlichkeitspsychologie, experimentelle Psychologie und Entwicklungspsychologie. Ihre Fragestellungen gelten vorzugsweise psychologischen Risiken.[12] Psychodynamische, insbesondere psychoanalytische Perspektiven auf die Differenz von subjektiv bewusstem und unbewusstem Sinn in den Aktivitäten von Netizens sind in der medien- und kommunikationswissenschaftlichen Digitalisierungsforschung (Zyoud et al. 2018) leider immer noch randständig.[13]

3. Kommunikationsgemeinschaften und Diskurspartner

Wenn wir nur natürliche Personen als Netizens zählen lassen wollen, dann sind Netizens eine Teilmenge der, wie die sprachpragmatische Philosophie Apels dies begreift, virtuell unbegrenzten Kommunikationsgemeinschaft aller wie menschliche Personen sprachfähigen Wesen. Will man der Teilmenge der Netizens zudem künstliche Akteure zusetzen, etwa ChatBots, die wie Netizens auftreten, oder könnten wir solche Akteure nicht mehr ausschließen, auch wenn wir wollten, dann bildet die Menge aller Netizens nur eine Schnittmenge mit der virtuell unbegrenzten Kommunikationsgemeinschaft aller wie menschliche Personen sprachfähigen Wesen. So zu denken ist natürlich anthropozentrisch, doch wüsste ich nicht, wie die Grenzziehung einer Kommunikationsgemeinschaft anders zu denken sein sollte als von innen nach außen. Ich sehe keinen Grund, daran nicht festzuhalten, dass wir, menschliche Personen, das Innen ausmachen.

Der philosophisch beste Grund, hieran festzuhalten, hat mit der Verschränkung von Vernunft, Sprache und Verantwortung zu tun: Dass wir, menschliche Personen,

11 Der Punkt wäre m.E. theoretisch am besten im Rahmen enaktivistischer Konzeptualisierungen von situierter menschlicher Handlungsmächtigkeit (Hutto et al. 2014; Drury/Tudor 2024) zu untermauern. Das kann ich an dieser Stelle nicht ausführen.

12 Als Sammelbegriffe für das entstehende Forschungsfeld *research on the psychology of the Internet and social media* hat sich das Label *Cyberpsychology* etabliert (Ancis 2020; Kirwan et al. 2024). Eine repräsentative Zeitschrift ist *Cyberpsychology, Behavior, and Social Networking*.

13 Siehe aber Johannsen 2018; Johansen/Krueger 2022; sowie die Beiträge im Kapitel III von Goodman/Clemente 2023; und in Grabska 2023 (besonders Löchl 2023).

nur unter unseresgleichen uns als in dem Sinne verantwortungsfähig denken können, der notwendig ist, um mit spezifischen normativen Ansprüchen zurechtzukommen, deren Zusammenspiel die Äußerungen, die wir im Medium menschlicher Rede machen, rationalisiert, sie beurteilbar, begründbar, kritisierbar und geltungsfähig macht. In diesem Punkt konvergieren zumindest die ansonsten unterschiedlichen sprachpragmatischen Perspektiven von Jürgen Habermas, Robert Brandom und Karl-Otto Apel. Vernunft

> »nimmt in der diskursiven Mobilisierung von Gründen eine explizite Gestalt an. Gute Gründe sind die Münze, in der sich Akte der Verständigung auszahlen; aus ihrer rational motivierenden Kraft speist sich das Ja und Nein der handelnden, erst recht der lernenden Subjekte«. (Habermas 2019: 379) »Rational practices, practices that include the production and consumption of reasons – the ›giving and asking for reasons‹ – […] must distinguish two sorts of normative status: a kind of commitment, undertaken by the assertional speech acts by which alone anything can be put forward as a reason, and a kind of entitlement, which is what is at issue when a reason is requested or required. This normative fine structure is inferentially articulated along three axes, defined by inheritance of commitment, inheritance of entitlement, and entailments according to the incompatibilities defined by the interactions of commitments and entitlements.« (Brandom 2000: 195)

Solange wir maschinelle Systeme, gleich ob sie kraft maschinellen Lernens menschliche Sprachmuster erlernen oder sogar Personalität mit technischen Mitteln simulieren,[14] nicht zur Verantwortung ziehen können wie Mitmenschen, können wir ihnen im Ernst auch keine Sprechakte zurechnen. Der springende Punkt ist nicht, »ob sie denken«, »intelligent sind« oder »Bewusstsein haben«, sondern ob sie verantwortungsfähig sind; was sie nicht sind.[15]

Gehen wir davon aus, dass die Einsicht trägt, dass menschliche Vernunft, Sprache und Verantwortung konstitutiv miteinander verschränkt sind, dann können wir die profunde praktische Bedeutsamkeit des Gedankens einer maximal inklusiven Gemeinschaft durch Sprachgebrauch auf einfache Weise erläutern. (Apel formuliert diesen Gedanken in terms einer »realen« und zugleich »idealen« und »virtuell unbegrenzten« Kommunikationsgemeinschaft.) Um einen *Grundgedanken* handelt es sich insofern, als ohne ihn nicht verständlich gemacht werden kann, wie wir mit

14 Zu ersterem siehe im vorliegenden Band den Beitrag von Krämer, zu letzterem die Beiträge von Kerrin Jacobs und Natalia Juchniewicz im vorliegenden Band. Ich weiß nicht, ob Juchniewicz ihr Konzept einer ver- und geteilten Verantwortung auch auf Sprecherverantwortung für Sprechakte ausdehnen möchte. Wenn ja, haben wir eine Kontroverse.

15 Siehe den Beitrag von Susanne Hahn im vorliegenden Band.

Äußerungen, die wir als Behauptungen verstehen, für das Mitgeteilte Allgemeingültigkeit beanspruchen können (z.B. für Tatsachenbehauptungen, dass sie wahr sind) und doch *zugleich* auch alle Gründe, die uns hier und jetzt zurecht von der Gültigkeit überzeugen, für vorläufig und revidierbar halten sollten. Mehr noch: Normativ und deshalb nicht nach Belieben verwerfbar ist der Grundgedanke der virtuell unbegrenzten Kommunikationsgemeinschaft insofern, als wir uns selbst sinngemäß entsprechend verstehen *sollen*, wenn wir argumentieren, d.h. wenn wir uns über bessere oder schlechtere Gründe einig werden *wollen*. Kurz gesagt: Der Gedanke ist grundlegend wichtig für die kommunikativ rationale Beteiligung an ernsthaft argumentativer Kommunikation (Apel 2011; Kettner 2016), also für einen im Ganzen unserer Kommunikationspraktiken für uns unverzichtbaren Teil.

Wenn wir die Praxis ernsthaft argumentativer Kommunikation, kurz: Diskurs, auch als ein Medium begreifen, können wir die spezifisch an diesem Medium Beteiligten in dieser Eigenschaft als *Diskurspartner* charakterisieren, so wie oben die am Hypermedium Internet Beteiligten als Netizens. Dann lässt sich die Frage stellen, ob und wie gut das Diskursmedium sich in andere Mediengestalten einbilden bzw. dort reproduzieren lässt, z.B. in digitalkulturellen, speziell solchen, die von informeller Kommunikation leben, wie die Sozialen Medien. (Klarerweise *können* wir im Hypermedium Internet ausgesprochen diskursfreundliche Mediengestalten kultivieren – am besten in hierauf spezialisierten, an diesen Zweck formal angepassten Rationalitätsumgebungen, wie in machen Online-Foren und Plattformen z.B. der Wissenschaftskommunikation –, und gewiss *können* wir es mit einiger Anstrengung auch im Medium der guten alten Email bzw. in der Praxis schriftlicher Korrespondenz.)

Ich möchte nun Apels Gedanken einer zumindest für *argumentative* Praktiken *konstitutiven* umfassenden Kommunikationsgemeinschaft etwas genauer erläutern, um Missverständnissen vorzubeugen. Im nächsten Abschnitt werde ich dann eigene Erfahrungen mit der kommunikativen Rationalität von Personen qua Netizens und qua Diskurspartner in Sozialen Medien darstellen.

Apels philosophische Behauptung, dass der Gedanke der virtuell unbegrenzten Kommunikationsgemeinschaft konstitutiv für Diskurs ist und unter Diskurspartnern nicht verworfen werden kann, hat einen einfachen und nicht wegzudiskutierenden Erfahrungsgehalt: Menschen, die die Praxis des Miteinanderargumentieren dafür einsetzen, angesichts von Meinungsverschiedenheiten erkennen zu wollen, wer Recht hat, wenn nicht alle gleichermaßen Recht haben können, müssen unterstellen, dass sich auf diesem Wege (d.h. via Argumentation) eine *für alle gleichermaßen* einsichtige Auffassung, wer denn wirklich Recht hat, herausstellen *könnte*, der niemand mehr ernsthaft widersprechen *könnte*. Wann immer Personen miteinander argumentieren, um Einsicht in das, was wirklich gilt, zu gewinnen, erfahren sie, dass jeder sich selbst so wie jeder jeden anderen in einer bestimmten Position und Beziehung verstehen muss, und dass dies eine eigentümliche Gemeinschaft-

lichkeit zwischen ihnen stiftet. Eigentümlich ist diese Gemeinschaftlichkeit insofern, als keiner in der Menge der aktuell Beteiligten sie auf die aktuelle Situation *dieser* Beteiligten *hier* und *jetzt* festlegen und begrenzen kann. Es ist die Gemeinschaftlichkeit von einsichtsorientierten Kritikern und Begründern von Geltungsansprüchen, die sich so nicht nur aktual, sondern zugleich auch hypothetisch, als *mögliche* Kritiker und Begründer von Geltungsansprüchen verstehen, also als Aktualisierer entsprechender Rollen. Zwar vergemeinschaftet diese, für ernsthaft argumentative Kommunikationspraxis (Diskurs) konstitutiv erforderliche und durch die so konstituierte Praxis zugleich reproduzierte Weise von Gemeinschaftlichkeit tatsächlich jeweils nur eine *endliche* Menge von Argumentierenden. Sie ist und bleibt aber, und das sollte ein jeder, der sich beteiligt, wissen, *unabgeschlossen* in dieser wie in jeder anderen bestimmten endlichen Menge von Argumentierenden. Die Gemeinschaftlichkeit, die zwischen argumentierenden Personen qua argumentierenden Personen besteht, ist und bleibt fortgesetzt offen für andere argumentierende Personen (Diskurspartner) und für andere strittige Geltungsansprüche.

Was Apels Formel der »virtuell unbegrenzten« Kommunikationsgemeinschaft erfassen soll, wäre also nicht nur im Sinne einer *Menge* von Elementen (wirkliche und mögliche Diskurspartner und Dissense) zu verstehen, sondern vielmehr als ein *Modus* von Vergemeinschaftung, d.h. eine interpersonelle Beziehung. Jede reale Kommunikationsgemeinschaft weiß sich (oder sollte zumindest sich wissen) als nur begrenzt inklusiv und doch zugleich auch als unbegrenz*bar* inklusiv, jedenfalls immer dann, wenn sie im Modus diskursiver Argumentation kommuniziert. Dass Diskurspartner als Beteiligte in wirklichen Kommunikationsgemeinschaften ihre Gemeinschaft zugleich als Beteiligung an einer »idealen« Kommunikationsgemeinschaft verstehen dürfen, ist also nichts Ominöses, sondern zunächst nichts weiter als ein Zutrauen, das unter kritisierend und begründend Argumentierenden zuhause ist, dass ihr Argumentieren, wenn alles richtig gemacht wird, zu geteilten Überzeugungen führen kann, die, weil und soweit sie auf mitteilbaren Einsichten gründen, in einem immerzu noch erweiterbaren Kreis von Argumentierenden immer weiter und immer wieder geteilt, geprüft und ggf. revidiert werden können.[16]

Die Ausdifferenzierung des Diskursmediums aus dem Großen und Ganzen unserer alltagspraktisch vertrauten Kommunikationspraktiken kann man sich so klarmachen: Diskurse entspringen Dialogen, diese dem geselligen Gespräch, der Un-

16 Die Aufrechterhaltung dieses Zutrauen hat diskurspraxisexterne kulturelle Voraussetzungen, die schwinden oder auch gezielt angegriffen werden können. Diese Tatsache ergibt einen philosophischen Grund (wenn es denn noch eines solchen bedarf), in der kulturellen Normalisierung von Praktiken, die der Generalisierung von Misstrauen gegen geltungsbeanspruchende Kommunikationspraktiken dienen, u.a. eine Subversion des Diskursmediums zu befürchten.

terhaltung unter Anwesenden. Eine notwendige Bedingung für *dialogische* Kommunikationspraxis ist die Möglichkeit, zwischen Gesprächspositionen von Rede und Gegenrede zu wechseln, zeitlich unbegrenzte Monologe wären dysfunktional. Wird sie zu *Diskussion*, so verteilt sie Rollen an ihre Teilnehmer, die Diskutanten, sowie eine rollenspezifische Verantwortung, dafür zu sorgen, dass die wichtigsten normativen Anforderungen des Argumentierens nicht konterkariert werden. Kultivieren wir Diskussion zu *Debatten*, werden diese Gesprächspositionen zu den Gesprächsrollen von Proponent/Opponent ausdifferenziert, und das Moment von Antagonismus, die Einheit von Kooperation und Konflikt, kann zum Streitgespräch bzw. zur *Kontroverse* gesteigert werden. Diskussionen (ggf. mit Debatten und Kontroversen) werden zu *Diskursen*, wenn alle Beteiligten in vorbehaltlos kommunikativen Argumentationshandlungen ein gemeinschaftlich geteiltes Ziel verfolgen: das Ziel, den *wirklichen* Wert von *mutmaßlich* hinreichend guten, tatsächlich aber *fraglich* gewordenen Gründen allgemeinverbindlich neu zu bestimmen, um auf diesem Wege eine nachvollziehbare Meinungsverschiedenheit aufzuheben.

4. Erfahrungen mit kommunikativer Rationalität in Sozialen Medien

Aktion Mensch, die größte deutsche Förderstiftung für gemeinnützige Projekte, lud 2002 mit einer mächtigen multimedialen Kampagne, die zentral auch ein für damalige Verhältnisse technisch aufwändiges offenes Diskussionsforum für Netizens organisierte, die Bevölkerung ein, sich an der seinerzeit aktuellen Diskussion über rasante Fortschritte in Medizin und Biotechnologie zu beteiligen, statt sie allein Wissenschaftlern und Politikern zu überlassen. Ich habe zusammen mit interessierten Studierenden die vermutete diskursiv mobilisierende Wirkung des Diskussionsforums im Rahmen eines normativen Verständnisses von deliberativer Demokratie einzuschätzen versucht (Kettner 2006): Wenn man die Aktion als einen Beitrag zur Organisation von Mit-Verantwortung für die Biopolitik einer deliberativen Demokratie betrachtet, lag ihre besondere Leistung in der öffentlich wahrnehmbaren Vervielfältigung von Fragen. Die Fragen wurden keiner expliziten Zensur unterworfen. Das gab kritischen Netizens z.B. die Chance, zu beobachten, welche Fragen zwar gestellt werden *könnten*, aber nicht gestellt werden, und Mutmaßungen darüber anzustellen, warum das so ist.

Die nicht von kommerziellen Betreibern vorgegebene, sondern von den Veranstaltern konfigurierte digitalkulturelle Mediengestalt führte bei einem nicht unbedeutenden Teil der Netizens zur Verbesserung ihrer Problemwahrnehmung, zur, wenn man so will, »Bewusstseinserweiterung« oder Vergrößerung des persönlichen und auch des öffentlichen Resonanzraums für Problemwahrnehmungen.

Allerdings enthielt der größte Teil der mitgeteilten Äußerungen (ca. 90 % im untersuchten Korpus) thematisch irrelevante Frotzeleien, Grobheiten und blan-

ken Unsinn. Eine plausible Teilerklärung für diesen Befund sahen wir in einer für informelle soziale-mediale Kommunikation vielfach belegten Schwäche: Wo durch abwesende oder allenfalls schwache Moderation ein, von technischen Beschränkungen und Back End Eingriffen unautorisierter Dritter einmal abgesehen, *unvermachteter* Spielraum für freie Äußerung geschaffen wird, wirken sich bereits wenige hochaktive Netizens, die bewährte informelle Regeln des fruchtbaren Diskutierens (von Paul Grice als »Konversationspostulate« beschrieben) missachten und von keiner Redaktion oder Moderation diszipliniert werden, verheerend auf die Diskussion aus. Hinzu kommt die technisch optionierte und kulturell inzentivierte Förderung eines albernen Versteckspiels mit Pseudonymen und wechselnden Adressierungen. Die dadurch bewirkte Tendenz zum Unterengagement in der Diskussion, zur Zerfaserung und zur Unverbindlichkeit kann allenfalls, so unser Eindruck, dort kompensiert werden, wo ein Thema (die »Frage«) so beschaffen ist, dass sie persönliche Erfahrung und emotionale Teilnahme mobilisiert. Je sachlicher und unpersönlicher aber ein Thema ist und dementsprechend auf Seiten der Netizens qua potentiellen Diskurspartnern den Willen und das Können voraussetzt, ggf. auch sachlich und distanziert zu diskutieren und zu debattieren, desto unwahrscheinlicher wird es, dass eine »kritische Masse« der beteiligen Netizens diesen Ansprüchen gerecht wird. Zusammenfassend fanden wir, dass mit der gewählten Mediengestalt der erklärte Zweck, Interesse an ethischen und politischen Fragen zu wecken, wo Fraglosigkeit war, zumindest nicht verfehlt und in Teilen sogar erreicht wurde. Erfolgsmindernd wirkte sich aus, dass ein Großteil der Netizens die angebotene Rationalitätsumgebung für andere Zwecke nutzen wollte (vor allem für rücksichtslose Selbstdarstellung) und mangels Moderation auch konnte, was an Sabotage heranreicht.

4.1 Einschränkung durch die Normalität struktureller Unverbindlichkeit

Im Jahr 2006 eröffnete Aktion Mensch eine explizit auf Diskurs angelegte Kommunikationsplattform namens *dieGesellschafter.de*, die bald Beteiligungszahlen im Millionen- und getätigte Äußerungen im fünfstelligen Bereich hatte. Anders als das vorige, auf Fragen abstellende Projekt, war das erklärte des zweiten, nicht nur die »Frage nach der Zukunft unseres Gemeinwesens aus der Reformrhetorik von Talkshows und Expertenrunden zurück in die Gesellschaft zu tragen« (so die Selbstbeschreibung), sondern auch Antworten zu ergründen: »Unter dem Horizont der ebenso grundlegenden wie offenen Frage *In was für einer Gesellschaft wollen wir leben?* gewinnen aktuelle politische und ethische Diskussionen Freiraum für neue Blickwinkel und Konzepte. Zugleich formuliert die Frage auch einen Maßstab, an dem sich diese Konzepte messen lassen müssen.«

Zusammen mit dem Soziologen Thomas Loer und einer studentischen Forschungsgruppe haben wir das Gesellschafterprojekt durchgängig begleitet. Zum

Gesellschafterprojekt gehörte ein kleines Moderatorenteam, das in die erwünschte größte informelle Äußerungsfreiheit der Netizens nur behutsam und nur im Falle starker Entgleisungen eingreifen wollte. Aus den Nöten der Moderatoren, ihre Rollenverantwortung zu präzisieren, konnten wir ebenso lernen wie aus der Analyse von Strängen der gespeicherter Äußerungen.[17]

Dabei verdichteten sich unsere am 1000-Fragen Projekt schon gemachten Beobachtung von Unverbindlichkeit weiter zu der Annahme einer strukturellen Unverbindlichkeit informeller themenoffener Online-Kommunikation. Dies sind einige Facetten der Unverbindlichkeit: (1) Themen mäandern, aus Meinungsverschiedenheiten entwickelt sich nur selten ein begrenztes und begrenzendes Thema im Sinne einer voranzubringenden Fragestellung. (2) Da nichts zur Themenzentrierung auffordert, wird die Kommunikation immer wieder durch an privater Expression interessierte Netizens auf idiosynkratische Abwege geführt, endet in Sackgassen oder kommt nicht über geselliges Gespräch hinaus. (3) Positionen werden geäußert ohne dass Bereitschaft, auf Kritik einzugehen, eingefordert wird; dieselbe Meinung kann immer wieder geäußert werden. (3) Kein Stand der Debatte und Kenntnis davon kann und darf erwartet bzw. muss verantwortet werden. (4) Beteiligung und Abwesenheit bleiben willkürlich.

Was macht die Unverbindlichkeit zu einer strukturellen? Einige Facetten: (1) Die – technisch unnötige – Erlaubnisnorm der Klarnamenvermeidung, Anonymisierung, Pseudonymisierung verankert Unverbindlichkeit in einer starken, von vielen Netizens habituell befolgten Konvention. (2) Die fortlaufende Aggregation des Stroms von Äußerungen erzeugt ein anwachsendes Archiv, das mangels resümierender Metadiskussionen keiner der Beteiligten noch vergegenwärtigen oder bei anderen Netizens voraussetzen kann, so dass zum einen willkürlich ad hoc Rückbezüge, zum anderen eine präsentistische Verlagerung von Bezugnahmen auf das je Tagesaktuelle und sture Wiederholung zu relativ rationalen Strategien werden, um in dieser Medienumgebung zurechtzukommen. (3) Die technisch angebotene Aufhebung der Synchronie der Unterhaltung erleichtert die Beteiligung, erschwert aber permanent die Navigation in ihr. (4) Die technisch angebotene Leichtigkeit, jederzeit eine neue Äußerung setzen zu können (hier: loszuschreiben) unterstützt die Einstellung, auf schon gemachte Äußerungen keine Rücksicht nehmen zu müssen. (5) Die technisch angebotene Möglichkeit, sich anonym und informell zu äußern, unterstützt die Einstellung, keine Rücksicht auf andere Netizens als verletzliche Individuen nehmen zu müssen.

17 Besonderen Dank an Magdalena Assmann, Pola Boehm, Boris Bugla, Paul Endres, Malte Härtig, Ann-Kathrin Löhr, Anne Ostermann, Gesine Stern, Melchior Walker und die Teilnehmer der *Philosophy and the Social Sciences Conference* 2007 in Prag, wo wir Projektergebnisse präsentierten.

Wir fanden in vielen »Diskussionssträngen« unbelegte Geschichten, nicht verfolgbare Quellenangaben, Verschwörungserzählungen (z.B. Islamisierung des Abendlandes), unlogische Behauptungen (Tautologien, widersprüchliche Konstruktionen, ungültige Schlüsse), argumentloses Nebeneinander inkompatibler Positionen, ausschweifende Phantasien (z.B. Weltrettung durch Weltherrschaft von Frauen), falsche Sachverhaltsfeststellungen (vor allem falsche Zahlen), heftige Entwertungen anderer Netizens, Anfeindungen, Freund-Feind-Denken, leidenschaftliche Empörung, Verballhornungen, Schwarz-Weiß-Denken, Gerüchte, Geschwätz. Unser Fazit war: Das Gesellschafterprojekt war als Versuch eines ethisch-politischen Selbstverständigungsdiskurses von Netizen-Citoyens kühn, aber gescheitert. Die Organisatoren nahmen die erfolglose Diskussionsplattform bald wieder vom Netz, das Moderatorenteam war frustriert, und mangels erlangter Ergebnisse blieb Resonanz in »alten« Leitmedien nahezu aus.[18]

4.2 Einschränkung durch strukturelle Unkooperativität

Wie kommt es zu der Dysfunktionalität relativ zu Erwartungen, es würden sich verantwortungsvolle Diskutanten-, Proponenten- und Opponentenrollen in der für das Gesellschafterprojekt gewählten Mediengestalt stabilisieren lassen? Um Erfahrungen zu sammeln, beteiligten sich einige Mitglieder der Forschungsgruppe als Netizen-Diskutanten und mischten sich pseudonymisiert ein. Wir lancierten Themen, kritisierten begründungsbedürftige, aber begründungslos gesetzte Positionen, setzen unsererseits begründungsbedürftige Positionen ohne Begründung, versuchten reflexiv die Form zu diskutieren, die der Gang der »Diskussion« an kritischen Stellen nahm, und verglichen die Folgen unserer online Interventionen kontrafaktisch mit Folgen, die wir für unsere Interventionen erwartet hätten, falls sie in der Rationalitätsumgebung einer überschaubaren, zum Zweck der Diskussion versammelten Kommunikationsgemeinschaft erfolgt wären.

Unter der Annahme struktureller Unverbindlichkeit, wie oben dargestellt, werden viele überraschende Befunde erklärlich. Normativ bewertbar als rationalitätsmindernd wird strukturelle Unverbindlichkeit, sofern wir uns auf Maßstäbe (Standards) für die Rationalität von Kommunikationspraktiken relativ zu den Zwecken, für die die Praktiken gut sein sollen, beziehen können. Dass *kooperativ* miteinander Gespräche zu führen ein grundierender Zweck von Diskussion ist, erscheint plausibel. So gesehen, dürfen wir die Gricesche Rekonstruktion von »Konversationsma-

18 Die Jahreschronik des Vereins notiert für 2006 lapidar, die »Aufklärungsinitiative« habe die Bevölkerung »aufgefordert, sich als aktive Gesellschafter zu begreifen und in die Diskussionen um die Zukunft der Gesellschaft einzuschalten. Dieser sehr breite Ansatz wird später zugunsten einer Fokussierung auf die Kernthemen der Aktion Mensch wieder aufgegeben« [https://www.aktion-mensch.de/ueber-uns/chronik/chronik-detail].

ximen«[19] als eine Teilrekonstruktion von kommunikativer Rationalität behandeln. Dann lässt sich die merkliche Minderung rationaler Diskussion (bis hin zum Zerfall) als Entkräftung von einigen bis allen Konversationsmaximen modellieren. Diese Entkräftung lässt sich aus dem Zusammenwirken von Eigenschaften der Medienumgebung und Personeigenschaften der Netizens erklären.

4.3 Einschränkung durch Marginalisierung von indirekter Rede

Die reale und zugleich ideale Kommunikationsgemeinschaft haben wir oben mit Apel so interpretiert, dass argumentative Kommunikation als eine Praxis deutlich geworden ist, die in sich das Diskursmedium ausbilden kann. Im Diskursmedium kann u.U. eine Rationalitätsumgebung entstehen, die Diskurspartner spezifisch dazu anhält, ihre geistigen Kräfte für Geltungsreflexion einzusetzen um, wie oben gesagt, intersubjektiv geteilte Auffassungen des Gültigen zu justieren, wo nötig. Zum Tragen kommt der Einsatz von Geltungsreflexion im Erkennen und seinerseits durch Gründe ausweisbaren Beurteilen der Überzeugungskraft von Gründen.

Dafür, dass Geltungsreflexion zum Tragen kommen *kann*, braucht es geeignete *sprachliche* (syntaktisch, semantische, pragmatische) Ressourcen. Ohne Sprache keine Geltungsreflexion. Ein notwendiges Minimum, mit dem das »Spiel des Anforderns und Angebens von Gründen« (Brandom 2000: 189–196) schon irgendwie läuft, gibt es wohl in allen kulturellen Wir-Gruppen mit gemeinsamer Umgangssprache, in der Warum-Fragen aufgeworfen und beantwortet werden können. Komplexere sprachliche Ressourcen und eine Redepraxis mit einem Repertoire, das diese auch zu nutzen versteht, können die Möglichkeiten von Geltungsreflexion aber erweitern und verbessern.

Die theoretische Betrachtung der Zusammenhänge von Redepraxis, Geltungsreflexion und Diskurs würde eine Auseinandersetzung mit der Theorie-Praxis der Rhetorik erfordern und muss an dieser Stelle unterbleiben, ebenso wie die Auseinandersetzung mit verschiedenartigen Möglichkeiten, Redepraxis als solche sprachphilosophisch zu durchdringen.[20] Stattdessen möchte ich an dieser Stelle nur auf eine einzige, für die Verbesserung von Geltungsreflexion relevante sprachliche Ressource aufmerksam machen.

Zu Repertoire-Elementen, die die Möglichkeiten von Geltungsreflexion im Diskursmedium steigern, würde ich Handlungsweisen wie die folgenden zählen:

19 Als deren allgemeinste formuliert Grice: »Make your conversational contribution such as is required, at the stage at which it occurs, by the accepted purpose or direction of the talk exchange in which you are engaged.« (Grice 1991: 26)

20 So kann Redepraxis etwa, wie in der klassischen Sprechakttheorie, intentionalistisch (Burkhardt 1990), universal- oder transzendentalpragmatisch, inferentialistisch (Berdini 2013), rationalistisch (Brandom 2000), oder, so der neueste Trend (Degen 2023), probabilistisch analysiert werden.

Gedankengänge allgemein kennzeichnen, eine Begründung geben, eine Bedingung ansprechen, Absicht und Zweck ausdrücken, eine Folge ausdrücken oder eine Folgerung ziehen, etwas einräumen, rechtgebend zustimmen (»Konsens«) oder widersprechen (»Dissens«), eine Meinung äußern, etwas feststellen, beurteilen oder bewerten, Gewissheit, Zweifel oder Vermutungen ausdrücken, eine Forderung erheben oder einen Vorschlag machen, auf etwas vor- oder zurückverweisen, etwas besonders hervorheben, u.a.m. Sobald Diskurspartner mit Sequenzen solcher Redehandlungen zurechtkommen müssen, erleichtert ein interessanter grammatischer Apparat die Kontrolle über diese Komplexität: indirekte Rede.

Die »Wiedergabe und Erörterung fremder Reden, des fremden Wortes [ist] eines der am weitesten verbreiteten und wesentlichsten Themen menschlicher Rede« (Günthner 2000: 1).[21] In direkter Rede, in der grammatisch der indikativische, gewissermaßen wirklichkeitsverbürgende Verb-Modus vorherrscht, bleiben etwaige Interferenzen zwischen der zitierten Äußerung und der Perspektive des zitierenden Sprechers unartikuliert. Nutzen Sprecher hingegen die indirekte Redewiedergabe, rekonstruieren (paraphrasieren, interpretieren) sie den Äußerungsinhalt aus ersichtlich ihrer eigenen Perspektive und geben nicht vor, die tatsächlichen Worte des Sprechers wiederzugeben, auf dessen Äußerungen sie sich beziehen, oder dessen Perspektive originalgetreu zu übernehmen (vgl. Günthner 2000 mit Literaturverweisen). Der grammatische Apparat der indirekten Rede macht es vergleichsweise leichter, die in der Sprechakttheorie sogenannten illokutionären Kräfte zu relationieren. (»Du behauptest, A habe Zweifel, ob C wirklich glaube, die Regierung betrügt uns.«) »Wenn man sich nicht ganz sicher ist, ob die Informationen tatsächlich stimmen oder die Quellen zuverlässig sind, ist der Konjunktiv immer eine gute Wahl« (Volodina 2023, mit sinnfälligen Beispielen).

Durch den (epistemischen) Nuancenreichtum indirekter Rede können Sprecher besser deutlich machen, dass sie z.B. an einer Aussage zweifeln oder andere Einstellungen zum Mitteilungsinhalt haben, während Sprecher mit indikativisch beherrschter direkter Rede das ihnen Mitgeteilte eher nur so wiederholen, wie andere sie geäußert haben.

Zwar sind die meisten Sozialen Medien technisch mit Emojis, Likes und weiteren Möglichkeiten ausstaffiert, die gewisse funktionale Äquivalente zu Code-Switching und anderen, grammatischen oder prosodisch-stimmlichen Möglichkeiten schaffen, mit deren Hilfe selbst innerhalb direkter Rede fremde Äußerungen nicht bloß direkt wiedergegeben, sondern auch inszeniert werden können. Durch solche Hilfsmittel kann ein zweiter, zitierender Netizen sein Selbstverhältnis zur Äußerung des ersten mitausdrücken. Die technisierbaren Möglichkeiten für Netizens bleiben aber nach Differenziertheit und Artikuliertheit m.E. weit hinter dem zu-

21 Günthner zitiert dabei M. Bachtin.

rück, was mit Formen indirekten Rede in Unterhaltungen unter Sprechern möglich ist.

Direkte Rede kann emotionale Tiefe und Authentizität vermitteln, indem sie genau zeigt, wie etwas gesagt wurde. Indirekte Rede kann in Situationen angemessen sein, wo direkte Rede als zu konfrontativ oder unhöflich angesehen werden könnte. Mittels indirekter Rede lassen sich Mitteilungen machen, ohne direkt zu konfrontieren, was in sensiblen Kommunikationssituationen sehr vernünftig sein kann. Auch ermöglicht indirekte Rede, eine komplexe Mitteilung zu kondensieren oder zu vereinfachen, was besonders nützlich ist, wenn lange mitteilungsreiche Sequenzen zusammengefasst werden sollen, z.B. um einen Diskussionsstand zu äußern. Durch gezielte Wechsel zwischen direkter und indirekter Rede können unterschiedliche Perspektiven und Stimmen innerhalb einer Erzählung oder Diskussion hervorgehoben werden, was das Verständnis der Ansichten und Gründe der Beteiligten vertiefen kann.

Wie häufig wird indirekte Rede in Sozialen Medien benutzt? Zum Beispiel in Tweets und Retweets auf X/Twitter? Generell gilt: Wie häufig Netizens von indirekter Rede in sozialen Medien Gebrauch machen, variiert stark je nach Kontext und Zweck der Kommunikation. Mit Bezug auf Twitter lässt sich vorsichtig verallgemeinern: Die meisten twitternden Netizens bevorzugen direkte Rede oder direkte Zitate in ihren Tweets, um Authentizität und Originalität zu vermitteln, besonders dann, wenn Meinungen oder Aussagen von Personen des öffentlichen Lebens geteilt werden. Das Zitieren von Tweets ist eine Form der direkten Rede, die in dieser Mediengestalt sehr häufig vorkommt. Zwar könnte die künstliche, technisch gesetzte Verknappung des Äußerungsumfangs auf X/Twitter zum Gebrauch indirekter Rede auffordern, um Kernpunkte zeichensparend, quasi ökonomisch rational mitzuteilen. Dem steht aber ein ästhetischer bzw. stilistischer, den aufgebauten Beteiligungsgewohnheiten innewohnender Druck entgegen, erkennbar *informell* ›rüberzukommen‹. Unter diesem Druck hat man einen wertrational guten Grund, indirekte Rede zu vermeiden. Die künstliche Verknappung trägt sicher auch zum hohen Stellenwert von Bildern, insbesondere Memen bei, die komplexe Sinnbezüge verdichten, – allerdings propositional unartikuliert, also nur sehr eingeschränkt diskursivierbar.[22]

Zusammenfassend: Dass in text- und bildbasierten Gestalten von Social Media indirekte Rede marginalisiert wird, halte ich für eine empirisch feststellbare Tatsache. Zu einem die Performanz kommunikativer Rationalität von Netizens verschlechternden Faktor wird diese Marginalisierung unter Umständen dann, wenn Netizens diskursiv anspruchsvolle Zwecke in solchen Medien verfolgen wollen. Dieses Werturteil erscheint mir richtig.

22 Zu Memen siehe den Beitrag von Kai Denker im vorliegenden Band.

4.4 Strukturelle Entkopplung von Weitergabeentscheidungen und Geltungsreflexion

Die »dialektische« Kommunikationsgemeinschaft sensu Apel ist vom Gespräch her gedacht.[23] Begrenzt ist jede reale dadurch, welche Menschen, die miteinander reden könnten, in Gestalt des Geprächs wirklich füreinander erreichbar sind oder es wären, wenn sie wollten. Insofern die kulturelle Reproduktion der Gattung ohne Vergesellschaftung in natürlichen Sprachen unmöglich wäre, war und ist jeder von uns »immer schon« Mitglied qua Sozialisation. Die Grundlage (nicht: Plattform) dieser *existenziellen* Mitgliedschaft erlaubt uns lebenslang die Erzeugung aller Sprechhandlungsbeziehungen, die jeder von uns unablässig auf- und abbaut, verknüpft oder löst. Das Zuhören, die Zuwendung von Aufmerksamkeit auf gesprächsweise Äußerungen, die nötig ist, um sie auf sich wirken zu lassen, ist die Basishandlung, die in der Redepraxis darüber entscheidet, ob Wechselrede, Wiedergabe, Weitergabe von Mitteilbarem erfolgen oder ausbleiben.

Eine in digitalkulturellen Mediengestalten dem vergleichbare Basishandlung ist das Anklicken zum Aufrufen von Äußerungen anderer Netizens, um sie auf sich wirken zu lassen. Von der Entscheidung zu dieser Basishandlung hängt ab, ob und welche der medial eingerichteten Optionen für die Fortsetzung genutzt werden. So muss ich mich z.B. als X/Twitter-Netizen, der an dort laufender ›Konversation‹[24] sich beteiligen will, entscheiden, (1) auf eine Äußerung (›Post‹) gar nicht zu reagieren; (2) durch Klick auf ein Herz-Symbol mein Gefallen zu einem bleibenden Bestandteil einer Äußerung zu machen; (3) die Äußerung, ob unverändert oder durch einen Kommentar angereichert, simultan weiterzugeben (Retweet/Repost) an alle, die sich bis auf weiteres entschieden haben, Äußerungen von mir beachten zu können, das sind: alle Netizens, die sich einmal zum Beitritt zur Menge meiner Follower entschieden und dies noch nicht rückgängig gemacht haben; (4) zu antworten, falls man zur Menge der Antwortberechtigten zählt, um meine Äußerung zu einem weiteren Element in der Sequenz von Äußerungen zu machen, zu der die mir angezeigte bereits gehört. Diese Medienumgebung realisiert Kommunikationsgemeinschaft für Netizens so, dass eine fluide Vernetzung ihrer Äußerungen zu online-Unterhaltung(en), deren variabler Öffentlichkeitscharaker die Beteiligten selbst bestimmen können, sehr leichtgemacht und kräftig angereizt wird.

Längst ist die Menge an Forschung jeglicher Art zu Twitter kaum noch zu überschauen (vgl. schon Antonakaki 2021). Unter dem Problemdruck anschwellender Des- und Falschinformation in politischen und anderen Öffentlichkeiten

23 Darin liegt bei Apel (2011) dem Anschein zum Trotz kein reduktionistischer Logozentrismus, da er im Subjekt der Rede auch die Leiblichkeit und organische Verkörperung mitdenkt (s. Molina Molina 2017).

24 Für Details siehe [https://help.twitter.com/de/using-x/x-conversations].

boomt insbesondere Forschung zur Tauglichkeit Sozialer Medien für offen oder verdeckt strategische Kommunikation für politische Zwecke, z.B. Propaganda, gleichsam die dunkle politische Kehrseite der schon lange erforschten Tauglichkeit für kommerzielle Zwecke wie Werbung und Marketing, um Kaufentscheidungen zu beeinflussen. Im Rahmen philosophischer Digitalisierungsforschung wäre es m.E. lohnend, Ideen für die, selbstverständlich nur interdisziplinär zu bewerkstelligende Erforschung der Tauglichkeit digitalkultureller Medien für diskursaffine Zwecke, etwa öffentliche politische Deliberation, vorzuschlagen.

Eine solche Idee möchte ich nun skizzieren. Sie betrifft den Nexus von Aufmerksamkeit und Weitergeben, der Interesse verdient, weil er für die Verbreitungsreichweite und -dynamik von Äußerungen wie eine Schleuse wirkt. Um die Idee in aller Kürze zu erläutern, muss ich schematisieren:

> U sei eine Äußerung im Sozialen Medium M; p(U) sei die Wahrscheinlichkeit, dass wenn U einen Netizen *ego* erreicht und *ego* U auf sich wirken lässt, *ego* U dann an mindestens einen anderen Netizen *alter* weitergibt. R seien *ego*s Beweggründe dafür, U an *alter* weiterzugeben. V(R) sei der bewertbare rationale Wert dieser Gründe.

Wenn z.B. der Netizen namens Darkness postet »L'Eurovision 2024 n'est que le reflet de cette décadence occidentale en phase terminale. #Eurovision2024« [plus TV-Screenshot], der zum Zeitpunkt t von 38.500 Netizens angeschaut, 465 mal geherzt, 0 mal beantwortet und von 191 Netizens weitergegeben wurde, (die über alle Zahlen im Bilde sind, da diese laufend aktualisiert und im Tweet angezeigt werden), aus welchen Gründen entscheidet Netizen Nr. 192 sich, die Äußerung an seine Follower weiterzugeben? Angenommen, wir brächten solche Gründe in Erfahrung, was würden wir von ihnen halten? Die schematisch dargestellte Medienphase von Aufmerksamkeitszuwendung zu Weitergabeentscheidung legt drei interessante Forschungsfragen nahe:

> Q1: Welche Faktoren bestimmen p(U)?
> Q2: Welche Faktoren bestimmen V(R)?
> Q3: Welche Faktoren bestimmen de facto das Verhältnis p(U)/V(R), welche *sollten* es?

Zu erwarten ist, dass *ego*s Beweggründe R in die Antworten auf Q1 eingehen, es sei denn, *ego* verbreite U ganz ohne Grund. Unter Menschen, die eigentlich immer ihre Gründe haben, wenn man sie nur richtig fragt, wäre das aber nur als Grenzfall denkbar. Auch ist zu erwarten, dass außer R noch andere Faktoren X, Y, Z eine Rolle spielen, die gründefern sind, d.h. die in *ego*s versprachlichbarem *knowledge why* (Gründe-

wissen) nicht repräsentiert sind.[25] Hier greift wieder, wie am Schluss von Abschnitt 2 bemerkt, dass auch qua Netizens die konkreten Individuen im ganzen Umfang ihrer Subjektivität engagiert sind, also auch mit ihrer unbewussten Abwehr, Selbsttäuschung und Rationalisierung (Giampieri-Deutsch 2012).

Nun zu Q2. Um auf Q2 zu antworten müssen wir jedenfalls in Erfahrung bringen, wie *ego* über U denkt. Das erfordert einen dialogischen Zugang, also hermeneutische Methoden. Wir werden zudem auf bestimmte Rationalitätskonzeptionen verweisen müssen, die in signifikanten Kommunikationsgemeinschaften hochgehalten werden, und müssen klären, mit welchen Rationalitätskonzeptionen *ego* sich identifiziert, und zudem, was *ego* diesbezüglich über *alter* annimmt. Welchen *rationalen* Wert ein Beweggrund hat, lässt sich nur relativ zu einer urteilend in Anschlag gebrachten Auffassung dessen, wie *wir* vernünftigeres Verhalten von weniger oder widervernünftigem unterscheiden *sollten*. Das könnte unter *homines oeconomici* anders ausfallen als unter *homines ludentes* und wieder anders unter *homines religiosi* usw.[26]

Den Pluralismus-Relativismus, der sich hier andeutet, könnten wir dadurch aufheben, dass wir entsubstanzialisiertere bzw. formalere Rationalitätskonzeptionen in Anschlag bringen, wie die Diskursrationalität. Mit welchem Recht, lässt sich aber erst dann begründen, wenn wir auch die Frage Q3 einbeziehen. Denn in Antworten auf Q3 müssen wir auf Praktiken und ihre Zweckdienlichkeit verweisen, und unsere Rationalitätskonzeptionen sollten zu den Praktiken passen, in denen wir unterwegs sind.[27] Wollen wir als Diskurspartner unterwegs sein, dann *sollte* Diskursrationalität, auch als Netizen unter Netizens, für unser medieninternes Handeln bedeutsam bleiben. Wirklich bedeutsam bleiben kann sie aber nur, wenn die medieninterne Rationalitätsumgebung von M die Diskursrationalität wirksam unterstützt. Das sollte sie zumindest dann, wenn jemand, z.B. eine gemeinnützige Einrichtung, M für diskursaffine Zwecke tauglich hält und einsetzen will. Wie das misslingen kann, habe ich oben am Gesellschafterprojekt veranschaulicht.

Forschung zu Q1 könnte robuste Verallgemeinerungen erbringen, wie die Medienumgebung M als eine Rationalitätsumgebung den Raum der Gründe formt, worin Netizens, wenn sie in M handeln, sich orientieren. Forschung zu Q2 und Q3 könnte erhellen, welchen Anteil Geltungsreflexion an Weitergabeentscheidungen in bestimmten Kontexten hat. Gestützt auf Erfahrungen aus dem Gesellschafterprojekt

25 Dass p(U) bspw. mit dem Ausmaß der emotionalen Aufladung von U steigt, darf bezüglich Twitter als empirisch gut belegt gelten (Stieglitz/Dang-Xuan 2014).

26 Damit signalisiere ich die theoretische Auffassung, dass Rationalitätskonzeptionen nicht freistehen, sondern in Ethosformen eingebettet sind. Zugunsten dieser Auffassung kann ich an dieser Stelle nicht argumentieren. Es genügt für den obigen Punkt die Annahme, dass es guten Sinn macht, mehrere Rationalitätstypen zu unterscheiden. Sogar innerhalb eines Typs können Unterschiede eintreten (Kettner 2012).

27 Auch diesen normativen Punkt kann ich hier nur formulieren, aber nicht weiter begründen.

(s. Abschnitt 4) möchte ich hypothetisch behaupten, dass informelle Kommunikation in Sozialen Medien die Wahrscheinlichkeit und die Gründe des Weiterverbreitens von Äußerungen strukturell von Geltungsreflexion entkoppelt und depotenziert. Würde sich diese These bestätigen, hätten wir eine weitere Teilerklärung für medienspezifische Einschränkungen kommunikativer Rationalität.

Das Gros der Forschung zur Weitergabe und Ausbreitung in Netzen ist quantitativ und prädiktiv: Mit Hilfe von probabilistischen Modellen versucht man, »to realistically portray a given microblogging service« (Sun/Liu 2023), um für p(U) und andere Disseminationsmaße eine gute Vorhersagefunktionen zu bilden, neuerdings verstärkt durch maschinelles Lernen. Philosophische Digitalisierungsforschung im Rahmen einer normativen Theorie von Gründen ist noch unerprobt, könnte aber m.E. innovative Forschungsdesigns entwickeln und womöglich sogar mit probabilistischen Ansätzen kooperieren, z.B. um Media Design zur Verbesserung der Kapazität für Geltungsreflexion, wo es nötig wäre, zu betreiben und gegen zweckwidrige Interventionen resilienter zu machen.[28]

5. Primärprozesshafte Kommunikation

In diesem letzten Abschnitt nehme ich den am Ende von Abschnitt 2 liegengebliebenen Faden noch einmal auf, die Tatsache nämlich, dass wir als Beteiligte am Austausch in Sozialen Medien nicht nur teilnehmen als die, als die wir in Mediengestalt uns zeigen, Netizens, sondern dass wir im ganzen Umfang unserer Subjektivität engagiert sind. Dagegengehalten ist die technizistische Metapher der »Schnittstelle« (Interface) eher missverständlich, da sie suggeriert, wir seien nur in einem Übergangsbereich aktiv. Aber der Geist in der Maschine ist Geist von unserem Geist. Natürlich ist auch das bloß metaphorisch formuliert und bedarf der Explikation. Dafür bieten sich Geisttheorien aus mindestens drei Paradigmen an: Hegels Theorie des subjektiven und objektiven Geistes, die sich durchaus sozialphilosophisch auslegen und auf Medien beziehen lässt; enaktivistische Geisttheorien innerhalb des großen Feldes der *Philosophy of Mind*; kognitive und psychoanalytische Theorien des bewussten und des bewusstseinsfernen Seelenlebens von Personen. Ich nutze in diesem Abschnitt nur letztere, um eine These zu plausibilisieren, die m.E. erhellende Neubeschreibungen vieler irritierender Phänomene in digitalkulturellen Wir-Gruppen erlaubt und für einige auch psychodynamische Teilerklärungen verspricht.

Die These lautet: Soziale Medien stellen für informellen Austausch Medienumgebungen dar, die für primärprozesshafte Geistestätigkeit der Netizens durchlässiger sind als die in anderen Mediengestalten (z.B. journalistischen), so dass pri-

28 Dass dies auch für Wissenschaftskommunikation in sozialen und sogar in Expertennetzwerken sehr nötig ist, hat sich während der Covid-Pandemie dramatisch gezeigt (Shahbazi 2023).

märprozesshafte Geistestätigkeit in primärprozesshafter Kommunikation öffentliche Ausdrucksmöglichkeiten gewinnt.

Mit dieser These greife ich Sigmund Freuds alte Unterscheidung von Sekundär- und Primärprozess wieder auf. Innerhalb der Geschichte der psychoanalytischen Theoriebildung gehörte diese Unterscheidung anfangs so zentral zum metapsychologischen Theoriekern wie die Unterscheidung von bewusster, bewusstseinsnaher (*vorbewusster*) und bewusstseinsentzogener (*unbewusster*) Intentionalität. Späterhin wurde sie von klinischen, d.h. besonders für die Belange der Behandlungspraxis brauchbaren Konzepten wie *Abwehr* und *Übertragung* überholt, während in der Weiterentwicklung der Metapsychologie Persönlichkeitsmodelle in den Vordergrund des theoretischen Interesses traten, wie das bekannte Strukturmodell von *Ich*, *Überich* und *Es*, sowie trieb- und später vor allem objektbeziehungstheoretische Konzepte.

Kurz gesagt: Freud beobachtete zwei verschiedenartig organisierte Produktionsweisen von Sinnzusammenhängen im Seelenleben, die am prägnantesten hervortreten im Vergleich einer realitätsorientierten Person im Vollbesitz ihrer geistigen Kräfte mit derselben, wenn sie träumt, fantasiert, oder neurotische Symptome entwickelt. »Wir haben erfahren, dass die Vorgänge im Unbewussten oder im Es anderen Gesetzen gehorchen als die im vorbewussten Ich. Wir nennen diese Gesetze in ihrer Gesamtheit den Primärvorgang im Gegensatz zum Sekundärvorgang, der die Abläufe im Vorbewussten, im Ich, regelt.« (Freud 1941: 86)[29] Sekundärprozesshaft sensu Freud ist Denken oder überhaupt Geistestätigkeit, soweit sie einen rational organisierten Eindruck macht, und primärprozesshaft, wenn Sinngehalte wie losgelöst von der Erfahrung und ohne Rücksicht auf die Realität (›Realitätsprinzip‹) sich unkontrolliert verdichten und verschieben; wenn Sinngehalte zusammenwachsen, wo sie ›eigentlich‹ nicht zusammengehören; wenn Gegensinniges widerspruchslos bleiben kann (Kohärenz ohne Konsistenz); wenn zeitliche Ordnung beliebig wird (Zeitlosigkeit); wenn die Annäherung an lustvolle affektive Zustände (Lustgewinn z.B. in Form phantasierter oder halluzinierter Erfüllung triebhafter Wünsche) zum beherrschenden Attraktor wird (›Lustprinzip‹) oder disruptiv durchbricht. Sehr vereinfacht könnte man sagen, primärprozesshafte Geistestätigkeit ist schnell und rücksichtslos unrealistisch, aber hemmungslos lustvoll.[30]

---

29 Primär vs. sekundär ist also nicht zeitlich oder ontogenetisch gemeint. Eine sprachphilosophisch reflektierte Rekonstruktion gibt Marcia Cavell (Cavell 2005, zum Primärprozess s. bes. 80–82). Die derzeit genaueste Explikation gibt Robert Holt (Holt 1989a; Holt 1989b; Holt 2009: 28–37). Wie das Konzept von verschiedenen Schulrichtungen rezipiert wird, beschreibt Leichsenring 2022. Für aktuelle Theorien dualer kognitiver Prozesse siehe Evans/Stanovich 2013. Diese mit der psychoanalytischen Unterscheidung zu vermitteln, wäre reizvoll, muss aber an dieser Stelle unterbleiben.

30 »The primary process is a joint function of wishfulness and unrealism. […] The more thought (and also affect and behaviour) can be characterized as an unrealistic seeking for immedi-

Robert Holt schlägt den Bogen vom höchstpersönlichen Seelenleben des Individuums zu den Praktiken kultureller Wir-Gruppen, die dessen Lebenswelt ausfüllen:

> »From the beginning of his life, a child is also exposed to culture, and to special child's subculture, which contains numerous crystallized and far-reachingly organized primary-process systems. Myths, legends, fairy stories, and other simple types of fiction incorporating recognizable forms of the primary process […]. Here may be a new horizon for functional anthropology: cultural styles of primary process – modes of magical and autistic thinking which will be meaningfully related to other themes and traits in the culture concerned.« (Holt 1989a: 274)

Holt unterstreicht auch die Nähe primärprozesshaften Denkens zu Kreativität, spontanen Einfällen und befreitem Sichgehenlassen, was die ältere psychoanalytische Kunsttheorie (Stokes 1974) ›Regression im Dienste des Ich‹ nannte:

> »Culture provides a number of special contexts and social roles, including art, science, and humor, in which ›regressive‹ thought is allowed and in fact encouraged because of the social value put on the results. Invoking such contexts helps the person maintain active control.« (Holt 2009: 75)

Es liegt auf der Hand, dass diese analytische Perspektive für unser Verständnis irritierender Phänomene wie Verschwörungsdenken, Affektenthemmung, Realitätsverleugnung etwas abwirft, gerade dann, wenn diese in Medienumgebungen der Sozialen Medien verstärkt auftreten. In der analytischen Perspektive auf Integrationsfiguren von sekundär- und primärprozesshafter Sinnverarbeitung kann z.B. die Design- und Rezeptionsästhetik solcher Medienumgebungen auf psychodynamische Affordanzen hin untersucht werden.

Allerdings muss zugunsten meiner These noch erklärt werden, wie solche Medienumgebungen überhaupt zu Katalysatoren oder zumindest Resonanzböden von primärprozesshafter Geistestätigkeit werden können. Eine verblüffend kurze Antwort wäre die Rückfrage, wie sie es denn für sekundärprozesshafte werden, wie wir es selbstverständlich annehmen. Eine längere Antwort, die ich oben angedeutet habe, hätte die Vorstellung zu destruieren, das Seelenleben eines Individuums sei präkulturell verfasst und weder von symbolischen Ordnungen noch von Medien berührt. Diese merkwürdige Vorstellung, die wir im Hinblick auf bewusstes Seelenleben doch sofort zurückweisen würden, findet aber oft noch einen gewissen Anklang, sobald es um das Unbewusste im Seelenleben geht. Mit der Rede von ›Geis-

---

ate gratification, the more it is to be considered primary process […]. And the more thought or behavior is organized by adaptive considerations of efficiency in the search for realistic gratification, the more it approximates the ideal of secondary process.« (Holt 1989b: 297f.)

testätigkeit‹ versuche ich die vertrackte Frage nach der einen richtigen Ontologie von psychischer Realität aufzuschieben.

Meine These, dass primärprozesshafte Geistestätigkeit in primärprozesshafter *Kommunikation* öffentliche Ausdrucksmöglichkeiten gewinnen kann und dies besonders gut in Medienumgebungen Sozialer Medien, ist dann ontologisch unproblematisch oder jedenfalls nicht problematischer als die Annahme, dass sie in vielen Formen künstlerischer Produktion besonders gut öffentliche Ausdrucksmöglichkeiten gewinnen kann.

Statt durch eine mentalistische Ontologie möchte ich meine These durch Würdigung einiger medien*technisch* fundierter Eigenschaften untermauern, die ich nun kurz noch anspreche.

Der Stand der Technik bietet (1) die Möglichkeit unbegrenzter Reproduzierbarkeit von Äußerungen, damit aber auch ihr totale Dekontextualisierung. Ihr Sinn wird dadurch radikal verschiebbar. (2) Die technische Möglichkeit unbegrenzter Speicherung und Reaktivierung ist eine funktionale Entsprechung zur primärprozesshaften Zeitlosigkeit. (3) Die technische Möglichkeit grenzenloser Konnektivität verführt zu erregenden Phantasien. (4) Die technische Möglichkeit, in Echtzeit oder zeitversetzt zu agieren nährt das Phantasma einer ewigen Gegenwart. (5) Die Möglichkeit von starken Knotenbildung im Netzwerk macht die Äußerungen einiger Netizens so einflussreich, dass diese nach Belieben Sinnzusammenhänge verschieben oder verdichten können, die von vielen geteilt werden, *weil* sie von vielen geteilt werden, ohne Rücksicht auf medienexterne Realität. (6) Die technische Möglichkeit von positiven und negativen Rückkoppelungen in Kommunikationsströmen verleiht den Kommunikationsnetzen eine Schwingungsfähigkeit, die für affektiv enthemmende Lawineneffekte (wie Shitstorms und Candystorms) sorgt. (7) Die technische Möglichkeit von intramedial gebildeten Kommunikationsgemeinschaften, sich gegeneinander abzuschirmen, schaltet Negation aus.[31] So kann Gegensinniges sich unangefochten behaupten. (8) Last not least: Die Möglichkeit der Klarnamenvermeidung, die wie ein Freibrief für Enthemmung und eine Senke für Verantwortungszuschreibungen wirkt.

31 Das wird in der Forschung als Bildung von Echokammern und Filterblasen behandelt. »Users tend to aggregate in communities of interest, which causes reinforcement and fosters confirmation bias, segregation, and polarization. This comes at the expense of the quality of the information and leads to proliferation of biased narratives fomented by unsubstantiated rumors, mistrust, and paranoia.« (Del Vicario et al. 2016: 558). Siehe auch Sasahara et al. 2021.

6. Fazit

Wie können wir aus Besonderheiten der digitalkulturellen Kommunikationspraktiken in Sozialen Medien systematische Einschränkungen für kommunikativ rationale Kommunikation erklären? Wir konnten einige Erklärungshypothesen entwickeln, die in Betracht der Komplexität der aufgeworfenen Fragestellung aber bestenfalls als Teilantworten gelten dürfen: die Normalität struktureller Unverbindlichkeit, strukturelle Unkooperativität, die Marginalisierung von indirekter Rede, die strukturelle Abkopplung von Retweet- und anderen Weitergabeentscheidungen von der Reflexion auf den Wert des Teilens des Mitgeteilten für Zwecke der Diskussion und des argumentativen Diskurses.

Ich habe von zwei Diskussions-Projekten in Sozialen Medien berichtet, die nicht erfolgreich waren, wenn man den Erfolgserwartungen Normen kommunikativer Rationalität zugrunde legt, wie sie in diskursfähigen Kommunikationsgemeinschaften für Zwecke von Diskussion und Argumentation befolgt werden sollten. Die Betrachtung dieser Misserfolge hatte eine pragmatistische normative Pointe: Die Rationalitätsstandards, die wir in unsere Rationalitätsurteile über eine bestimmte Praxis investieren, *sollen prima facie* zu den jeweiligen Zwecken passen, für deren Verfolgung eine bestimmte Praxis den passenden Spielraum gewährt. Das hat Konsequenzen für die Selbstkritik von Kritik. Wenn rationale Erwartungen an Medienkommunikation enttäuscht werden, muss geklärt werden, wie rational diese Erwartungen selbst sind.

Instruktiv waren die Misserfolge der Diskussions-Projekte auch deshalb, weil sie weder aus mangelhafter Digitaltechnik noch aus der systematischen Verzerrung von Kommunikation durch offen oder verdeckt strategisch ausgespielte Interessen von kommerziell interessierten Unternehmen erklärt werden konnten.

Auf der Grundlage der ersten vier Teilantworten auf die komplexe Fragestellung wurde ein Forschungsdesign vorgeschlagen, das die in der empirischen kommunikationswissenschaftlichen Literatur erprobten quantitativen Ansätze, Weitergabeentscheidungen mit Hilfe probabilistischer Modelle vorherzusagen, ergänzen und u.U. auch korrigieren könnte, nämlich durch Artikulation des jeweiligen Raums der Gründe, in welchem die Beteiligten ihre Entscheidungen treffen. Um der allerdings begrenzten Bewusstheit und Artikulationsfähigkeit von Menschen für ihre Beweggründe wiederum Rechnung zu tragen, wurde dieser tendenziell rationalistische Ansatz mit einem weiteren Vorschlag für einen psychodynamischen Begriffsrahmen ergänzt, der in einer – noch zu entwickelnden – kritischen Massenpsychologie unserer gegenwärtigen, vom Hypermedium Internet geprägten Kommunikationsmedienumgebung eine wichtige Rolle spielen sollte: Das psychoanalytische Konzept von Primär- und Sekundärprozess erscheint für Forschungsperspektiven auf Kommunikation in Sozialen Medien besonders relevant, weil einige der medientechnisch beschreibbaren Eigenschaften dieser Mediengestalt den psy-

chologisch beschreibbaren Eigenschaften von primärprozesshaftem Denken funktional entsprechen und dadurch eher entgegenkommen, als sekundärprozesshaftem, in weitestem Sinne rational kontrolliertem Denken, Fühlen und Handeln.

Wenn sich die begründete Vermutung erhärten lässt, dass bestimmte Mediengestalten vorzugsweise primärprozesshaftem Denken Spielraum gewähren, wäre eine für die philosophische Digitalisierungsforschung interessante, auch politisch wichtige weiterführende Frage die nach den besonderen Machtverhältnissen, die diese Medienpraktiken umgeben und durchwirken: Wer kann solchen Spielraum für welche Zwecke wie erfolgreich in Dienst nehmen, mit den folgsamen Spielenden zusammen und auch gegen die widerstrebenden?

Literatur

Aichner T.; Grünfelder, M.; Maurer O.; Jegeni, D. (2021): Twenty-Five Years of Social Media. A Review of Social Media Applications and Definitions from 1994 to 2019, in: *Cyberpsychology, Behavior, and Social Networking*, 24(4), 215–222.

Ancis, J.R. (2020): The Age of Cyberpsychology: An Overview, in: *American Psychological Association* [doi.org/10.1037/tmb0000009].

Antonakaki, D.; Fragopoulou, P.; Ioannidis, S. (2021): A survey of Twitter research. Data model, graph structure, sentiment analysis and attacks, in: *Expert Systems with Applications*, 164, Art. 114006. [https://doi.org/10.1016/j.eswa.2020.114006].

Apel, K.-O. (2011): Die Logos-Auszeichnung der menschlichen Sprache. Die philosophische Tragweite der Sprechakttheorie, in: Ders., Paradigmen der Ersten Philosophie, Berlin: Suhrkamp, 92–137.

Bateman, J.A. (2021): What are digital media? in: *Discourse, Context & Media*, 41. [https://doi.org/10.1016/j.dcm.2021.100502].

Bayer, J.B.; Trieu, P.; Ellison, N.B. (2020): Social Media Elements, Ecologies, and Effects, in: *Annual Review of Psychology*, 71, 471–497.

Berdini, F. (2013): Speech acts and normativity. A plea for inferentialism, in: *Esercizi Filosofici*, 8(2), 71–88.

Brandom, R.B. (2000): Articulating reasons. An introduction to inferentialism, Cambridge (MA): Harvard University Press.

Burkhardt, A. (1990) (Hg.): Speech Acts, Meaning and Intentions. Critical Approaches to the Philosophy of John R. Searle, Berlin: De Gruyter.

Caliandro, A. (2017): Digital Methods for Ethnography. Analytical Concepts for Ethnographers Exploring Social Media, in: *Journal of Contemporary Ethnography*, 47(5), 551–557.

Cavell, M. (2006): Becoming a Subject, Oxford: Clarendon Press.

Degen, J. (2023): The Rational Speech Act Framework, in: *The Annual Review of Linguistics*, 9(5), 519–540.

Del Vicario, M.; Bessi, A.: Zollo, F.; Petroni, F.; Scala, A.; Caldarelli, G. et al. (2016): The spreading of misinformation online, in: *Proceedings of the National Academy of Science U.S.A.*, 113(3), 554–559.

Drury, N.; Tudor, K. (2024): Radical enactivism. A guide for the perplexed, in: *Journal of Theoretical and Philosophical Psychology*, 44(1), 1–16.

Evans, J.B.T.; Stanovich, K.E. (2013): Dual-Process Theories of Higher Cognition. Advancing the Debate, in: *Perspectives on Psychological Science*, 8(3), 223–241.

Freud, S. (1941): Abriss der Psychoanalyse, in: Ders., Gesammelte Werke, Bd. 17, Frankfurt a.M.: Fischer Verlag, 63–138.

Giampieri-Deutsch, P. (2012): Bewusste Gründe, nicht-bewusste Gründe, in: Nida-Rümelin, J.; Özmen, E. (Hg.), Welt der Gründe [*Deutsches Jahrbuch Philosophie*, 4], Hamburg: Felix Meiner, 406–416.

Giesecke, M. (2002): Von den Mythen der Buchkultur zu den Visionen der Informationsgesellschaft. Trendforschungen zur kulturellen Medienökologie, Frankfurt a.M.: Suhrkamp.

Goodman, D.M.; Clemente, M. (Hg.) (2024): The Routledge International Handbook of Psychoanalysis, Subjectivity, and Technology, New York: Routledge.

Grabska, K.; Mauss-Hanke, A.; Palußek, U.; Stakelbeck F. (Hg.) (2023): Virtuelle Berührung zersplitternde Realität. Zur Psychoanalyse von Digitalisierung und Internetkultur, Gießen: Psychosozial-Verlag.

Grice, P. (1989): Studies in the Way of Words, Cambridge (MA): Harvard University Press.

Günthner, S. (2000): Zwischen direkter und indirekter Rede, in: *Zeitschrift für germanistische Linguistik*, 28(1), 1–22.

Habermas, J. (2022): Überlegungen und Hypothesen zu einem erneuten Strukturwandel der politischen Öffentlichkeit, in: Ders., Ein neuer Strukturwandel der Öffentlichkeit und die deliberative Politik, Berlin: Suhrkamp, 9–68.

Holt, R. (1989a): The development of the primary process. A structural view, in: Ders., Freud reappraised. A fresh look at psychoanalytic theory, New York: Guilford Press, 253–279.

Holt, R. (1989b): The present status of Freud's theory of the primary process, in: Ders., Freud reappraised. A fresh look at psychoanalytic theory, New York: Guilford Press, 280–301.

Holt, R. (2009): Primary process thinking. Theory, measurement, and research, Lanham (MD): Jason Aronson.

Hutto, D.; Kirchoff, M.E.; Muyn, E. (2014): Extensive enactivism. Why keep it all in?, in: *Frontiers in Human Neuroscience*, 8, Art. 706.

Janich, P. (2006): Was ist Information? Kritik einer Legende, Berlin: Suhrkamp.

Johannsen, J. (2018): Psychoanalysis and Digital Culture. Audiences, Social Media, and Big Data, New York: Routledge.

Johanssen, J.; Krueger, S. (2022): Media and Psychoanalysis. A Critical Introduction, London: Karnac Books.

Kettner, M. (2006): 1000 Fragen zur Bioethik. Zur Organisation von Mitverantwortung für Biopolitik in der deliberativen Demokratie, in: Heidbrink, L.; Hirsch, A. (Hg.), Verantwortung in der Zivilgesellschaft. Zur Konjunktur eines widersprüchlichen Prinzips, Frankfurt a.M.: Campus Verlag, 189–217.

Kettner, M. (2012): Gute Gründe für und in Konzeptionen ökonomischer Rationalität, in: Nida-Rümelin, J.; Özmen, E. (Hg.), Welt der Gründe [*Deutsches Jahrbuch Philosophie*, 4], Hamburg: Felix Meiner, 231–245.

Kettner, M. (2016): Der Raum der Gründe und die Kommunikationsgemeinschaft der Begründer, in: Quante, M. (Hg.), *Deutsches Jahrbuch für Philosophie*, 8, Hamburg: Felix Meiner, 637–655.

Kirwan, G.; Connolly, I.; Barton, H.; Palmer, M. (2024): An Introduction to Cyberpsychology, London: Routledge.

Löchl, E. (2023): Subjekt und Medium in der digitalen Welt. Psychoanalytische Erkenntnismöglichkeiten und -grenzen, in: Grabska, K.; Mauss-Hanke, A.; Palußek, U.; Stakelbeck F. (Hg.), Virtuelle Berührung zersplitternde Realität. Zur Psychoanalyse von Digitalisierung und Internetkultur, Gießen: Psychosozial-Verlag, 39–64.

Leichsenring, F. (2022): Primär- und Sekundärprozess, in: Mertens, W. (Hg.), Handbuch psychoanalytischer Grundbegriffe, Stuttgart: Kohlhammer, 5. Aufl., 726–729.

McLuhan, M. (1994): Understanding Media. The Extensions of Man, Cambridge (MA): The MIT Press.

Molina Molina, L. (2017): El cuerpo como apriori del conocimiento científico y el giro hacia la facticidad de la Física contemporánea. Un diálogo de K.-O. Apel con y contra M. Heidegger, in: *Daimon. Revista Internacional de Filosofia*, 457–466. [doi.org/10.6018/daimon/268591].

Ronzhyn, A.; Cardenal, A.S.; Batlle Rubio, A. (2023): Defining affordances in social media research. A literature review, in: *New Media & Society*, 25(11), 3165–3188.

Sasahara, K.; Chen, W.; Peng, H. et al. (2021): Social influence and unfollowing accelerate the emergence of echo chambers, in: *Journal of Computational Social Science*, 4, 381–402.

Schwarzenegger, C.; Koenen, E.; Pentzold, C.; Birkner, T.; Katzenbach, C. (Hg.) (2022): Digitale Kommunikation und Kommunikationsgeschichte. Perspektiven, Potentiale, Problemfelder, Berlin: SSOAR. [https://www.ssoar.info/ssoar/handle/document/79536].

Schmidt, J.H.; Taddicken, M. (2022): Handbuch Soziale Medien, Wiesbaden: Springer VS.

Schönhagen, P.; Meißner, M. (2021): Kommunikations- und Mediengeschichte. Von Versammlungen bis zu den digitalen Medien, Köln: Herbert von Halem Verlag.

Scolari, C.A. (2012): Media Ecology. Exploring the Metaphor to Expand the Theory, in: *Communication Theory*, 22(2), 204–225.

Shahbazi, M.; Bunker, D.; Sorrell, T.C. (2023): Communicating shared situational awareness in times of chaos. Social media and the COVID-19 pandemic, in: *Journal of the Association for Information Science and Technology (JASIST)*, 74(10), 1185–1202.

Stieglitz, S.; Dang-Xuan, L. (2014): Emotions and Information Diffusion in Social Media. Sentiment of Microblogs and Sharing Behavior, in: *Journal of Management Information Systems*, 29(4), 217–248.

Stokes, A. (1974). Primary Process, Thinking and Art, in: *Contemporary Psychoanalysis*, 10, 327–342.

Sun, W.J.; Liu, X.F. (2023): Deep attention framework for retweet prediction enriched with causal inferences, in: *Applied Intelligence*, 53, 24293–24313.

Volodina, A. (2023): Warum muss man bei indirekter Rede den Konjunktiv benutzen?, in: IDS Grammis, [https://grammis.ids-mannheim.de/fragen/6829].

Zyoud, S.H.; Sweileh, W.M.; Awang, R. et al. (2018): Global trends in research related to social media in psychology. Mapping and bibliometric analysis, in: *International Journal of Mental Health Systems*, 12, Art. 4. [https://doi.org/10.1186/s13033-018-0182-6].

Philosophie der Künstlichen Intelligenz

Ein strukturierter Überblick

Vincent C. Müller

Abstract: *This paper presents the main topics, arguments, and positions in the philosophy of AI at present (excluding ethics). Apart from the basic concepts of* intelligence *and* computation*, the main topics of artificial cognition are* perception, action, meaning, rational choice, free will, consciousness, *and* normativity. *Through a better understanding of these topics, the philosophy of AI contributes to our understanding of the nature, prospects, and value of AI. Furthermore, these topics can be understood more deeply through the discussion of AI; so we suggest that »AI Philosophy« provides a new method for philosophy.*

Keywords: *AI philosophy; philosophy of AI; cognition; artificial intelligence; meaning*

1. Thema und Methode

1.1 Künstliche Intelligenz

Der Begriff *Künstliche Intelligenz* wurde nach dem »Dartmouth Summer Research Project on Artificial Intelligence« von 1956 populär, dessen Ziele wie folgt formuliert wurden:

> »Die Studie geht von der Vermutung aus, dass jeder Aspekt des Lernens oder jedes andere Merkmal der Intelligenz im Prinzip so genau beschrieben werden kann, dass eine Maschine in der Lage ist, ihn zu simulieren.« (McCarthy et al. 1955: 1)[1]

Dies ist das ehrgeizige Forschungsprogramm, das davon ausgeht, dass menschliche Intelligenz oder Kognition als regelbasierte Berechnung über eine symbolische Repräsentation verstanden oder modelliert werden kann, so dass diese Modelle getes-

1 Alle in diesem Beitrag vorkommenden Übersetzungen zitierter fremdsprachiger Literatur stammen von Matthias Kettner. [Anm. der Hrsg.]

tet werden können, indem sie auf verschiedenen (künstlichen) Computern ausgeführt werden. Im Erfolgsfall würden die Computer, auf denen diese Modelle laufen, künstliche Intelligenz aufweisen. KI und Kognitionswissenschaft sind zwei Seiten derselben Medaille. Dieses Programm wird gewöhnlich als *klassische KI* bezeichnet:[2]

a) KI ist ein Forschungsprogramm zur Entwicklung intelligenter computerbasierter Agenten.

Die von John Searle eingeführten Begriffe *Starke KI* und *Schwache KI* stehen in der gleichen Tradition. *Starke* KI bezieht sich auf die Idee, dass »der entsprechend programmierte Computer wirklich ein Geist ist, in dem Sinne, dass von Computern, die die richtigen Programme erhalten, buchstäblich gesagt werden kann, dass sie verstehen und andere kognitive Zustände haben.« *Schwache* KI bedeutet, dass KI lediglich mentale Zustände simuliert. In diesem schwachen Sinne »besteht der Hauptwert des Computers für die Erforschung des Geistes darin, dass er uns ein sehr mächtiges Werkzeug an die Hand gibt.« (Searle 1980: 353).

Andererseits wird der Begriff »KI« in der Informatik häufig in einem Sinne verwendet, den ich als *technische KI* bezeichnen möchte:

b) KI ist eine Sammlung von Informatikmethoden für Wahrnehmung, Modellierung, Planung und Handlung (Suche, logische Programmierung, probabilistisches Schließen, Expertensysteme, Optimierung, Steuerungstechnik, neuromorphes Engineering, maschinelles Lernen usw.). (Görz et al. 2020; Pearl/Mackenzie 2018; Russell 2019; Russell/Norvig 2020).

Es gibt auch eine Minderheit in der KI, die dafür plädiert, dass sich die Disziplin auf die Ziele von a) konzentriert, während die derzeitige Methodik unter b) beibehalten wird, meist unter dem Namen *Artificial General Intelligence* (AGI).[3]

Das Vorhandensein der beiden Traditionen (klassisch und technisch) führt gelegentlich zu Vorschlägen, dass wir den Begriff »KI« nicht verwenden sollten, weil er starke Behauptungen impliziert, die aus dem Forschungsprogramm a) stammen, aber sehr wenig mit der eigentlichen Arbeit unter b) zu tun haben. Vielleicht sollten wir lieber von »maschinellem Lernen« oder »entscheidungsunterstützenden Maschinen« oder einfach von »Automatisierung« sprechen (wie im Lighthill-Bericht von 1973 vorgeschlagen: Lighthill 1973). Im Folgenden werden wir den Begriff der »Intelligenz« klären, und es wird sich zeigen, dass es ein einigermaßen kohärentes Forschungsprogramm der KI gibt, das die beiden Traditionen vereint: Die *Erzeugung intelligenten Verhaltens durch Rechenmaschinen*.

2 Ein Beispiel dafür: Dietrich 2002. Der klassische historische Überblick ist Boden 2006.

3 Die AGI-Konferenzen werden seit 2008 organisiert.

Diese beiden Traditionen bedürfen nun einer Fußnote: Beide wurden weitgehend unter dem Begriff der *klassischen KI* entwickelt, was hat sich also mit dem Übergang zum *maschinellen Lernen* (ML) geändert? ML ist eine traditionelle (konnektivistische) Rechenmethode in neuronalen Netzen, die keine Repräsentationen verwendet. (Rosenblatt 1957; Buckner (forthcoming); Garson/Buckner 2019; LeCun et al. 2015) Seit ca. 2015, mit dem Aufkommen von massiver Rechenleistung und massiven Daten für tiefe neuronale Netze, hat sich die Leistung von ML-Systemen in Bereichen wie Übersetzung, Textproduktion, Spracherkennung, Spiele, visuelle Erkennung und autonomes Fahren dramatisch verbessert, so dass sie in einigen Fällen dem Menschen überlegen ist. ML ist jetzt die Standardmethode in der KI. Was bedeutet dieser Wandel für die Zukunft der Disziplin? Die ehrliche Antwort lautet: Wir wissen es noch nicht. Wie jede Methode hat auch ML ihre Grenzen, aber diese Grenzen sind weniger restriktiv, als man viele Jahre lang dachte, denn die Systeme zeigen eine nichtlineare Verbesserung – mit mehr Daten können sie sich plötzlich deutlich verbessern. Ihre Schwächen (z. B. Überanpassung, kausale Schlussfolgerungen, Zuverlässigkeit, Relevanz, Blackbox) können denen der menschlichen rationalen Entscheidung recht nahe kommen, insbesondere wenn »prädiktive Verarbeitung« die richtige Theorie des menschlichen Geistes ist (s. unten Abschnitte 4.1, 6).

1.2 »Philosophie der KI« und Philosophie

Eine Möglichkeit, die Philosophie der KI zu verstehen, ist, dass sie sich hauptsächlich mit drei kantischen Fragen beschäftigt: Was ist KI? Was kann KI tun? Was sollte KI sein? Ein wichtiger Teil der Philosophie der KI ist die *Ethik* der KI, aber wir werden diesen Bereich hier nicht diskutieren.[4]

Traditionell befasst sich die Philosophie der KI mit einigen ausgewählten Punkten, bei denen Philosophen etwas über KI zu sagen haben, z. B. zu der These, dass Kognition eigentlich Berechnung ist oder dass Computer sinnvolle Symbole als solche verarbeiten können.[5] Eine Überprüfung dieser Punkte und der entsprechenden Autoren (Turing, Wiener, Dreyfus, Dennett, Searle, u.a.) würde zu einer fragmentierten Diskussion führen, die nie ein Bild des Gesamtprojekts ergibt. Es wäre so, als würde man eine Menschheitsgeschichte im alten Stil anhand einiger weniger »Helden« schreiben. Außerdem wird in dieser Sichtweise die Philosophie der KI von ih-

4 Im vorliegenden Band sind die Texte in der Rubrik »Verantwortungsverhältnisse« dafür einschlägig.

5 Es gibt nur sehr wenige Überblicksartikel und keine aktuellen. Siehe Carter 2007; Copeland 1993; Dietrich 2002; Floridi 2003; Floridi 2011. Einiges von dem, was Philosophen zu sagen hatten, kann als Unterminierung des Projekts der KI angesehen werden, vgl. Dietrich et al. 2021.

rer Cousine, der Philosophie der Kognitionswissenschaft, getrennt, die wiederum eng mit der Philosophie des Geistes verbunden ist (Margolis et al. 2012).

Im Folgenden versuche ich einen anderen Weg: Wir betrachten die *Komponenten eines intelligenten Systems*, wie sie sich in der Philosophie, der Kognitionswissenschaft und der KI darstellen. Eine Möglichkeit, solche Komponenten zu betrachten, ist, dass es relativ einfache Tiere gibt, die relativ einfache Dinge tun können, und dann können wir uns zu komplizierteren Tieren »hocharbeiten«, die außer diesen einfachen Dingen noch mehr tun können. Ein schematisches Beispiel: Eine *Fliege* wird immer wieder gegen das Glas stoßen, um zum Licht zu gelangen; eine *Kobra* wird verstehen, dass sich hier ein Hindernis befindet, und versuchen, es zu umgehen; eine *Katze* wird sich vielleicht daran erinnern, dass sich hier beim letzten Mal ein Hindernis befand, und sofort einen anderen Weg einschlagen; ein *Schimpanse* wird vielleicht erkennen, dass das Glas mit einem Stein zerbrochen werden kann; ein *Mensch* wird vielleicht den Schlüssel finden und die Glastür aufschließen … oder aber das Fenster nehmen, um hinauszukommen. Um sich mit der Philosophie der künstlichen Intelligenz zu befassen, brauchen wir also ein breites Spektrum an Philosophie: Philosophie des Geistes, Erkenntnistheorie, Sprache, Werte, Kultur, Gesellschaft, u.a.m.

Außerdem ist die Philosophie der KI in unserem Ansatz nicht nur »angewandte Philosophie«: Es geht nicht darum, dass wir eine Lösung in der Werkzeugkiste des Philosophen bereithalten und sie »anwenden«, um Probleme der KI zu lösen. Das philosophische Verständnis selbst ändert sich, wenn man den Fall der KI betrachtet: Es wird weniger anthropozentrisch, weniger auf unseren eigenen menschlichen Fall konzentriert. Ein tieferer Blick auf Konzepte muss normativ von der Funktion geleitet werden, die diese Konzepte erfüllen, und diese Funktion kann besser verstanden werden, wenn wir sowohl die natürlichen Fälle als auch den Fall der aktuellen und möglichen KI betrachten. Dieses Papier ist somit auch ein »proof of concept« für die Philosophie durch die begriffliche Analyse von KI: Ich nenne dies KI-Philosophie.

Ich schlage also vor, die Frage vom Kopf auf die Füße zu stellen, wie Marx gesagt hätte: Wenn wir KI verstehen wollen, müssen wir uns selbst verstehen; und wenn wir uns selbst verstehen wollen, müssen wir auch KI verstehen!

2. Intelligenz

2.1 Der Turing-Test

»Ich schlage vor, die Frage ›Können Maschinen denken?‹ zu untersuchen« schrieb Alan Turing zu Beginn seines Aufsatzes in der führenden philosophischen Zeitschrift *Mind* (Turing 1950). Das war 1950, Turing war einer der Gründerväter des Computers, und viele Leser des Aufsatzes werden damals noch nicht einmal von

solchen Maschinen gehört haben, denn es gab nur ein halbes Dutzend Universalcomputer auf der Welt (Z3, Z4, ENIAC, SSEM, Harvard Mark III, Manchester Mark I) (s. Anonym 1950). Turing erklärt kurzerhand, dass die Suche nach einer Definition des Begriffs »Denken« sinnlos sei, und schlägt vor, seine ursprüngliche Frage durch die Frage zu ersetzen, ob eine Maschine erfolgreich ein »Nachahmungsspiel« spielen könne. Dieses Spiel ist unter dem Namen »Turing-Test« bekannt geworden: Ein menschlicher Befrager wird über »Teleprinting« mit einem anderen Menschen und einer Maschine verbunden, und wenn der Befrager die Maschine nicht von dem Menschen unterscheiden kann, indem er ein Gespräch führt, dann sagen wir, dass die Maschine »denkt«. Am Ende des Aufsatzes kommt Turing auf die Frage zurück, ob Maschinen denken können, und sagt: »Ich glaube, dass sich am Ende des Jahrhunderts der Wortgebrauch und die allgemeine gebildete Meinung so sehr verändert haben werden, dass man von denkenden Maschinen sprechen kann, ohne mit Widerspruch zu rechnen.« (Turing 1950: 442) Turing schlägt also vor, unseren alltäglichen Begriff des »Denkens« durch einen operativ definierten Begriff zu ersetzen, einen Begriff, den wir mit einem Verfahren testen können, das ein messbares Ergebnis hat.

Turings Vorschlag, die Definition des Denkens durch eine operative Definition zu ersetzen, die sich ausschließlich auf das Verhalten stützt, passt in das intellektuelle Klima der damaligen Zeit, in der der Behaviorismus noch eine dominierende Kraft war: In der Psychologie ist der Behaviorismus ein *methodologischer* Vorschlag, der besagt, dass die Psychologie zu einer echten wissenschaftlichen Disziplin werden sollte, indem sie sich auf überprüfbare Beobachtungen und Experimente stützt, anstatt auf subjektive Selbstbeobachtung. Angesichts der Tatsache, dass der Geist anderer eine »Black Box« ist, sollte die Psychologie zur Wissenschaft von Reiz und Verhaltensreaktion, von Input-Output-Beziehungen werden. Die frühe analytische Philosophie führte zu einem *reduktionistischen Behaviorismus*. Wenn die Bedeutung eines Begriffs seine »Überprüfungsbedingungen« sind, dann *bedeutet* ein mentaler Begriff wie »Schmerz« lediglich, *dass* die Person zu einem bestimmten Verhalten bereit ist.

Ist der Turing-Test über beobachtbares Verhalten eine nützliche Definition von Intelligenz? Kann er unsere Rede von Intelligenz »ersetzen«? Es ist klar, dass es intelligente Wesen geben wird, die diesen Test nicht bestehen, zum Beispiel Menschen oder Tiere, die nicht tippen können. Man kann also mit Fug und Recht behaupten, dass Turing das Bestehen des Tests nur als hinreichende Voraussetzung für Intelligenz ansah, nicht als notwendige Voraussetzung. Wenn also ein System diesen Test besteht, muss es dann intelligent sein? Das hängt davon ab, ob Sie glauben, dass Intelligenz nur intelligentes Verhalten ist, oder ob Sie glauben, dass wir für die Zuschreibung von Intelligenz auch die interne Struktur betrachten müssen.

2.2 Was ist Intelligenz?

Intuitiv betrachtet ist Intelligenz eine Fähigkeit, die intelligentem Handeln zugrunde liegt. Welches Handeln intelligent ist, hängt von den Zielen ab, die verfolgt werden, und vom Erfolg beim Erreichen dieser Ziele – denken Sie an die oben erwähnten Beispiele tierischen Verhaltens. Der Erfolg hängt nicht nur vom Agenten ab, sondern auch von den Bedingungen, unter denen er agiert, so dass ein System mit weniger Möglichkeiten, ein Ziel zu erreichen (z. B. Nahrung zu finden), weniger intelligent ist. In diesem Sinne lautet eine klassische Definition: »Intelligenz misst die Fähigkeit eines Agenten, Ziele in einem breiten Spektrum von Umgebungen zu erreichen.« (Legg/Hutter 2007: 402) Hier ist Intelligenz die *Fähigkeit, flexibel Ziele zu verfolgen*, wobei Flexibilität mit Hilfe unterschiedlicher Umgebungen erklärt wird. Dieser Intelligenzbegriff aus der KI ist ein *instrumenteller* (bezogen auf Zielerreichung) und normativer Begriff von Intelligenz, in der Tradition der klassischen Entscheidungstheorie, die besagt, dass ein rationaler Agent immer versuchen *sollte*, den erwarteten Nutzen zu maximieren (siehe Abschnitt 6).[6]

Wenn die KI-Philosophie Intelligenz als relativ zu einer Umgebung versteht, dann kann man, um mehr Intelligenz zu erreichen, entweder den Akteur oder die Umgebung verändern. Der Mensch hat beides in großem Umfang durch das getan, was als »Kultur« bekannt ist: Wir haben nicht nur ein ausgeklügeltes Lernsystem für Menschen geschaffen (um den Agenten zu verändern), sondern auch die Welt physisch so gestaltet, dass wir unsere Ziele in ihr verfolgen können; um zu reisen, haben wir z. B. Straßen, Autos mit Lenkrädern, Karten, Straßenschilder, digitale Routenplanung und KI-Systeme geschaffen. Das Gleiche tun wir jetzt für KI-Systeme, sowohl für das lernende System als auch für die Veränderung der Umgebung (Autos mit Computerschnittstellen, GPS usw.). Indem wir die Umwelt verändern, werden wir auch unsere Wahrnehmung und unser Leben verändern – vielleicht auf eine Art und Weise, die sich zu unserem Nachteil auswirkt.

In den Abschnitten 4-9 werden wir uns mit den wichtigsten Komponenten eines intelligenten Systems befassen, doch zuvor werden wir den Mechanismus der KI erörtern: die Berechnung.

6 Z.B. Simon 1955; Thoma 2019. Siehe auch den neo-behavioristischen Vorschlag in Coelho Mollo 2022.

3 Berechnung (»Computation«)

3.1 Der Begriff des Rechnens

Die Maschinen, auf denen KI-Systeme laufen, sind »Computer« oder »Rechner«, so dass es für unsere Aufgabe wichtig sein wird, herauszufinden, was ein Computer ist und was er prinzipiell tun kann. Eine damit zusammenhängende Frage ist, ob die menschliche Intelligenz vollständig oder teilweise auf Berechnungen zurückzuführen ist. Wenn sie vollständig auf Berechnungen zurückzuführen ist, wie die klassische KI angenommen hatte, dann scheint es möglich zu sein, diese Berechnungen auf einem künstlichen Computer nachzubilden.

Um zu verstehen, was ein Computer ist, ist es nützlich, sich die Geschichte der Rechenmaschinen in Erinnerung zu rufen – ich sage »Maschinen«, denn vor ca. 1945 war das Wort »Computer« oder »Rechner« eine Bezeichnung für einen Menschen, der einen bestimmten Beruf hat, für jemanden, der Berechnungen durchführt. Diese Berechnungen, z.B. die Multiplikation zweier großer Zahlen, werden durch ein mechanisches Schritt-für-Schritt-Verfahren durchgeführt, das, wenn es einmal vollständig ausgeführt ist, zu einem Ergebnis führt. Solche Verfahren werden »Algorithmen« genannt. 1936 schlug Alan Turing als Antwort auf Gödels »Entscheidungsproblem« vor, dass der Begriff »etwas berechnen« dadurch erklärt werden könnte, »was eine bestimmte Art von Maschine tun kann« (genau wie er vorschlug, den Begriff der Intelligenz im »Turing-Test« zu operationalisieren). Turing skizzierte, wie eine solche Maschine aussehen würde, mit einem unendlich langen Band als Speicher und einem Kopf, der Symbole von diesem Band lesen und darauf schreiben kann. Diese Zustände auf dem Band sind immer spezifische diskrete Zustände, so dass jeder Zustand von einem Typ aus einer endlichen Liste ist (Symbole, Zahlen, u.a.), also zum Beispiel entweder der Buchstabe »V« oder der Buchstabe »C«, nicht etwa ein bisschen von jedem. Mit anderen Worten, die Maschine ist »digital« (nicht analog).[7] Etwas Entscheidendes kommt hinzu: In der »universellen« Version der Maschine kann man das, was der Computer tut, durch weitere Eingaben *verändern*. Mit anderen Worten: Die Maschine *kann* so *programmiert werden*, dass sie einen bestimmten Algorithmus ausführt, und sie speichert dieses Programm in ihrem Speicher.[8] Ein solcher Computer ist ein Universalcomputer, d.h. er kann jeden beliebigen Algorithmus berechnen. Es sollte erwähnt werden, dass auch weiter gefasste Begriffe der Berechnung vorgeschlagen wurden, z.B. analoges Rechnen und Hypercomputing (Piccinini 2021; Shagrir 2022; Siegelmann 1995; Siegelmann 1997).

7 Negroponte 1995. Siehe auch Haugeland 1985: 57; Müller 2013.

8 Gödel 1931; Turing 1936. Das ursprüngliche Programm ist skizziert in Hilbert 1900. Siehe z.B. Copeland et al. 2013.

Es stellt sich auch die Frage, ob das Rechnen eine reale Eigenschaft physikalischer Systeme ist, oder eher nur eine nützliche Art, diese Systeme zu beschreiben. Searle hat gesagt: »Die elektrischen Zustandsübergänge sind der Maschine immanent, aber die Berechnung liegt im Auge des Betrachters.« (Dodig-Crnkovic/Müller 2011; Searle 2004: 64) Wenn wir eine antirealistische Sichtweise der Berechnung annehmen, dann ändert sich die Situation radikal.

Genau dieselbe Berechnung kann auf verschiedenen physischen Computern durchgeführt werden und eine unterschiedliche Semantik haben. Es gibt also drei Beschreibungsebenen, die für einen bestimmten Computer besonders relevant sind: (a) die *physische Ebene* der tatsächlichen »Realisierung« des Computers, (b) die *syntaktische Ebene* des berechneten Algorithmus und (c) die *symbolische Ebene* des Inhalts, dessen, was berechnet wird.

Physikalisch gesehen kann eine Rechenmaschine aus allem gebaut werden und jede Eigenschaft der physikalischen Welt nutzen (Zahnräder, Relais, DNA, Quantenzustände usw.). Dies kann als Verwendung eines physikalischen Systems zur Kodierung eines formalen Systems angesehen werden (Horsman et al. 2014). Tatsächlich wurden alle Universalcomputer mittels großer Mengen von Schaltern gebaut. Ein Schalter hat zwei Zustände (offen/geschlossen), also arbeiten die darauf basierenden Computer mit zwei Zuständen (ein/aus, 0/1), sie sind *binär* – dies ist eine Designentscheidung. Binäre Schalter können leicht zu »Logikgattern« kombiniert werden, die auf Eingaben in Form der logischen Verknüpfungen in der booleschen Logik (die ebenfalls zweiwertig ist) reagieren: NOT, AND, OR, usw. Wenn sich solche Schalter in einem Zustand befinden, der *syntaktisch* als 1010110 verstanden werden kann, dann könnte dies *semantisch* (nach den derzeitigen ASCII/ANSI-Konventionen) den Buchstaben »V«, die Zahl »86«, einen hellgrauen Farbton, einen grünen Farbton usw. darstellen.

3.2 »Computationalismus«

Wie wir gesehen haben, ist die Vorstellung, dass im Berechnen die Ursache für die Intelligenz natürlicher Systeme, z.B. des Menschen, zu finden ist und zur Modellierung und Reproduktion dieser Intelligenz verwendet werden kann, eine Grundannahme der klassischen KI. Diese Auffassung ist häufig mit der Ansicht gekoppelt (und durch sie motiviert), dass menschliche mentale Zustände funktionale Zustände sind und dass diese funktionalen Zustände die eines Computers sind: »Maschinenfunktionalismus«. Diese These wird in den Kognitions- und Neurowissenschaften oft als Selbstverständlichkeit vorausgesetzt, ist aber in den letzten Jahrzehnten auch erheblich kritisiert worden.[9] Die Hauptquellen für diese Ansicht sind

9 Edelman 2008; Miłkowski 2018. Zur Diskussion: Harnad 1990; Scheutz 2002; Shagrir 1997; Varela et al. 1991.

die Begeisterung für die universelle Technologie des digitalen Rechnens sowie frühe neurowissenschaftliche Befunde, die darauf hindeuten, dass menschliche Neuronen (im Gehirn und im Körper) ebenfalls in gewisser Weise binär sind, d.h. entweder senden sie ein Signal an andere Neuronen, sie »feuern«, oder sie tun es nicht. Einige Autoren verteidigen die *Physikalische Symbolsystemhypothese*, d.h. den Computationalismus, sowie die Behauptung, dass nur Computer intelligent sein können (vgl. Boden 2006: 1419ff.; Newell/Simon 1976: 116).

4. Wahrnehmung und Handlung

4.1 Passive Wahrnehmung

Es mag überraschen, dass die Überschrift dieses Kapitels Wahrnehmung und Handlung verbindet. Aus der KI und der Kognitionswissenschaft können wir aber lernen, dass die Hauptfunktion der Wahrnehmung darin besteht, Handeln zu ermöglichen; ja, dass die Wahrnehmung eine Art von Handeln *ist*. Das traditionelle Verständnis von Wahrnehmung in der Philosophie ist die *passive* Wahrnehmung, bei der wir uns selbst beobachten, wie wir die Welt beobachten, und zwar in dem, was Dan Dennett das *kartesische Theater* genannt hat: Es ist, als ob ein kleiner Mensch in meinem Kopf säße, der die Außenwelt durch unsere Ohren hört und durch unsere Augen beobachtet (Dennett 1991: 107). Diese Vorstellung ist letztlich absurd, vor allem weil sie voraussetzen würde, dass noch ein weiterer kleiner Mensch im Kopf dieses kleinen Menschen sitzt. Und doch wird in der philosophischen Literatur ein Großteil der Diskussion über die menschliche Wahrnehmung so behandelt, als wäre sie etwas, das in meinem Kopf passiert.

Da ist zum Beispiel das 2D-3D-Problem beim Sehen; das Problem, wie ich die visuelle Erfahrung einer dreidimensionalen Welt durch ein zweidimensionales Wahrnehmungssystem erzeugen kann (die Netzhaut ist eine zweidimensionale Schicht, die unsere Augäpfel von innen bedeckt). Es muss doch einen Weg geben, die visuellen Informationen in der Netzhaut, dem Sehnerv und den optischen Verarbeitungszentren des Gehirns zu verarbeiten, um diese dreidimensionale Erfahrung zu erzeugen. Aber so geht es nicht wirklich zu.[10]

4.2 Aktive Wahrnehmung

Tatsächlich entsteht der dreidimensionale Eindruck durch eine Interaktion zwischen mir und der Welt (im Falle des Sehens durch die Bewegung meiner Augen und meines Körpers). Es ist besser, die Wahrnehmung in Anlehnung an den Tastsinn zu

10 Für eine Einführung in die Vision siehe O'Regan 2011: Kap. 1–5.

betrachten: Berühren ist etwas, das ich *tue*, um die Weichheit eines Gegenstandes, die Beschaffenheit seiner Oberfläche, seine Temperatur, sein Gewicht, seine Biegsamkeit usw. zu erfahren. Ich tue dies, indem ich handle und dann die Veränderung des sensorischen Inputs wahrnehme. Das nennt man eine Wahrnehmungs-Handlungs-Schleife: Ich tue etwas, das die Welt verändert, und verändere damit die Wahrnehmung, die ich habe.

Es wird nützlich sein zu betonen, dass dies auch bei der Wahrnehmung meines eigenen Körpers geschieht. Ich weiß nur deshalb, dass ich eine Hand habe, weil meine visuelle Wahrnehmung der Hand, die Propriozeption und der Tastsinn übereinstimmen. Wenn das nicht der Fall ist, ist es ziemlich einfach, mir das Gefühl zu geben, dass z.B. eine Gummihand meine eigene Hand ist – dies ist als die »Gummihand-Illusion« bekannt. Wenn eine Handprothese in geeigneter Weise mit dem Nervensystem eines Menschen verbunden ist, kann die Wahrnehmungs- und Handlungsschleife wieder geschlossen werden, und der Mensch wird sie als seine eigene Hand empfinden.

4.3 Prädiktive Verarbeitung und Verkörperung

Diese Sichtweise der Wahrnehmung hat kürzlich zu einer Theorie des »prädiktiven/vorhersagenden Gehirns« (predictive brain) geführt: Das Gehirn wartet nicht passiv auf Eingaben, sondern ist *immer aktiv an* der Handlungs-Wahrnehmungs-Schleife beteiligt. Es erstellt *Vorhersagen* darüber, wie der sensorische Input in Anbetracht meiner Handlungen sein wird, und gleicht diese Vorhersagen dann mit dem tatsächlichen sensorischen Input ab. Der Unterschied zwischen den beiden ist etwas, das wir zu minimieren versuchen, was als »Prinzip der freien Energie« bezeichnet wird (Clark 2013; Clark 2016; Friston 2010).

In dieser Tradition ist die Wahrnehmung eines natürlichen Agenten oder auch eines KI-Systems etwas, das eng mit der physischen Interaktion des Körpers des Agenten mit der Umwelt verbunden ist; die Wahrnehmung ist somit eine Komponente der verkörperten Kognition. Ein nützlicher Slogan in diesem Zusammenhang ist »4E-Kognition«, der besagt, dass Kognition *verkörpert* ist; sie ist in eine Umgebung mit anderen Agenten *eingebettet*; sie ist eher *enaktiv* als passiv; und sie ist *ausgedehnt* (»extended«), d.h. sie findet nicht nur im Kopf statt (Clark/Chalmers 1998; Clark 2003; Newen et al. 2018). Ein Aspekt, der eng mit der 4E-Kognition zusammenhängt, ist die Frage, ob Kognition beim Menschen grundsätzlich repräsentational ist und ob Kognition in der KI repräsentational sein muss (siehe Abschnitt 5).

Verkörperte Kognition wird manchmal als empirische These über die tatsächliche Kognition (insbesondere beim Menschen) oder aber als These über die geeignete Gestaltung von KI-Systemen und manchmal auch als Analyse dessen, was Kognition ist und sein muss, dargestellt. In letzterem Verständnis würde eine nicht verkörper-

te KI zwangsläufig bestimmte Merkmale der Kognition vermissen lassen (Dreyfus 1972; Pfeifer/Bongard 2007).

5. Bedeutung und Repräsentation

5.1 Das Argument des Chinesischen Zimmers

Wie wir oben gesehen haben, beruht die klassische KI auf der Annahme, dass der entsprechend programmierte Computer tatsächlich ein Geist *ist* – mit dieser Annahme kennzeichnete John Searle die *starke KI*. In seinem berühmten Aufsatz »Minds, Brains and Programs« stellte Searle das Gedankenexperiment des »Chinesischen Zimmers« vor (Searle 1980). Das Chinesische Zimmer ist ein Computer, der wie folgt aufgebaut ist: Es gibt einen geschlossenen Raum, in dem John Searle sitzt und ein großes Buch in der Hand hält, das ihm ein Computerprogramm mit Algorithmen vorgibt, wie die Eingabe zu verarbeiten und die Ausgabe zu liefern ist. Was er nicht weiß, ist, dass die Eingabe, die er erhält, ein chinesischer Text ist, und dass die Ausgabe, die er liefert, sinnvolle chinesische Antworten oder Kommentare zu dieser sprachlichen Eingabe darstellen. Die Ausgabe, so die Annahme, ist von der eines kompetenten chinesischen Sprechers nicht zu unterscheiden. Und doch versteht Searle in diesem Raum kein Chinesisch und wird mit dem Input, den er erhält, auch nicht Chinesisch lernen. Daraus schließt Searle, dass *Berechnungen für Verstehen nicht ausreichen*. Es kann keine starke KI geben.

In der weiteren Erörterung seines Arguments des Chinesischen Zimmers geht Searle auf zwei erwartbare typische Entgegnungen ein: Die *System-Antwort* akzeptiert zwar, dass Searle gezeigt hat, dass keine einfache Manipulation der Person im Raum diese Person in die Lage versetzen wird, Chinesisch zu verstehen, wendet aber ein, dass die Manipulation von Symbolen doch vielleicht das *umfassendere System*, von dem die Person nur ein Teil ist, in die Lage versetzen wird, Chinesisch zu verstehen. Steckt in Searles Argument also vielleicht ein Fehlschluss vom Teil aufs Ganze? Dieser Einwand wirft allerdings die Frage auf, warum man denken sollte, dass das Gesamtsystem Eigenschaften aufweist, die der algorithmische Prozessor selbst nicht hat.

Eine Möglichkeit, auf diese Herausforderung mit dem Vorschlag einer bestimmten Systemveränderungen zu antworten, nennt Searle die *Roboter-Antwort*. Sie räumt ein, dass das größere System, so wie es beschrieben ist, zwar kein Chinesisch versteht, aber nur weil dem System etwas fehlt, was Chinesisch sprechende Menschen haben, nämlich eine kausale Verbindung zwischen den Worten und der Welt. Wir müssten also Sensoren und Effektoren zu unseren Computer hinzufügen, die für die notwendige kausale Verbindung sorgen würden. Searle entgegnet auf diesen Vorschlag, dass die Eingabe von Sensoren für den Searle im Inneren des

Zimmers »nur noch mehr Chinesisch« wäre; sie würde kein weiteres Verständnis liefern, tatsächlich hätte Searle keine Ahnung, dass die Eingabe von einem Sensor stammt (Cole 2020; Preston/Bishop 2002).

5.2 Rekonstruktion

Ich denke, wir können den Kern des Arguments des Chinesischen Zimmers als eine Erweiterung der folgenden Beobachtung Searles betrachten:

> »Niemand würde annehmen, dass wir Milch und Zucker durch eine Computersimulation der formalen Abläufe bei der Laktation und der Photosynthese erzeugen können, aber wenn es um den Geist geht, sind viele Menschen bereit, an ein solches Wunder zu glauben.« (Searle 1980: 424)

Der Kern des Arguments lässt sich dann so rekonstruieren:

1. Ein System, das nur syntaktische Manipulationen vornimmt, kann keine Bedeutungen erfassen.
2. Ein Computer nimmt nur syntaktische Manipulationen vor.
3. Also kann ein Computer keine Bedeutungen erfassen.

In Searles Terminologie hat ein Computer *nur eine Syntax* und *keine Semantik*; den Symbolen in einem Computer fehlt die Intentionalität (Gerichtetheit), die der menschliche Sprachgebrauch hat. Am Schluss seines Aufsatzes fasst er seine Position zusammen:

> »›Könnte eine Maschine denken?‹ Die Antwort lautet natürlich: Ja. Wir sind genau solche Maschinen. […] Aber könnte etwas denken, verstehen und so weiter, allein kraft dessen, dass es ein Computer mit der richtigen Art von Programm ist? […] die Antwort ist: Nein.« (Searle 1980: 422)

5.3 Berechnungen, Syntax und Kausalkräfte

Wenn man Searles Argument auf diese Weise rekonstruiert, stellt sich die Frage, ob die Prämissen wahr sind. Mehrere Kommentatoren haben argumentiert, dass Prämisse 2 falsch ist, weil man das, was ein Computer tut, als sinnvolle Reaktion auf sein Programm verstehen müsse (McCarthy 2007; Boden 1988: 97; Haugeland 2002: 385). Ich bin der Meinung, dass dies ein Irrtum ist, denn der Computer *folgt* diesen Regeln nicht, er ist lediglich so konstruiert, dass er diesen Regeln *entsprechend handelt*, wenn seine Zustände von einem Beobachter entsprechend interpretiert werden.[11] Abgese-

11 Vgl. schon das Argument bei Wittgenstein 1960[1953]: §§ 82–86, 198, 217 usw.

hen davon hat jeder tatsächliche Computer, jede physische Realisierung eines abstrakten Algorithmus-Prozessors, sehr wohl kausale Kräfte, er kann mehr als bloß syntaktische Manipulationen durchführen. Er kann zum Beispiel das Licht an- oder ausschalten.

Das Argument des Chinesischen Zimmers hat die Aufmerksamkeit in der Sprachphilosophie weg von Konventionen und Logik hin zu den Bedingungen gelenkt, unter denen ein Sprecher das meint, was er sagt (Sprecherbedeutung), oder überhaupt etwas meint (Intentionalität); es hat neue Diskussionen angeregt, insbesondere über die Rolle, die *Repräsentationen* in der Kognition spielen, und über die Rolle des Rechnens mit Repräsentationen (Searle 1984; Searle 2004).

6 Rationale Wahl

6.1 Normative Entscheidungstheorie (MEU)

Ein rationaler Akteur nimmt die Umwelt wahr, findet heraus, welche Handlungsoptionen bestehen, und trifft dann die beste Entscheidung. Genau darum geht es in der Entscheidungstheorie. Sie ist eine normative Theorie darüber, wie ein rationaler Akteur angesichts des ihm zur Verfügung stehenden Wissens handeln *sollte* – und keine deskriptive Theorie darüber, wie rationale Akteure tatsächlich handeln *werden*.

Wie sollte also ein rationaler Akteur entscheiden, welche die bestmögliche Handlung ist? Er bewertet die möglichen Ergebnisse jeder Wahl und wählt dann die beste aus, d.h. diejenige, die den höchsten subjektiven Nutzen hat, d.h. Nutzen aus der Sicht des jeweiligen Akteurs. Man beachte, dass rationale Entscheidungen in diesem Sinne nicht notwendigerweise egoistisch sind. Es könnte durchaus sein, dass der Akteur dem Glück einer anderen Person einen hohen Nutzen beimisst und daher rational eine Handlungsweise wählt, die den Gesamtnutzen, wie er diesen selber sieht, durch das Glück dieser anderen Person maximiert. In realen Situationen weiß der Akteur in der Regel nicht, wie die Ergebnisse bestimmter Entscheidungen aussehen werden, so dass er unter Unsicherheit handelt. Um dieses Problem zu überwinden, wählt der rationale Akteur die Handlung mit dem *maximalen erwarteten Nutzen* (MEU), wobei der Wert einer Wahl gleich dem Nutzen des Ergebnisses multipliziert mit der Wahrscheinlichkeit des Eintretens dieses Ergebnisses ist. Man denke an die rationalen Erwartungen, die man hat, wenn man bei bestimmten Glücksspielen oder Lotterien mitmacht.

Komplizierter sind die Fälle von Entscheidungen, wo die Rationalität der je bestimmten Wahl von den nachfolgenden Entscheidungen *anderer Akteure* abhängt. Solche Fälle werden oft mit Hilfe von »Spielen« beschrieben, die zusammen mit anderen Akteuren gespielt werden. In solchen Spielen ist es oft eine erfolgreiche

Strategie, mit anderen Akteuren zu kooperieren, um den subjektiven Nutzen zu maximieren.

Im Diskurs der künstlichen Intelligenz ist es üblich, KI-Agenten als rationale Agenten im beschriebenen Sinne zu betrachten. So bemerkt beispielsweise Stuart Russell:

> »Kurz gesagt, ein rationaler Agent handelt so, dass er den erwarteten Nutzen maximiert. Die Bedeutung dieser Schlussfolgerung kann gar nicht hoch genug eingeschätzt werden. In vielerlei Hinsicht ging es bei der künstlichen Intelligenz vor allem darum, herauszufinden, wie man rationale Maschinen bauen kann.« (Russell 2019: 23)

6.2 Ressourcen und rationale Handlungsfähigkeit

Es ist nicht der Fall, dass ein rationaler Agent *tatsächlich* immer die perfekte Option wählt. Das liegt vor allem daran, dass ein solcher Agent damit zurechtkommen muss, dass seine Ressourcen begrenzt sind, insbesondere Informationsspeicherung (Datenspeicher) und Zeit (bei den meisten Entscheidungen ist Zeit eine kritische Größe). Die Frage ist also nicht nur, was die beste Wahl ist, sondern auch, wie viele Ressourcen ich für die Optimierung meiner Wahl aufwenden sollte; wann sollte ich aufhören zu optimieren und anfangen zu handeln? Dieses Phänomen wird als *eingegrenzte Rationalität* (bounded rationality) oder *begrenzte Optimalität* bezeichnet und verlangt in der Kognitionswissenschaft eine *ressourcenrationale* Analyse (Lieder/Griffiths 2020; Russell 2016: 16ff.; Simon 1955: 99; Wheeler 2020). Außerdem gibt es keine feststehende Menge diskreter Optionen, aus denen man wählen kann, und so muss ein rationaler Akteur nicht nur über seine Mittel nachdenken, sondern auch über seine Ziele (siehe Abschnitt 9).

Die Tatsache, dass (natürliche oder künstliche) Akteure bei ihren Entscheidungen mit begrenzten Ressourcen umgehen müssen, ist für das Verständnis der Kognition von enormer Bedeutung. In der Philosophie wird dies oft nicht in vollem Umfang gewürdigt – selbst in der Literatur über die Grenzen der rationalen Wahl scheint man oft der Meinung zu sein, es wäre irgendwie »falsch«, Heuristiken zu verwenden, die Voreinstellungen (biases) enthalten, sich von der relevanten Umwelt »anschubsen« zu lassen (nudging), oder die Umwelt für »erweiterte« oder »situierte« Kognition zu nutzen.[12] Eigentlich wäre es jedoch irrational, nach perfekten kognitiven Verfahren zu streben, ganz zu schweigen von kognitiven Verfahren, die in jeder Umwelt perfekte Ergebnisse liefern.

12 Kahneman/Tversky 1979; Kahnemann 2011; Thaler/Sunstein 2008, vs. Kirsh 2009.

6.3 Rahmungsproblem(e)

Das ursprüngliche sogenannte Rahmungsproblem (frame problem) der klassischen KI bestand darin, wie das Überzeugungssystem eines Akteurs nach einer erfolgten Handlung *aktualisiert* werden kann, ohne alles anführen zu müssen, was sich *nicht* geändert hat. Dies erfordert eine Logik, in der sich die Schlussfolgerungen ändern können, wenn eine Prämisse hinzugefügt wird – eine nicht-monotone Logik. (Shanahan 2016) Über dieses eher technische Problem hinaus gibt es ein philosophisches Problem der Aktualisierung von Überzeugungen nach einer Handlung, das von Dennett popularisiert wurde und die Frage aufwirft, wie man herausfinden kann, was relevant ist und wie weit der Rahmen für *Relevanz* gezogen werden sollte. Wie Shanahan bemerkt, ist »Relevanz ganzheitlich, ergebnisoffen und kontextabhängig«, aber logische Schlussfolgerungen sind es nicht (Dennett 1984a; Shanahan 2016).

Es gibt eine sehr allgemeine Version des Frame-Problems, die von Jerry Fodor formuliert wurde. Er vergleicht es mit »Hamlets Problem: wann man aufhören soll zu denken«. Und er meint, dass »modulare kognitive Verarbeitung *ipso facto* irrational […] ist, weil weniger als alle relevante und verfügbare Evidenz einbezogen wird« (Fodor 1987: 140f.; Sperber/Wilson 1996). Fodor macht damit auf das Problem aufmerksam, dass man, um eine logische Schlussfolgerung, insbesondere eine Abduktion, durchzuführen, schon entschieden haben muss, was überhaupt als relevant gelten soll. Er scheint jedoch die Tatsache zu unterschätzen, dass man sich nicht um *alles* kümmern kann, was relevant und verfügbar ist (denn unsere Rationalität ist eingegrenzt). Es ist derzeit unklar, ob das Rahmungsproblem ohne fragwürdige Annahmen über Rationalität formuliert werden kann. Ähnliche Bedenken treffen die Behauptung, Gödel habe die tiefen Grenzen von KI-Systemen aufgezeigt (Koellner 2018a; Koellner 2018b; Lucas 1996). Womöglich beinhaltet Intelligenz doch mehr als nur instrumentelle Rationalität.

6.4 Kreativität

Entscheidungen, die mit *Kreativität* zu tun haben, werden oft für etwas gehalten, das über alles Mechanische hinausgeht und daher für eine bloße Maschine unerreichbar ist. Der Begriff des »schöpferischen Schaffens« hat in unserer gesellschaftlichen Praxis erhebliches Gewicht, insbesondere wenn diese Schöpfung durch geistige Eigentumsrechte geschützt ist – und KI-Systeme *haben* Musik, Malerei und Texte geschaffen oder mitgeschaffen. Es ist überhaupt nicht klar, ob es einen Begriff von Kreativität gibt, der ein Argument gegen maschinelle Kreativität liefern würde. Ein solcher Begriff müsste zwei Aspekte miteinander verbinden, die in einem Spannungsverhältnis zu stehen scheinen: Einerseits scheint Kreativität eine Ursächlichkeit zu implizieren, die den Erwerb von Wissen und Techniken einschließt (man

denke an J.S. Bach, wie er eine neue Kantate komponiert), andererseits soll Kreativität so etwas wie ein nicht-verursachter, nicht-vorhersehbarer Einsichtsfunke sein. Es ist gar nicht klar, ob ein solcher Begriff von Kreativität überhaupt formuliert werden kann oder sollte (Boden 2014; Colton/Wiggins 2012; Halina 2021). Vielleicht ergibt sich eine plausible Erklärung von Kreativität, wenn wir davon ausgehen, dass es bei Kreativität darum geht, sich zwischen verschiedenen Räumen der Relevanz zu bewegen, ähnlich wie beim Rahmungsproblem.

7. Freier Wille und Kreativität

7.1 Determinismus, Kompatibilismus

Das Problem, das gewöhnlich unter der Überschrift »freier Wille« behandelt wird, ist die Frage, wie physische Wesen wie Menschen oder KI-Systeme so etwas wie einen freien Willen haben können. Die übliche Einteilung möglicher Positionen im Diskurs über den freien Willen lässt sich in Form eines Entscheidungsbaums darstellen. Die erste Verzweigung ist die Frage, ob der *Determinismus* wahr ist, d.h. die These, dass alle Ereignisse verursacht werden. Die zweite Verzweigung ist, ob der *Inkompatibilismus* wahr ist, d.h. die These, dass es keinen freien Willen gibt, wenn der Determinismus wahr ist.

Die als *harter Determinismus* bekannte Position besagt, dass der Determinismus tatsächlich wahr ist und es deshalb so etwas wie Willensfreiheit nicht gibt – dies ist die Schlussfolgerung, die die meisten seiner Gegner zu vermeiden versuchen. Die Position, die als *Libertarismus* bekannt ist (nicht im politischen Sinne), stimmt zu, dass der Inkompatibilismus wahr ist, fügt aber hinzu, dass der Determinismus nicht wahr ist und wir daher frei sind. Die als *Kompatibilismus* bekannte Position besagt, dass Determinismus und freier Wille miteinander vereinbar sind und es daher durchaus sein kann (und wohl tatsächlich auch so ist), dass der Determinismus wahr ist *und* der Mensch einen freien Willen hat.

Daraus ergibt sich eine kleine Matrix von Positionen:

	Inkompatibilismus	*Kompatibilismus*
Determinismus	harter Determinismus	optimistischer/pessimistischer Kompatibilismus
Nicht-Determinismus	Libertarismus	[keine beliebte Option]

7.2 Kompatibilismus und Verantwortung in der KI

Wenn ich sage, dass ich etwas aus freien Stücken getan habe, bedeutet das in erster Näherung, dass es *an mir lag*, dass ich die *Kontrolle* hatte. Dieser Begriff von Kontrolle lässt sich erläutern, indem man sagt, ich hätte anders handeln können als ich es getan habe, insbesondere hätte ich anders handeln können, wenn ich *mich* anders *entschieden* hätte. Und dass ich mich anders entschieden hätte, wenn ich andere *Vorlieben* oder *Kenntnisse* gehabt hätte (z.B. hätte ich diese Fleischbällchen nicht gegessen, wenn ich eine Abneigung gegen Schweinefleisch hätte und wenn ich gewusst hätte, dass die Bällchen Schweinefleisch enthalten). Der entsprechende Freiheitsbegriff beinhaltet also eine *epistemische Bedingung* und eine *Kontrollbedingung*.

Ich handle also frei, wenn ich gemäß meinen Präferenzen (meinem subjektiven Nutzen) handle. Aber warum habe ich diese Präferenzen? Wie schon Aristoteles wusste, unterstehen sie nicht meiner willlentlichen Kontrolle, ich könnte nicht einfach *beschließen*, andere Präferenzen zu haben und sie dann haben. Harry Frankfurt hat allerdings deutlich gemacht hat, dass ich Präferenzen oder Wünsche *zweiter Ordnung* haben kann, d.h. ich kann präferieren, andere Präferenzen zu haben als die, die ich tatsächlich habe (z.B. könnte ich es mögen, Fleischbällchen nicht zu mögen). Dass ich meine Präferenzen durch rationales Denken außer Kraft setzen kann, nennt Frankfurt den *Willen*, und dieser ist eine Bedingung dafür, dass ich eine Person bin. Näherungsweise kann man also sagen, *frei zu handeln bedeutet, so zu handeln, wie ich mich entscheide; mich so zu entscheiden, wie ich es will; und so zu wollen, wie ich es vernünftigerweise vorziehe, zu wollen* (Dennett 1984b; Frankfurt 1971).

Die Debatte läuft darauf hinaus, dass der Begriff des freien Willens bei KI oder Menschen die Funktion hat, persönliche *Verantwortung* zu ermöglichen, und nicht, eine *Ursache* zu bestimmen. Die eigentliche Frage lautet: Unter welchen Bedingungen ist ein Akteur für seine Handlungen *verantwortlich* und *verdient es*, dafür gelobt oder getadelt zu werden? Dies gilt unabhängig davon, ob wir frei von kausaler Determination handeln; diese Art von Freiheit bekommen wir nicht und brauchen wir auch nicht.

Zwischen »Optimisten« und »Pessimisten« gibt es eine weitere Debatte darüber, ob Menschen diese Bedingungen tatsächlich erfüllen (insbesondere, wieweit sie wirklich ihre Präferenzen kausal hervorbringen können) und daher zu Recht für ihre Handlungen verantwortlich sind und Lob oder Tadel *verdienen* – und ob Belohnung oder Bestrafung dementsprechend hauptsächlich zukunftsorientierte Gründe haben sollten (Dennett/Caruso 2018; Mele 2006; Pink 2004; Strawson 2004). Was KI-Systeme betrifft, so hat das Nichtvorhandensein von Verantwortung Konsequenzen für ihren Status als moralische Akteure, für die Existenz von »Verantwortungslücken« und für die komplexe normative Frage, welche Art von

Entscheidungen wir Systemen überlassen sollten, die nicht verantwortlich gemacht werden können. (Müller 2021; Simpson/Müller 2016; Sparrow 2007)[13]

8 Bewusstsein

8.1 Bewusstheit und phänomenales Bewusstsein

Zunächst ist es sinnvoll, zwei Arten von Bewusstsein zu unterscheiden: *Bewusstheit* und *phänomenales Bewusstsein*. Bewusstheit ist die Vorstellung, dass ein System kognitive Zustände auf einer Basisebene hat (z.B. spürt es Wärme) und auf einer Metaebene Zustände hat, in denen es sich der Zustände auf der Basisebene bewusst ist. Diese Bewusstheit bzw. dieser Zugang beinhaltet die Fähigkeit, sich an die kognitiven Zustände auf der Basisebene zu erinnern und sie zu nutzen. Dies ist der begriffliche Sinn von »bewusst« im Unterschied zu »unbewusst« oder »unterbewusst«. Und für ein mehrschichtiges KI-System scheinen diese Unterscheidungen auch machbar zu sein.

Mit Bewusstheit geht oft, aber nicht notwendigerweise, einher, dass der kognitive Zustand auf der Basisebene sich für das Subjekt auf eine bestimmte Weise *anfühlt*. Dies wird philosophisch als *phänomenales Bewusstsein* bezeichnet: dies, wie mir die Dinge *erscheinen* (griechisch *phaínetai)*. Dieser Begriff des Bewusstseins lässt sich wahrscheinlich am besten mit Hilfe von zwei klassischen philosophischen Gedankenexperimenten erklären: der Fledermaus und der Farbenwissenschaftlerin.

Angenommen, Sie und ich essen beide Schokoladeneis. Dann kann ich immer noch nicht wissen, wie das Eis für Sie schmeckt, und ich würde es auch dann nicht wissen, wenn ich alles über das Eis, über Sie, über Ihr Gehirn und Ihre Geschmacksknospen wüsste. *Wie* es für Sie schmeckt, ist etwas, das für mich epistemisch unzugänglich ist, ich kann es niemals wissen, selbst wenn ich alles über die physische Welt wüsste. Genauso wenig kann ich je wissen, wie es sich anfühlt, eine Fledermaus zu sein (Nagel 1974; Nagel 1987: Kap. 3).

Eine ähnliche Pointe zur Frage des nicht Wissbaren macht Frank Jackson in dem vieldiskutierten Artikel »Was Mary nicht wusste« (Jackson 1986). In seinem Gedankenexperiment soll Mary eine Person sein, die in ihrem Leben noch nie etwas Farbiges gesehen hat, die aber eine perfekte Farbenwissenschaftlerin ist: Sie weiß alles, was es über Farbe zu wissen gibt. Eines Tages kommt sie aus ihrer schwarz-weißen Umgebung heraus und sieht zum ersten Mal Farbe. Es scheint, als ob sie in diesem Moment etwas Neues lernt.

Das Argument, das hier vorgebracht wird, scheint für einen geistig-physikalischen *Dualismus* von *Substanzen* oder zumindest *Eigenschaften* zu sprechen: Ich könn-

13 Siehe auch den Beitrag von Susanne Hahn im vorliegenden Band. [Anm. der Hrsg.]

te alles Wissen der Physik haben, ohne das Wissen der phänomenalen Erfahrung zu haben, also ist die phänomenale Erfahrung kein Teilgebiet der Physik. Wenn der Dualismus wahr ist, dann schaut es so aus, dass wir nicht hoffen dürfen, mit der richtigen physikalischen Technologie, wie z.B. der KI, phänomenales Bewusstsein zu erzeugen. In der Gestalt des *Substanzdualismus*, wie ihn Descartes und ein Großteil des religiösen Denkens angenommen haben, ist der Dualismus heute allerdings unpopulär: Die meisten Philosophen gehen von einem Physikalismus aus, der besagt, dass »alles physisch ist«.

Eine ganze Reihe von Argumenten gegen die Reduktion mentaler auf physische *Eigenschaften* werden diskutiert, so dass man wohl mit Fug und Recht behaupten kann, dass der *Eigenschaftsdualismus* eine große Anhängerschaft hat. Dieser wird oft mit dem Substanzmonismus zu einer Version der »Supervenienz des Mentalen auf dem Physischen« kombiniert, d.h. zu der These, dass zwei Entitäten mit denselben physischen Eigenschaften auch dieselben mentalen Eigenschaften haben müssen. Einige Philosophen haben diese Beziehung zwischen dem Eigenschaftsdualismus und der Möglichkeit eines künstlichen Bewusstseins in Frage gestellt. So behauptet David Chalmers, dass »die physikalische Struktur der Welt – die genaue Verteilung von Teilchen, Feldern und Kräften in der Raumzeit – logisch mit der Abwesenheit von Bewusstsein vereinbar ist, so dass das Vorhandensein von Bewusstsein eine weitere Tatsache über unsere Welt ist«. Trotz dieser Behauptung unterstützt er den Computationalismus und meint: »starke künstliche Intelligenz ist wahr: Es gibt eine Klasse von Programmen, bei denen jede Implementierung eines Programms dieser Klasse bewusst ist.« (Chalmers 1999: 436; Chalmers und Searle 1997; Davidson 1970)

Bedeutsamer aber als die Diskussion über Dualismen ist das Verständnis der *Funktion* von Bewusstsein in KI-Systemen oder bei natürlichen Agenten: Warum ist das phänomenale Bewusstsein beim Menschen so, wie es ist? Wie könnten wir feststellen, ob ein System Bewusstsein hat? Könnte es einen Menschen geben, der physisch genauso gebaut wäre wie ich, aber kein Bewusstsein hätte (ein »philosophischer Zombie«) (O'Regan 2011)?

8.2 Das Selbst

Die persönliche Identität ist für Menschen vor allem deshalb bedeutsam, weil sie eine Voraussetzung für die Zuweisung von Verantwortung ist (siehe Abschnitt 6.4): Um Schuld oder Lob zuzuweisen, muss ich in einem bestimmten Sinne *dieselbe Person* sein in wie derjenige, der die betreffende Handlung ausgeführt hat. Es gehört zu unserem Selbstverständnis, dass es ein Leben in der Vergangenheit gibt, das meines ist, und nur meines – wie das möglich ist, ist als die »Frage der Persistenz« bekannt. Die Standardkriterien dafür, dass ich dieselbe Person bin wie der kleine Junge auf dem Foto, sind meine *Erinnerung* daran, dieser Junge zu sein, und die *Kontinuität meines Körpers* über die Zeit. Wir Menschen neigen dazu zu glauben, dass *Erinne-*

rung oder *bewusste Erfahrung* oder *geistige Inhalte die* Kriterien für persönliche Identität sind, weshalb wir uns auch vorstellen können, unseren Tod zu überleben oder in einem anderen Körper zu leben (Metzinger 2009; Olsen 2019).

Was also ist ein »Teil« dieses beständigen Selbst? Abgesehen von philosophischen Phantasien und neurologischen Raritäten[14] gibt es heute keinen Zweifel mehr daran, was »Teil von mir« ist und was nicht – ich arbeite ständig daran, diese persönliche Identität aufrechtzuerhalten, indem ich prüfe, ob die verschiedenen Sinne übereinstimmen, z. B. versuche ich, nach der Türklinke zu greifen, ich sehe, wie meine Hand die Klinke berührt, ich kann sie fühlen … und dann sehe ich, wie sich die Tür öffnet, und spüre, wie meine Hand sich nach vorn bewegt. Das ist etwas ganz anderes als ein Computer: Die Komponenten der Standard Von-Neumann-Architektur (Eingabesystem, Speicher, Direktzugriffsspeicher, Prozessor, Ausgabesystem) können sich im selben Gehäuse befinden oder meilenweit voneinander entfernt sein, sie können sogar in mehrere Komponenten aufgeteilt sein (z. B. bei Off-Board Arbeitsprozessen an rechenintensiven Aufgaben) oder in Räumen wie der »Cloud« gespeichert sein, die nicht durch einen physischen Ort definiert sind. Und das ist nur die Hardware, die Software steht vor ähnlichen Problemen, so dass ein beständiges und abgegrenztes Selbst auszubilden keine leichte Aufgabe für ein KI-System wäre. Es ist auch gar nicht klar, ob es überhaupt eine Funktion für ein Selbst in der KI gibt und welche Konsequenzen für die Zuschreibung von moralischem Handeln und Behandeltwerden das hat.

9. Normativität

Kehren wir kurz zu den Fragen der rationalen Wahl und der Verantwortung zurück. Stuart Russell meint, dass »die KI das Standardmodell übernommen hat: Wir bauen optimierende Maschinen, wir geben ihnen Ziele vor, und los geht's.« (Russell 2019: 172) Nach diesem Verständnis ist die KI ein Werkzeug, und wir müssen ihr die Ziele vorgeben, die sie erreichen soll. Die KI verfügt ausschließlich über *instrumentelle Intelligenz*, um die vorgegebenen Ziele zu erreichen. Zur *allgemeinen Intelligenz* gehört jedoch auch eine metakognitive Reflexionsfähigkeit, welche Ziele für mein jetziges Handeln relevant sind (Nahrung oder Unterkunft?) und eine Reflexion darüber, welche Ziele man verfolgen sollte (Müller/Cannon 2022). Eine der vielen offenen Fragen ist, ob ein nicht-lebendes System »echte Ziele« in dem Sinne haben kann, der für Handlungsentscheidungen und Verantwortung erforderlich ist, d.h. Ziele, die für das System einen subjektiven Wert haben und die das System reflektierend als

14 Z.B. »Der Mann, der aus dem Bett fiel« in (Sacks 1985) oder die Betrachtung des Menschen als Superorganismus, basierend auf dem menschlichen Mikrobiom.

wichtig erkennt. Ohne eine solche Reflexion über Ziele wären KI-Systeme keine moralischen Agenten und es könnte keine »Maschinenethik« geben, die diesen Namen verdient. Ähnliche Überlegungen gelten für andere Formen der normativen Reflexion, z.B. in der Ästhetik und der Politik. Diese Diskussion in der KI-Philosophie deutet darauf hin, dass der normativen Reflexion eine basale Funktion im kognitiven System zukommt, ob beim Menschen oder bei KI-Systemen.

Literatur

Anonym (1950): Digital Computing Newsletter, in: Office of Naval Research, Mathematical Sciences Division: Washington (DC), 2(1), 1–4.

Boden, M.A. (2014): Creativity and artificial intelligence. A contradiction in terms?, in: Paul, E.S.; Kaufman, S.B. (Hg.), The philosophy of creativity. New essays, Oxford: Oxford University Press.

Boden, M.A. (1988): Computer models of the mind. Computational approaches in theoretical psychology, Cambridge: Cambridge University Press.

Boden, M.A. (2. Auflage 2006): Mind as machine. A history of cognitive science, Oxford: Oxford University Press.

Buckner, C. (forthcoming): From deep learning to rational machines. What the history of philosophy can teach us about the future of artificial intelligence, New York: Oxford University Press.

Carter, M. (2007): Minds and computers. An introduction to the philosophy of artificial intelligence, Edinburgh: Edinburgh University Press.

Chalmers, D.J. (1999): Précis of The Conscious Mind, in: *Philosophy and Phenomenological Research*, LIX(2), 435–438.

Chalmers, D.J.; Searle, J. (1997): Consciousness and the philosophers. An exchange, in: *New York Review of Books*, 15.05.1997. [https://www.nybooks.com/articles/1997/05/15/consciousness-and-the-philosophers-an-exchange/] (Zugriff: 25.05.2024).

Clark, A. (2003): Natural born cyborgs. Minds, technologies, and the future of human intelligence, Oxford: Oxford University Press.

Clark, A. (2013): Whatever next? Predictive brains, situated agents, and the future of cognitive science, in: *Behavioral and Brain Sciences*, 36(6), 181–204.

Clark, A. (2016): Surfing uncertainty. Prediction, action, and the embodied mind, New York: Oxford University Press.

Clark, A.; Chalmers, D.J. (1998): The extended mind, in: *Analysis*, 58(1), 7–19.

Coelho Mollo, D. (2022): Intelligent Behaviour, in: *Erkenntnis*, 89, 705–721.

Cole, D. (2020): The Chinese room argument, in: Zalta, E.N.; Nodelman, U. (Hg.), Stanford Encyclopedia of Philosophy, Stanford (CA): The Metaphysics Research Lab. [http://plato.stanford.edu/entries/chinese-room/] (Zugriff: 22.05.2024).

Colton, S.; Wiggins, G.A. (2012): Computational creativity: The final frontier?, in: *Frontiers in Artificial Intelligence and Applications*, 242, 21–26.

Copeland, J.B. (1993): Artificial intelligence. A philosophical introduction, Oxford: Blackwell.

Copeland, J.B.; Posy, C.J.; Shagrir, O. (Hg.) (2013): Computability. Turing, Gödel, Church, and Beyond, Cambridge (MA): MIT Press.

Davidson, D. (1970): Mental Events, in: Foster, L.; Swanson, J. (Hg.), Experience and Theory, Amherst (MA): University of Massachusetts Press, 137–149.

Dennett, D.C. (1984a): Cognitive wheels. The frame problem of AI, in: Hookway, C. (Hg.), Minds, machines, and evolution. Philosophical studies, Cambridge: Cambridge University Press, 129–152.

Dennett, D.C. (1984b): Elbow room. The varieties of free will worth wanting, Cambridge (MA): MIT Press.

Dennett, D.C. (1991): Consciousness explained, New York: Little, Brown & Co.

Dennett, D.C.; Caruso, G.D. (2018): Just deserts, *Aeon*, 1, 1–20.

Dietrich, E. (2002): Philosophy of artificial intelligence, in: *The Encyclopedia of Cognitive Science*, 203–208.

Dietrich, E.; Fields, C.; Sullins, J.P.; Van Heuveln, B.; Zebrowski, R. (2021): Great philosophical objections to artificial intelligence. The history and legacy of the AI wars, London: Bloomsbury Academic.

Dodig-Crnkovic, G.; Müller, V.C. (2011): A dialogue concerning two world systems: Info-computational vs. mechanistic, in: Dodig-Crnkovic, G.; Burgin, M. (Hg.), Information and computation. Essays on scientific and philosophical understanding of foundations of information and computation, Boston: World Scientific, 149–184.

Dreyfus, H.L. (2. Auflage 1992): What computers still can't do. A critique of artificial reason, Cambridge (MA): MIT Press.

Edelman, S. (20008): Computing the mind. How the mind really works, Oxford: Oxford University Press.

Floridi, L. (2011): The philosophy of information, Oxford: Oxford University Press.

Floridi, L. (Hg.) (2003): The Blackwell guide to the philosophy of computing and information, Oxford: Blackwell.

Fodor, J.A. (1987): Modules, frames, fridgeons, sleeping dogs, and the music of the spheres, in: Garfield, J.L. (Hg.), Modularity in knowledge representation and natural-language understanding, Cambridge (MA): The MIT Press, 25–36.

Frankfurt, H. (1971): Freedom of the will and the concept of a person, in: *The Journal of Philosophy*, LXVIII(1), 5–20.

Friston, K.J. (2010): The free-energy principle. A unified brain theory?, in: *Nature Reviews Neuroscience*, 11, 127–138.

Garson, J.; Buckner, C. (2019): Connectionism, in: Zalta, E.N.; Nodelman, U. (Hg.), Stanford Encyclopedia of Philosophy, Stanford (CA): The Metaphysics Research Lab. [https://plato.stanford.edu/entries/connectionism/] (Zugriff: 25.05.2024).

Gödel, K. (1931): Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme I, in: *Monatshefte für Mathematik und Physik*, 38, 173–198.

Görz, G.; Schmid, U.; Braun, T. (5. Auflage 2020): Handbuch der künstlichen Intelligenz, Berlin: De Gruyter.

Halina, M. (2021): Insightful artificial intelligence, in: *Mind and Language*, 36(2), 315–329.

Harnad, S. (1990): The symbol grounding problem, in: *Physica D*, 42, 335–346.

Haugeland, J. (1985): Artificial intelligence. The very idea, Cambridge (MA): MIT Press.

Haugeland, J. (2002): Syntax, semantics, physics, in: Preston, J.; Bishop, M. (Hg.), Views into the Chinese room. New essays on Searle and artificial intelligence, Oxford: Oxford University Press, 379–392.

Hilbert, D. (1900): Mathematische Probleme, in: Nachrichten von der Königlichen Gesellschaft der Wissenschaften zu Göttingen, Math.-Phys. Klasse, 3, Göttingen: Lüder Horstmann, 253–297.

Horsman, C.; Stepney, S.; Wagner, R.C.; Kendon, V. (2014): When does a physical system compute?, in: *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Science*, 470 (2169), 1–25.

Jackson, F. (1986): What Mary didn't know, in: *Journal of Philosophy*, 83, 291–295.

Kahneman, D.; Tversky, A. (1979): Prospect theory. An analysis of decision under risk, in: *Econometrica*, 47, 263–291.

Kahnemann, D. (2011): Thinking fast and slow, London: Macmillan.

Kirsh, D. (2009): Problem solving and situated cognition, in: Robbins, P.; Aydede, M. (Hg.), The Cambridge handbook of situated cognition, Cambridge: Cambridge University Press, 264–306.

Koellner, P. (2018a): On the Question of Whether the Mind Can Be Mechanized, I: From Gödel to Penrose, in: *Journal of Philosophy*, 115(7), 337–360.

Koellner, P. (2018b): On the question of whether the mind can be mechanized, II: Penrose's new argument, in: *Journal of Philosophy*, 115(9), 453–484.

LeCun, Y.; Bengio, Y.; Hinton, J. (2015): Deep learning, in: *Nature*, 521(7553), 436–444.

Legg, S.; Hutter, M. (2007): Universal intelligence. A Definition of machine intelligence, in: *Minds and Machines*, 17(4), 391–444.

Lieder, F.; Griffiths, T.L. (2020): Resource-rational analysis. Understanding human cognition as the optimal use of limited computational resources, in: *Behavioral and Brain Sciences*, 43, e1, 1–60.

Lighthill, J. (1973): Artificial intelligence. A general survey, in: Science Research Council (Hg.), Artificial intelligence. A paper symposion, London: Science Research

Council. [http://www.chilton-computing.org.uk/inf/literature/reports/lighthill_report/p001.htm] (Zugriff: 25.05.2024).

Lucas, J.R. (1996): Minds, machines and Gödel. A retrospect, in: Millican, P.J.R.; Clark, A. (Hg.), Machines and Thought, Oxford: Oxford University Press, 103–124.

Margolis, E.; Samuels, R.; Stich, S. (Hg.) (2012): The Oxford handbook of philosophy of cognitive science, Oxford: Oxford University Press.

McCarthy, J. (2007): John Searle's Chinese room argument. [http://www-formal.stanford.edu/jmc/chinese.html] (Zugriff: 06.10.2007).

McCarthy, J.; Minsky, M.; Rochester, N.; Shannon, C.E. (1955): A proposal for the Dartmouth summer research project on artificial intelligence. [http://www-formal.stanford.edu/jmc/history/dartmouth/dartmouth.html] (Zugriff: 25.05.2024).

Mele, A.R. (2006): Free will and luck, Oxford: Oxford University Press.

Metzinger, T. (2009): The ego tunnel. The science of the mind and the myth of the self, New York: Basic Books.

Miłkowski, M. (2018): Objections to computationalism. A survey, in: *Roczniki Filozoficzne*, LXVI(8), 1–19.

Müller, V.C. (2013): What is a digital state?, in: Bishop, M.J.; Erden, Y.J. (Hg.), The Scandal of Computation – What is Computation? – AISB Convention 2013, Hove: AISB, 11–16. [http://www.aisb.org.uk/asibpublications/convention-proceedings] (Zugriff: 25.05.2024).

Müller, V.C. (2020): Ethics of artificial intelligence and robotics, in: Zalta, E.N.; Nodelman, U. (Hg.), Stanford Encyclopedia of Philosophy, Stanford (CA): The Metaphysics Research Lab. [https://plato.stanford.edu/entries/ethics-ai/] (Zugriff: 25.05.2024).

Müller, V.C. (2021): Is it time for robot rights? Moral status in artificial entities, in: *Ethics & Information Technology*, 23(3), 579–587.

Müller, V.C.; Cannon, M. (2022): Existential risk from AI and orthogonality. Can we have it both ways?, in: *Ratio*, 35(1), 25–36.

Müller, V.C. (forthcoming): Can machines think? Fundamental problems of artificial intelligence, New York: Oxford University Press.

Nagel, T. (1974): What is it like to be a bat?, in: *Philosophical Review*, 83(4), 435–450.

Nagel, T. (1987): What does it all mean? A very short introduction to philosophy, Oxford/New York: Oxford University Press.

Negroponte, N. (1995): Being digital, New York: Vintage.

Newell, A.; Simon, H. (1976): Computer science as empirical enquiry. Symbols and search, in: *Communications of the Association of Computing Machinery*, 19(3), 113–126.

Newen, A.; Gallagher, S.; De Bruin, L. (2018): 4E Cognition. Historical Roots, Key Concepts, and Central Issues, in: Newen, A.; De Bruin, L.; Gallagher, S. (Hg.), The Oxford Handbook of 4E Cognition, Oxford: Oxford University Press.

O'Regan, K.J. (2011): Why red doesn't sound like a bell. Understanding the feel of consciousness, New York: Oxford University Press.

Olsen, E. (2019): Personal identity, in: Zalta, E.N.; Nodelman, U. (Hg.), Stanford Encyclopedia of Philosophy, Stanford (CA): The Metaphysics Research Lab. [https://plato.stanford.edu/entries/identity-personal/] (Zugriff: 25.05.2024).

Pearl, J.; Mackenzie, D. (2018): The book of why. The new science of cause and effect, New York: Basic Books.

Pfeifer, R.; Bongard, J. (2007): How the body shapes the way we think. A new view of intelligence, Cambridge (MA): MIT Press.

Piccinini, G. (2021): Computation in physical systems, in: Zalta, E.N.; Nodelman, U. (Hg.), Stanford Encyclopedia of Philosophy, Stanford (CA): The Metaphysics Research Lab. [https://plato.stanford.edu/entries/computation-physicalsystems/] (Zugriff: 25.05.2024).

Pink, T. (2004): Free will. A very short introduction, Oxford: Oxford University Press.

Preston, J.; Bishop, M. (Hg.) (2002): Views into the Chinese room. New essays on Searle and artificial intelligence, Oxford: Oxford University Press.

Rosenblatt, F. (1957): The Perceptron. A perceiving and recognizing automaton (Project PARA), in: *Cornell Aeronautical Laboratory Report*, 85(460/461), 1–29.

Russell, S. (2016): Rationality and intelligence. A brief update, in: Müller, V.C. (Hg.), Fundamental issues of artificial intelligence, Cham: Springer, 7–28.

Russell, S. (2019): Human compatible. Artificial intelligence and the problem of control, New York: Viking.

Russell, S.; Norvig, P. (4. Auflage 2020): Artificial intelligence. A modern approach, Upper Saddle River: Prentice Hall.

Sacks, O. (1985): The Man Who Mistook His Wife for a Hat, and Other Clinical Tales, New York: Summit Books.

Scheutz, M. (Hg.), (2002): Computationalism. New directions, Cambridge: Cambridge University Press.

Searle, J.R. (1980): Minds, brains and programs, in: *Behavioral and Brain Sciences*, 3, 417–457.

Searle, J.R. (1984): Intentionality and its place in nature, in: Ders. (Hg.), Consciousness and language, Cambridge: Cambridge University Press, 77–89.

Searle, J.R. (2004): Mind. A brief introduction, Oxford: Oxford University Press.

Shagrir, O. (1997): Two dogmas of computationalism, in: *Minds and Machines*, 7, 321–344.

Shagrir, O. (2022): The nature of physical computation, New York: Oxford University Press.

Shanahan, M. (2016): The frame problem, in: Zalta, E.N.; Nodelman, U. (Hg.), Stanford Encyclopedia of Philosophy, Stanford (CA): The Metaphysics Research Lab. [https://plato.stanford.edu/archives/spr2016/entries/frame-problem/] (Zugriff: 25.05.2024).

Siegelmann, H.T. (1995): Computation beyond the Turing limit, in: *Science*, 268(5210), 545–548.

Siegelmann, H.T. (1997): Neural networks and analog computation. Beyond the Turing limit, Basel: Birkhäuser.

Simon, H. (1955): A behavioral model of rational choice, in: *Quarterly Journal of Economics*, 69, 99–118.

Simpson, T.W.; Müller, V.C. (2016): Just war and robots' killings, in: *The Philosophical Quarterly*, 66(263), 302–322.

Sparrow, R. (2007): Killer robots, in: *Journal of Applied Philosophy*, 24(1), 62–77.

Sperber, D.; Wilson, D. (1996): Fodor's Frame Problem and Relevance Theory, in: *Behavioral and Brain Sciences*, 19(3), 530–532.

Strawson, G. (2004): Free will, in: Craig, E. (Hg.), Routledge Encyclopedia of Philosophy Online. [https://www.rep.routledge.com/articles/thematic/free-will/] (Zugriff: 25.05.2024).

Thaler, R.H.; Sunstein, C. (2008): Nudge. Improving decisions about health, wealth and happiness, New York: Penguin.

Thoma, J. (2019): Decision Theory, in: Pettigrew, R.; Weisberg, J. (Hg.), The open handbook of formal epistemology, PhilPapers Foundation, 57–106.

Turing, A. (1936): On Computable Numbers, with an Application to the Entscheidungsproblem, in: *Proceedings of the London Mathematical Society*, 2(42), 230–265.

Turing, A. (1950): Computing machinery and intelligence, in: *Mind*, LIX, 433–460.

Varela, F.J.; Thompson, E.; Rosch, E. (1991): The embodied mind. Cognitive science and human experience, Cambridge (MA): MIT Press.

Wheeler, G. (2020): Bounded Rationality, in: Zalta, E.N.; Nodelman, U. (Hg.), Stanford Encyclopedia of Philosophy, Stanford (CA): The Metaphysics Research Lab. [https://plato.stanford.edu/archives/fall2020/entries/bounded-rationality/] (Zugriff: 25.05.2024).

Wittgenstein, L. (1960[1953]): Philosophische Untersuchungen, in: Ders., Schriften I, Frankfurt a.M.: Suhrkamp, 279–544.

IV
Machtverhältnisse

Die beflissene Willfährigkeit vor den Oberflächen des Digitalen

Brauchen die digitalen Wirklichkeiten ein neues Konzept von Macht?*

Rainer Adolphi

Abstract: *This article shows how a complex sensorium for heteronomy and oppression and for social pathologies has emerged over a long period of cultural (and semantic) development and as the result of historically concrete experience – and how, to an increasing degree, the digitalization of reality is transforming, numbing and distorting precisely these forms of reflective consciousness. To this end, the article focuses on questions at several distinct levels: In what do the new forms of ›power‹ actually consist that our previously developed sensorium is now failing to register? What impact is living in digitally mediated realities in which subjects want, but are also forced to participate, having on their mentalities? How do the various forms of our previous sensorium (together with their developmental paths and mediating authorities) come to be robbed of their power? What are the defining features of the mediality of the technical in the age of* digital *technology and its (pragmatic, unconscious) omnipresence in our surrounding environment (for instance: living in surfaces; increasing iconification; fantasies of power paired with the conviction of impotence; a new understanding of history etc. – in sum, a loss of consciousness of agency)? And not least the power of language – how the transformed realities and the resulting lack of consciousness are structurally reproduced in a new language, in thought, in the processes of understanding (and at present, unfortunately, still to a large extent also in philosophical reflection).*

Keywords: *digital technology as the environment of life; power (anonymous); critical consciousness (disappearance); space of action; mediality of technology; surfaces (fusion); anthropology of the technical; blackboxing; iconification*

* In Dankbarkeit sei dieser Beitrag der Erinnerung an Hans Poser (1937–2022) gewidmet.

I Sensorien für ›Macht‹ (Geschichte des kritischen Bewusstseins)

Digitalisierung ist ein Geschehen. Mit den Prozessen von Digitalisierung verändert sich das soziale Sein. Es verändern sich die gesellschaftlichen Verhältnisse, die sozialen Bezüge unter den Menschen, die Alltagswelt und nicht zuletzt die individuellen Lebensverfasstheiten wie -gestimmtheiten. Ein Technisches, eine neue Epoche des menschenerschaffenen Technischen, verändert alles, und dies binnen kürzester Spannen. Wir rechnen schon nicht mehr nur in Zeiträumen von Generationen. Es ist ein Aktualzeit-Geschehen, vor Augen sichtbar und in seinen Gestalten spürbar. Wo, wie und wieweit eingetreten, Digitalisierung schafft Wirklichkeiten, objektiv wie subjektiverseits.

Mit dem Geschehen verändert sich auch der geschichtliche Horizont. Das sich überschlagende Neue lässt Utopien, mit denen der Prozess an seinen Anfängen begleitet wurde, nur noch als rührend naiv anmuten,[1] zugleich andererseits wird der gegenwärtige reale Stand doch immer neu mit Perspektiven von SciFi-Zukünften illustriert. Und alles, was *vor* dem neuen digitalen Zeitalter gedacht wurde über die Bedeutung von technischen Erfindungen und Errungenschaften und deren Implementierung, scheint wie hinter einer Zeitmauer. – Davon unmittelbar erfasst ist auch ein zentrales Strukturmoment kritischer Verständigung: nach Verhältnissen von ›Macht‹ zu fragen – ›Macht‹ zu bezeichnen, zu erklären (oder zu rechtfertigen), zu kritisieren. Auf der Höhe der Entwicklungen zu sein, entscheidet sich dementsprechend in vielem daran, inwiefern das Kategorienfeld als Verständnis hier modifiziert werden muss. Brauchen wir ein neues Konzept, um Macht-Verhältnisse zu denken, und zu analysieren? Reichen die bisherigen Konzepte nicht mehr genügend aus im Angesicht neuer Wirklichkeiten einer digitaltechnisch gewordenen Welt?

Das wäre an sich nichts Ungewöhnliches. Auch das Denken und allgemeiner die theoretische Reflexion über ›Macht‹ haben ihren geschichtlichen Erfahrungsbezug, historisch und kulturell. Alles Bisherige, der Hintergrund unserer heutigen Verständigungen, zeigt sich als Etappen, und diese haben zu tun mit einem jeweiligen neuen Problemerfordernis bzw. Problemdruck. – Schematisierend entflochten, und ohne Berechnung auf Vollständigkeit: In den Verhältnissen der antiken Welt stand weithin die *Macht-über-mich-selbst* im Fokus, am deutlichsten (und bis heute wirkungsmächtigsten) dabei in der stoischen Tradition und ihrem ›Individualismus‹ angesichts des als unselig und durch kein Engagement rettbar empfundenen sozialen und vor allem politischen Kosmos, dem ›Individualismus‹ des *Rette-sich-*

1 Um nur exemplarisch vier einflussreiche Ausformungen zu nennen: Haraway 1985; Moravec 1988; Weiser 1991; Barlow 1996.

wer-kann! angesichts der Bedrängungen durch die Realitäten der Mitwelt und überhaupt deren Fluiden – modern gesprochen, die Fokussierung auf die Macht der ›Lebenskunst‹. Mit dem Problem des politischen Zentralismus, und in der Folge auch der neuzeitlichen Gestalt des Problems von Kirche-und-Staat, wurde der Begriff der ›Macht‹ zur Frage, woher die *Souveränität* sich leitet, Souveränität im Gefüge vieler Kräfte- und Interessenpole, die letzte Souveränität. Dieser Begriff (und seine weitere Tradition) hatte, zumal solange für sich allein, strukturell stets etwas Absolutistisches, auch Totalitaristisches: letzte Macht über alles. Dieser Begriff von ›Macht‹ war darin vor allem ein Legitimations-Begriff, ›Macht‹ qua Legitimität des Monopols auf die Instrumente und Zeichen der Hoheit. Mit der Empirisierung der neuzeitlichen Wissenschaft trat dann zugleich die Bedeutung der anwendenden Technik ins Zentrum. Wo Technik ›methodisch‹ wird, und mit der Suche nach probatem Gesetzes-Wissen, wurde das, dass *Wissen* Macht ist und verschafft, zu einer neuen konzeptionellen Leitdimension: Wissen, anderweitige Kräfte und Macht-Potenziale zu nutzen. Macht ist dies darin (negativ) gegen Schicksal – und generell: gegen aus Unkenntnis die Bereitschaft zur Hinnahme gegebener Verhältnislagen – und (positiv) zur Ausweitung menschlicher Potenz wie Möglichkeitsräume.

Weitere Etappen, strukturell Problemerfahrungen auf Gestalten des in sich Reflexiven hin, kamen hinzu. Mit der Ideologiekritik seit dem 19. Jahrhundert, sei es die marxsche oder die nietzschesche Tradition, wurde die Bedeutung von Macht als *Deutungsmacht* bewusst. Zugleich dann wurde, im Bewusstsein der immer weiter voranschreitenden Arbeitsteilung und Differenzierung, ferner der Besonderung von Lebensformen wie -ideen, die Integrationsaufgabe einer Ordnung offenkundig; nach Überwindung allgemeiner Klassenkampf-Modelle wurde erstmals vom gesellschaftlich Pluralistischen gesprochen[2] – und die Rolle *partizipatorischer Macht* offen markiert, Macht gleichsam im Horizontalen sozialer Formen. Und wie neuzeitlich die *Macht des Subjekts* gedacht wurde, sei es als Macht per Vernunft oder als Kooperation oder als Widerstand (bzw. anarchisch), gehört mit den modernen Emanzipationsbewegungen der Topos mit dazu, dass für sozialen, kulturellen, gesellschaftlichen Fortschritt auch die ›*Machtfrage*‹ gestellt werden muss. Schließlich ein letzter Eintrag in das Kategorienfeld von ›Macht‹ wäre, was man als Facetten von *subjektloser Macht* kennzeichnen kann: die Macht-Effekte bei Massen-Phänomenen (Bildung von Massen, Massen-Verhalten etc.), Macht in der fest-stellenden und zuteilenden Diskursivierung des Lebens (Diskursivierung zu Sachverhalten, entstanden-herrschende Regel-Praktiken, was als ›normal‹ gilt und in welchen Differenzierungen bzw. Rubrizierungen, allgemein die herrschenden Normgestalten), ›systemische‹ Macht wie aus der funktionalen Ausdifferenzierung und Eigenlogik der Prozesse, die Gesellschaftliches ausmachen. –

2 So Laski 1917: 1–26 [Kap. I].

Für bisherige Wirklichkeiten mithin ein komplexes Sensorium für Aspekte von ›Macht‹. Der Begriff ist nicht *einer*. Die Bedeutungen heben einander nicht auf. Herausgebildet hat sich ein Gefüge kritischer Verständigung; zusammen formen die verschiedenen Dimensionen einen Stand des Auskultivierten. Die einzelnen Bedeutungen und die jeweiligen zugehörigen Kategorienfelder wie Diskursformen tragen normative Ansprüche – auch was legitime Gestalten sozialer Verhältnisse sind, und Ansprüche als soziale und politische Geschichtsagenda, wenn Betreffendes vorenthalten ist –, und zugleich bekommen Ohnmachtserfahrungen eine Sprache, eine Artikulation. Empfundenes, Drückendes muss nicht mehr nur im Dumpfen bleiben.

Digitalisierung und die mit ihr geschaffenen veränderten Wirklichkeiten sind eine andere Erfahrungslage als das, woraus die bisherigen Konzepte von ›Macht‹ hervorgegangen sind. Das Gefüge des Erworbenen zu sehen, gibt hier die erforderliche geschichtliche Verfremdung, reflektierend nicht mitgerissen zu werden von dem Geschehen. Geschichtliches Bewusstsein braucht dies denk- und diskursgeschichtliche Bewusstsein. – Eine andere Erfahrungslage: Denn schon überhaupt dass Technisches mitspielt, scheint in den bisherigen Dimensionen nur so vorgesehen, dass es entweder eine (individual-) subjektiv *intendierte Instrumentalität* verkörpert oder *funktional* ist in (nichttechnischen) Prozessen. Und mit dem Technischen, das in den Entwicklungen der Digitalisierung zunehmend zur Lebensumwelt selbst geworden ist, haben die Formen der Gegebenheiten sich entscheidend gewandelt gegenüber dem, was für bisherige Erfahrungen Handlungsraum und Orientierungen waren; hinzu kommt die diesem neuen Technischen innewohnende Fähigkeit, mit seinen Etablierungen einschneidend *Fakten* zu schaffen für Wirklichkeiten und deren Möglichkeiten. Alte Handhaben der Verständigung greifen da, zumindest in bisherigen Perspektiven der Diskurse, offenbar in der Tat immer weniger, werden diffus, verlieren ihre Sicherheit. Gegensätzlichste Bewertungen leiten sich nun aus ihnen, und zur Grundlage gegensätzlichster Geschichtsperspektiven. Die Wirklichkeit setzt in Teilen ihre kritische Verständigung und Gestaltung außer Kraft. Das faktisch Eintretende generiert seine eigenen, dominanten Narrative.

Aus dem Tiefenhorizont heraus, was für noch nicht digitalisierte Lebenswirklichkeiten schon einmal erreicht war, möchte das Folgende – am Problem der Thematisierung von ›Macht‹ – zu umreißen versuchen, weshalb für die Zeitalter-Geschehnisse der Digitalisierung eine Verständigung offenbar so schwierig ist, um Differenzierungen zu tragen und kritisch-reflektierende Potenziale.[3] Dazu soll zunächst zurückgegangen werden zu einem Punkt der Verzweigung. Er zeichnet sich dadurch aus, dass nach ihm sozialwissenschaftliche Perspektiven wie Analysen und

3 Zum Feld, wie in Bezug auf allgemein Technik bisher ihre ›Macht‹ und Machtauswirkungen verstanden sowie kritisch reflektiert wurde, vgl. Hubig 2015.

(›kulturtheoretische‹ oder allgemein-normative) Theoreme über die Ideenform von Prozessen sich tendenziell getrennt haben.

II Fügsamkeiten in Herrschendes (Eine neue Problemaktualität eines klassischen soziologischen Theorems von Max Weber)

Es gibt wenige Themen des Denkens, bei denen verschiedenste Richtungen sich auf ein selbes gemeinsames Leitzitat beziehen. Bei ›Macht‹ ist dies seit einhundert Jahren in Max Weber: »Macht bedeutet jede Chance, innerhalb einer sozialen Beziehung den eigenen Willen auch gegen Widerstreben durchzusetzen, gleichviel worauf diese Chance beruht.« (Weber 1980[1922]: 28) Das war die Bestimmung für ein bestimmtes Wissenschaftsprogramm gewesen, und in einer bestimmten historischen Entwicklungsepoche, mit dem Blick dieser Wirklichkeiten. Seitheriges ist gewissermaßen ein Sichabarbeiten an diesem Zitat. Dabei wird indes die zweite Hälfte der Bestimmung i.Allg. weggelassen – um dafür gleich mit einem jeweiligen eigenen Konzept einzusetzen.

Diese zweite Hälfte ist genauer doppelseitig; beides sucht sich zu entbinden von dem Permissiven im alltagsweltlichen, vorwissenschaftlichen Verständnis. Das eine davon ist zumeist immerhin noch vage bewusst. Weber argumentiert hier für ein Programm, das sich im Weiteren genau nicht am Konzept der ›Macht‹ spezieller festmacht, sondern am Konzept der ›Herrschaft‹. Denn »der Begriff ›Macht‹ ist soziologisch amorph« (ebd.). Das meint, über Spezifischeres – das Wer? (und Wer-wie-beschaffen?), Inwiefern?, Wie?, Worüber?, Wozu?, usw. – lässt sich damit theoretischerseits fast nichts differenzieren, und auch empirisch neigten die Kategorien dazu, Prozesse zu konfundieren und zu vieles in eine Schublade zu packen oder zu parallelisieren. Verschiedenstes, von einem unmittelbaren affektregenden oder manipulativen Verhältnis zwischen zwei Menschen bis zu einer Vorrangverteilung aus einer bestimmten Gunst der Stunde bzw. Lage,[4] kann sich als Macht ausprägen, situativ und ohne dass es betreffende soziologische Strukturen hätte (oder haben müsste). Das »soziologisch [A]morph[e]«, wie Weber es als eine drohende Problematik einer zur permissiven Verwendung verführenden Rede von ›Macht‹ anmerkt, wären denn nur Beschreibungen – Beschreibungen eines eben in Blick Genommenen. Mit welchen Theorie-Elementen das Material dann erschlossen ist, wäre allzu schnell unausgewiesen.

Das ist darum die eine Seite: ›Macht‹, als Denk- und Kategorienfeld einer Analyse, muss so weit wie möglich in einer betreffenden Form von *Herrschaft* konkre-

4 »Alle denkbaren Qualitäten eines Menschen und alle denkbaren Konstellationen« können jemanden zur situativen Macht erheben (Weber 1980[1922]: 28f.).

tisierbar sein. Und ›Herrschaft‹ bedeutet dabei in einer allgemeinen Weise das Soziologische einer in eine dauerhafte Form des Darauf-eingestellt-seins gekommenen Asymmetrie in irgend Verhältnissen des Lebens. Zu dieser einen Seite gehört jedoch zugleich die komplementäre: das Nicht-Aufbegehren, die Ein- und Unterordnung, die Dienstbarkeit. Es sind die Verhaltensgeprägtheiten, einem Anspruch auf Asymmetrie – einer Funktions- und konkreterweise Handlungszuweisung an mich, wie manifest oder indirekt immer sie sich artikulieren bzw. an mich kommen möge – Folge zu leisten. Weber fasst dies in einem altertümlich anmutenden, auch sonst in der Wissenschaftssprache ganz unterminologischen Begriff als »Fügsamkeit«. Er macht sich dabei das Ganze zunutze, was im Deutschen hier mit anklingt und in diesem Wort versammelt ist.

Es ist nicht nur das Sich-Unterordnen, Sich-Beugen; es ist ebenso das Sich-Zusammenfügen von etwas – hier im Fall sozialer Formen sowohl horizontal, d.h. das Sich-Zusammenfügen zu einer Gruppierung der Gehorchenden bzw. Ausführenden oder Nicht-dawider-Aufbegehrenden, als auch vertikal, durch diese Verteilung von Wollenssetzungsautorität (von Wenigen/Einzelnen) und Willfährigkeit (der Vielen) ein agierendes Ganzes zu bilden. Im Weiteren schwingt mit das In-einander-greifen, sozial im Sinne eines wechselseitigen Auf-einander-eingestellt-seins – hier das Eingestellt-sein, dass von bestimmten Führungspolen, Entscheidungseliten oder ›institutionellen‹ Organen (bzw. in Gestalt von deren Funktionsrepräsentanten) eine bindend imperativische Anweisung (oder Festlegung einer zu nehmenden Faktizität des Handlungs- resp. sozialen Raums) ausgeht,[5] und umgekehrt das Eingestellt-sein, dass resp. in welcher Weise die Massen als Gehorchende (Hinnehmende/Ausführende) sich betreffend verhalten oder reagieren werden. Nicht zuletzt schließlich ist es das meist weitgehende Habituelle oder Habituell-geworden-sein des betreffenden Verhaltens, wie es in dem Suffix »-samkeit« zum Ausdruck kommt.

Dies ist eine fruchtbare theoriestrategische Entscheidung. Das Potenzial dieser Entscheidung, wesentlich auch die Seite des einer Macht bzw. Herrschaft Unterliegenden mit einzubeziehen – die Prozesse auf *dieser* Seite –, liegt darin, sehen zu können, wie auch die »Fügsamkeit« ihre *Gründe* und (gewordene/bewirkte) Geformtheiten hat. Dies darf nicht abgeschoben werden auf irgendein bequemes Theorem von einer ›natürlichen‹, gar anthropologisch begründeten Autoritätsanerkennung oder Theorem von einer prinzipiellen, sozialcharakterlichen Passivität bzw. einem herrschenden Sozialfatalismus. In Gestaltungen sozialer Formationen verstetigen sich die »Fügsamkeiten« denn auch gemeinhin zu einem ebenfalls qualitativ soziologischen Handlungssachverhalt: zu dem, was Weber dann im Konkreten unter der analytischen Kategorie der »Disziplin« von Hinordnungs- bzw. Wil-

5 Und dass die große Masse der solcherart Eingestellten auch untereinander sich durch ihre Gemeinsamkeit ihrer Subordination bzw. Nichtaufbegehren ›verstehen‹.

ligkeits-Gruppen und insofern Massen gefasst hat (vgl. Weber 1980[1922]: 28f. u. öfter). Dass subjektseitige *Gründe* (Interessen, Überzeugungen, mentale Dispositionen u.a.) mitspielen, gilt dabei erst recht bei starker Dynamik der Verhältnisse – weshalb die Bindungen der »Fügsamkeiten« nicht löchrig werden – und gilt, insofern neuartige Macht-Formen entstehen. Eine in Gestalten von Herrschaft geronnene Macht ist stets wesentlich auch ein Sich-beherrschen-*lassen*; und von dem her mag es, zumal unter Kräften starker Veränderung, auch so sein, dass das Imperativische (plus der Subjektspol, von dem das auszuführende ›Wollen‹ komme) weithin ›nur‹ imaginiert ist oder von irgendwoher eine betreffende Imagination induziert ist.

Man muss nicht behaupten, dass in Webers definitorischer Bestimmung, wie er sie für sein soziologisches Theorieprojekt fixiert hat, alles vollkommen klar ist. Für die Prozesse der akzelerierenden Überformung aller Verhältnisse durch Digitalisierung ist es sicher noch nicht gänzlich zureichend. Bereits angesichts der Wirklichkeiten seiner eigenen Zeit war auch Weber selbst bezeichnenderweise nicht ganz eindeutig. Technisches etwa verstand er generell noch rein instrumentell. Man wird darum nicht bei Weber stehenbleiben können. Doch, in Webers Bestimmung äußert sich eine Problemintuition. Und was immer sonst man ihm nachsagen will, so gilt gleichwohl allgemein, er hatte *Theoriebewusstsein*: Theoriebewusstsein angesichts der Unschärfen von Alternativen.

Dass »der Begriff ›Macht‹ […] soziologisch amorph« ist, muss keine entsprechend negative Bindung sein auch für eine kritische und philosophische Reflexion. Vielmehr, es gibt einen Hinweis. Dazu gilt es den Blick der Folgerung umzudrehen: Kritische und philosophische Reflexion zu Macht und zumal spezifisch zu einem Strukturstadium von Macht wird *nicht ohne* die Qualifizierung der betreffenden Formen von *Herrschaft* in differenzierter Weise ausführbar sein.[6] Zu dem aber gehört: den Intentionalitätszusammenhang zu bezeichnen, das Eingefügtsein bzw. Gekoppeltsein, sich als viele so zu verhalten, wie im ›Wollen‹ des machthabenden Pols gesetzt ist; zu bezeichnen, wie das Betreffende sich zur Herrschaft aufgeschwungen, eingerichtet und in sich verstetigt hat – was wohl niemals möglich ist ohne (bezeichenbare) Strukturen, Ausführungstechniken, vermittelnde Mitwirkende sowie intern gewisse Vorstellungen einer (herausgestrichenen oder doch akzeptablen) ›Berechtigung‹ dazu; und eben auch die Ursachen und Weise der entgegenkommenden

6 Grenzen dessen, was es von Webers Projekt-Ansatz aufzunehmen gilt, liegen indes *nicht* darin, dass man ihn voluntaristisch o. dgl. interpretieren müsste. Ungeachtet des Theorierahmens und auch Webers Redeweise (»Wille«, »soziale Beziehung«, »Befehl« usw.) ist das Potenzial seiner Überlegung gerade hier bei diesem Themenbereich ›Macht‹ nicht einseitig im Sinne der üblichen Etiketten wie ›handlungstheoretisch‹ (bei gezeichneter Großalternative Handlung vs. System) oder ›verstehende Soziologie‹ (gegenüber Erklärungen von Gesellschafts-Prozessen und darin Kausalursachen).

Fügsamkeiten zu bezeichnen, weshalb die Beziehungen nicht auf spürbarer Gewalt beruhen oder auf ›natürlicher‹ gewordener Autoritätsanerkennung oder Vorbild-Hinaufblicken. Weber hat damit Bedingungen auch an ein für veränderte Strukturen von Macht nachrückendes Verständnis vorgezeichnet. Von dem mit Weber als Problem Offengehaltenen aus zeigt sich das Neue umso deutlicher.

III Neue Weisen von Macht (Enteignung von Möglichkeitsraum und Möglichkeitshorizonten)

›Macht‹ war, im Bisherigen, elementar ein Emanzipationsbegriff. Die Entfaltung der Reflexion über ›Macht‹, in den Bedeutungsdimensionen, gehört in jeder ihrer Etappen zu einem Diskurs, sich über Grenzlinien und über Dominiertwerden durch Fremdes zu verständigen – Berechtigtes wie Erwünschtes zu rechtfertigen und Bestehendes zu kritisieren.[7] Die Problemerfahrung, die hinter ihm stand, war stets eine, von anderem beherrscht zu werden – eingetreten oder drohend oder imaginiert –, Erfahrung, von Möglichkeiten abgehalten zu werden. Sie hat sich in Gestalt eines je entstandenen Bewusstseins der Fesselungen und Gefährdungen dem Kosmos von Denken und Handeln eingeschrieben. Im Sensorium, zunehmend und darin in der Vielschichtigkeit sich entwickelnd, wurden in diesem Sinne Horizonte von Widerstandsmöglichkeiten zu Bestandteilen des Verständnisses und in Begriffsfeldern der sozialen und psychologischen Sprache markiert.

Dass Veränderungshinsichten wie ›Digitalisierung‹ als solches bestimmte, vollends neuartige Formen von Macht bringen könnten, bei denen bestehende Sensorien und Begriffsverständnisse, ein bisheriger Stand, nicht mehr genügend greifen sollten, könnte sich hier als weniger plausibel herausstellen – oder: an anderer Stelle gelagert zu sein –, als im Vorverständnis eines Alltagsgefühls sofort erwartet. Entfaltet sich doch auch die oft metaphorisch beschworene ›Macht der Algorithmen‹, eingeschlossen der Digitalisierungs-Fortschritt und der Digitalisierungs-Druck, noch allemal – jedenfalls bisher – im Rahmen menschlicher Praxen und dessen, was ihnen darin zugeteilt wird. Wie alle menschliche Lebensgestaltung haben auch Digitalisierungen eine Seite des (scheinbar indifferenten) Instrumentellen – ein von Menschen ›Erfundenes‹ und Entwickeltes (sowie Implementiertes) und ein durch das Faktum der Benutzung Anerkanntes –, und wie alles Technologische wären sie darin vorderhand ›nur‹ eine Effektivierung, Vereinfachung und Verdichtung

7 Das gilt im Letzten auch für die vielerlei Thematisierungsgestaltungen in der Bedeutungsdimension subjektloser Macht – kritische Sensorien und Kategorien wie Argumentformen zu bilden gegenüber einer Dominanz des Anarchisch-Ungefügten, kontingent Destabilisierenden usw.

richtiger und falscher Handlungspläne und Praxen – oder in einem Transformationsmodell: Stufen eines allgemeinen Prozesses der ›Rationalisierung‹ –, aber per se kein qualitativ neues Stadium.

Nichtsdestoweniger sind auch Gefühle hier eine Realität. Dass für das Eingespielte des kritischen Bewusstseins das Neue der digitalisierten Wirklichkeiten einen weiteren Problemdruck bedeuten könnte, sollte man nicht leichtfertig als Subjektivismen beiseitewischen. Man tut gut daran, in beide Richtungen, Verlängerung der instrumentellen Handlungsgestaltung und Lebensgestaltung oder aber neuer Äon, vorsichtiger anzusetzen. Im Sinne des von Max Weber Eröffneten wäre zuerst zu fragen, inwiefern in zunehmend digitalisierten Wirklichkeiten sich etwas gewandelt hat, das ein Beherrschtwerden *ist*, Außenbestimmtheit von (persönlichen, sozialen und politischen) Möglichkeitsräumen und Möglichkeitsvorstellungen – und im Weiteren sich zum Beispiel zu verselbständigen droht –, aber doch mentalerseits evtl. nicht adäquat und unambivalent einbefasst ist von den bisherigen, in der Geschichte der Problemerfahrungen herausentwickelten Sensorien des kritischen Bewusstseins. Darum noch einmal, nun konkret: Braucht es für neue Verhältnisse und Prozesse von Macht ein verändertes, weitergehendes *Konzept* von ›Macht‹? – Vor hoher Theorie sei dazu der Blick in die Breite gesetzt, eine Verortung. Alles Offenkundige, was es dabei auch schon in die Talkshows, ins Feuilleton der Lebensberatung und auch in die Gesetzgebungsüberlegungen gebracht hat, sei dafür erst einmal zurückgestellt.

Das Phänomenale lässt sich zu vier Problemfeldern gruppieren. In einer Welt, in der Digitaltechnisches bestimmend geworden ist in allen Prozessen, gibt es (1.) augenfälligerweise *neue ökonomische Macht*: Monopole oder Quasi-Monopole in einer neuen Dimension, einem jeweiligen neuen Bereich von in praxi fast unabdingbar Erforderlichem – von Geräten, Services, Dienstleistungen, Informationen, was für die Teilhabe am Sozialen (und in einem weiteren Sinne: die innergesellschaftliche Kommunikation) gebraucht wird. Gerade *durch* die rasante Erweiterung von Möglichkeiten, was das Digitaltechnische erbringt, ergibt sich die faktisch zu Monopolen sich entwickelnde Macht daraus, das dafür Erforderliche zur Verfügung bieten zu können. Es gibt diese Monopol-Macht in einem Ausmaß, das im Früheren des 19. und 20. Jhs. im Kapitalismus oder Staat schon längst zur Zerschlagung, Aufsichtsreglementierung, Einschränkung von Marktbeherrschung und geschützter Patent- wie Copyrighthoheit geführt hätte (›im öffentlichen Interesse‹); man hätte diese Parallelmacht nicht zugelassen. Nun aber erschweren dies schon die geballte Aggregation der interessierten Nutzer*innen, und dass etwas Benötigtes (oder als lebensnotwendig Gedachtes, oft auch Vorgespiegeltes) ein Proprietäres ist, welches das allgemeine soziale Leben dabei gerade immer stärker in seinen Kosmos verstrickt und an sich bindet.[8] Die heutige Strukturlage manifestiert sich in der Hilflo-

8 Die Abhängigkeiten gelten ja auch für die Instanzen der Öffentlichkeit.

sigkeit der Institutionen, der Hilflosigkeit einer Gegenmacht des Gesellschaftlichen und Politischen.

Macht, die sich primär aus dem Ökonomischen etabliert, besteht ferner durch das neue Herangreifen an unsere Entscheidungen als ökonomischer Subjekte: durch *Tracking* und zielgenaue Lockungen – und gekaufte ›Influencer‹, die sich wie unsere Freunde darstellen, sowie ›Experten‹ des sozialen Erfolgs und der persönlichen ökonomischen Lebensoptimierung – unser ökonomisches Verhalten anzuheizen sowie zu steuern. Alles changiert zum Werbeumfeld, bei dem entsprechend potente Akteure sich einkaufen können in den Zugang zu unseren privatesten Bedarfs- und Interessenkontexten, in die Wahrnehmungsbegleitung unserer persönlichen Horizonte.[9] Alle wollen scheinbar nur unser Bestes.

Parallel bildet sich zudem im allgemeinen ökonomischen Marktprozess neuartige Macht, indem das, was bisher, d.h. nachdem traditionale lokale Monopole und Verbindlichkeiten überstiegen waren, als sich selbst zum letztlichen Nutzen austarierende *Pluralität* (Pluralität aller Seiten der Wirtschaft) jedenfalls leidlich funktioniert hatte, in vielem den entstandenen neuen globalen Monopolen oder Quasi-Monopolen von *Plattform-Strukturen* weicht. Gegen die bisherigen Weisen der Märkte, gegen die pluralistischen Beziehungen der Akteure des Handels mit Waren wie Dienstleistungen kam es in vielem zur monopolistischen Verdichtung von Plattform-Macht, die Bedingungen diktiert, zu einer schnellen Marktverdrängung von früheren Wegen und Vermittlungsakteuren führt und meist auch erhebliche Anteile der Gewinne für sich abschöpft. – Und zu den neuen Weisen von ökonomischer Macht gehört nicht zuletzt auch das, dem alles unterworfen ist: die neue *Un*sicherheit – im persönlichen ökonomischen Stand viel direkter abhängig geworden zu sein von der *Irrationalität* bei der Vervielfachung der vermeint nur ökonomischen Rationalität des Systemgeschehens im Großen, Irrationalität aus der Vervielfachung und Beschleunigung der scheinbar rein ökonomischer Rationalität folgenden Kleinschritte. Es ist, selbst ohne Kriege oder Naturkatastrophen, die *Un*sicherheit unseres im Leben erworbenen Wohlstands und der Vorsorge, Unsicherheit durch die Anfälligkeit und Überhitzung der Finanzmärkte im Zustand der digitalisierten Welt. Und was hier für die Einzelnen gilt, dem sind in gleicher Weise ebenfalls die sozialen Verbände und auch die Staaten unterworfen.

9 Zu dieser Macht gehört, dass wir *gecatcht* werden durch unsere geglaubte Schlauheit: bei gezeichnetem Zeit- und Konkurrenzdruck einen persönlichen Vorteil, eine einmalige Chance nicht zu *versäumen* (›Sichere dir schnell …!‹, ›Verpassen Sie nicht …!‹, ›Nur noch heute!‹, ›Gewinne …!‹, ›Habe es als Erster …!‹, usw.). Appelliert wird an die ›Schnäppchen‹-Mentalität in allen Dingen, die sich dabei herausgebildet hat. – Verstärkend steht das gezeichnete allseitige Bedrohungs- und Möglichkeiten-Szenario (ebenfalls die Affekte in Bezug auf ein Negatives ansprechend: ›Riskieren Sie nicht …!‹).

In der bestehenden Welt der zunehmend durch Digitaltechnisches geprägten Wirklichkeiten gibt es (2.) neue Macht als *neuartige Weisen von Disziplinarmacht*. Darunter fällt nicht nur die manifeste Kontrolle, die unter dem Zeichen, uns *vor einander* zu schützen – vor den heimlichen gesellschaftsgefährdenden Subjekten im Sozialkörper –, in jedwedem Staat sich immer weiter ausweitet.[10] Die Angst vor dem Mitmenschen, die in einem anonym gewordenen gesellschaftlichen Raum sich ausbreitet, die geschwundene Form der Vertrauenserfahrungen, verschafft dem in vielem auch die Akzeptanz, die implizite Zustimmung, die die Kontrolle als Allgemeinfaktum des Gesellschaftlichen soweit nicht als Repression empfinden lässt. Neue Disziplinarmacht erwächst ebenso in Gestalt des überall angesammelten Wissens über uns – selbst das Privateste, das wir, ohne dass durch offene Gewalt oder psychologische Manipulation, preisgeben oder das unschwer zu einem Profil abzuschöpfen ist.[11]

Vor allem jedoch ist es die schleichend sich einstellende Macht zur *Konformität*, welche sich schlicht durch die (weithin vorgegebenen) Modi-der-Teilhabe und das allseitige begierige Teilhaben-Wollen – zum Teil freilich auch ein Teilhaben-Müssen – zur Wirkung bringt. Alles wird zum Markt, zum Marktmäßigen. Aus Möglichkeiten, die uns das Digitaltechnische eröffnet für die Gestaltung unseres sozialen Orts, wird schnell die Selbst-Performanz, ja der Zwang dazu. Aus der Assistenz der Verwirklichung wird der Vampir meines Lebens – in immer weiter gehenden Bereichen die Zumutung, die digitalen Möglichkeiten mit dem Leben meiner Person zu füttern.[12] Das Ich muss den Bedingungen und Kriterien einer betreffenden Markt-Konformität (und nicht zuletzt dem überwältigenden Zeit-Takt dabei) nachkommen. Mitzumachen getrieben wird es durch die Ängste des Ausgeschlossen-seins, des Nicht-mithalten-könnens (sowie umgekehrt Phantasien des schnellen Aufstiegs, der Selbstbeförderung in eine höhere soziale Welt, sei es eine des Prestiges, eine der ›Insider‹ oder die der Erwachsenen).

Die Konformitätseffekte werden verstärkt durch die Beeinflussung, die *durch* Wissen und Bilder erfolgt: durch deren schiere Mengenmacht, die – und sei der Sachverhalt oder die Community relativ klein – zeichnet, was ›normal‹ ist. Auch in diesem Wissen und den Bildern liegt Disziplinarmacht. Denn zugleich ist es ein unbestimmter, offener, unendlicher Marktplatz von Verkörperungen des Anspruchs

10 Das heutige China ist davon nur das extreme Beispiel.

11 Auch in der verzerrenden Gestalt: Angesichts der Menge von erreichbaren und mit irgendeiner digitalen Quelle augenscheinlich belegbaren Informationen ist faktisch die Beweislast umgekehrt. Im Zweifelsfall muss stets der betroffene Einzelne belegen oder plausibilisieren, dass etwas *nicht* (oder nicht in relevantem Maße) zutrifft; oder muss die Schwelle, sich dem Zudrängen zur Wehr zu setzen, immer weiter erniedrigen (oder das, was ihm*ihr ›etwas ausmacht‹, dass das über ihn*sie geglaubt wird).

12 Das betrifft nicht nur die sozialen Medien. Sie sind da nur ein, freilich wichtiger Bereich davon.

auf Wertwichtigkeit und Bedeutsamkeit-für-andere. In ihm nimmt entsprechend die Abhängigkeit, Resonanz *zu bekommen* – Resonanz spüren zu wollen, ersatzweise auch, sie sich vorzustellen –, überhand, und mit der Offenheit und Unbestimmtheit zählt dabei in der Allgemeinheit immer stärker die pure Quantität, nicht eine Qualität. Es herrscht dann: die Quote. Und die Menschen internalisieren die Quote.[13]

Schließlich kontinuiert sich der Konformitätsdruck durch eine kennzeichnende Verdrängungsmacht, mit der – selbstverstärkend die Zugehörigkeiten der Einfügung in die neuen Strukturgegebenheiten des Digitalisierten – die Altinstanzen der Sozialisation zum Teil *ent*machtet werden. Die klassischen gesellschaftlichen Instanzen von Erziehung, politischer Gemeinschaft, ›öffentlichem Leben‹, ›Kirche‹, Wissenschaft sowie auch ein Großteil von sozialen Zusammenschlüssen und Engagement, die Instanzen, welche in modernen Verhältnissen – entsprechend dem kritischen Bewusstsein für berechtigte Autorität und ›Macht‹-in-falschen-Händen – unter dem normativen Ideal (Funktionsideal) stehen, nicht eigentlich Konformitäten zu disziplinieren, sondern von den Konformitätsvorgegebenheiten von Herkunftsmilieus und Sondergemeinschaften *zu befreien*, werden bedrängt durch neue Communities der Sozialisation und des Rollen-Verhaltens. Es haben sich Parallel- und Alternativformen gebildet, die subjektiv ›einfacher‹ erscheinen und leichtere/schnellere Anerkennungsresonanz versprechen. Mit ihren Einflüssen, mit denen sie neben das etablierte Gesellschaftliche treten, spielen sie eine Anziehung aus, die, was immer sonst, jedenfalls nicht pluralistisch ist, sondern ist, ein Gruppen-›Wir‹ vorzustellen – häufig gerade wieder fragmentarisiert, und durch Abgrenzungen sich bestimmend.

Als weiteres grundlegendes Problemfeld stehen (3.) die – weithin indirekten – neuen Macht-Relationen, die mit der *Vervielfachung der ›Akteure‹* bzw. Akteurszellen kommen. Es ist eine wesenhafte *Veränderung des Handlungsraums überhaupt*. Dies gilt für hinzukommende ›künstliche Akteure‹ – Datenmengen analysierende Algorithmen, Sachverhaltsrubrizierung durch elektronische Mustererkennung, Codes von Programm-Verhalten, Bots, vernetzte Automatismen, Entscheidungs-Vorgaben, usw.[14] –, aber auch für humane Subjekte in ihrem Agieren in den oder vermittelt durch die digitalisiert technischen Systeme und die dadurch formatierten diskretisierten Einzelakte. Der Handlungsraum geht mehr und mehr hinaus über den Kreis der identifizierbaren, in vielfältigen Kommunikationsbezügen sich

13 Das gilt nicht nur für Heranwachsende (in ihrer Orientierungssuche und Suche nach ihrem anerkannten Platz in einer Gemeinschaft) oder sonst Menschen mit wenig ›handgreiflicher‹ sozialer Interaktion und Rückmeldung. Es erwächst schlicht überall dort und darin, wo mangels anderer erlebter Orientierung, und dazu zählt auch deren Stabilität, dies überall stehende, uns umstellende Wissen und diese Bilder Dominanz bekommen.

14 Dabei macht es für diesen allgemeinen Sachverhalt vorderhand keinen Unterschied, in welchen Anteilen das Digitale der ›künstlichen Akteure‹ hineinprogrammiert ist durch menschliche Subjekte oder ›selbstlernend‹ ist.

bekannt – einschätzbar, gar vertrauensvoll – machenden und ›verantwortlichen‹ Humansubjekte (sowie wirtschaftlichen, ›weltanschaulichen‹, gouvernementalen Organisationen). Statt Handeln von komplexen Wer-Charakteren (Einzelne, Gruppen, Organisationen, Institutionen) kommt es zunehmend zu einer Zersplitterung in einzelne Verhaltenszüge; und die werden dabei nach dem Muster von *rational choice* interpretiert. Umgekehrt werden nicht komplexe Motivationshintergründe erkannt (oder gesucht, erwogen), sondern der Grund der Akte wird mit einem betreffenden *Funktions*zusammenhang gleichgesetzt. Vervielfachung der ›Identitäten‹ und Diffuswerden des Handlungsraums sind zwei Seiten derselben Wandlung.

Verantwortung muss dabei oft nachsimuliert werden und in die Intentionalitätsverhältnisse wie -horizonte humaner Subjekte nachimplementiert werden. Und aufseiten des jeweilig digitaltechnisch Neuen, Hinzukommenden, auch wenn noch zurückverfolgt werden kann, wer (und wie) es ›ins Spiel‹, in den Stand der Praktiken gebracht hat, bleiben im Konkreten die Akteure dabei weithin im Dunkeln. Alles, was nicht als manifeste (Fremd-) Interessen oder Machthandeln erkennbar ist – oder sekundär diskursiviert ist in dieser Weise –, erscheint als das Je-Funktionsadäquate in eben einer Handlungsumwelt, d.h. deren Faktizitäten, ansonsten neutral (neutral Instrumentelles), ohne Voraussetzungen. Statt der meist im Dickicht verschwindenden und unbelangbaren menschlichen Akteure der Entwicklung, die etwas an Funktionsorte bringen und kontrollieren bzw. steuern, erscheinen Veränderungen (und deren Tribut, Forderungen, Folgen) als das Funktionsbessere, Funktionseffektivere, -schnellere, -komfortablere, -erweiternde: ein sachnotwendiger ›Fortschritt‹, Schritte eines sachlogisch, funktionslogisch Notwendigen.

Als eine eigene, zusätzliche Dimension des Phänomenalen steht schließlich das, was man formelhaft unter (4.) der *Macht-der-Geschichte* fassen kann. Alles, was *nicht* dem Gang des – scheinbar – sich selbst prozessierenden ›Fortschritts‹ folgt, *nicht* mit der Entwicklung des beständig Noch-weiter-gehend-Neuen mitgeht und sich dem jeweilig Avanciertesten assimiliert, ist in den digitaltechnisch gewordenen Wirklichkeiten in einem hohen – und meist zu spät kommenden – Maße begründungsbedürftig.[15] Das betrifft Waren-Produkte, Geräte, Leistungen und auch Menschen mit ihren Fertigkeiten, Kompetenzen, Ideengehalten, normativen Kriterien und den vertrauten (und innersozial vertrauensbildenden) Umwelten des Gewohnten. Die Seiten sind vertauscht. Die alte Errungenschaft, das bisher begründet Bestehende und Sozial-Geteilte hat kein inneres Recht zu existieren. Und oft sind schon längst Fakten geschaffen – Fakten neuer Umwelten des Handlungsraums und des Lebens –, bevor auch nur bewusst wird, worauf zu achten und was zu reglementieren bzw. sozial zu flankieren wäre. Vielleicht zum ersten Mal hat

15 Einzelne Retro-Tendenzen wie im Kulturellen, zum Beispiel als prestigeträchtiges Sich-Herausheben aus der Masse die Wiederentdeckung der Analogfotografie oder der Schallplatte, ändern daran nichts.

der Topos von der ›Macht der Geschichte‹ wirklich eine massive reale Bedeutung. Die Macht der Geschichte ist die Macht der Verdrängung; und die Macht, dass alles sich an der Spitze zu sammeln drängt, jedes, was nicht an der Spitze der Entwicklungsveränderungen mit dabei ist, ins Hintertreffen gerät (oder sich zumindest so empfindet). –

Neue ökonomische Macht, neue Disziplinarmacht, neue Macht-Relationen durch ›künstliche‹ Akteurspole und die Zersplitterung der Handlungswelt in einzelne Verhaltenszüge, neue Macht-der-Geschichte: sie alle sind hier eine Macht, die *anonym* ist, strukturell wie auch in ihren Prozessen anonym. Einzelne wie Gruppen sind höchstens Exekutoren solcher betreffenden Macht; sie bekommen nicht ihrerseits dadurch besondere, digitalcharakteristische Macht, sind vielmehr ihrerseits ebenfalls den Bedingungen dieser Macht unterworfen, nur gleichsam mit verschieden günstigem Los dabei. Interessenträger und Bevorteilte bei diesen neuartigen Formen von Macht müssen auch nicht machthungrig oder bes. skrupellos sein. Das Außen-Bestimmtsein in Möglichkeitsräumen und Möglichkeitsvorstellungen, in dem ein Beherrscht-*werden* sich vollzieht, hat in diesen Formen strukturell kein Gesicht.[16]

Einen Teil bei diesem Vierfachen, das beträfe vor allem die negativ gerichteten Effekte, könnte man wohl mit ganz klassischen Konzepten von ›Macht‹ zumindest beschreiben – in Einzelphänomenen als Entmächtigung alter Sozial- und Subjekterrungenschaften. Für weitergehende Thematisierungsperspektiven und theoretische Folgerungen, solange man nicht bewusst darauf verzichten will, den Typus des mit Digitalisierung kommenden Neuartigen als solchen zu begreifen, reicht dies jedoch nicht. Je mehr es zu fragen gilt, was die »Fügsamkeiten« entstehen lässt und ausmacht, kommen die klassischen Konzepte von ›Macht‹ hier an theorieprinzipielle Grenzen.

IV Der Druck auf das Subjektive (Die Prozesse von Ermächtigung und Entmächtigung)

Bereits bei Max Weber war gesehen, wie es typologisch ganz verschiedene Weisen von Gründen sein können, die Fügsamkeiten bewirken. Neben dem schlichten Hineinwachsen in eine Sozialwelt der von den anderen Subjekten geübten Praktiken und zugewiesenen Positionen, neben mithin erfahrenen Erwartungen und dann stummer Gewohnheit in einem bestehenden Es-ist-so und So-verhält-mansich, spielen primär in vielem auch abgeschätzte eigene Vorteile des Mitmachens (oder übergroße Nachteile bei Nicht-Einfügung), habitualisiert zu gelernter kalkulierter Anpassung mit, ferner fehlende oder nicht (bzw. nicht mehr) bewusste

16 Bzw. bekommt dies nur in Gestalt von eventuellen sekundären (Verursacher-) Projektionen.

Alternativen. Und zumeist relativ nur begrenzt ist der Faktor ausschlaggebend, den klassische (›philosophische‹) Theorien der normativen Begründung sozialer Formen und Ordnungswelten zum Zentrum setzen: ein ›Ethisch‹-Normatives, die gewisse Einstimmung in das Rechte und die ›Legitimität‹ einer Herrschaft bzw. allgemein Asymmetrie.[17]

Dies, was Weber in Bezug auf politisch verfasste Ordnungen theoretisch typisiert hat, ist noch nicht das Muster für auch die Gegebenheiten in durch Digitaltechnisches geprägten Handlungswelten (und zumal deren sich beschleunigende Selbstveränderung). Was Weber, und hierin exemplarisch für die meisten soziologischen Thematisierungen von einstigen ›philosophischen‹ Fragestellungen, für den Horizont des Zwecks seiner universalen Theorie nicht mit einbezogen hat, sind: die Herausentwicklung eines *kritischen Sensoriums* in Kulturen (einschl. der Sensibilitäten, die in sprachlichen Begrifflichkeiten und Differenzierungen der Verständigung geronnen sind), die material-mentalitätshistorischen Prozesse, dass in Praktiken und dem Umgang mit betreffenden, die Lebensumwelt ausmachenden Gerätschaften offenkundigerweise zugleich bestimmte Selbstbeschreibungen und Ideale generiert werden, und überhaupt den Faktor, was die Subjekte über ihr Handeln und sich als Handelnde (sowie Motive) denken.[18] Das bringt eine Unbestimmtheit der Reflexion. Und in genau diesem Bereich vollziehen sich ausschlaggebende der Auswirkungen des Geschehens der Digitalisierung für das Soziale. Die mit Digitalisierungs-Entwicklungen kommenden veränderten Lebenswirklichkeiten bedeuten in der Tat eine neue Konstellation. Es ist denn nicht einfach die Erweiterung von Verhaltensanpassung wie kritischer Bewusstheit angesichts von Prozessen von Macht und Herrschaft, sondern wesentlich auch strukturell neu bei Handlungs-Verhältnissen und Bewusstseins-Formen, bes. bei den Quellen der Fügsamkeiten.

Dem erforderlichen Blick steht gerade die Erfolgsgeschichte bisheriger Herrschaftskritik und Emanzipationsentwicklungen, deren Muster, entgegen. Ausgehend von den bisherigen Problemerfahrungen mit Macht und der Fremd-Bestimmtheit von Möglichkeitsräumen malt gemeinhin die Kritik der Verhältnisse, in denen die digitaltechnischen Veränderungen sich ausprägen, vom Gewohnten her ein Bild, dass sie Herrschaft sind in dem bzw. über das, was den unterliegenden Subjekten *vorenthalten* ist, ein Drängen zu einem nur Begrenzten, das diese allenfalls bei Strafe von Sanktionen übersteigen könnten – diese Negativmacht ausagiert von den Big Playern wie auch den kleinen Betreibern der digitaltechnischen Systeme und ihres beständigen Entwicklungsdrucks. Diese Diskursivierung sucht ein bestimmtes ungutes, jedenfalls unsoziales, polarisiertes Wollen auf Seiten der die Herrschaft Ausübenden namhaft zu machen. Sie denkt in Wollen-gegen-Wollen. Es

17 Vgl. Weber 1980[1922]: 122f., 20.

18 Zudem ist es eine Typisierung im Blick auf weithin relativ statische Ordnungstraditionen, nicht eigentlich für dynamische Verhältnisse, gar solche beständiger rasanter Umbrüche.

ist ein Modell antagonistischer Akteursintentionalitäten. Die *subjektiven Horizonte und Einstellungen* all der großen Massen derer, die in den Handlungswirklichkeiten dieser sich verändernden Welt leben, mit dem immer mehr hinzukommenden Digitaltechnischen umgehen, sind indes nicht unberührt. – Es ist der Blick auf dies, der die starken Kräfte der Fügsamkeiten verständlicher macht. Sie formen sich hier daraus, dass auch die Herrschaft im Wesentlichen an anderem ansetzt als bei dem, was (bisher) Macht bei einer politischen Ordnung ist.

(1) Auf die Subjekte gesehen, ist das Technische, das Digitalisierung bringt, und ihr allgegenwärtiges In-Reichweite-kommen in der normalen Lebenswelt *Er*mächtigung und – genau damit – zugleich Prozesse von *Ent*mächtigung. Denn einerseits ist es die Lockung mit dem, bestehende Macht-Verhältnisse, Macht von Natur und in Gesellschaft, umzudrehen (oder unterlaufen zu können). Teilzuhaben an den Errungenschaften des Digitaltechnischen sieht für die Einzelnen wie auch soziale Unternehmungen und die Gesellschaft so aus, als sei es keine Herrschaft – allenfalls eine vernachlässigbare strategische Paktierung mit Erfindungseignern und Lieferern –, sondern als würde man vor allem Herrschaft *bekommen*: erweiterte (instrumentelle) Herrschaft über bisher Bedrängendes[19] und zu neuen Handlungsräumen; und als werde dies erstmals in einer nicht von angestammten gesellschaftlichen Macht-Positionen und Privilegien dominierten Weise möglich, fair und demokratisiert. Neben dem Instrumentellen ist es vor allem sozial auch die Aura des Egalitären – dass jede*r es bedienen bzw. handhaben kann (nach kleiner Einübung),[20] das Aufbrechen der Geschichte mit ihren angesammelten strukturellen Chancenungleichheiten, die Befreiung von den Einschränkungen aufgrund von Herkunft, Status, Geschlecht, Bildung, Alter.

In der Perspektive all der Masse derer, denen sich dies neue Instrumentelle bietet, erscheint es als die endliche Einlösung des modernen Subjekts. Die Handelnden, sie alle, befähigt zu machen zu Möglichkeiten und Kontrolle des je Eigen(st)en, Kontrolle auch über das, was bisher das Emanzipative der modernen Ideen eingeschränkt hatte, erscheint hier als die Erreichbarkeit dessen, was in der Geschichte nur Utopie war: eine Welt der Freien und Gleichen, in Entscheidung nach ehrlichen Mehrheiten, und je um je Wahl des unvoreingenommen Besten (geschichtlich Besten). – Das ist denn das eine. Die mit dem immer weiteren Digitaltechnischen durchwobenen Wirklichkeiten bringen Gegebenheiten der dem Verfügen offen scheinenden Handlungswelt, die subjektiverseits so aussehen wie Selbstmacht und Könnens-Macht, nun exponentiell erweitert. Und dieser Rahmen des Lebens scheint darin die *Erfüllung* all dessen, was die aus bisherigen Problemerfahrungen erwachsenen Sensorien als die Aspekte und Strukturen zu Bewusstsein gebracht haben, die widrigenfalls den Widerstand gegen Zumutungen des Sicheinstellend-Be-

19 Darunter auch Gegebenheiten des Lebens, angefangen bei Zeit.

20 Und dass dies Leichte auch ein entscheidender Zweck ist.

stehenden hervorrufen würden: Erfüllung für die Befähigung zu Selbstmacht über *mein eigen* Bestes (Wissen über mich, meinen Körper, meine skalierte Leistung und Möglichkeiten; Lebensempfehlungen; Kenntnis der Pluralität von Optionen); Mittel, um mich abzuheben von der Masse (Abhebung in Konsum, in stark affektiv besetzten Genuss-Vorlieben (Musik, Filme usw.) und in Performing); das Überall-mich-mit-sichtbar-machen-können sowie -dürfen; die Souveränität (gleich allen anderen) im – als selbstverständliches Sozialfaktum angesehenen – Wettlauf um gegenseitige Übermächtigung und Übertrumpfung unter den sozialen Akteuren; das *Spielen* mit Optionen und in erdachten Szenarien;[21] die Unabhängigkeit, Bindungen und Verbindlichkeiten relativ leicht und des Öfteren zu wechseln (auch eigene Bindung an mein jeweilig vorhergehendes ›Ich‹); und nicht zuletzt das, präsent zu sein – und Kontrolle auszuüben – unabhängig von Zeit und Raum.

Diese anmutende Ermächtigung lebt aus dem Selbstverständnis neuzeitlicher Subjektivität. Der Sog der Einfügung in die digitaltechnischen Lebensumwelten pfropft darauf auf, zapft dies gewissermaßen an. (2) Die gezeichneten Möglichkeiten durch digitaltechnische Transformationen speisen dabei, und damit beginnen die Gegeneffekte, schon als solches auch Macht-*Phantasien*. Macht-Phantasien gehen über jeden Zustand tatsächlicher subjektiver Beherrschung nun stets weit hinaus, gehen aller Entwicklung immer voran. Sie schaukeln sich auf mit jeder neuen Erfindung/Entwicklung im Digitaltechnischen oder schon deren Ankündigung, und sie entziehen sich einem kritischen Empfinden, d.h. ob man etwas Betreffendes wirklich können sollte.[22] Vorstellungen, die mit dem neuzeitlichen Subjektivitäts-Gedanken, als theoretischem, historisch unbestritten verbunden waren[23] – nicht nur Unabhängigkeit von Diktaten und vorgegebenem Eingeordnet-sein, sondern Suprematie (oder Absolutheit) der (reifen, erwachsenen, welterfahrenen) menschlichen Reflexions-Bewusstheit oder Gestaltungs-Möglichkeiten der ihre Vernunft ergreifenden ›Menschheit‹ gegenüber Natur (Naturanteile in allem), übrigen Lebewesen, Sozialität sowie bisherigem hegemonialem ›Göttlichen‹ –, changieren dann leicht zu einer Bewusstseinswelt der Omnipotenz. Omnipotenz-Vorstellungen, wie

21 Zu den Prozessen der Ermächtigungen und ihrer Strahlkraft gehört auch das heute hohe Sozialprestige von: ›Kreativität‹ – dass wir durch die durch das Digitaltechnische gegebenen Möglichkeiten unsere ›Kreativität‹ ›entdecken‹, verwirklichen, ausleben könnten.

22 Und im Zweifelsfall erforderlicher Abwägung oder ›Kasuistik‹: sich der Reflexion zu entziehen, bis zu welcher Schwelle einer hinzunehmenden Folge oder eines in Kauf zu nehmenden ›Preises‹ und Schwelle der Vorsicht und Absicherung, zumal gegen zu spät merkliche Eigendynamik es gelten müsste, dass eine visionierte Macht-Möglichkeit nicht besser bewusst versagt, begrenzt oder zurückgehalten werden sollte.

23 Und die auch manche extreme Utopie genährt haben – vor allem in früheren Jahrhunderten, nämlich bis zum *Erreichen* der Möglichkeiten sowie größerer Unternehmungen der tatsächlichen realen Umsetzung und mithin Erfahrungen damit.

sie ansonsten die infantile Welt magischen Verhaltens ausmachen bzw. Zeichen entsprechender Regression sind, werden hier selbst im kleinsten Ego angefacht.[24]

Besonders aber machen sich die Einwirkungen auf direkt die Sensorien geltend, Einwirkungen, die zustande kommen durch eine Besonderheit der digitaltechnischen Welt, was es sonst nur umgekehrt in ideologischen Weltbildern der Anstrengung für eine *zukünftige*, neue Menschheit (oder ›Wir‹) gibt: dass die Einfügung in die neu kommenden Formen und Praktiken – und die dazu gehörenden Anpassungsforderungen, Zumutungen und Folgen – mit meinem eigenen wahren Willen identisch ist, meine Kollaboration für eine bessere, menscheneigentlichere, glücklichere Welt. Dies bringt Verzerrungen des möglichen kritischen Begleitbewusstseins. Gerade weil aus subjektiver Binnensicht in den Akten und Vorstellungen die Empfindung vorherrscht, als komme hier – in Ermächtigung und Erweiterung – allein *ihre* Intentionalität zum Austrag, wird das, was einst die Quelle und Basis der herausentwickelten kritischen Sensorien und des fallweisen Widerstands war, gelähmt: Die performative Selbstdeutung digitaltechnischen Handelns, des digitaltechnisch vermittelten oder assistierten Handelns, Agierens in diesen Lebensumwelten verschmilzt für die einzelnen, mit Gegebenheiten umgehenden Subjekte mit dem, was einst die herausentwickelte Bastion des *Reflektierens*[25] war. Sich zu verstehen als wollendes, selbstbestimmtes und auch verantwortendes, als reflektierendes und normativ einforderndes und auch selbstkritisches Subjekt, dieser Gedanke der Emanzipation von aller vorgegebenen und ohne Zustimmung herrschenden Fremd-Macht, ist in Breite eine Errungenschaft des neuzeitlichen Menschen. Hier aber, zu leben in den Welten des Digitaltechnischen und umzugehen mit den Mitteln des Möglichgewordenen, ist die Ermächtigung – die schnellere, erleichterte, entlastende Erreichung eines Intendierten und der Zweckziele – durch Einfügung und Mithandeln zugleich Entschwinden einer Potenz, weil Entmächtigung der begleitenden Errungenschaft des distinkten normativen Empfindungsvermögens.[26] *Das Gefühl wird amorph*. Der ›Subjekt‹-Gedanke verliert sein zugehöriges Sensorium für die Phänomene.

Von diesen Prozessen der subjektiven Horizonte und Einstellungen aus zeigt sich auch vieles Begleitende von Digitalisierung qua Geschehen in seiner ganzen Tragweite. Einiges Offenkundige, von dem manches auch schon verschiedentlich

24 Für kollektive Akteure, Betriebe und Verbands-Gruppierungen verselbständigt sich dies weniger, denn die Rückmeldung der harten Realitäten der Erreichung ihrer Ziele bleibt ihnen (außer bei punktuellen ideologischen Verblendungen oder Fanatismus) auch in der Epoche des Digitaltechnischen zu manifest. – Die neuzeitlichen, bisherigen Macht-Phantasien waren die des *Gemeinschaftlichen*, des *kollektiven Subjekts* (Gesellschaft, bis hin zur ›wir Menschheit‹).

25 Oder ggf. auch Refugium einer pragmatischen *reservatio mentalis*.

26 Mit der Potenzierung der Ermächtigungen wird sukzessive auch das Bewusstsein verdrängt, dass das wirklich volle (Selbst-)Verständnis, wie es als neuzeitliche Subjektivität sich heraus-

angemerkt worden ist, drückt die Energien des subjektiven Bewusstseins zusätzlich in einseitige Gestalt. Alles ist ein Druck auf das Subjektive, aus den Möglichkeiten zur Ergreifung der Mittel – und komplementär der Stummstellung der Diskursivierung, worin das Mitgehen im ›Fortschritt‹ von rationalnotwendig-sachdeterminiert erscheinenden Entwicklungen reflektiert wäre – ein Müssen werden zu lassen: So zunächst überhaupt, dass in entscheidenden lebensweltlichen Anwendungen bzw. Implementierungen alles *zu* schnell[27] an die damit Umgehenden kommt – Informationen, Entscheidungen, Updates etc., aber auch Sozialkontakte –, als dass das subjektive Bewusstsein und die subjektiven Verhaltensweisen nicht nur mitgezogen würden mit dem Strömen; es gibt zu geringe Lücken in der Zeit, als dass das Subjektive nicht weitestgehend nur funktionsadäquates Re-agieren wäre. Ebenso zweitens, dass der in alles hereinkommende Prozess der digitaltechnischen Transformationen mit seinem beginnenden Eintreten seine eigenen Standards setzt. Die Auswirkungen sind die genannten Ängste, je um je als einzelne, begrenzte Akteure *aus*geschlossen werden zu können von dem Großen, zum nur noch Objekt herabsinken zu können, Ängste, Erreichtes zu *verlieren*, weil alle anderen (resp. deren große Mehrheiten) mitgehen. Es ist der, und sei es subtile, unbewusste, Druck des ›Was passiert mir sonst, wenn ich *nicht* Gleiches betreibe (oder mir zu eigen mache) oder noch stärker?!‹ – alle rüsten auf in der Ermächtigungsmöglichkeit, machen sich beständig besser. Als ein dritter maßgeblicher Bereich solchen Drucks wirkt herein, dass in einer Welt des bestehenden Digitaltechnischen dies, in seiner Faktizität, im Konkreten – mitsamt Praktiken, (Ziel-)Ideen, Sozialverhalten und -empfinden – zugleich für alle Nachwachsenden (oder auch die Orientierung-Suchenden) eine *neue Basissozialisation* ausmacht, gegen die vormalige Subjekt-Ideale einer Erziehung zur Mündigkeit oder von ›Gebildet‹-sein allenfalls hinterherrennen können. Und schließlich, ergänzend von der anderen Seite her, den Gegenhalt gegen all solchen Druck schwächend, gibt es weithin nichts ›Religiöses‹ (Religiös-Transzendentes) oder Menschenbilder mehr, was mit dem eingetretenen Geschehen nicht kompatibel wäre. Alte Erbschaften eines irgend Absoluten sind entschwunden, wurden zum Teil schon Jahrzehnte vor Eintreten der Phase des neuen Digitaltechnischen einem Relativismus-Bewusstsein preisgegeben. Auch das ist eine flankierende Realität.

(3) Unter diesem mehrfachen Druck bildet sich im Ganzen ein ambivalenzverdrängendes Zustands-Bild, und das heißt auch: Geschichts-Bild. Darin ist subjek-

gebildet hat, auch noch die andere Seite hat: nicht nur das im Sinne des (gesellschafts- und ökonomietheoretischen) ›Liberalismus‹ und der mentalistischen Letztbegründung egologische (oder solipsistische), um sich kreisende, rationalitätskalkulierende, strategische, kompetitive Subjekt ist, sondern zugleich mit Ansprüchen an Empfindungsvermögen, innerliche Steigerung und Differenziertheit sowie an tiefere Sozialitätsbezogenheit.

27 Bzw. umgekehrt: in ihren Mengen *zu viel*.

tiv ein zunehmender *Verlust des Kontrafaktischen* eingetreten. Der Möglichkeitssinn kennt dann nur noch die Richtung des faktisch schon Eingetretenen, als Erwartung oder Vision der instrumentellen Ausweitung, der weiteren Potenzierung. Das Geschehen der Digitalisierungen, ihr unablässiger Fortgang, erscheint so sehr als ein Eines und Gesamtes, dass – zumindest im normalen lebensweltlichen Bewusstsein – kaum mehr differenziert werden kann, keine differenzierenden Reflexionen des Was-wäre-wenn…, keine grundsätzlich andersgearteten Szenarios mehr, anders denn die faktischen Ermächtigungs-Ideen und -Ideale, und keine Achtsamkeit gegen Analogieverallgemeinerungen. Es ist keine Alternative (Geschichts-Alternative) mehr vor Augen, deren reales Pendant nicht im Niedergang wäre.

Mit allem ist, psychologisch, der Spielraum eingeschränkt, bei den Ausprägungen von Digitalisierungsprozessen ein Bewusstsein zu entwickeln (oder auch: zu bewahren), dass es sich nicht um ein zu erfüllendes Sein handelt, sondern um Sachverhalte der Wertung – um individuelle wie gesellschaftliche Wertungsmöglichkeiten. Das Ineinander von Ermächtigung und Entmächtigung wird dann vollends zur hingenommenen Tatsache der humanen Existenz überhaupt. Durch den Umgang mit den digitaltechnischen Möglichkeiten, und indem diese eine neue Lebenswelt bilden, kommt es schleichend zur Aushöhlung des Gefühls für: Gesellschaftlichkeit, für die Dynamik von Struktur-Prozessen (›systemische‹ Realitäten), allgemein für Bedingtheiten durch Etabliert-Herrschendes. Alles wird verprivatisiert, alles sich selber oder anderen Einzelnen zugeschrieben, ansonsten der untrüglichen neutralen Rationalität der digitaltechnischen Systeme (bei konfliktweise dann wiederum eigentümlich schnell abgeschoben auf eine punktuelle Zufalls-Irrationalität ›der Systeme‹). Über das tatsächlich Klein-Machende hinaus kommt es so zugleich im Elementaren zu Empfindungen gerade der Macht*losigkeit*.[28] Empfindungen fehlender Aktions- und Gestaltungsmöglichkeit bilden sich, neben der manifesten Gewichteverteilung bei den einhergehenden Konfligierungen,[29] dabei im Wesentlichen auch bei entsprechenden *Narrativen*.

---

28 Das Klein-Machende ist nicht nur, dass – von der schieren Potenz der Datenprozessierungs-Leistungen noch ganz abgesehen – gegenüber den (digital-) technischen Instrumenten und ihren Möglichkeiten bzw. Angeboten von Möglichkeiten ein Gefühl des menschlichen Ungenügens sich einstellt und wächst, sondern ebenso in den Horizonten der Sozialität. Der Vorstellung etwa, *gesehen* zu werden (in all den Möglichkeiten, sich digital präsent zu machen in einem Forum von Wichtigkeits-Gemeinschaft), der Vorstellung, ein *Wer* zu sein, steht der unendliche Vergleich gegenüber – und mit ihm die strukturelle Stresskondition, sich nur in idealisierter Gestalt zeigen zu können und trotzdem mit dem gegen das eigene Sein und Leben, d.h. was man ›aus sich gemacht hat‹, gerichteten Verdacht, schlecht abzuschneiden, oder jedenfalls nicht gut genug.

29 Bis hin zu den klassischen liberalen Wohlfahrts-Verständnissen des Politischen: ›Die Menschen *wollen* es doch!‹, ›Wir können es als Verbesserung ihnen doch nicht vorenthalten!‹, etc.

Das sind die sich einschleichenden Narrative, dass es real keine Möglichkeiten der grundsätzlich ausscherenden Gestaltung gäbe; oder je noch nicht entschieden werden könne, ob und wie ein eingreifend-korrigierendes oder -flankierendes Handeln nötig wäre. Es gibt kein kämpferisches Narrativ, keine bedeutsamen Gegennarrative gegen den Fortgang des Faktischen oder Narrative, was positive Errungenschaften der Geschichte, Errungenschaften der Auseinandersetzung mit der sozialen Conditio humana sind, die es zu bewahren und zu sichern gilt, die aber strukturell zu entschwinden begonnen haben. Macht-Phantasien und Ohnmachtsglaube gehören auf eigentümliche Weise zusammen. Gerade hier, wo man eigentlich viel mehr und schneller bewirken könnte als bei anderen Problemherausforderungen wie Klima, Ungerechtigkeit, globale Ungleichheit oder Demographie, bringt die Subjektivität der Zeitalterwahrnehmung sich selbst zur Untätigkeit, im Persönlichen wie im Kollektiven. – Um dies angemessen zu begreifen, gilt es von den Fügsamkeiten zurück zu gehen auf das, wie das Technische gedacht werden muss, um diese Einwirkungen aufs Subjektive verstehen zu lassen: zurück zu dem, was technischerseits diese Bereitschaften generiert bzw. antreibt in den Wirklichkeiten des Digitalisierten.

V Verwoben ins Technische (Transformation der Umwelten)

In den klassischen Reflexionen über Technik, wo immer sie über die Geschichtsszenarien und großen Existenzfragen von Segnung-oder-Fluch?, Selbstverwirklichung-vs.-Hybris, Entlastung-oder-Entfremdung?, Entfaltung-oder-Verkümmerung-des-Lebens? (bzw. Hilfe zu freien Kapazitäten wie Räumen – für Entwicklung der Humanität, Kultur, zwischenmenschliche Moral – oder Feind wahrer Kultur?) etc. sich herausentwickelt haben, ist bereits in der ersten, frühen Phase als wesentliches Merkmal technischer Artefakte gesehen, dass sie in ihrer Nutzung bzw. Handhabung mit dem Aktor verschmelzen.[30] Ein human oder gesellschaftlich Gewünschtes verbessernd oder erweiternd, verändert in den (bisherigen) Verhältnissen von Handlungssubjekt und Welt das technische Artefakt die Umwelt-Relation und vor allem die subjektive ›Umwelt‹-Schwelle. Das macht zugleich den Übergang, dass aus Instrumentellem ein Medium wird, Zonen des Medialen sich anreichern. – Was für Technisches allgemein gilt, macht sich bei Digitaltechnischem in besonderem Maße geltend.[31] Welche theoretischen Perspektiven dies

30 So seit den technikphilosophischen Zeitalterdiagnosen von Ernst Kapp (Kapp 1877).

31 Faktoren und Momente des Prozesses, sowohl die Einwirkungen auf die subjektiven Horizonte und Haltungen (s. vorigen Abschn. IV) als auch hier im Folgenden, haben sicher nicht in allen Feldern in schlechthin gleicher bzw. analoger Weise statt. Da müsste man in ferne-

haben könnte, dazu sollen hier zwei entscheidende Aspekte in den Blick gebracht werden.

[A.]

Die Erweiterungen des (instrumentellen) Könnens, die mit technischen Errungenschaften, mit Entdeckungen, Entwicklungen und Implementierungen kommen, führen im Digitaltechnischen, noch unabhängig von den genannten innersubjektiven Erfahrungsentkopplungen wie Macht-Phantasien und Ohnmachtsglaube, schon in Hinsicht der quantitativen Parameter zu spezifischen Verschiebungen der Schwellen. Dies sowohl objektiverseits, im überschnell passierten Verändern von realweltlichen Gegebenheiten, auch möglichen Zukünften – möglicherweise nun eine für humane Lebenswelten *zu* große und *zu* unmittelbare Effektivität – als auch subjektiv, in Visionen des Könnens, ohne Sensibilitäten für Schwierigkeiten, Risiken, Folgeketten, Grenzen, soziale Rücksichtnahmen – möglicherweise ist mit den digitaltechnisch kommenden Erweiterungen zu viel in die Hände eines Einzelnen gebracht; und nicht zuletzt drittens in einer keineswegs unbedenklich großen *Vernetzung* der Aktionspole – unter nichtidealen Bedingungen und bes. in Krisen (Krisen, die neu sind, nicht in Analogie zu schon einmal gelernten Bewältigungsstrategien) bleiben möglicherweise keine genügend großen untangierten Bereiche resp. Ressourcen, um ›besonnen‹ zu reagieren: menschlicherseits dann die Prozesse irrationalitätsanfälligen *Massen*-Verhaltens,[32] digitaltechnisch die Dominoeffekte.

Dies geht hinaus über das generelle ›Kulturkritische‹, die einhergehenden Tendenzen von Verdinglichung und Vernutzung der Welt und des Lebens. Und zugleich sind diese schwerwiegenden neuen Gegebenheiten nur ein Teil des zu Reflektierenden. Denn es tritt in mehrfacher Weise auch ein dem Technischen als solchem in seinen Verschmelzungsprozessen innewohnendes *Unbewusst-Werden* ein. Das vollzieht sich hier in besonderer Tragweite, insofern das genuin Digitaltechnische gemeinhin gerade nicht (bzw.: schon lange nicht mehr) als weltmäßig materielles Gerät heraussticht bzw. im Vordergrund steht, durch seine sinnfällig erfahrene Wuchtigkeit

---

ren Untersuchungsschritten wohl unterscheiden und spezifizieren. – Ein Besonderes im Ganzen ist der Bereich der Medizin, in dem auch bei allen Entwicklungen durch neue digitaltechnische Möglichkeiten nach wie vor dieselben Ziele (Therapierung usw.) bleiben, zudem die klassischen Praktiken der Kollegialität (kollegiale Pluralität), gleichzeitig hohes Maß an Nicht-Wissen und Besonderheit des Einzelfalls. Gleichwohl herrschen heute auch dort die gleichen – durch die Wirklichkeiten der Digitalisierung gekommenen oder zementierten – blickverengenden Denk- und Sprechweisen, Reflexionstopoi und Argumentformen, die es zu überwinden gälte (s. unten Abschn. VI).

32 Das in vielem noch immer gültige Muster solcher Para-Intentionalität hatte schon Le Bon 1895 analysiert.

sich präsent macht, vielmehr sich in seinen Funktionen gerade nutzungskomfortabel unsichtbar machen will; und insofern es im meisten auch nicht mehr beliebig beiseitegelegt werden kann, wieder herausgenommen oder ausgeklammert werden kann – nicht mehr beliebig gewechselt werden kann zwischen bisherigen Praktiken (und deren ›erdenden‹, die Widerständigkeit der Realität spürbar machenden Erfahrungen), womit vieles ebenso, nur nicht so schnell oder effektiv zu bewerkstelligen wäre, und dem hinzugekommenen (Digital-)Technischen. Die Prozesse der Verschmelzungen sind hier besonders unauffällig, die Effekte besonders suggestiv.

Die verschmelzungsbedingten, zunehmend sich einstellenden Unbewusstheiten, gegen die sukzessive keine ›natürlichen‹ Erfahrungen mehr Halt und Orientierung bieten, betreffen hier: das Gefühl für mich selbst – d.h. ohne die durch (Digital-)Technisches hinzukommenden, anverwobenen Extensionen des ›Informationen‹-Bekommens, Empfindens, Könnens und Tuns – bzw. Gefühl für die Zurüstung und Formung meiner selbst, um in Schnittstellen Lebewesen/Technik funktional ineinanderzugreifen; ›Gesellschaft‹ und Gesellschaftsbedingtheiten – trotz (oder oft gerade wegen) Vernetzung und beständig sich potenzierender ›Kommunikation‹ zieht sich alles auf ein *Single-end*-Bewusstsein zusammen;[33] wie (über die Aspekte des Quantitativen hinaus) Horizonte, Richtungen und Erwartungen von ›Möglichem‹ sich verändern; Unbewusstheiten, wie erreichte Fähigkeiten und Möglichkeiten ihrerseits weitere neue Ergänzungen, Flankierungen, Steigerungen zu erfordern scheinen bzw. dies prozessieren – der Selbsterweiterungsdrang der zu eigen gemachten installierten Systeme, der einmal vorherrschend gewordenen Geräte und Programm-Linien; Unbewusstheiten bzgl. geschaffener realer Veränderungen und Irreversibilitäten in Lebenswelten und sozialen Formen – trotz (oder oft gerade aufgrund) der Rasanz und Unablässigkeit tendenziell immer weniger Bewusstheit über die mit Einführung und Etablierung einer bestimmten digitaltechnischen Entwicklung geschaffenen Fakten;[34] umgekehrt Unbewusstheiten im Sinne eines schwindenden Gefühls für meine Bedürfnisse – aus Wünschen wird, dass Befriedigungen erwartet oder vorausgesetzt werden, d.h. nur noch ›Besitzstands‹-Bewusstsein; ferner eine sich breitmachende existenzielle Taubheit oder Dumpfheit, eigene Lebensstimmungen nicht mehr zu spüren;[35] hinzu dann die *Entwicklungs-*

33 Auf der Gegenseite auch Unbewusstheiten bezüglich nicht-menschlicher Natur.

34 Ein ganz eigener Punkt wären die Unbewusstheiten bezüglich dem, wie sehr man – als Einzelne wie als Gesellschaften und gesellschaftliche Organe – *abhängig* wird (oder geworden ist), und sei es ›nur‹ emotional bzw. in Verfahrensgewohnheiten: was meist nur dann kurzzeitig aufblitzt und man es wirklich an sich heranlässt, wenn etwas *nicht* oder nicht im gewohnten flüssigen Ablauf das Erwartete bewerkstelligt (oder erkennbar einem böswilligen Angriff von außen, durch ein Fremdsystem ausgesetzt ist).

35 Zu erwägen wäre auch, ob z.B. bei Kindern und Jugendlichen eine – trotz (oder gerade wegen) all des und relativ besonders für sie erreichbar Gewordenen – sich ausbreitende *Lebensstimmung eines stummen Unglücklichseins* ursächlich ebenfalls, obwohl von Familienverhältnissen

Unbewusstheiten – für Verluste, für Schwinden *bisheriger* Möglichkeiten, bzw. wie Bisheriges auch Errungenschaften waren (von den Prozessen bedingte zunehmende Unbewusstheiten über Errungenschaften der alten, nun vergehenden Welt); und nicht zuletzt Unbewusstheiten bzgl. Materialität(en) überhaupt – angefangen bei Server-›Farmen‹, Energieverbrauch, CO_2-Bilanz etc.[36]

Welche Herausforderungen dies für eine Theorie bedeutet, exemplifiziert sich wohl nirgends so wie bei dem Klassiker der Medientheorie, Marshall McLuhan. Um als wesensmäßig medientheoretisches Konzept nicht den Boden zu verlieren, um nicht im Vagen von nur eigenen hohen Begrifflichkeiten im Schwimmen zu bleiben – und die Absicht geht auf Konkretes, auf zivilisationshistorische Analyse von spezifischen Stadien der (neuzeitlichen) realen Geschichte des Medialen –, bleibt das Konzept zum einen noch offen gebunden an die *anthropologische Untermauerung* des alten, aus der Philosophie hervorgewachsenen Denkens. Technik und ihr Mediales sind in allgemein-anthropologischer Perspektive der Offenheit und Plastizität des Menschen – nicht als geschichtliche Anthropologie – angesetzt, angesetzt als die durch Inventionsfähigkeit geschaffenen Erweiterungen unserer Möglichkeiten als Lebewesen und von deren Gemeinschaftsbildungen.[37] In diesem Sinne wäre jede technische Verlagerung des Bezugs zu einer ›Umwelt‹, jedes durch Technisches in die ›Hände‹ bzw. Reichweite des Menschen Gekommene, wo nicht nur kontingent endemisch ausgebildet, eben Eines innerhalb von universalgeschichtlichen Stufen oder Schichten: zu modellieren als ein zunehmenderweise erweitertes und im Prinzip von uns nach unseren Vorstellungen ebenso geschaffenes wie beherrschtes *Hinaus*ragen in die Welt. Es ist rein vom Menschen aus gedacht, und nach durchaus alten Vorstellungen von Subjektivität dabei. Aller eventuelle Aspekt von *Hinein*ragen – in ein ›Material‹, eine Materialität – oder von veränderter verlagerter Grenze bzw. Schwelle ist davon überblendet. So allgemein wie das Anthropologische noch, so allgemein ist in dieser medientheoretischen Tradition auch das Universalgeschichtliche.

Während denn, vom Ansatz her und aufs Ganze, für Kritik oder gar die konkrete Bezeichnung von Macht (Macht über Subjekte) da dieser Allgemeinheit wegen eigentlich kein Raum – Zwischenraum – bleibt, sieht McLuhan gleichwohl die in dieser Entwicklung der technischen Zivilisation sich einstellende Bewusstlosigkeit

---

bis Weltpolitik auch vieles andere hereinspielt, mit der neuen Gegebenheit, in Lebensräumen und -formen des Digitalisierten zu leben, zu tun hat. Vgl. etwa für die USA die Umfrage des Pew Research Center von 2019 (d.h. noch vor dem zweifellos zusätzlichen Faktor der großen Covid-19-Pandemie 2020ff.) Horowitz/Graf 2019.

36 Natürlich auch: durch problematischen Bergbau gewonnene, erforderliche metallene Rohstoffe für die Herstellung der Equipments; oder am entgegengesetzten Ende die Entsorgung der durch die hohen Neuerungsraten in riesigen Mengen anfallenden Altgeräte.

37 So auch der Untertitel von McLuhans epochemachendem Werk Understanding Media (McLuhan 1994[1964]): The Extensions of Man.

darüber, was unser jeweiliges Eigenselbst ist und wo vielmehr das mediale Instrument beginnt. Unsere menschlichen *Sinne* verschmelzen dergestalt mit den verlängernden, immer weiter ›nach außen‹ hereinverwobenen Instrumentalitäten, dass es mit den Etappen des Technisch-Medialen zunehmend zur Betäubung (»numbness«, »narcotic«, »anestetic«: vgl. McLuhan 1994[1964]: 6, 42–47, 62–65) bzgl. dieser technischen Extremitäten und der in *ihnen* sowie ihren Prozessen liegenden Vorgaben und Imperative kommt. Wenn dabei ein Mal auch explizit von einem »prison without walls« (McLuhan 1994[1964]: 20) die Rede ist, so mag man darin einen Vorblick auf eine doch mögliche kritische Perspektive finden: in der Entwicklung unseres technischen Zustands nicht schnell genug den immer neuen Erweiterungen und dadurch Narkotisierungen hinterherzukommen, nicht schnell genug mit der Erwirkung eines (sekundär herauszubildenden) *Bewusstseins* über diese Verschmelzungszonen oder -bereiche. Das ändert indes im Blick auf Theorie nichts an der grundlegenden anthropologisierenden Bestimmung von Technik und ihren medialen Gestaltungen sowie Etappen.

[B.]

Dass das Allgemein-Anthropologische nicht in gleicher Weise die gegenläufige Perspektive zur Seite hat, die Perspektive des Bewusstseins-Widerstands gegen die Konditionierungen durch gesellschaftlich sich etablierende technisch-mediale Formen des Lebens, gegen die vegetativen Assimilierungen, vegetativen Prägungen, bleibt denn eine nicht gänzlich geklärte Ambivalenz dieser klassischen medientheoretischen Konzeption. Digitaltechnisch-Mediales potenziert dabei die theoriekonzeptionellen Herausforderungen. Dass *digital*technische Wirklichkeiten eine besondere Weise von Verschmelzungen mit sich bringen, weist noch deutlicher auf die Grenze alter anthropologischer Modellierungen hin, einer ›Dialektik‹ gerecht zu werden, die man allgemein als die ›Dialektik‹ des Vom-Menschen-aus und des zu erwägenden eventuellen Gegen-den-Menschen bezeichnen könnte. Der phänomenale Ort dafür ist, dass digitaltechnische Wirklichkeiten nicht nur, wie alle Technik, neue Eindring-Tiefen schaffen – Eindring-Tiefen gedanklich wie kausal –, auch ein immer weiteres Heranrücken von Verfügbarkeiten. Sondern sie schaffen auch, im Umgehen mit ihnen bzw. was sie als Umgehen vorsehen, für welches Umgehen sie eingerichtet sind, in einer besonderen, hinzukommenden Weise neue *Oberflächen*. Und die Oberflächen bekommen im Zustand des *Digital*technischen eine sich verselbständigende Macht.

Das Eindringen und die Verfügbarkeiten betreffen im Digitaltechnischen entscheidenderweise rationale Relationalitäten. So: das – möglich gewordene – Eindringen bei Datenmengen überhaupt und das Herausdestillieren von verfügbarkeitsrelevanten Verallgemeinerungen, (wahrscheinlichen) Kausalitäten, Beziehungen sowie Entscheidungsparametern; bei Strukturierung zu relevanten

Mustern, und für (›identifizierende‹) Rubrizierung oder Zuordnung von konkretem Besonderem; bei zu kalkulierenden Szenarios und bei Strategischem innerhalb tiefer Möglichkeitsketten und komplexen Verzweigungen oder Bewertungsbilanzen; beim Überspringen von in der Empirizität der Lebenswelt sonst Separiertem, wie dem Zugleich von mehrerlei wahrnehmbaren Räumen oder Zeiten oder Realitäten (z.B. in *augmented reality*);[38] bei – statistisch sowie prognostisch – Mengen-Verhalten und menschenweltlich bei Erscheinungen im Blick auf Massen; ebenso umgekehrt in Richtung von Faktoren von Kontingenz (›Zufall‹), bei Transmutation von Risiko-Konstellationen und -Szenarien in berechenbare (oder als berechenbar erscheinende) Werte bzw. Funktionsgrößen; bei Vor-Entscheidungen und Assistenz-Formen, bei denen ›das Richtige‹ schon gesetzt ist, ohne dass der einzelne die digitaltechnischen Vorwegnahmen (noch) wahrnehmen kann oder nach eigener Kognition, Wertung, Bewusstheit und Reflexion selbst entscheiden könnte – Entmündigung (oder positiv akzentuiert: Obhutnahme) der menschlichen Wirklichkeitserfahrung und ihrer Intelligenz, weil sie mutmaßlich falsch oder unzweckhaft (oder: zu langsam) sein könnten; und nicht zuletzt zu nennen, bei Eindringen allgemein in Privatheit, den Wall der Privatheit, wo bisher nur Personales hereinkonnte – Eindringen durch nutzbar machbares Wissen über den einzelnen, d.h. möglich gewordenes Eindringen gegenüber menschlichen Individuen, in ihren Wesens- und Identitätsbereichen;[39] u.a.m.

Oberflächen und Oberflächen-Bezüge ergeben sich, wo hier das Technische seine digitalelektronische ›Intelligenz‹ mit der der handhabenden Nutzer (oder Setzer der Zwecke und Ziele) koordinieren resp. koordiniert bekommen muss. Für dies ließen sich drei Gedankenstücke einer in gewisser Weise zu McLuhan komplementären medientheoretischen Konzeption einbringen. – (1) Das Technische, das sich im Geschehen der Digitalisierung entwickelt und was es für menschliche Welt-Verhältnisse bedeutet, ist nur einseitig erfasst, wenn es klassisch als Werkzeug – Zweck-Instrument, Ermächtigungen erweiternd – oder was man mit dem betreffend Technischen machen kann modelliert ist.[40] Was darüber hinausgehend begriffen werden muss, dazu legt sich nahe, zunächst Technisches qua Werkzeuge (im weitesten Sinne), (materielle) Geräte, Maschinen (Einzelmaschinen, zu punktuellen Bewerkstelligungszwecken, dabei gemeinhin Wirkungsverstärkungen oder allgemeine außer-menschliche Energie-Quellen nutzend) und *Apparate* zu unterscheiden.[41]

38 Auch dass allgemein die Generierung des digitaltechnisch Hervorgebrachten oder Erweiterten in Kopräsenz erscheint, für die Wahrnehmenden nicht mehr als Prozess (oder Prozess-Verhältnisse), sondern als virtuelle Gleichzeitigkeit, d.h. Ganzheit.

39 Darunter auch Wissen über ihr Unbewusstes oder was sie für sich behalten wollen, etwa durch digitaltechnisches Eindringen zu Fremderkenntnis ihrer Emotionen.

40 Darunter zählt auch das Anthropologische wie bei McLuhan.

41 Dass es exakte Grenzen gäbe, muss man dabei wohl nicht unterstellen.

Digitaltechnisches ist nicht die einzige Gestalt von Apparaten. Es gibt Apparatehaftes auch außerhalb bzw. vorher, so etwa auf intellektuelle und maschinengerätehafte Techniken zurückgreifende wie den Verwaltungsapparat einer Gesellschaft oder den ›Nachrichten‹-Apparat von Zeitung, Rundfunk, Fernsehen, auch den ›Unterhaltungs‹- und ›Freizeit‹-Apparat zur Führung des Lebens. Doch mit dem Digitaltechnischen findet sich, dass apparatehafte Technik auch im Kleinen (und bis zum Unsichtbaren) sowie in der persönlichen Lebensumwelt gegenwärtig wird. Der Stand, den das in den frühesten Anfängen der werkzeuginstrumentellen Macht beginnende Technische des Menschen heute angenommen hat, ist die für das Digitaltechnische charakteristische allseitige Entwicklung zur Apparate-Welt.[42]

Apparate sind, über Werkzeuge, (materielle) zweckbesondere Geräte und Maschinen hinaus, Entwicklungen dahin, dass sie zu Gerät- oder Sachsystemen einer Ganzheit werden, die eine spezifische Außen-Seite ihres Handhabens und ihres Tätigungsergebnisses haben. Apparate liefern ermächtigungspräsentierende Ergebnisse, die, weil ihre Prozesse einer eigenen Operations-Ratio, eigenen technisch optimal prozessierbaren Logizität folgen, in einem bestimmten symbolischen Code gestaltet sind, der dann wiederum überhaupt menschenverstehbar ist oder gemacht werden kann;[43] dieser überlagert sich desto mehr der ›natürlichen‹ Welterfahrung und Widerständigkeit der Welt, zumal Umwelt, je mehr er sich als Verfügen über die Wirklichkeit präsentiert – als die Wirklichkeit selbst, nur eben jetzt durcherkannt, vereignet und beherrscht. Apparate, ferner, sind dabei Entwicklungen dahin, ihre Komplexität – die immer größere innere Komplexität, ebenso was sie dabei intern alles an ›Informationen‹ und Bezügen heranziehen – zunehmend zu verdecken. Auf Vielfältiges applizierbar, wirken Apparate andererseits durch den symbolischen Code vereinheitlichend. Apparat-Strukturen entindividualisieren das Subjekt und den ›Eigentums‹-Zusammenhang seiner Tätigungen mit dem Gerät- bzw. Sachsystem – die Handhabung wird durch die Operationsmöglichkeiten vorgegeben, und wer immer den Apparat (operations-›richtig‹) handhabt, der Apparat liefert ein formgerecht-universelles Ergebnis-für-jedermann. Es ist ein Egalitarismus der Handhabung – Offenheit und Beliebigkeit des Wer – und ebenso auf der gegenüberliegenden Seite ein Egalitarismus der Ergebnisse – ein jedes *ein* Exemplar,

42 Einen ersten Ansatz dazu, noch vor dem Stadium der (bzw. aller neueren) Entwicklungen des *Digital*technischen, mag man bei V. Flusser eingebracht finden: Flusser 1991[1983]; Flusser 1985.

43 Es sind denn, außer in den frühen Anfängen der Informationstechnik, wo die Handhabung durch reine Techniker*innen/Programmierer*innen selbst erfolgte und insofern das Zweite allenfalls aus pragmatischen Gründen dienlich schien, genauer jeweils *zwei* Code-Systeme: das betreffende digital*technische* ›innere‹ Code-System, d.i. der digitalelektronische (›Maschinen‹-) Code der technisch optimal prozessierbaren Logizität des Apparats, und das Symbolisie*rende* eines symbolischen Codes an der Außen-Seite von Bedienung und Ergebnis.

das sich in den Kosmos der schon bestehenden Ergebnisse einfügt, alles durch die codierte Gestalt potenziell einander zugeordnet.[44]

(2) Durch Diversifizierung der Nutzung erreichten IT-Wissens, durch Anwendungs-Entwicklungen und vor allem die gerätetechnischen Miniaturisierungen gibt es Dinge von apparatehaftem Charakter zunehmend im lebensweltlichen Nahbereich und in Funktionen für persönliche Praktiken der Lebensführung.[45] Technisches ist nicht mehr bloß *in* der Umwelt, sondern indem immer mehr Lebensbezüge dadurch vermittelt (und bisherige Weisen ersetzt) sind, wird das Digital-Apparatehafte zunehmend zur Umwelt selbst. Umwelt ist es nicht nur, weil es in vielem heute mit drin ist, d.h. Umwelt materiell-räumlich, sondern mehr noch zeitlich und in den Lebensrelevanzen, Umwelt in Gestalt der Menge an Zeit, in der es ein dominierendes Mittel bei Lebensbezügen ist[46] und die Menschen sich mit ihm – und über es mit ihren Lebenswichtigkeiten – beschäftigen und schon dadurch anderem, Bisherigem immer weniger Platz, d.h. zeitliche Möglichkeiten bleibt. Das Digital-Apparatehafte totalisiert immer mehr die betreffenden Aufmerksamkeiten, allem voran die Aufmerksamkeitsweisen, verdrängt andere Bezüge, gräbt sie ab, trocknet sie aus. Und Umwelt ist es vor allem auch in Hinsicht auf die subjektive Wahrnehmung und das Agieren. Sosehr das meiste, im Umgehen damit, weiterhin als je besondere Einzelnutzungen und ›Objekte‹ sich darstellt, sind es doch inzwischen zusammengewachsene Gesamtumwelten, objektiv durch zunehmende Vernetzung und subjektiv durch die schiere Übermenge dessen, was man sie hat übernehmen lassen. Das Digital-Apparatehafte verschiebt für die einzelnen die Schwellen zu dem, was ihrem Wissen, Handeln, Bewusstheit und Reflektieren ›Umwelt‹ ist.[47] Was ›hinter‹ den Schwellen des (digitalelektronisch gewirkten und präsentierten) Wissens-von und des Interagierens liegt, steht als Faktum der Welt. An den jeweiligen Code und den Modus, wie die einzelnen – durch die Repräsentanz auf der Außen-Seite des Apparatehaften – ihren Bezug darauf haben, formt sich durch die humane Plastizität eine Adaption, wie vormals an naturale, soziale und (geistig-)kulturelle Umwelt.

---

44 Dies alles ungeachtet der (Wunsch-) Freiheiten, was ein einzelner alles kann bzw. könnte mit dem Apparat – *Freiheiten* des Kann bzw. Könnte, womit Apparate für persönliche Anwendung/Nutzung gemeinhin locken.

45 Dies verstärkt sich auch schlicht äußerlich, indem die Digitaltechnik dadurch zu Massenprodukten wird – und durch die entsprechende einschneidende Verbilligung wiederum noch weiter in die Bereiche des persönlichen Lebens hereinkommt.

46 Zugleich im mobilen Notebook oder Smartphone ein Universalgerät.

47 Dies auch in seiner stillen Weise: Wo aus der manifesten Umwelt etwas subtrahiert, *heraus*genommen ist, was in digital-apparatehaften Prozessen heinzelmännchenartig ausgeführt wird (wie die unsichtbare Dienerschaft-Parallelwelt in den Schlössern des Absolutismus) – Apparate, die man *nicht* merken soll.

Auch wenn in bisherigen Lebenswelten vieles ebenfalls habituierte Gewohnheiten, passive Übernahme, Reagieren, Reflex, oft auch Imitation war, keineswegs also die idealisierte klare Bewusstheit und die Rationalität des freien Geistes (bzw. aus Eigenmacht kommende rationale Gestaltung), ist doch der grundlegende qualitative Unterschied eingetreten, dass das, worauf die Subjekte sich beziehen, bisher entweder universale Positivitäten waren (›Natur‹, in ihrer ganzen auch Irregularität) oder vor allem andere Menschen (in ihrer personhaften Komplexität) und Soziales, und das Menschliche wie Soziale dabei mit Normativem der rechten Gestaltung bzw. Veränderung des Verhältnisses. Die Schwelle des Digital-Apparatehaften ist jedoch, dass in diesem Medialen an dem Kontakt mit der Außen-Seite wir etwas an ein gedachtes Ausführungs-Prozessieren übergeben; und umgekehrt neutrale Informationen und Entscheidungskriterien bekommen – unser *eigener* Mentalraum durch eine neutrale Ergänzungs-Instrumentalität einfach vergrößert und vereffektiviert. Der Anwendungs-Nutzung des Digitalen ist eine neutrale Ratio unterstellt. Die Apparate ›wollten‹ nichts (anders als die Umwelt anderer Menschen und normativ geladener sozialer Formen); was sie an ihren Oberflächen sehen, erkennen oder steuern lassen, wäre so rein die (objektiv-logische) Ausführung eines gegebenen Funktions-Auftrags. Und ebenso wenig wie einen eigenen, mehr als operativ-verzwecklichenden Einfluss auf das Inhaltliche ihrer präsentierten Output-Erscheinungen oder auf (unautorisierte) Steuerungen oder Vernetzungen lassen sie in ihrer Apparatehaftigkeit bemerkbar werden, inwiefern von dem, was ›hinter‹ den Oberflächen nicht in unseren persönlichen Händen ist, etwas in den Händen *anderer* ist, also nicht lediglich die reine Rationalität einer Maschinenlogik – von der Subjektivität der Entwickler*innen und Programmierer*innen über die Interessen der proprietären Halter*innen von Software (Betriebs-Software), Hardware und digitaler Infrastruktur bis zu Kontrolle, böswilliger Manipulation oder Abschöpfen von Daten.

Wenn das Digital-Apparatehafte die Umwelt geworden ist, in der man sich, teilhabend am Arbeits- wie am sozialen Leben vorfindet und mit der man umgeht, ja durch die Macht der Verhältnisse umzugehen genötigt ist, assimiliert der Mensch sich unwillkürlich; die Beziehung zu *ihren* Positivitäten ist absolut.[48] Es gibt, wo nicht aus vormaligen Umwelten und deren Erfahrungen heraus Kriterien

48 Eine Psychologie des Verhältnisses zu den digital-apparatehaften Oberflächen wird wohl auch empirisch konstatieren können, dass dabei, dies verstärkend, oft andere Kräfte und Weisen von Verschmelzungen zugleich inkorporiert sind – archaische wie auch allgemein anthropologisch bedingte: magische Verschmelzungen; libidinöse Verschmelzungen; einen überkommende Verschmelzungsgefühle mit der ›Natur‹ (›romantisch‹ oder ästhetisch oder schwärmerisch); Verschmelzungen mit der herrschenden Macht selbst; kosmische Verschmelzungen (*unio*, mit dem Sein überhaupt); Verschmelzungen mit Rollen, oder einer Funktion, die man innehat bzw. zugewiesen bekommen hat; animistische (oder spezifische ›totemistische‹) Verschmelzungen.

der Distanz und des möglichen Widerstands, da tendenziell kein Sensorium für die massiven Begleitbetäubungen, solange dies Apparatehafte offenbar reibungslos und effektiv funktioniert. – Gerade deswegen bedarf es aber, dies im Unterschied zu den vormaligen Gestalten von Umwelten, in der heutigen Reflexionslage auch einer Differenzierung. Das betrifft hier zentral die Rede von Black-box-Strukturen, *blackboxing*, Black-box-Effekten. In den Verständigungen und Diskursen droht dies sonst unverkennbar von einem anzeigenden Problembegriff – pragmatisch-heuristische Kategorialität für das opake apparatehafte ›Dahinter‹ oder ›Drinnen‹ bei den uns Menschen zugewandten Oberflächen-Seiten eines digitaltechnischen Prozessierens – zu einem pauschalen Narrativ zu werden, und als anscheinend einziges Problem: ein Narrativ ähnlich wie die alten ›kulturkritischen‹ Figuren von Verdinglichung, von Segnung-oder-Fluch, Entlastung-oder-Entfremdung usw. (s.o.) bzw. an deren Stelle.

Die Herausforderung für die Reflexion ist, für das ›Umwelt‹-mäßige des Lebens den Denkraum des Wie-eine-Natur und des Sozialen und des (Geistig-)Kulturellen offen zu halten, die Herausforderung, hier nicht gleicherweise zu nivellieren wie jene alten Figuren. Dazu ist

– zuerst sicher zu unterscheiden die relative Intransparenz für die große Menge: die Intransparenz-Schwelle von programmierten Bedien-Oberflächen, die für all jene besteht, die ein Digital-Apparatehaftes ›nur‹ benutzen resp. bedienen, gegenüber denen, die technischerseits diese betreffende Oberfläche und das dazugehörige apparatehafte Modul entwickelt, eingerichtet/programmiert oder die dies wiederum mit Vorgaben beauftragt haben und entsprechend betreiben (oder vertreiben). Dies ist im Letzten Ausdruck der mit dem Digitaltechnischen gekommenen neuen Zwei-Klassen-Gesellschaft, ein Black-box-Gefälle der intellektuellen Souveränität in der digitalisiert gewordenen Gesellschaft.[49]

49 Davon noch einmal abzuheben wäre die pragmatische Black-box-Schwelle, die es selbstredend auch beim digitaltechnischen Stadium gibt: ein Funktionieren ›im Prinzip‹ verstehen/nachvollziehen zu können, sich aber, solange etwas läuft und seinen Einsatzzweck offenbar erfüllt, persönlich nicht einarbeiten und sich damit beschäftigen zu wollen. – Allerdings muss man doch unterscheiden. Die Aura des Geheimnisvoll-Fremden, die fast alle neue (höhere) Technik seit jeher umgeben hat und bei vielen zu einem ersten ›Fremdeln‹ und vielleicht Skepsis geführt hat, die erst dann der habituellen Benutzung gewichen sind, hat sich zu einer im digitaltechnischen Stadium neuen Qualität verschoben. Obwohl schon wie aus einer ganz anderen Epoche erscheinend, liegt es noch nicht lange zurück, dass man mit leidlich guter Schulbildung ›im Prinzip‹ (und ungeachtet aller Patent-Geheimhaltungen) das Funktionieren so gut wie aller Technik, mit der man in Berührung kam, verstehen/nachvollziehen konnte (selbst Atomkraftwerke, Mondlandung, Herstellungsverfahren in der chemischen Industrie usw.) – man aber pragmatisch vor Ort nur ein genügendes Knowing-how wissen will, einem dies aus *eigener* Entscheidung und zur eigenen Entlastung genügt.

– Eine dagegen ganz andere Dimension von Black-box-Strukturen ist die, die dem digitalelektronisch-algorithmischen Prozessieren des Apparats als solchen innewohnt, nämlich wenn dessen digitalelektronisches System (so wie in heute avancierten Entwicklungen) zu gewissen eigenständigen Informationsgewinnungs- und Entscheidungs-Strategien freigegeben, d.h. darauf hin eingerichtet ist und ihm eine ›eigene (eigenentwickelte) Intelligenz‹ ermöglicht ist. Die Effekte sind die bei ›selbstlernenden‹ Systemen und der Gestalt von ›künstlichen neuronalen Netzen‹, und deren *Tiefen-Intransparenz* ist auch für die Entwickler*innen die Schwelle gegenüber dem Apparatehaften und seinem Funktionieren: die Eigenrationalität, die sich aus dem – aus unabsehbar hohen Mengen und Wegen von Optimierungsanstößen erfolgenden – intern generierten prozeduralen Selbstum- und -ausbau des Systems ausformt und einrichtet als Zustand seiner Operativität, seines Funktionierens. Dies kann nur (abzweckungseffektiv) ›trainiert‹, d.h. berichtigt und nachjustiert werden, genauer, wenn generierte funktionsfalsche Muster und ein Bias *offenkundig* geworden sind. Es verliert aber dadurch im Prinzipiellen nichts von seiner Black-box-Bedeutung innerhalb der apparatehaft gewordenen Umwelt.[50]
– Und schließlich drittens gilt es differenzierend abzuheben die angesichts der Übermengen an Unüberschaubarem und zugleich Effektiv-Funktionierendem sich einstellende subjektive Ansicht – und dann Erwartungsvoraussetzung –, dass es ohnehin so sei, dass alles und überall nur Oberflächen seien. Dann gäbe es (außer bei altphilosophischen idealistischen Subjektivitäts-Träumer*innen) auch keine Ansprüche mehr[51] – *alles* so grundlegenderweise nur die manifesten Gestalt-Seiten einer Tiefengenerierung, dass in solcherart Umwelt genau auch gar keine begründeten normativen Kriterien bestünden, d.h. außer bei einem System eben den Kriterien seines Funktionierens und des rechten Umgehens damit.

50 Die Tiefen-Intransparenz – opak für *jede* dem jeweiligen System äußere Rationalität, sei sie menschlich oder auch andere KI – bleibt entweder überhaupt weitesthin; oder es ist, auch wo vorausgeschaut eine begleitende (Entwickler*innen-) Menschenverstehbarkeit rückgefordert wird (*explainable AI* / XAI), im Konkreten aber Behinderung der Effektivität. Maximal effektive Ausbildung einer ›selbstlernenden‹ inneren Rationalität und umfassend erklärende Menschenverstehbarkeit seiner algorithmischen Gestalt dabei ist nicht beides zugleich zu haben.

51 Stattdessen bildet sich, bes. im privaten Sektor bzw. von dem aus sich ausbreitend, vielerorts gerade gegenteilig ein Kult der Oberflächen heraus – das spielerisch-fluktuierende Sich-bewegen-können auf dem Positiven der eben faktischen Oberflächen, und sozial ein Kult des Nomadenhaften, Rückkehr von den Fesseln des Sesshaften. Das steht dann als neue Grundeinstellung des Lebens und zur Welt.

(3) Mit Digital-Apparatehaftem umzugehen sind andere Lebensumwelten als das jahrmillionenalt zum Menschlichen dazugehörende bisherige (relativ direkte) Umgehen mit einem humanen Gegenüber und sozialen Formen (plus äußerer Natur-Wirklichkeit).[52] Die Positivität dieser Umwelten wird zusätzlich verstärkt durch eine weitere mit bes. dem Digitaltechnischen kommende Veränderung des Charakters der Oberflächen, deren Reflexion der Theorie bedarf. Denn die Masse an zwischen Mensch und jeweiligen digitalelektronischen (apparatehaften) Programmgesamtheiten zu koordinierenden ›Informationen‹ ist – getrieben durch die (Ermächtigungs-) Möglichkeiten dazu – so angewachsen, dass sie immer mehr in der *dichten Weise von Ikonischem* codiert sind. Mit der Welt der Apparate hat in unserem Bezug zur Wirklichkeit (und woraus sich grundlegende ›Weltbilder‹ formen) in signifikanter Weise sich der Anteil des Ikonischen potenziert. Visuelles aber hat per se eine unmittelbarere und erheblich höhere Macht, einen zu vereinnahmen oder überzugreifen auf einen, und Ikonisches zeichnet sich innerhalb dessen aus durch Bedeutungs-Einheiten (einzelner distinkter ikonischer Codierungen, ikonischer Objektivierungen) und bedeutungshafte Figurationen, die unseren menschlichen Geist intuitiv ergreifen – als Ganzheiten – und die zudem die Aufbereitung zu dieser Code-Gestalt ins Dunkel drücken. Der ikonische Modus ist die große Vereinfachung; in ihm sind Bedeutungen *verdichtet*, meist hoch verdichtet, und der Bezug installiert eine tiefgehende Unmittelbarkeit der Erfassung, des Eingehens auf Seiten des menschlichen Geistes. Die bisherige Dominanz der Schrift, ihrer Bedeutungen, der in ihrer Grammatik mit bereitgestellten Logik und ihres schritt-linearen analytischen Charakters ist einhergehend mit der Ausbreitung des Digitaltechnischen zunehmend ersetzt, zum Teil sind die Lebensumwelten überhaupt ikonisch geprägt.

Im Modus des Ikonischen ist die Koordination von menschlicher Person und ›Intelligenz‹ eines apparatehaften Digitalelektronischen kaum offen für Distanzmöglichkeit der Reflexion oder Selbstkritik. Sondern es herrschen geformte einfache Habitualitäten, umwelteigentümliche Traditionen von Bedeutungs- und Interaktionsgewohnheiten stärker vor als in Mündlichkeit, Schriftlichkeit und sozialer Interaktion; auch rein Reaktives hat sich tiefer eingegraben.[53] Je weiter die Bedeutungswelt des *Ikonischen*, des ikonisch Verdichteten, desto intensiver sind die humanen Beteiligten dem Sog der Oberflächen ausgeliefert, dem einfachen – einnehmenden – Bildlichen, einer flächigen Vereinnahmung des Geistes in den Lebensumwelten der digitalisierten Gerätschaften und Kommunikation. Ein neuer Holismus

52 Vom Taktilen und Haptischen im rein Funktionalen noch ganz abgesehen – der glatten metallischen oder Kunststoff-Materialität der Oberflächen (und in der Bedienung für zumeist einfache, häufigst polar-binäre Optionen).

53 Dies schon allein durch den hohen Takt des Apparats sowie der heranströmenden Interaktionen bzw. Interaktionsaufforderungen.

des Bildlichen lässt auch in dieser Hinsicht Wirklichkeiten wieder opak werden und stück-punktuell: zu leben Ablauf für Ablauf in ikonisch-holistischen *Szenen* (Szenen von Figurationen) und deren Typisierungen. –

Justierungen also an klassischen technik- und medienphilosophischen Traditionen sind nötig. Aber es lässt sich zugleich schon ersehen, dass auch die Entwicklung der Digitalisierung von immer mehr Lebensbezügen und gesellschaftlichen Feldern noch allemal *Technik* bleibt. Von der Problematik des systematischen Unbewusst-Werdens (s. [A.]) und diesen drei genannten erforderlichen Gedankenstücken – in Schlagworten: Apparatehaftigkeit, neuartige Lebensumwelten, Ikonifizierung – aus kann sich der Blick denn auch erweitern, vergrundsätzlicht ins Konzeptionelle einer Theorie.

Mit der Entwicklung zur Welt des Digitalisierten ist es zum ersten Mal so, dass ein *Technisches* ein ganzes *Kultur*-Stadium bestimmt. In einer aus der allgemeinen Geistes- und Kulturphilosophie heraus spezialisiert hervorgewachsenen Technikphilosophie und -soziologie, später zudem betreffenden Ethik, hatte sich die Reflexion der in der Neuzeit immer manifesteren Realitäten, ein Kosmos der technischen Artefakte geworden zu sein, eingerichtet, dabei in langer Tradition mit dem Nachhall anthropologischer und geschichtsphilosophischer Muster. Muss dies deshalb heute wieder zurück zu einer Programmatik, die allgemein von ›Kultur‹ aus denkt (und wie dann auch die Umwelthaftigkeit, weil nun im Digitalen so umfassend, eine Prägung wie ›Kultur‹ wäre)? Muss das Denken wieder zurück in die Nachfolge einer Theorie-Gestalt und der Theorie-Stelle einer allgemeinen Kulturphilosophie?

Gerade weil diese Frage heute so prinzipiell wieder im Raum steht, gilt es einstweilen jedoch vorsichtig zu sein: vorsichtig, nicht allzu schnell vorauszusetzen, dass nach jahrhundertelangem Kontinuum der neuzeitlichen gesellschaftlichen Technik, ihrer Entdeckungen, Entwicklungen und Implementierungen, mit dem digitalelektronischen Stadium dies nun so qualitativ anders ist,[54] dass es aus der begleitenden Reflexion, ›Technik‹ als Wesensvermögen menschlicher Wesen und ihrer Gesellschaften zu denken, herausführt – dass die ganze Tradition bisheriger Auseinandersetzungen, die in Thematisierungsformen geronnen sind, tendenziell allenfalls nachrangig würde, nur Spiegel des Vormaligen der Welt (sozialen, gesellschaftlichen, individuellen, menschheitlichen Welt) wäre. Vorsicht gilt es zu wahren, um nicht Einsichten und Theorembausteine, zu denen es wiederum vor allem schon hoch reflektierte Diskussionen gibt, in denen sie sich bereits geläutert und differenziert haben, bloß einfach zu verlieren. Das bliebe sonst leicht nur die betriebsame Flucht nach vorn, die hier das Entweichen in die Unbestimmtheiten ist. Von ›Kultur‹

54 Gleiche Vorsicht gälte für die Verallgemeinerung, dass die digitalelektronisch gewordene Welt ›nur‹ das Ausziehen der Linie einer (mit Husserl zu sprechen) ›Urstiftung‹ des neuzeitlich operationalistischen Verständnisses des Geistes und von Intelligenz sei.

zu reden, dieser Thematisierungs-Zuschnitt, bleibt immer leicht zu allgemein. – In diesem Bewusstsein ein abschließender Abschnitt in gegensinniger Perspektive.

VI Der Druck auf das System der Reflexion (Versuchungen der Theorie in einer noch unentschiedenen Lage)

Die Geschichtserwartung, die vor einem Jahrhundert, aus den Erfahrungen der großen Technisierung der Gesellschaft mit und nach dem 1. Weltkrieg heraus exemplarisch Ernst Jünger formuliert hatte, ist so nicht eingetreten. Dass das neuzeitliche (›bürgerliche‹) Subjekt eine veraltende Gestalt sei und wieder abtrete aus der Geschichte, war ihm Diagnose wie zugleich Hoffnung auf eine neue Dynamik der Geschichte gewesen, gegen die Kristallisation einer bürgerlichen Endzeitzivilisation, das Lebloswerden ihrer Mentalität (Jünger 1932). Doch der gezeichnete Prozess, dass der im Verlauf der Neuzeit mit ihren Emanzipationsideen vereinzelte Mensch durch die Technisierung wieder zu einem Kollektiv werde – in Industrialisierung Verschmelzung mit der Maschine, Zucht durch die Maschine und darüber Verschmelzung zu einem Gesamtkörper unter einer kollektiven, in den technischen Gerätschaften materialisierten großen Aufgabe und Projekt –, ist nur in einer Zwischenepoche machtvoll geworden, dazwischentretend in den gesellschaftlich-politischen Totalitarismen des 20. Jahrhunderts, die auch technologische Utopien waren.

Zurückgekehrt in Dauer ist indes nicht der Triumph des neuzeitlichen Subjekts mit seinen errungen-herausgebildeten normativen Standards und Sensorien. Dieses zeigte sich in der Tendenz vielmehr verunsicherter denn je. Im 20. Jahrhunderts, angesichts der großen Katastrophen, seine Schwäche oft beschworen, hat dieses Subjekt heute dagegen mit der ganz anderen Revolution der digitalisierten Welt auf stille Weise eine Absorption seiner Kräfte durch ein Technisches erfahren. Verändert hat sich Analoges, was Jünger für die Verschmelzung zum gefügten Kollektiv – durch die Materialität der Maschinen – erwartet hatte. Auch das Technische der digitalisierten Epoche verkörpert sich in Verschmelzungen, aber es ist nicht der durch sein Arbeiten sich vermassende Mensch, sondern eine Verschmelzung gerade des Einzelnen mit der persönlichen Digitalumgebung – wie er von ihrem Apparatehaften persönlich in Anspruch gezogen wird (und sich aus seinen Ermächtigungserwartungen in Anspruch nehmen lässt) – und weithin bis in seine privatesten Lebensbereiche. Statt der naturwüchsigen, geradezu ›darwinistischen‹ Verdrängungen durch die Herrschaft eines neuen Typus von Mensch-sein, die Gegenkonzeptionen wie die von Jünger visionierten, ist es die Selbsttransformation der Einzelnen selbst; dabei aber ebenfalls mit Verdrängungen, der Verdrängung anderer, relativierender Erfahrungsbereiche. Die Kräfte und Folgen dessen bewusst zu machen, ist als Theorie keine weniger herausfordernde Aufgabe als einst die Suche

nach einer der Tragweite angemessenen Reflexion des unter Gegebenheiten einer technisch gewordenen Wirklichkeit sich vermassenden Menschen.

Als das große Geschehen unserer gegenwärtigen Geschichte verändert Digitalisierung Handlungsräume und Möglichkeitsvorstellungen und auch das Denken. Ihre fortschreitenden Prozesse, so darf man das Bisherige vielleicht resümieren, erscheinen wie ein Empowerment des Subjekts zu bisher nicht Erreichbarem oder Erlaubtem, aber zugleich haben sie für die Subjektgestalt dabei Seiten einer genötigten, schleichend vollzogenen inneren Schwächung, Fragilität durch das Leben in und mit solchen Umwelten. Es kommt zu einer Welt des Subjekts, in der dieses, um unter die mit der Digitalisierung kommenden neuen Mächte (s. Abschn. III) gefügt zu bleiben, gar nicht explizit beherrscht werden muss: weil es von sich her so weit ›außer sich‹ gekommen ist in seinen Habitusformen, dass es als ›Subjekt‹ – oder auch wo als Objekt von statistischen Relationen und Zusprechungen oder als Objekt der (externen) Einflussnahme – sich selbst zur Integration, Kompatibilität und Fügsamkeit bringt. – Zu den Hürden für die Reflexion, die damit kommen und die es zu gewärtigen gilt, hier die Skizze eines Rahmens.

(1) Wenn das Materiale der Gerätschaften, und das Phänomenale, was es bedeutet, in einer solchen Lebensumwelt zu existieren und damit umzugehen, nicht gesehen, bilden sich die Positivitäten der Formen der herrschenden Welt und ihrer Verhältnisse in den Theorien ab. Zwischen allgemeiner ›Geschichtstheorie‹ des neuen Zeitalters – in Differenz zum vordigitalisierten Zustand des Lebens und der sozialen Formen – und konkreten Phänomentheoremen bleibt dann eine Lücke. Dort werden die Thematisierungsverständnisse anfällig für bestimmte Muster. Das heutige Theoriedenken findet sich entsprechend umstellt von Versuchungen. – Dies beginnt schon in einem Topos der Reflexionshaltung, Topos der Distanzierung. Wie einst seit Jünger mit der Kritik am Geschichtsanspruch des neuzeitlichen Subjekts auch das nur noch Ironisieren, ja die Häme über die moralischen und ›humanistischen‹ Vorstellungen der Kultur der liberalen Bürger-Individualität verbunden war, so sind auch heute wieder die Verständigungen und Reflexionskonzepte in Analogie oft flankiert durch eine Selbstdemontage der überhaupt normativen Ausrichtung der Theorieperspektive. Es kommt zu einer Atmosphäre derselben Argumente wie einst: dass das ganz neuartige Geschehen der Digitalisierung aller Lebens- und Gesellschaftsverhältnisse anderwärts mit lediglich pathetisch aufgeladenen ›moralischen‹ und alt-modernen ›humanistischen‹ Empfindungs- wie Denkmustern gefasst sei. Verabschieden müsse man sich von dem, mit ›Moralischem‹ gegen die Zeichen der Zeit anzutreten. Nüchtern-realistisch müsse das große kybernetische Geschehen der informationellen Revolution durchdacht werden, anstatt nicht herauszukommen aus den alten Idealismen, einer Moderneromantik und ihrer Subjektivitätsduselei, all dieser bloßen Kammerdiener-Perspektive angesichts des großen Neuen und seiner Macht. Die kursierenden Theorien sind auffällig gekennzeichnet durch oft eine große Selbstgerechtigkeit. Allem voran dieser Selbstgerechtigkeit, al-

les Unbehagen und Kritik als nur veraltete Mentalitäten und Verklärungen eines Bisherigen zu verdächtigen, gilt es zu wehren.

Doch auch die Reflexionsformen der Verständigungen und Debatten bewegen sich weithin im Rahmen von voraussetzungsvollen Mustern, in denen Wesentliches, was als Probleme hereinkommen kann, schon vorentschieden ist. Wie bei jedem neuen (oder in Blick gekommenen) Problemfeld, werden *die Verhältnisse unter den Thematisierungsformen und Wissenschaften neu verhandelt*. Hier ist bezeichnend eine Haltung der ›Arbeitsteilung‹, wie sie auch sonst heute oft sich eingerichtet hat, um in einer internalisierten Kritik schon grundsätzlich dem vorzubauen, nicht alte ›philosophische‹ Programmformen und Ansprüche (in diesem Fall: geschichtsphilosophische, subjekttheoretische und sozialphilosophische) zu reproduzieren. Separiert sind für eine Grundlegungstheorie Weisen des Rückzugs aufs Allgemeine als *Formales*: Kategorien zu entwickeln, Bezüge zu explizieren, mögliche Gestaltformen zu differenzieren usw., doch gezielt keine geschichts-, lebenswelt- oder gesellschaftsmateriale Deutungen zu geben oder mit zu integrieren, sondern dies an Anwendungen, in anderen Wissenschaften oder Zeitalterverständigung, zu delegieren. Dies ist eine herrschende intellektuelle Strategie.

Das findet sich heute etwa überall dort, wo ›Digitalisierung‹, als Prozess, als das vornehmlich Technische (sowie die gesellschaftlichen Implementierungen usw.), wie (und wieweit) Lebenswichtigkeiten, Soziales, Ökonomisches, Politisches, Technisches selbst und Wissenschaftliches nun in einer jeweilig digitali*sierten* Weise statthaben bzw. dies können, verstanden, d.h. angesetzt ist, von dem abgehoben das Grundlegende in einer Dimension des Wesens der Digitalität – Digitalität strukturell in Abhebung zu aller vormaligen Weise von Wirklichkeit der Welt (oder des Seins) – begriffen werden müsse.[55] Oder, der Rückzug des Theorieverständnisses aufs Allgemeine als Formales findet sich in der Weise, alles, auch das Stadium des Digitalen, als intelligente Bildungen unseres seit jeher symboli*sierenden* Geistes – und die Effektivierungen, wenn es Symbolisierungen sind, die (wie in allem Rechnenden) in formalen Operationen angewandt und systematisch auskultiviert werden können – zu verstehen, d.h. anzusetzen: eine Universalität, wo historische und kulturell spezifische Formen und Stadien höchstens rubrikhaft gegeneinander typisiert werden können. – In beiderlei Weise, die Kehrseite des solcherart

55 Der Theorie-Gehalt, im Ertrag innerhalb des Gefüges der angesetzten ›Arbeitsteilung‹, ist dann analog zu – in den bisherigen Wirklichkeiten – dem der universalen ›Systemtheorie‹ und ihren Modellen: eine allseitige *Beschreibungsbegrifflichkeit* sowie Thematisierungsfiguren oder -schemata zu fundieren. – Als zwei breit rezipierte Konzeptualisierungsmodelle innerhalb der aktuellen Diskussion seien nur genannt: die von L. Floridi begründete Theorie der »infosphere« als neuer Dimension der Wirklichkeit (Floridi 2014); und die von F. Stalder begründete Theorie der »Digitalität« (Stalder 2016) – letztere Medienphilosophie allgemein philosophisch bei J. Noller weiter ausgebaut zur Ontologie (Ontologie der Virtualität: Noller 2022).

aufs Universellste sich Beschränkenden ist, dass es zu nicht Wenigem oder Unerheblichem kommt, was die Haltungen einer ›Arbeitsteilung‹ und entsprechender Abstrahierungs-Thematisierungen nur als beiläufige Aspekte (oder nur als eben allgemein Formales, welches immer irgendwie dazugehöre) betrachten können: Gestalten von *Macht*; das, was es mit den Subjekten macht (außer den Ermächtigungen und erweiterten Eindringtiefen), wenn sie sich in den Faktizitäten solcher Lebensumwelten bewegen oder bewegen müssen; das Technische selbst sowie die Materialität des Gerätehaften; und nicht zuletzt ebenso die Dynamik der (jeweiligen) Herausformung, und dass es sich in den Wirklichkeiten ja allemal um *gemischte* Systeme handelt, in denen es das Nicht-Digitalisierte (oder weniger Digitalisierte) eben auch noch gibt – angefangen beim ›alten‹ Menschen, alten Habitualitäten, alten Praktiken, bisherigen Rechtsnormen –, mit dem das neu Entstandene parallel zumindest mittelfristig zusammen existieren muss. Betroffen sind von der mit den Prozessen kommenden intellektuellen Konstellation, dass die Verhältnisse unter den Thematisierungsformen und Wissenschaften neu verhandelt werden, denn signifikant die Soziologie und auch die Psychologie, für die es in jenen Konzepten augenscheinlich keinen rechten Ort mehr gibt bzw. dies nicht vorgesehen ist. Das ist der Druck auf das System der Reflexion.

Das Einschneidendste aber, wie, wenn Realitäten der neuen Prozesse und Wirkungen nicht gesehen, das faktisch sich Einrichtende sich in den Theorien abbildet – strukturell zu wenig Theorie-Abstand besteht –, ist das, wie fast wie in einer Gruppenkonformität oder -zwang auch im Konkreten der Kreis der Reflexionsweisen sich auf prägende Modi und Horizonte verengt hat.[56] Im Binnen des Geschehens, aus den Selbstperspektivierungen in dem installierten Neuen, werden dem Denken bestimmte *Reflexionsformen* nahegelegt oder bekommen entscheidende zusätzliche Präferenz.[57]

– Das sind[58] im handlungstheoretischen und normativen Ansatz (dabei auch im Verhältnis von humaner und digitaltechnischer Intelligenz): ganz grundlegend überhaupt konsequentialistische Ansätze und der Kreis ihres Denkens; weiterhin konkret utilitaristische Modellierungen, und *rational choice* (usw.). Entwicklungs- und lerntheoretisch andererseits ist es für menschliche wie künstliche Intelligenz eine eigentümliche Renaissance behavioristischer Verständnisse und Argumentationsfiguren. Auch gibt es einen starken Sog, flankierend das

56 Ohne dass man dies angesichts der Disparatheit der Zugänge und Themen schon ein ›Paradigma‹ nennen könnte.

57 Dabei sei einmal abgesehen von den Extremen: den technizistischen Zukunftsvisionen, ganz gleich, ob mit positiven oder negativen Wertungen dabei.

58 Alles hier nur in Stichworten (und ohne bibliographische Spezifizierungen und Differenzierungen), im Sinne der beabsichtigten Skizze.

allgemeine Konzept von Akteurselbst sowie Handeln (und Reflektiertheit) nicht so hoch anzusetzen, sondern, gerichtet gegen gemutmaßte ›mentalistische‹ Großerzählungen, in Tendenz an *Verhalten* (und darin generierte operativ fungierende Selbstmodelle für die Binnenreglementierung) zu binden. Und fast durchweg ist Intelligenz (abgesehen vom Operativen), ihre ›Welt‹ und Orientierungen, unverkennbar als konstruktivistisch verstanden.

– Sodann ist den Reflexionen ein Präjudiz für Neutralisierung gesellschaftlicher, lebensweltlicher und psychologischer Prozesse eingeschrieben, rein von ›dem Menschen‹ aus denkend: das Digitaltechnische ist als verfügbare Assistenz (Assistenz unserer menschlichen Vermögen und Aktionszwecke bzw. -aufgaben) verstanden; oder als das, als Intelligenz einfach eine weitere – nun sehr effektive – Ausweitung unserer Vermögen durch Ankoppelung derer der digitaltechnischen Systeme zu haben.[59] Und das meiste zieht – auch wo gemeinhin implizit, schlicht durch Nichtthematisierung, nichts, dass hier etwas anderes zu erwägen wäre – den Gedanken heran, der schon im Stadium am Anfang des 20. Jahrhunderts als *cultural lag* formuliert wurde (Ogburn 1922): dass nicht das sich entwickelnde Technische (hier Digitaltechnische) als solches ein Problem sein könne, sondern wie ›man‹ (d.h. die betreffend zeitgenössische Population) damit *umgeht*; und dass, wenn dies in größerer Masse in unangemessener, problembringender Weise geschehe, es i.Allg. daran liege, dass ›die‹ Menschen (hier nun zumal in ihrer Gestalt als die Einzelnen, in Bewusstsein, Mentalität und/oder Praktiken) und ihre in den gesellschaftlichen Mächten geronnenen Normen nur noch nicht mit der technischen Entwicklung und den dadurch neuen Handlungsräumen Schritt gehalten hätten. Wenn Probleme und drohende Pathologien in der Welt der digitalisierten Wirklichkeiten in den Blick kommen, sind die Reflexionen auch gerade heute schnell mit dem bei der Hand, dass die Ursachen auf der Seite des *Umgehens* damit, bei den von Mensch und Gesellschaft relativ noch nicht bewältigten Fehlanpassungen zu suchen seien.[60]
– Und um nur noch ein Letztes zu nennen in dieser Reihe, so wird in Hinsicht auf das, dass die Veränderungen des Geschehens der Digitalisierung wesentlich auch Fakten schaffen, i.Allg. das universelle Erfolgsmodell der Technikbewertung (Technology Assessment, seit den 1960er Jahren) fortgeführt: für alles

59 Wenn in Einseitigkeit ausgearbeitet, wird dann aus der Konzeption des ohnehin allemal »extended mind« (Clark/Chalmers 1998; Clark 2008) des Menschen die Begründung – und Grundrechtfertigung – der *extended evaluation* und *extended decision*.

60 Für das wird dann der Pädagogik (in Schulsystem und Volkspädagogik) eine weitere Aufgabe auferlegt (›Medienkompetenz‹, Ethik in den digitalen Interaktionen, ›Lebenskunst‹ der neuen Wirklichkeiten, usw.), d.h. es wird der *Pädagogik* (bzw. indirekt der Selbstpsychologisierung) zugewiesen.

›Gremien‹ und deren nüchterne Erörterung einzurichten – Gremien, deren Expertise, noch vor irgendwie an wiederum Öffentlichkeit, in primärem Prozess adressiert ist an die Institutionen der *bestehenden* politischen bzw. bereichsorganisatorischen Mächte. Solcherweise etwas in Erwartung der Objektivierung delegieren zu können, das hat in der Digitalisierungs-Konstellation in besonderem Maße eine Entsorgungsfunktion. Getragen ist die ganze Perspektive von der Einstellung, dass die Diskurse ›der‹ Gesellschaft (in Gestalt ihrer Gruppierungen-Vertreter und Senior-›Experten‹) ausreichen; und dass dann auf dem Wege der ›Aushandelung‹ (und danach einer ermächtigten Institution zur Bewirkung und Kontrolle) das bestmögliche Erforderliche zustande kommt.

(2) Muster mithin der Thematisierungsverständnisse. Was alles nicht infrage gestellt ist, ist vermutlich, wie stets bei Verständigungen, das Hartnäckigste. Daran hat indes entscheidend mit Teil die Selbstverstärkung, die die Diskursivierungen durch das erfahren, was die *Sprache* ist, mit der Digitalisierung und die mit ihr kommenden Veränderungen durchdacht werden. Bestimmtes Sprachliches kommt den Verengungen entgegen. Das sind nicht nur die Begriffssprachen der genannten Theorie- und Argumentationsformen. Sondern es bilden sich spezifische Sprachlichkeiten, die die Prozesse begleiten, in Alltagslebenswelt wie auch Wissenschaft ab. Sie haben sich dem ganzen Denken eingeschrieben, als die Sprache, in der man Gestaltwirklichkeiten und Erfahrungen, den Entwicklungsgang und normative Einordnungen fasst. Neben den sich bildenden Formen von neuartiger ökonomischer Macht, neuer Disziplinarmacht, neuen Macht-Relationen überhaupt (durch ›künstliche‹ Akteurspole und die Zersplitterung der Handlungswelt in einzelne Verhaltenszüge) und neuer Macht-der-Geschichte (s. Abschn. III) – hinzu dem, bei dem metaphorisch eine ›Macht‹ zu bezeichnen ist, wie Macht der Oberflächen, Macht des Visuellen und Ikonischen – kann man dies vielleicht als eine volle weitere, fünfte Form neuer Macht, die mit den Prozessen der Digitalisierung in ihrem historischen und gesellschaftlichen Kontext kommt, fassen: Macht in Gestalt der *Herrschaft der Sprache*, Herrschaft der Vokabulare und Argumentwendungen.

Dies dabei charakteristisch von zwei Seiten her. Zum einen wirkt die alte ›Community‹-Ideologie der Anfänge der heutigen IT-Revolution fort – die Vorstellungen des partizipatorischen Peer-to-Peer, des Zusammenwirkens unter einer Vision und des Bottom-up überhaupt, gegen die alten Gesellschaftlichkeiten und politische Sphäre gerichtet (sowie auch gegen die angestellten Großexperten des öffentlichen Sektors, einschl. denen der institutionalisierten Wissenschaft). Die Bestände dieser alten Aura werden auch zum Teil gezielt angezapft, werden instrumentalisiert, bes. von Interessen der neuen ökonomischen Macht, um Bindungen an *ihre* Systeme zu

bewirken, an ihre jeweiligen Entwicklungslinien auf dem Markt.[61] Zugleich zum andern finden die Verständigungen und die intellektuellen Reflexionen, wo sie anheben, sich schon immer in einem Feld vor, in dem das meiste schon gedeutet *ist*: sprachlich geronnen durch die Visionen und Entwicklungen der Digitaltechnik selbst – und diese erfolgt zu immer übermächtigeren Anteilen aus den Vorreitergestaltungen der Industrie, aus den großen (und ggf. auch den aufstrebenden) Firmen heraus, d.h. seitens der Akteure der Verteilung und Ausweitung eines Markts, mit Interessen an ökonomischen Segmenten, Branding und Bindung (auch innerhalb der jeweiligen Entwickler*innen-Gruppen). Schon im Technischen rennt die nicht von der Industrie betriebene Forschung inzwischen oft hinterher, erst recht in der Reflexion der Technik. Die Hoheit über die Sprache liegt in vielem bei den Interessenten des Einflusses, des Verkaufs dieser Produkte und der Sicherung von Marktanteilen. Der öffentliche Sektor und seine Wissenschaften sind strukturell zum Nachzügler geworden. Schrittmacher der Innovation oder Richtungsgeber der Entwicklungen ist, wohl zum ersten Mal bei einer ›Grundlagenforschung‹, nicht mehr das Öffentliche, das Gesellschafts-Allgemeine. Zudem gibt es den Druck, dass eine Reflexionskompetenz wie die der ›Philosophie‹ allem voran und als Dringlichstes die Diskurse zu *moderieren* habe, die *bestehenden* Diskurse mit ihren herrschend herausgebildeten Positionen, Perspektiven und Normkriterien. Das absorbiert vieles. Diese Aufgabenzuteilung schwächt per se die Möglichkeiten *kritischer* Reflexion. Der allgemeine Verständigungs- und Konsensbedarf überstrahlt alles. Gerade auch das, was man bei den neuen Wirklichkeiten, die mit den Prozessen der Digitalisierung kommen, noch nicht weiß.

Sprache, zusammen mit Bildlichem, ist das Unbewussteste. Dass es weithin keine eigenen Kategorien und eigenes Idiom gibt, die dem Charmierenden des ›Digitalisierungs‹-Sprechs etwas Nachdenkend-Sperriges entgegensetzten, und dass die Entwicklungen auch so hochdynamisch sind, dass umgekehrt wenig Kritik der Vokabulare (sowie vollends der Rhetorik) entstehen konnte, ist darum keine Nebensächlichkeit. Die alten Formen der Geschichts- und Sozialitätsverständigungen über Momente von ›Macht‹ danken ab oder verblassen, und den kursierenden herrschenden Narrativen gelten schon die Kriterien einstiger Wirklichkeitsreflexion, vom einstigen Verständnis von ›Ethik‹ und ›gutem Leben‹ ganz zu schweigen, oft nur noch als Götzen einer vormaligen akademischen Gesellschafts- und Subjekttheorie und eines vormaligen ›bürgerlichen‹ Lebensentwurfs, belastet mit all der einstigen Schwere und Tiefenglaube. In einer auch sonst zunehmend undurchschaubaren Lebenswelt bietet sich das sich einrichtende Digitalisierte gerade als *leichte* Umwelt an.

61 Die Sprachen und Vokabulare einer einstmals visionierten alternativen Lebensform – welche Vorstellungen diese Sprachlichkeiten transportieren bzw. in Assoziationen aktivieren – fungieren dann als Vermittler von beabsichtigtem (Produkt-) Image und (Produkt-) Lifestyle.

Brauchen wir ein neues förmliches Konzept, um in bestehenden Wirklichkeiten Verhältnisse von Macht zu denken sowie zu analysieren? Das lässt sich vielleicht momentan nicht einmal mit einem einfachen Ja oder Nein beantworten, nicht, wenn damit etwas gänzlich Eigenständiges gemeint wäre, so wie bei den großen Bedeutungsdimensionen, die sich in der Geschichte herausgeformt haben (s. Abschn. I). Vermutlich führte dies nur in zu hohe Theorie. Es bleibt dabei: Für die Zeitalter-Geschehnisse der Digitalisierung ist eine Verständigung, die Differenzierungen zu tragen vermag, schwierig. Man wird sich dem stellen müssen. Was jedoch notwendig geboten ist, sind Reflexionen, die die Potenziale der errungenen, auskultivierten Sensorien aus den Problemverständigungen der Geschichte nicht preisgeben, vielmehr stärken, auch und gerade gegen den Sog der im Heutigen kursierenden *Theoreme*; und Reflexionen, die zugleich auch mögliche Deformationen (oder Selbstdeformationen) von Sensorien innerhalb der unmittelbaren subjektiven Wahrnehmung und Empfindung der neuen Phänomenwirklichkeiten in Rechnung zu stellen vermögen.

Das wäre eine sich als Reflexion emanzipierende bewusste Offensivhaltung. Bei der Frage der Fügsamkeiten anzusetzen, und die Thematisierungspotenziale von Soziologie und Psychologie wieder integral einzubringen, dürfte dafür ein fruchtbares Programm sein. Mögen die Ansprüche früherer Jahrzehnte zu groß gewesen sein und mit zu hoher Universalität des gezeichneten Normativen, das heutige Denken ist, wo nicht wilde futuristische Visionen sich ihm verselbständigen, oft noch eigentümlich defensiv.

Literatur

Barlow, J.P. (1996): A Declaration of the Independence of Cyberspace. [https://www.eff.org/cyberspace-independence] (Zugriff: 08.03.2024).

Clark, A.; Chalmers, D. (1998): The extended mind, in: *Analysis*, 58(1), 7–19.

Clark, A. (2008): Supersizing the Mind. Embodiment, Action, and Cognitive Extension, Oxford: Oxford University Press.

Floridi, L. (2014): The Fourth Revolution. How the Infosphere is Reshaping Human Reality, Oxford: Oxford University Press.

Flusser, V. (1991[1983]): Für eine Philosophie der Fotografie, Göttingen: European Photography.

Flusser, V. (1985): Ins Universum der technischen Bilder, Göttingen: European Photography.

Haraway, D. (1985): Manifesto for Cyborgs. Science, Technology, and Socialist Feminism in the 1980s, in: *Socialist Review*, 80, 65–108.

Horowitz, J.M; Graf, N. (2019): Most U.S. Teens See Anxiety and Depression as a Major Problem Among Their Peers. [https://www.pewresearch.org/social-tren

ds/2019/02/20/most-u-s-teens-see-anxiety-and-depression-as-a-major-problem-among-their-peers/] (Zugriff: 08.12.2023).

Hubig, C. (2015): Macht der Technik (Die Kunst des Möglichen III. Grundlinien einer dialektischen Philosophie der Technik), Bielefeld: transcript Verlag.

Jünger, E. (1932): Der Arbeiter. Herrschaft und Gestalt, Hamburg: Hanseatische Verlagsanstalt.

Kapp, E. (1877): Grundlinien einer Philosophie der Technik. Zur Entstehungsgeschichte der Cultur aus neuen Gesichtspunkten, Braunschweig: Georg Westermann.

Laski, H.J. (1917): Studies in the Problem of Sovereignty, New Haven: Yale University Press.

Le Bon, G. (1895): Psychologie des foules, Paris: Alcan.

McLuhan, H.M. (1994[1964]): Understanding Media. The Extensions of Man, Cambridge (MA): The MIT Press.

Moravec, H. (1988): Mind Children. The Future of Robot and Human Intelligence, Cambridge (MA): Harvard University Press.

Noller, J. (2022): Digitalität. Zur Philosophie der digitalen Lebenswelt, Basel: Schwabe.

Ogburn, W.F. (1922): Social Change with Respect to Culture and Original Nature, New York: B. W. Huebsch.

Stalder, F. (2016): Kultur der Digitalität, Berlin: Suhrkamp Verlag.

Weber, M. (1980[1922]): Wirtschaft und Gesellschaft. Grundriß der verstehenden Soziologie, Tübingen: Mohr.

Weiser, M. (1991): The computer for the 21th century, in: *Scientific American*, 265(3), 94–104.

Digitalisierung als Prozess

Der philosophische Blick auf die Möglichkeit allmählicher Disruption

Armin Grunwald

Abstract: *This article analyzes the phenomenon of »gradual disruptions« at a societal level. These are upheavals with considerable to dramatic damage potential that do not occur unexpectedly and suddenly, but build up gradually until they finally lead to the sudden disruption of familiar constellations. This phenomenon can be observed in many areas of processes of digital transformation. The construction of »digital twins« of practices from the analogue world plays a special role in transformation through digitalization. This leads to considerable acceleration effects and increases both the possibility and the risk of disruptive change. The thesis is that the digital transformation that is currently being pursued and permitted is gradually creating and consolidating dependencies that are also gradually becoming more vulnerable. Dependencies on digital technology infrastructure that have become total are latent disruptions that can be described in epistemic, communicative, ethical and pragmatic dimensions. Particularly affected are practices that depend on the preservation of slow deliberation, correctable learning processes and trust in responsible design, for instance, democratic governance. The universal need to adapt to digital-technological dependency means a loss of future in the sense of a space that is amenable to creative shaping. This critical philosophical thesis counters the popular optimistic narrative that digitalization is the key to opening up the future.*

Keywords: *digital transformation; disruption; digital twin; technological dependency; loss of future*

1. Digitalisierung als Narrativ und Prozess

Die Digitalisierung überformt, so die gängige Redeweise, praktisch alle Bereiche der Lebensgestaltung, individuell wie kollektiv. Gemeinsam mit innovativen Nutzungsideen und Geschäftsmodellen eröffnet der digitaltechnische Fortschritt in rascher Folge neue Handlungsoptionen (z.B. Neugebauer 2018): Mustererkennung durch Big Data Analytik, Beschleunigung von Innovationsprozessen, individuali-

sierte Dienstleistungen, Roboter als künstliche Assistenten, lernende Algorithmen, autonome Entscheidungssysteme (ADM), selbst fahrende Autos, neue Formen der Schaffung von Kunstwerken und Texten, immer bessere Simulation menschlicher Fähigkeiten, und vieles mehr. Visionäre Erzählungen von erheblicher Reichweite über die bereits sichtbaren Effekte der Digitalisierung hinaus verleiten dazu, sie als Epochenbruch zu verstehen.

Der Begriff der Digitalisierung dient dabei auf zwei Ebenen unterschiedlichen Zwecken. Einerseits geht es deskriptiv um die Beschreibung empirisch beobachtbarer Phänomene, etwa den Einsatz digitaler Werkzeuge in Arbeitswelt, Freizeitgestaltung und öffentlicher Kommunikation sowie die Folgen dieses Einsatzes. Andererseits stellt er ein dominantes *Narrativ* zeitgenössischer Diagnostik mit überschießenden und teils visionären, teils fatalistischen und teils normativen Intentionen dar wie etwa in den Formulierungen eines ›digitalen Determinismus‹ (Mainzer 2016), der Rede von der ›digitalen Revolution‹, trans- und posthumanistischen Ideen der Überwindung defizitärer menschlicher durch eine vermeintlich perfekte digitaltechnische Zivilisation (Loh 2018; Grunwald 2019a) sowie politische Botschaften, die Gesellschaft müsse sich ›fit machen für die Digitalisierung‹.

Im Fokus dieses Beitrags[1] steht die Exploration des philosophischen Blicks auf die Digitalisierung als Prozess in spezifischer Hinsicht. Die zunehmende Ausstattung des individuellen und kollektiven Lebens, zusehends aber auch der gebauten Umwelt in Gebäuden, Ortschaften und Infrastrukturen mit digital funktionierenden Sensoren, Datenspeichern und Algorithmen ermöglicht diesen Prozess und stellt sozusagen die Grammatik dafür bereit. Digitalisierung *als Prozess* bezieht sich im Folgenden auf die auf diese Weise ermöglichte Transformation gesellschaftlicher Zusammenhänge, so etwa in Bezug auf Demokratie, durch die Umstellung der Wirtschaft auf eine Daten- und Wissensökonomie, neue Mensch/Maschine-Verhältnisse (Ethikrat 2023), Veränderungen der Arbeitswelt (Börner et al. 2018), der öffentlichen Kommunikation und in der Selbst- und Weltwahrnehmung von Menschen (Grunwald 2021).

Das wesentliche, die transformative Leistung erst ermöglichende Element der Digitalisierung liegt in der digitalen *Verdopplung der Welt*. Der analogen Welt aus Materie und Energie wird eine digitale Welt aus Daten, Modellen und Algorithmen zur Seite gestellt, in der Datenabbilder die Gegenstände der analogen Welt als ihre so genannten ›digitalen Zwillinge‹ in gewissen Hinsichten repräsentieren sollen. Konsumprofile sind genauso Elemente des digitalen Zwillings von Menschen wie verfügbare medizinische Daten, Daten aus Überwachungskameras oder Bewegungsprofile aus Handydaten. Diese Zwillinge sind speicher- und kopierbar, durch Algo-

1 Der vorliegende Text führt einen früheren Artikel (Grunwald 2019b) weiter und hat von Diskussionen in Workshops des CAIS-Projekts *Philosophische Digitalisierungsforschung* (2019–2022) profitiert.

rithmen zur Mustererkennung nutzbar, jedenfalls solange keine Regulierung dagegensteht, und durch Suchbefehle nach bestimmten Eigenschaften recherchierbar. Die dabei gewonnenen Erkenntnisse können in die analoge Welt rückübertragen und genutzt werden, etwa für individualisierte Werbung. Eine der Visionen vieler Digitalfirmen ist es, möglichst vollständige digitale Zwillinge aller analogen Objekte zu erzeugen, diese im Hintergrund mit schnellen Algorithmen auszuwerten und die Ergebnisse in der analogen Welt für geschäftliche oder politische Zwecke zu nutzen. Diese Operationen in der Welt der digitalen Zwillinge sind im Vergleich zur Auswertung in der analogen Welt dramatisch beschleunigt. Der Wegfall vieler Dämpfungsfaktoren der analogen Welt, etwa durch die räumlichen und zeitlichen Dimensionen vieler Prozesse, führt zu erheblichen Beschleunigungseffekten und erhöht sowohl Möglichkeit als auch Risiko disruptiver Veränderung. Dieser Thematik vor allem gelten die folgenden Überlegungen.

Philosophie und Ethik werden zur Orientierung in der und für die Digitalisierung nachgefragt. Einerseits geht es um Analysen zu und Antworten auf konkrete ethische Herausforderungen wie etwa in Computerethik, Datenethik, Internetethik oder Maschinenethik (Misselhorn 2018), behandelt häufig auch in Ethik-Kommissionen mit spezifischem Auftrag (z.B. Ethik-Kommission 2017). Andererseits stehen grundsätzliche philosophische Fragen in der Diskussion, so etwa zu Selbstverständnis und Zukunft des Menschen angesichts Künstlicher Intelligenz mit schnell wachsenden Fähigkeiten (z.B. Mainzer 2016; Nida-Rümelin/Weidenfeld 2018). Philosophische Herausforderungen der Digitalisierung erschöpfen sich also nicht in Aufgaben für Angewandte Ethik, sondern eröffnen auch Anfragen an handlungstheoretische, demokratietheoretische, bewusstseinsphilosophische, anthropologische und technikphilosophische Reflexion.

Im Folgenden behandele ich spezifische Konstellationen, die im Zuge der raschen Digitalisierung immer wieder Gegenstand wissenschaftlicher, philosophischer und öffentlicher Debatten sind. Das Thema, in dieser Formulierung vielleicht paradox klingend, sind *allmähliche Disruptionen* auf gesellschaftlicher Ebene. Damit sind Umbrüche mit erheblichem bis dramatischem Schadenspotential gemeint, die nicht unerwartet und plötzlich auftreten wie eine weltweite Pandemie oder ein Angriffskrieg, sondern die sich allmählich aufbauen, bis sie schließlich zum Zerbrechen vertrauter Konstellationen führen (Abschn. 2). Hintergrund ist die Beobachtung, dass zentrale Krisenphänomene der Gegenwart wie Klima- und Umweltprobleme, allmählich einkehrende und sich verfestigende Abhängigkeiten mit ebenso allmählicher Vulnerabilitätssteigerung und die Krise der Demokratien nicht plötzlich hereinbrechen, sondern sich mit vielen Vorzeichen langsam aufbauen und erst allmählich große bis dramatische Ausmaße annehmen. Dieser Typus potentieller und allmählicher Disruption lässt sich, so die These, in vielen Bereichen der Digitalisierung als Prozess erkennen (Abschn. 3). Schließlich werden Anfragen *allmählicher Disruption* an die Philosophie in den Blick genommen (Abschn. 4).

2. Disruption als Begriff der Zeitdiagnostik

Disruption ist in den letzten ca. zehn Jahren zu einem vielfach verwendeten Begriff geworden. Obwohl die Wortherkunft auf eher unangenehm klingende Bedeutungen verweist (lt. disrumpere = platzen, zerbrechen, zerreißen), geschieht dies einerseits in positiver Intention. So stehen disruptive Innovationen als technologische Sprünge oder paradigmatische Wechsel von Geschäftsmodellen hoch im Kurs der Innovationspolitik, vielfach motiviert durch die Veränderung wirtschaftlicher Verhältnisse in der Digitalisierung. Im Gegensatz zum inkrementellen, auf allmählichen Produktverbesserungen beruhenden Innovationsgeschehen zielt Disruption auf grundsätzliche Umwälzungen, in denen teils über Jahrzehnte bestehende Marktverhältnisse in kurzer Zeit umgestürzt werden. Den Gewinnern (oft »Disruptoren« genannt) stehen neue, oft globale Marktchancen offen. Auch gänzlich neue Märkte können entstehen, wie immer wieder im Rahmen der digitalen Transformation, etwa in den *social media*. Konkreter politischer Ausdruck dieses Denkens ist die im Jahre 2019 von der deutschen Bundesregierung gegründete *Bundesagentur für Sprunginnovationen* (SPRIND), die mit einem erheblichen Budget disruptive Technologie und Innovationen fördern soll, um die deutsche Volkswirtschaft zu stärken.

Der Beginn dieser Karriere des Disruptionsbegriffs liegt in der Theorie disruptiver Technologie (Bower/Christensen 1995). Diese wurde rasch auf den Bereich der Innovation ausgeweitet (z.B. Danneels 2004), indem auch von bahnbrechenden und disruptiven Innovationen gesprochen wird. Die Erwartungen an disruptive Innovation sind teils erheblich (de la Vera/Ramge 2021). Eine zentrale Annahme darin ist die fortschreitende Beschleunigung aller Innovationsvorgänge, die Regulierung und vorausschauende Politik obsolet werden lassen. Freilich sind sowohl die Begriffsbestimmungen als auch Voraussetzungen und Erwartungen umstritten (Gans 2017): »›Disruption‹ is a business buzzword that has gotten out of control. Today everything and everyone seem to be characterized as disruptive – or, if they aren't disruptive yet, it's only a matter of time before they become so«. In dieser Kritik verflacht der Begriff der Disruption zu einem Synonym für Erfolg, was freilich an der bislang erfolgreichen Begriffskarriere nichts geändert hat.

Andererseits werden seit einigen Jahren auch *Krisenphänomene* in den begrifflichen Kontext der Disruption gestellt. Vor allem gelten Corona-Pandemie und Ukraine-Krieg als disruptive Ereignisse. Beide haben eine lange Zeit der weitgehenden Stabilität, jedenfalls im Globalen Norden, beendet und zeigen, nach verbreiteter Diagnose, den Übergang in eine Zeit der permanenten Krise an. Auf diese Weise wird der Begriff der Disruption zur Bezeichnung des Zerbrechens stabiler gesellschaftlicher Zustände mit der nicht selten befürchteten Folge dramatischer Auswirkungen eingesetzt. Indiz dafür ist die verbreitete Kommunikation katastrophischer Narrative in der öffentlichen Debatte, vor allem die Sorge vor einem Atomkrieg, der

Klimawandel als Ende der Bewohnbarkeit der Erde und das nachlassende Vertrauen in Demokratie in vielen Staaten.

In beiden erwähnten Bedeutungen meint Disruption den plötzlichen Abbruch vertrauter Konstellationen als, je nach Kontext, erwünschtes oder befürchtetes Ereignis. Stabilitätserwartungen, Kontinuitätsannahmen und Planungssicherheiten zerbrechen und lassen die Aussichten auf Zukunft in einem unsicheren Licht erscheinen. Das Platzen, Zerbrechen und Zerreißen der lateinischen Wortherkunft (s.o.) weist semantisch auf die Zeitform mehr oder weniger plötzlicher, abrupt auftretender Ereignisse hin. So gesehen erscheint die Rede von der *allmählichen* Disruption zunächst begrifflich falsch, widersinnig oder zumindest paradox.

Der nähere Blick erlaubt eine Differenzierung. Semantisch zeigen sich im Begriff der Disruption zwei Bedeutungsanteile: zum einen das *Zerbrechen* bislang stabiler Verhältnisse, zum anderen die *Schnelligkeit* dieses Zerbrechens. Während der erste Bedeutungsanteil dem Begriff etymologisch inhärent ist, kann der zweite flexibler gehandhabt werden. Zeitskalen des Zerbrechens sind dehnbar. So wird beispielsweise die Erfindung des Buchdrucks im späten Mittelalter gerne als disruptiv dargestellt – historisch gesehen erstreckte sich diese Disruption über viele Jahrzehnte der Diffusion in die damaligen Gesellschaften hinein. Zerbrechen und Abbruch können sich auch, und darauf kommet es mir jetzt an, *allmählich* über längere Zeiträume aufbauen, sich durch schwache Signale ankündigen und erst im zeitlichen Verlauf zu nur scheinbar plötzlich auftretenden Disruptionen im Sinne eines qualitativen Bruchs führen.

So wurde nach dem Beginn des Ukraine-Kriegs vielfach festgestellt, dass dieser nicht ohne Vorwarnung begann, sondern eine Vorgeschichte in Form russischer Angriffe auf Gebiete der ehemaligen Sowjetunion hatte. Viele Beispiele für Disruption mit erkennbaren, wenngleich oft nicht erkannten Vorzeichen sind aus der technischen Welt bekannt, vor allem Materialermüdung und Verschleiß. Die tägliche Belastung vieler technischer Objekte wie z.B. von Keilriemen in altmodischen Automobilen oder Brückenbauwerken, führt allmählich zu Verschleiß und Degradierung. Lange funktionieren sie dennoch verlässlich, bis der Verschleiß ein Ausmaß erreicht, bei dem das Bauteil von einem auf den anderen Moment ausfällt, dass also in dem gewählten Beispiel der Keilriemen reißt oder die Brücke einstürzt. Ein Beispiel aus der Klimadebatte sind die so genannten Kipp-Punkte (*tipping points*, s. Gladwell 2000). Bei weiterer Erwärmung könnten selbstverstärkende Rückkopplungseffekte einsetzen, die in kurzer Zeit dramatische Folgen, also eine disruptive Wirkung hätten. Das Disruptive ist in Vorgängen dieser Art also in inkrementellen und nur schwer erkennbaren Prozessen angelegt, kann lange unerkannt bleiben und dem frühzeitig intervenierenden und vorbeugenden Eingriff entgehen, jedoch potentiell weitreichende und plötzlich eintretende Folgen erzeugen. Die *Tragik* derartiger allmählicher Entwicklungen ist, so könnte man in leicht existenzialistischer

Emphase sagen, dass sich im inkrementellen Verlauf schwerwiegende Disruptionen zwar schleichend ankündigen, aber dann abrupt vollziehen können.

Mit dieser semantischen Differenzierung wird im Folgenden die Möglichkeit *allmählicher Disruption* in der Digitalisierung als Prozess betrachtet. Es soll dabei um mögliche Entwicklungen mit Schadens- oder sogar katastrophischem Potential gehen, nicht um die Frage nach (von manchen Akteuren) erwünschter disruptiver Innovation.[2]

3. Disruptive Potentiale der digitalen Transformation

Der teils sehr schnelle Erfolg vieler Entwicklungen im Rahmen der Digitalisierung, beispielsweise der globalen Expansion der *social media* innerhalb weniger Jahre ungefähr ab 2010 oder gegenwärtig die schnelle Diffusion von KI-Anwendungen, führt zu Verschiebungen in vielen Bereichen. Hierzu gehören etwa Mensch/Technik-Verhältnisse, Verantwortungsverteilungen, industrielle Produktion, Sicherheitspolitik und Überwachung, öffentliche Kommunikation und politische Meinungsbildung, Solidarität und Wettbewerb, Arbeitsmarkt- und Arbeitswelt sowie Freizeitverhalten und Medienkonsum. Einige dieser Verschiebungen bilden den Kern der teils aufgeregten öffentlichen Debatte zur Digitalisierung und vieler weitreichender Befürchtungen darin (Grunwald 2019). Diese Verschiebungen werden im Folgenden unter dem Aspekt potentieller und allmählicher Disruption diskutiert. Dabei lassen sich zwei häufig ineinander verflochtene Perspektiven unterscheiden: Verschiebungen in Mensch/Technik-Verhältnissen (3.1 und 3.2) sowie in gesellschaftlichen Konstellationen (3.3-3.5).

3.1 Normierung menschlichen Handelns

Nach dem gängigen Verständnis sollen Technik und darauf aufbauende Dienstleistungen als Mittel für menschliche Zwecke dienen, Bedürfnisse befriedigen und Probleme lösen, um, so die Erzählung seit der Aufklärung, die Optionenvielfalt menschlichen Handelns zu erweitern und die Emanzipation des Menschen zu befördern. Dies ist jedoch nur die halbe Wahrheit. Denn während Technik menschliche Handlungsoptionen erweitert, führt sie simultan zu Anpassungsnotwendigkeiten unterschiedlichster Art bis hin zum Zwang (Grunwald 2022).

2 Freilich bedürfte auch diese Richtung einer philosophischen Kritik, ist doch die mangelnde ex ante-Einschätzbarkeit potentieller Folgen eine notwendige, mit dem begrifflichen Kern disruptiver Innovation verbundene Begleiterscheinung, was ernsthafte Fragen an die Verantwortbarkeit dieses Innovationstypus motiviert.

Technische Systeme strukturieren und regulieren menschliches Handeln, etwa durch Bedienungsanleitungen, Vorschriften und Benutzeroberflächen. In vielen Feldern ist dies trivial, wenn etwa in der Nutzung eines Spatens zum Umgraben bestimmte körperliche Bewegungen erforderlich sind oder wenn zur Bedienung einer Waschmaschine die Bedienungsanleitung zu beachten ist. Digitale Technik ändert jedoch subtil menschliches Handeln und Verhalten, möglicherweise ohne, dass dies bemerkt wird. Die Debatte um ›Software als Institution‹ (Orwat et al. 2010) hat darauf aufmerksam gemacht, dass Softwaresysteme regulierende Kraft haben können, z.B. durch die Regelung von Transaktionen oder von Zugangs- und Nutzungsrechten. So strukturieren privat geführte *social media*-Plattformen die öffentliche Kommunikation, sortieren Suchmaschinen mit von Privatfirmen entwickelten Algorithmen die Weltwahrnehmung ihrer Nutzer und strukturieren Online-Plattformen Geschäftsprozesse und Crowd-Sourcing.

Spezifische Anpassungsnotwendigkeiten entstehen im Zusammenwirken autonomer Software- und Robotiksysteme mit Menschen. In der Industrie 4.0-Welt, in der Roboter mit Menschen in der industriellen Produktion zusammenarbeiten sollen, muss aus Funktionalitäts- wie auch Sicherheitsgründen eine missverständnisfreie Kommunikation an diesen Schnittstellen gewährleistet werden, ähnlich beim autonomen Fahren im Mischverkehr mit menschlichen Verkehrsteilnehmern. Die Forderung

> »Um eine effiziente, zuverlässige und sichere Kommunikation zwischen Mensch und Maschine zu ermöglichen und Überforderung zu vermeiden, müssen sich die Systeme stärker dem Kommunikationsverhalten des Menschen anpassen und nicht umgekehrt erhöhte Anpassungsleistungen dem Menschen abverlangt werden« (Ethik-Kommission 2017: 13)

ist leicht zu erheben und anthropologisch nachvollziehbar. Jedoch steht zu befürchten, dass es im realen Ablauf nicht so kommen wird, sondern dass die fortschreitende Digitalisierung das menschliche Handeln allmählich nach den Anforderungen technischer Systeme und technischer Kommunikation reguliert und normiert. Trotz allen Bemühens um menschliche Autonomie und Wahrung der Wahlfreiheit könnte es zum Gegenteil kommen, nämlich zum schleichenden und unbemerkten Verlust von Freiheiten. Als Indiz für diese Vermutung wird vor allem der Sog des Sicherheitsdenkens angeführt, wie etwa beim autonomen Fahren. Im Rahmen eines ›Schiefe-Ebene-Arguments‹ könne das Postulat der Sicherheit in der Folge ihrer technischen Durchsetzung menschliche Freiheit letztlich komplett aushebeln. Auch wenn

> »[…] es dem Leitbild des mündigen Bürgers widersprechen [würde], würde der Staat weite Teile des Lebens zum vermeintlichen Wohle des Bürgers unentrinn-

> bar durchnormieren und abweichendes Verhalten sozialtechnisch bereits im Ansatz unterbinden wollen« (Ethik-Kommission 2017: 20),

bedeutet das dennoch nicht, dass sich das Leitbild in der Wirklichkeit durchsetzt. Ähnlich verhält es sich in der digitaltechnisch ermöglichten und durchsetzbaren Überwachung im privaten und öffentlichen Bereich, die immer wieder mit Sicherheitsargumenten gegenüber Freiheitsargumenten begründet wird. Die »allmähliche Disruption« in diesem Feld wäre ein unbemerktes Hinübergleiten in eine Welt, in der das Sicherheitsinteresse des Staates in der Hierarchie der abzuwägenden Aspekte ganz nach oben wandert und es in der Folge zu immer weiteren digitaltechnisch durchgesetzten Normierungen des menschlichen Handelns kommt, die letztlich zum Abschied vom freiheitsorientierten Individualismus hin zu einem gelenkten Kollektivismus führen könnten.

3.2 Abhängigkeit als latente Disruption

Moderne Gesellschaften sind bereits heute vollständig vom reibungslosen Funktionieren kritischer Infrastrukturen wie z.B. der Stromversorgung abhängig (Petermann et al. 2011). Dies gilt in zunehmendem Maß auch für digitale Infrastrukturen. Bei einem Ausfall des Internet würden Finanztransaktionen unmöglich, würde die Weltwirtschaft zusammenbrechen, wäre keine mediale Kommunikation mehr möglich, würde die medizinische Diagnostik vieler etablierter Verfahren beraubt, würden die internationalen Logistikketten stillstehen, und vieles mehr. Mit der zunehmenden Einführung von ADM-Systemen (*automated decision-making*) entsteht eine Abhängigkeit von KI-gesteuerten Systemen, die gemeinsam mit deren *black box*-Charakter und Intransparenz, aber auch aufgrund des psychologischen *automation bias* (Ethikrat 2023) eine zunehmende Abhängigkeit von diesen Systemen in entscheidungsrelevanten Kontexten wie Polizei und Sozialwesen bedeutet.

Die allmähliche Verdrängung des Bargelds ist ein aktuelles Beispiel für die Ambivalenz technischer Infrastrukturen. War zunächst der bargeldlose Zahlungsverkehr als Erleichterung für Wirtschaft und Privatpersonen eine *zusätzliche* Option neben dem Bargeld, findet ein allmählicher Übergang zu einer Welt ohne Bargeld statt (TAB 2020). Bargeld wird allmählich verdrängt, teils durch Konsumentenverhalten und Bequemlichkeit, teils durch Anreize und Regulierung aus Politik und Wirtschaft unter der Argumentation, so könnten Schwarzmarkt und Schwarzarbeit unmöglich gemacht werden, in der Pandemie verstärkt durch das Argument der prophylaktischen Kontaktlosigkeit. Sobald sich der bargeldlose Zahlungsverkehr vollständig durchgesetzt haben sollte, wie dies in einigen Ländern bereits weitgehend der Fall ist, ist die Wahlfreiheit der Bezahlpraxis verschwunden und bei Ausfall des Internets wäre kein Einkauf oder Zahlungsverkehr mehr möglich. War die bargeldlose Zahlung zunächst eine zusätzliche Option und erhöhte die Wahlmöglichkeiten, wurde

sie allmählich dominant, schließlich aufgrund des Verschwindens des Bargelds alternativlos und damit zum Zwang mit der Kehrseite der Abhängigkeit.

Abhängigkeiten sind für sich keine Disruption, aber sie tragen deren Keim in sich. Total gewordene Abhängigkeiten sind *latente Disruptionen*. Als Disruptionen auf Abruf bauen sie sich über wachsende Abhängigkeiten allmählich auf, können aber im Ernstfall, wenn z.B. die digitalen Techniken nicht mehr reibungslos funktionieren würden, abrupt eintretende und möglicherweise katastrophale Folgen haben. Auf ihr unbegrenztes reibungsloses Funktionieren zu setzen und Funktionsfähigkeit und Stabilität moderner Gesellschaften davon abhängig zu machen, ist allerdings eine Wette »ums Ganze« im Sinn von Hans Jonas (Jonas 1979) und entsprechend ethisch problematisch. Unerwartete Hacker-Ereignisse, ein Zusammenbruch der staatlichen Ordnung oder schwere wirtschaftliche Turbulenzen könnten auch Infrastrukturen wie das Internet betreffen und im schlimmsten Fall dysfunktional machen. Auch wenn es meist schwer, wenn nicht unmöglich ist, den Zeitpunkt in der allmählich ablaufenden Entwicklung zu erkennen, ab dem die vollständige Abhängigkeit einsetzt, dürfte dieser Punkt in Bezug auf viele digitale Infrastrukturen und Plattformen längst überschritten sein – was bedeutet, dass moderne Gesellschaften bereits im Modus dieser latenten Disruption sind.

Obwohl digitale Infrastrukturen als Mittel zu menschlichen Zwecken aufgebaut werden, verschiebt sich das Zweck/Mittel-Verhältnis allmählich: aufgrund steigender Abhängigkeit von digitaler Technik geraten Menschen in die Notwendigkeit, alles zu tun, um diese in gutem Zustand zu erhalten. Die technische Infrastruktur wird vom Mittel zum Zweck. Dieses allmähliche Umschlagen von Freiheit und Machtverhältnissen entspricht der Hegelschen Dialektik von Herr und Knecht (reformuliert nach Grunwald 2019a: 17):

> »Ein Herr hat einen Knecht. Dieser Knecht muss alles für den Herrn tun. Dadurch verlernt der Herr die lebensnotwendigen Dinge. Der Herr wird abhängig vom Knecht, und schließlich wird aus dem Knecht der eigentliche Herr. Der Herr muss dann dafür sorgen, dass es dem Knecht gut geht. Fatal daran ist: Der Übergang vom Herrn zum Knecht geschieht unmerklich.«

3.3 Verlust der Zukunft

Digitale Techniken gelten als Synonym für Zukunft, ähnlich wie die Kernenergie im damals voller Optimismus so genannten Atomzeitalter der 1950er und 60er Jahre. Jedoch operieren digitale Techniken grundsätzlich auf Basis vergangener Daten. So bilden die digitalen Zwillinge (s.o.) immer nur eine Welt von gestern ab, z.B. indem Kundenprofile ausschließlich auf Basis vergangener Kaufakte und Konsumprozesse erstellt werden können. Digitale Zwillinge bilden grundsätzlich nur die *Vergangenheit* ihrer analogen Originale ab. Big-Data-Technologien können nur vergangene

Daten auswerten und ebenso vergangene Muster erkennen. KI-Systeme können nur an Daten aus der Vergangenheit trainiert werden, da Daten aus der Zukunft nicht verfügbar sind. Auch wenn mit Hilfe von KI und *Big Data* versucht wird, quantitative Prognosen zu erstellen, basieren diese auf Mustererkennung anhand vergangener Daten. Durch den unabdingbaren Datenbezug ist digitale Technik unentrinnbar auf vergangene Verhältnisse fixiert. Wenn Datensätze, digitale Zwillinge und durch KI aufgedeckte Korrelationen und Muster für Zukunftsaussagen genutzt werden, werden vergangene Verhältnisse auf die Zukunft übergewälzt, ihr sozusagen übergestülpt. Zukunft als ein zumindest teilweise offener Raum alternativer Pfade und Möglichkeiten wird durch eine datenbasierte Verlängerung der Vergangenheit ersetzt.

Die Digitalisierung bzw. einige ihrer Bereiche könnten auf diese Weise konservativ werden, indem sie Zukunftsgestaltung nicht an neuen Ideen, sondern an alten Daten ausrichten. Angesichts vielfacher anthropologischer Bestimmungen des Menschen als Wesen mit Zukunft und der Fähigkeit der vergegenwärtigenden Reflexion möglicher Zukünfte (z.B. Kamlah 1973), die nicht nur Verlängerung der Vergangenheit, sondern auch kreative Neuschöpfung im offenen Raum vieler Möglichkeiten sein und sogar kontrafaktischen und utopischen Charakter tragen kann, kann es hier zu einer allmählichen Disruption kommen, in der die grundsätzliche Offenheit der Zukunft zugunsten einer datengetriebenen Orientierung an der Vergangenheit in den Hintergrund tritt oder ganz verschwindet.

Verstärkt werden kann diese allmähliche Entwicklung durch den grassierenden digitalen Determinismus (Mainzer 2016). Danach ist die Digitalisierung von einer eigendynamischen Entwicklung gekennzeichnet und könne nicht nach Werten oder gesellschaftlichen Zielen gestaltet werden. Sie fahre wie ein Zug mit hoher Geschwindigkeit, den man weder aufhalten noch in seiner Richtung beeinflussen könne. Eine andere rhetorische Form beschreibt sie mit (vermeintlichen) Sachzwang-Argumenten und (ebenso vermeintlichen) Alternativlosigkeiten als ein unausweichliches Naturereignis wie etwa einen Tsunami oder ein Erdbeben. In der deterministischen Perspektive bliebe Mensch und Gesellschaft nur die Anpassung. Anpassung jedoch bedeutet ebenfalls einen Verlust von Zukunft im Sinne eines gestaltungsoffenen Raumes.

In Hinblick auf allmähliche Disruption ist hier das Phänomen selbsterfüllender Narrative zu erwähnen: wenn sehr viele Menschen davon überzeugt sind, dass die Digitalisierung eigendynamisch und nicht gestaltbar abläuft, dann werden mögliche Interventionen und Gestaltungsoptionen gar nicht erst wahrgenommen. Das technikdeterministische Narrativ würde durch diese fatalistische Haltung bestätigt und gewänne weitere Überzeugungskraft.

Der digitale Determinismus ist weder theoretisch noch empirisch haltbar, worauf vielfach hingewiesen wurde (z.B. bereits Ropohl 1982). Wissenschaftliche Erkenntnis und gesellschaftliche Haltungen oder Trends sind jedoch oft nicht im

Einklang miteinander, sondern im Widerspruch. Auch wenn von wissenschaftlicher, z.B. sozialkonstruktivistischer Seite immer wieder auf die Gestaltungsoffenheit der Digitalisierung und auf alternative Entwicklungspfade hingewiesen wird, auch wenn praktische Ansätze der Gestaltung wie etwa das Value Sensitive Design (vgl. Grunwald 2015) exploriert und erprobt werden, impliziert das nicht bereits einen Einfluss auf gesellschaftlicher Ebene. Dort kann es trotz wissenschaftlicher und philosophischer Widerlegung zum allmählichen Verlust der Zukunft als Gestaltungsraum kommen.

3.4 Verantwortungsdiffusion

An den ständig neu entstehenden Schnittstellen zwischen Menschen und digitalen Systemen werden Zuständigkeiten neu verteilt. Automatisierte bzw. autonome Entscheidungssysteme (ADM-Systeme), industrielle Produktion in der Kooperation zwischen Menschen und Robotern in der Industrie 4.0 und das autonome Fahren sind Beispiele. Einige soziologische Modelle sprechen angesichts dieser Konstellationen von verteilter Handlungsträgerschaft zwischen Mensch und Technik (Latour 2005; Rammert/Schulz-Schaeffer 2002), auch philosophische Theorien nehmen eine zwischen Mensch und Maschine verteilte Verantwortung an (Floridi 2016). Sie können für bestimmte Phänomene auf der empirischen Ebene durchaus einen Erkenntnisgewinn bedeuten. Über Verantwortungszuschreibung im normativen Sinne sagen sie jedoch nichts aus, da Verantwortung handlungstheoretisch nur Akteuren mit intentionalem Handlungsvermögen zukommen kann. Digitale, auch KI-gestützte Systeme verfügen nicht über Intentionen, sondern führen komplexe mathematische und statistische Operationen auf der Basis von Daten durch. Die Möglichkeit der Verantwortungszuschreibung und -trägerschaft bleibt daher, jedenfalls gegenwärtig und in der absehbaren Zukunft, Menschen vorbehalten (Ethikrat 2023).

Die Lokalisierung von Verantwortung und ihre Zuschreibung an spezifische Akteure wird allerdings durch Digitalisierung komplex. Zwar verbleiben Entscheidungen und damit Verantwortung beim Menschen, zusehends jedoch auf eine digital vermittelte Weise. Digitale Systeme und ihre Hersteller schieben sich zwischen intentional handelnde Menschen und realweltliche Effekte. Verantwortung wandert von individuellen Autofahrern oder, im Falle von militärischen Drohnen, von Soldaten zu Personen und Institutionen im Hintergrund, zu Firmen, Programmierern, Managern, Geheimdiensten, Generälen oder Regulierungsbehörden:

> »Die dem Menschen vorbehaltene Verantwortung verschiebt sich bei automatisierten und vernetzten Fahrsystemen vom Autofahrer auf die Hersteller und Betreiber der technischen Systeme und die infrastrukturellen, politischen und rechtlichen Entscheidungsinstanzen.« (Ethik-Kommission 2017: 11)

Auf Basis der vorliegenden Erfahrungen mit Verantwortungszuschreibung in komplexen und arbeitsteiligen Zusammenhängen, beispielsweise in großen Unternehmen, erscheint die Aufgabe, Verantwortung bis hinein zu Schuld- und Haftungsfragen konkret festzulegen, als zwar anspruchsvoll aber machbar. Mit der Komplexität digital zwischen Menschen und Digitaltechnik verteilter Zuständigkeiten steigen allerdings die Gefahr einer allmählichen ›Verantwortungsdiffusion‹ ins Nichts und das Risiko intentionaler Verantwortungsverschleierung.

Eine wichtige Rolle hierbei spielt ein psychologischer Effekt, der spezifisch für digitale und insbesondere KI-Systeme ist: der »*automation bias*« (Saldar et al. 2020; vgl. auch Ethikrat 2023). Empirisch wurde gezeigt, dass viele Menschen algorithmisch erzeugten, auf großen Datenmengen beruhenden und mit KI-unterstützten Entscheidungsverfahren berechneten Ergebnissen stärker vertrauen als Menschen, vermutlich aufgrund von Objektivitätsunterstellungen gegenüber mathematischen Verfahren und einem Subjektivitätsverdacht gegenüber Menschen. Damit wird Verantwortung – zumindest unbewusst – den Algorithmen als »Quasi-Akteuren« zugeschrieben. Auch bei einer rechtlich, handlungstheoretisch und ethisch reflektierten Zuschreibung von Verantwortung in Entscheidungsprozessen, in denen ein KI-System normativ strikt auf *Entscheidungsunterstützung* begrenzt und menschliche Entscheider die Entscheidung treffen müssen, könnten KI-Systeme auf diese Weise allmählich in die Rolle der »eigentlichen« Entscheider geraten und damit menschliche Verantwortung substantiell entleeren, sie zu einer bloß formalen Hülle degradieren.

Im Zusammenwirken mit dem Blackbox-Charakter von komplexen KI-Systemen könnten hier zusätzlich schleichend Abhängigkeiten von intransparenten und unverständlichen technischen Systemen entstehen, während die menschliche Urteilskraft aufgrund des Vertrauens in das technische System und mangelnder Nutzung der eigenen Fähigkeiten verkümmert. (Bainbridge 1983 spricht diesbezüglich von »ironies of automatization«.) Die allmähliche Disruption bestünde hier in der Kombination des Verlierens menschlicher Fähigkeiten in Bezug auf Urteilskraft und kritisches Denken mit zunehmender Abhängigkeit von den digitaltechnischen Entscheidungsverfahren, letztlich also von den Werten und Interessen der hinter diesen stehenden menschlichen Akteure in den großen Digitalkonzernen. Wenn in digitalen Beratungsangeboten, so etwa bei Gesundheits-Apps, algorithmischer Rechtsberatung oder in Finanzgeschäften nicht transparent ist, wem hier das Vertrauen entgegengebracht wird, worauf sich dieses gründet und wer welche Verantwortung trägt, können allmählich undurchschaubare und zunehmend intransparente Konstellationen entstehen, welche die Bedingungen der Möglichkeit ethischer Reflexion und nachvollziehbarer Verantwortungszuschreibung untergraben würden.

3.5 Abbruch von Reflexions- und Lernmöglichkeiten

Beschleunigung ist Teil des kapitalistischen Wirtschaftssystems. Sie setzt Kreativität und Innovation frei, vor allem durch Wettbewerb. Beschleunigung ist ein vielfach im Kontext der Digitalisierung diskutiertes Phänomen, war jedoch bereits vor hundert Jahren ein beachtliches Thema:

> »Domestic life, political institutions, international relations and personal contacts are shifting with kaleidoscopic rapidity before our eyes. We cannot appreciate and weigh the changes; they occur too swiftly. We do not have time to take them in. No sooner we begin to understand the meaning of one such change than another comes and displaces the former.« (Dewey 1931: 54)

Die Erhöhung der Rechengeschwindigkeit, die Möglichkeit, Millionen von Optionen in kürzester Zeit durchzurechnen, die Verknüpfung kreativer Ressourcen über das Internet und die Beschleunigung von Datentransfer und Kommunikation, vieles vermittelt und weiter beschleunigt durch digitale Zwillinge, verkürzen die Innovationszyklen. Wie eingangs schon erwähnt, zieht daher seit einigen Jahren die ›disruptive Innovation‹ als extreme Beschleunigung Faszination auf sich. Sie ist das Gegenteil allmählicher Innovationsprozesse und denkt die digital ermöglichte Beschleunigung im Extrem.

Allerdings kennt die klassische Wirtschaftstheorie auch zerstörerischen Wettbewerb. Die Beschleunigungsspirale ist nicht beliebig weit überdrehbar, sondern in Gefahr, die menschlichen und natürlichen Ressourcen zu übernutzen, aus denen sie sich speist. Eine Sorge in Bezug auf die Digitalisierung als Prozess bezieht sich auf negative und möglicherweise ruinöse Folgen immer weiterer Beschleunigung, insbesondere zur Frage, ob und wann die Beschleunigung wichtige Bedingungen der Reflexion grundsätzlich unterminieren könnte. Dies würde in Widerspruch mit den Prinzipien der Aufklärung, den Erkenntnissen der Technikfolgenabschätzung (Grunwald 2022) und den Anforderungen an nachhaltige Entwicklung stehen:

> »[Für nachhaltige Entwicklung] sind institutionelle Bedingungen zu entwickeln, die eine über die Grenzen partikularer Problembereiche und über Einzelaspekte hinausgehende Reflexion von gesellschaftlichen Handlungsoptionen ermöglichen.« (Kopfmüller et al. 2001: 305)

Reflexivität meint die vorausschauende und vorsorgende Befassung mit Folgen von Handlungen und Entscheidungen auf den verschiedensten Ebenen (Beck 1986), impliziert also die Antizipation dieser Folgen bereits *vor* der Ausführung von Handlungen sowie die Berücksichtigung der Ergebnisse der Reflexion in den subsequenten Entscheidungsprozessen. Sie setzt das sorgfältige Bedenken und Beraten, das Ab-

wägen von Alternativen, die Suche nach dem rechten Maß und ethisch legitimierten Kriterien voraus. All dies benötigt in doppelter Weise Zeit: zum einen für die Beratungs- und Erwägungsprozesse selbst und zum anderen zur Umsetzung der Ergebnisse in praktische Entscheidungen. Allmähliche Disruption würde hier bedeuten, dass mit dem Argument kapitalistischen Wettbewerbs die gesellschaftlichen Strukturen und Fähigkeiten zur Reflexion schleichend ausgehöhlt würden. Im Narrativ eines innovationsorientierten Fatalismus unter dem Primat wettbewerblichen Denkens kann man sich Reflexion nicht mehr leisten, da ansonsten die Konkurrenz schneller ist und Marktvorteile gewinnt.

Unbegrenzte Beschleunigung stößt damit nicht nur an Grenzen der Ressourcen-Verfügbarkeit und menschlicher Gewöhnung, sondern auch an Grenzen von Vernunft und Verantwortung. Denn wie die Semantik der Disruption schon besagt, bringt Disruption es mit sich, dass über ihre Folgen im Vorhinein wenig oder nichts gewusst werden kann, da prospektives Wissen nur aufgrund von Kontinuitäten gewonnen werden kann. Komplett diskontinuierliche Vorgänge wären Sprünge in das komplett Unbekannte. Was uns berechtigt, der grenzenlosen Beschleunigungsrhetorik zu widersprechen, ist die Tatsache, dass wir Menschen als *zoon politicon* und als moralisches Lebewesen auf Nachdenken, Beratung und Dialog angewiesen sind.

4. Philosophische Anfragen an potentiell disruptive Entwicklungen

Die genannten Beispiele verweisen auf die Phänomenologie allmählicher Disruptionen bzw. allmählicher Entwicklungen mit Disruptionspotential. Es zeigen sich übergreifende Muster vor allem in Bezug auf ihre (1) Epistemologie, (2) Beurteilung, (3) Pragmatik und (4) Kommunikation.

(1) Epistemologische Dimension: wie allmähliche Disruption erkennen?

Zur Phänomenologie allmählicher, sich schleichend aufbauender Entwicklungen gehört ihre meist schlechte Erkennbarkeit. Dies ist vor allem in ihren frühen Phasen eine Herausforderung, wenn die Datenlage schlecht ist und bestenfalls schwache Signale erkennen lässt. Aufgrund der schwachen Evidenz dieser Daten und fehlender Sensibilität für das erst langsam entstehende Disruptionspotential kann es schwer sein, systematische Forschung zur Abklärung des Sachverhalts zu motivieren und entsprechende Budgets zu mobilisieren. Vieles bleibt zunächst tendenziell spekulativ. Bloß *mögliche* Entwicklungen mit Disruptionspotential wechselwirken in einer unbekannten Zukunft mit anderen bloß *möglichen* Entwicklungen auf eine ebenfalls unbekannte Weise, sodass eine epistemologische Gemengelage hoher Komplexität und Unsicherheit entsteht. Erst im Fortschreiten der Entwicklung, also näher am Eintreten einer Disruption, wird die epistemische Evidenz größer,

weil die Effekte sichtbarer und das wissenschaftliche Verständnis der Zusammenhänge tiefergehend werden. Im Wissenszuwachs über den Klimawandel als einer derartigen allmählichen Disruption war dies in den letzten vierzig Jahren immer wieder zu beobachten.

Entsprechend ist in frühen Phasen möglicher Disruption der Auftrag von Hans Jonas (Jonas 1979) nicht oder kaum einlösbar, belastbares Wissen über die ›Fernwirkungen‹ der Digitalisierung zu gewinnen. Valides Folgenwissen mit der Möglichkeit konsequentialistischer Beurteilung gibt es zwar zu vielen Einzel- und Teilfragen der Digitalisierung, zu technischen Innovationen, Produkten, Dienstleistungen, Geschäftsmodellen und Regulierungsoptionen, jedoch noch kaum über langfristige Folgen allmählicher Verschiebungen. Erkenntnistheoretische Herausforderungen richten sich hier vor allem an die Epistemologie der Korrelation angesichts der Möglichkeiten, mit selbst digitalen Mitteln der Modellierung, des Data Mining und der KI schwachen Signalen disruptiver Entwicklungen möglichst belastbar auf die Spur zu kommen.

(2) Ethische Dimension: wie einordnen und beurteilen?

Die epistemologische Gemengelage hat unmittelbare Folgen für die Bewertung und Einordnung der nur allmählich sichtbar werdenden Entwicklung. Die wissenschaftsimmanent naheliegende Schlussfolgerung, dass Ethik sich zurückhalten solle, bis besseres Wissen verfügbar ist (Nordmann 2007), verbietet sich angesichts der hohen Relevanz möglicher Disruption. Angesichts begrenzter Ressourcen müssen verschiedene langsam ablaufende Entwicklungen oder befürchtete Ereignisse miteinander verglichen und nach Dringlichkeit abgestuft werden. Priorisierungen und Dringlichkeitseinschätzungen, hinter denen normative Kriterien und Relevanzen stehen, z.B. hinsichtlich des Leitbilds nachhaltiger Entwicklung, hängen jedoch mit der Evidenz des Wissens zusammen. Ein bloßer Verdacht reicht auch dann nicht für eine hohe Priorisierung mit z.B. daraus resultierender Ressourcenallokation, wenn der Verdacht im Falle seiner Bewahrheitung in eine nach anerkannten Maßstäben verhängnisvolle und auf jeden Fall abzuwendende Entwicklung führen würde. Hier kommt es also zu schwierigen Aufgaben der Bewertung der Lage und ihrer Einordnung im Vergleich mit anderen Entwicklungen.

Entsprechend ist die Befassung mit allmählichen Verschiebungen im Rahmen der Digitalisierung der Angewandten Ethik methodisch vorgelagert. Sie kann als *explorative Philosophie* bezeichnet werden (Grunwald 2010), die sich technikphilosophisch, anthropologisch und gesellschaftstheoretisch den teils spekulativen Einschätzungen dieser Entwicklungen sowie möglichen Folgen für Mensch und Gesellschaft widmet, um kommende Debatten in begrifflicher, konzeptioneller und methodischer Hinsicht *vorzubereiten*. Analog zur hermeneutischen Erweiterung der Technikfolgenabschätzung (Grunwald/Hubig 2018) steht die Exploration

derjenigen Quellen technik-, sozial- und wirtschaftsethischer Fragen und Themen im Vordergrund, die mit allmählichen Verschiebungen in Gesellschaft und im Mensch/Technik-Verhältnis korrelieren und die eine bessere ethische Einordnung erlauben, z.B. durch Vergleich mit anderen Feldern.

(3) Pragmatische Dimension: wie handeln?

In frühen Phasen möglicher Disruption kommt es zu Fragen nach Konsequenzen für das Handeln zwischen proaktiv intervenierender Prävention und dem Abwarten auf verbesserte Datenlagen und klarere Diagnosen. Hier ist das aus der Technikfolgenabschätzung (Grunwald 2022) bekannte Dilemma (Collingridge 1980) zu beachten: Zwar ist in frühen Phasen prinzipiell der weitere Gang der Dinge noch weit offen und daher besser zu beeinflussen als in den späteren, wenn die Konstellation durch Pfadabhängigkeiten bereits stark verfestigt ist. Allerdings ist dann das erforderliche Folgenwissen über die unter dem Verdacht allmählicher Disruption stehenden Entwicklungen zwangsläufig hochgradig unsicher oder fehlt ganz. Statt belastbarer Prognosen oder wenigstens plausibler Szenarien liegen üblicherweise nur mehr oder weniger spekulative Erwartungen oder auch Befürchtungen vor, deren epistemischer Gehalt oft nicht gut einschätzbar ist (Grunwald 2013).

In dieser Hinsicht stellen sich allmähliche Disruptionen als Radikalisierung, ja Extremform des Collingridge-Dilemmas dar: Vor dem disruptiven Ereignis ist nichts oder kaum etwas über die Folgen bekannt, so dass keinerlei vorsorgende oder proaktive Handlungen aktiviert werden können, danach jedoch ist das Ereignis geschehen und es ist zu spät für vorsorgende Maßnahmen. Es käme nur noch nachträgliche Reparatur in Frage. Das oben erwähnte Gebot der Reflexivität wäre somit verletzt. Angesichts dieser Situation stellt sich die Frage, wann die Evidenz eines Verdachts hinreichend groß ist, um intervenierende Maßnahmen zu legitimieren, Budgets zu mobilisieren, ggf. Freiheiten durch Regulation einzugrenzen etc. Das war die zentrale Konfliktthematik in den ersten Jahrzehnten der Debatten zum Klimawandel.

(4) Kommunikative Dimension: wie sprechen?

Die schlechte Erkennbarkeit allmählicher Entwicklungen und die Schwierigkeiten ihrer Bewertung haben Folgen auch für ihre Kommunikation. Geringe Evidenz des Wissens macht Kommunikation anfällig für Ideologie und Spekulation (Grunwald 2002). Auf der einen Seite kommt es zu Verharmlosung und Abwiegeln mit dem Verweis, man solle doch warten, bis bessere Daten vorliegen und sich die Evidenz erhärtet hat, statt vorschnell Ressourcen für Maßnahmen zu verschwenden. Auf der anderen Seite werden schwache Signale als harte Entwicklung verstanden und in die Zukunft extrapoliert, mit oft dramatischen Ergebnissen bis hin zu Befürchtungen

einer raschen Disruption. Gegenseitige Vorwürfe um Übertreibung, Ideologie, Spekulation, Verharmlosung und Schönrednerei, Leichtsinn, Verantwortungslosigkeit oder permanente Bedenkenträgerei sind die Folge. Während der Corona-Pandemie ließen sich diese Kommunikationsprobleme mannigfaltig beobachten. Immer wieder schien es zwischen dramatisierender Übertreibung auf der einen und Verharmlosung der Risiken auf der anderen Seite kaum noch einen Weg der vermittelnden Vernunft zu geben.

5. Wider den Fatalismus

Das verbreitete Unbehagen an vielen Aspekten der Digitalisierung in der öffentlichen und insbesondere intellektuellen Debatte kontrastiert auf merkwürdige Weise mit dem oft schnellen und durchschlagenden Markterfolg neuer digitaler Angebote und Dienstleistungen. Hier könnte von einem gespaltenen öffentlichen Bewusstsein gesprochen werden: Auf der einen Seite werden Komfort und Bequemlichkeit digitaler Applikationen fast blindlings wertgeschätzt, andererseits werden Sorgen über die damit verbundenen Entwicklungen geäußert. Der Ursprung dieses Unbehagens liegt nach den vorgetragenen Überlegungen in der Kombination von zwei methodisch voneinander unabhängigen, in der Sache freilich verbundenen Einstellungen. Das Zusammenwirken technikdeterministischer Muster (s.o. Abschnitt 3.3) mit Extrapolationen allmählicher Verschiebungen in die Zukunft führt zu Fragen des Typs, wohin denn das alles führen solle und wer das noch beeinflussen könne. Die damit verbundene fatalistische Tendenz stellt in Frage, ob eine notwendige Bedingung für die Möglichkeit einer Ethik der *Digitalisierung als Prozess* überhaupt noch erfüllt ist, nämlich das Vertrauen, diesen Prozess durch ethische Reflexion mitgestalten zu können.

Ethik der Digitalisierung als Prozess muss, wenn sie ihrer Aufgabe nicht nur als akademische Disziplin, sondern auch in der öffentlichen Debatte wahrnehmen will, dieses Unbehagen (s.o. Abschnitt 3) als Sorge und damit *als Ressource* verstehen: als Quelle notwendiger Sensibilisierung auf dem Weg in eine immer stärker digitalisierte Welt. Es besteht zunächst Aufklärungsbedarf in zwei Richtungen. Zunächst muss der ›digitale Determinismus‹ (Mainzer 2016) überwunden werden. Aus digitaltechnisch ermöglichten Potentialen wird nicht in determinierter Weise eine zukünftige Gegenwart, sondern hierüber befinden Entscheidungen auf den unterschiedlichsten Ebenen, die wertbehaftet und der ethischen Explikation und Kritik zugänglich sind (van de Poel 2009). Statt vorauseilender Anpassung an die vermeintlich eigendynamische Entwicklung der Digitalisierung geht es um ihre Gestaltung im Hinblick auf ein ethisch reflektiertes, gesellschaftliches Wollen. In der Digitalisierung *als Prozess* müssen Mitgestaltungsmöglichkeiten eingefordert und umgesetzt werden. Des Weiteren ist Aufklärungsarbeit dahingehend zu leisten, dass Vi-

sionen und Extrapolationen Erzählungen in der Immanenz der Gegenwart sind, aber keine Tatsachen aus der Zukunft beschreiben (Luhmann 1990; Grunwald 2013).

Philosophie der Digitalisierung bedarf daher der Kooperation mit den empirischen Sozialwissenschaften, der Rechts- und Politikwissenschaft, der Psychologie und der Technikfolgenabschätzung. Die durch die Digitalisierung als Prozess erzeugten normativen Unsicherheiten hängen, jedenfalls was die in diesem Beitrag betrachteten allmählichen Disruptionen betrifft, untrennbar mit empirisch fundierten Einschätzungen ihrer Relevanz, Ausprägung und Dramatik zusammen. Zwar ist es wichtig und richtig, einen ›digitalen Humanismus‹ (Nida-Rümelin/Weidenfeld 2018) oder ›digitale Mündigkeit‹ (Grunwald 2019a) zu fordern – es darf aber mit der Forderung nicht sein Bewenden haben.

Literatur

Bainbridge, L. (1983): Ironies of Automatization, in: *Automatica*, 19(6), 775–779.

Beck, U. (1996): Das Zeitalter der Nebenfolgen und die Politisierung der Moderne, in: Beck, U.; Giddens, A.; Lash, S. (Hg.): Reflexive Modernisierung. Eine Kontroverse, Frankfurt a.M.: Suhrkamp, 19–112.

Börner, F.; Kehl, C.; Nierling, L. (2018): Chancen und Risiken mobiler und digitaler Kommunikation in der Arbeitswelt, Berlin: Büro für Technikfolgen-Abschätzung beim Deutschen Bundestag. [www.tab-beim-bundestag.de/de/pdf/publikationen/berichte/TAB-Arbeitsbericht-ab174.pdf] (Zugriff: 25.04.2019).

Bower, J.L.; Christensen, C.M. (1995): Disruptive Technologies. Catching the Wave, in: *Harvard Business Review*, 69, 19–45.

Collingridge, D. (1980): The Social Control of Technology, New York: Pinter.

Danneels, Erwin (2004): Disruptive Technology Reconsidered. A Critique and Research Agenda, in: *Journal of Product Innovation Management*, 21(4), 246–258.

de la Vera, R.L.; Ramge, T. (2021): Sprunginnovation. Wie wir mit Wissenschaft und Technik die Welt wieder in Balance bekommen, Berlin: Econ.

Dewey, J. (6. Auflage 1931): Science and Society, in: Ders. (Hg.): The Later Works 1925–1953, Carbondale (IL): Southern Illinois University Press, 53–63.

Ehrenberg-Silies, S.; Peters, C.; Wehrmann, C.; Christmann-Budian, S. (2022): TAB2020 Welt ohne Bargeld – Veränderungen der klassischen Banken- und Bezahlsysteme. TAB Kurzstudie Nr. 2, Berlin: Büro für Technikfolgenabschätzung beim Deutschen Bundestag [file:///C:/Users/wb9296/Downloads/TAB-Kurzstudie-ks002-1.pdf].

Ethik-Kommission autonomes und vernetztes Fahren (2017): Endbericht. [https://www.bmvi.de/SharedDocs/DE/Publikationen/DG/bericht-der-ethik-kommission.pdf?__blob=publicationFile] (Zugriff: 23.3.2022).

Ethikrat – Deutscher Ethikrat (2023): Mensch und Maschine. Herausforderungen durch Künstliche Intelligenz, Berlin [https://www.ethikrat.org/fileadmin/Publikationen/Stellungnahmen/deutsch/stellungnahme-mensch-und-maschine.pdf].

Floridi, L. (2016): Faultless responsibility. On the nature and allocation of moral responsibility for distributed moral actions, in: *Philosophical Transactions of the Royal Society*, 374, 20160112 [https://doi.org/10.1098/rsta.2016.0112].

Gans, J. (2017): The disruption dilemma, Cambridge (MA): The MIT Press. [Announcement [https://mitpress.mit.edu/9780262533621/the-disruption-dilemma/] (Zugriff: 22.03.2023)].

Gladwell, M. (2000): The Tipping Point. How Little Things Can Make A Big Difference, New York u.a.: Little, Brown and Company.

Grunwald, A. (2002): Zwischen Präventionsnotwendigkeiten und Alarmismus: Problemwahrnehmungen in der Nachhaltigkeitsdiskussion, in: Ministerium für Umwelt und Verkehr Baden-Württemberg (Hg.): Kommunikation über Umweltrisiken zwischen Verharmlosung und Dramatisierung, Stuttgart/Leipzig: Hirzel, 87–101.

Grunwald, A. (2010): From Speculative Nanoethics to Explorative Philosophy of Nanotechnology, in: *NanoEthics*, 4(2), 91–101.

Grunwald, A. (2013): Modes of Orientation Provided by Futures Studies. Making Sense of Diversity and Divergence, in: *European Journal of Futures Studies*, 15:30 (2014) [https://doi.org/10.1007/s40309-013-0030-5].

Grunwald, A. (2015): Technology assessment and design for values, in: van den Hoven, J.; Vermaas, P.E.; van de Poel, I. (Hg.), Handbook of Ethics, Values, and Technological Design. Sources, Theory, Values and Application Domains, Dordrecht: Springer, 67–86.

Grunwald, A. (2017): Abschied vom Individuum – werden wir zu Endgeräten eines global-digitalen Netzes? In: Burk, S.; Hennig, M.; Heurich, B.; et al. (Hg.): Privatheit in der digitalen Gesellschaft, Berlin: Duncker&Humblot, 35–48.

Grunwald, A. (2019a): Der unterlegene Mensch. Die Zukunft der Menschheit im Angesicht von Algorithmen, Robotern und Künstlicher Intelligenz, München: RIVA-Verlag.

Grunwald, A. (2019b): Digitalisierung als Prozess. Ethische Herausforderungen inmitten allmählicher Verschiebungen zwischen Mensch, Technik und Gesellschaft, in: *Zeitschrift für Wirtschafts- und Unternehmensethik*, 20(2), 121–145.

Grunwald, A. (3. Auflage 2022): Technikfolgenabschätzung. Einführung, Baden-Baden: Nomos.

Grunwald, A. (2023): Disruption in Zeitlupe, in: Brand, C.; Meisch, S.; Frank, D.; Ammicht Quinn, R. (Hg.), Ich lehne mich jetzt mal ganz konkret aus dem Fenster. Festschrift für Thomas Potthast, Tübingen: Tübingen Library Publishing, 447–456.

Grunwald, A. (Hg.) (2021): Wer bist du, Mensch? Transformationen menschlicher Selbstverständnisse im wissenschaftlich-technischen Fortschritt, Freiburg: Herder.

Grunwald, A.; Hubig, C. (2018): Technikhermeneutik. Ein kritischer Austausch zwischen Armin Grunwald und Christoph Hubig, in: Friedrich, A.; Gehring, P.; Hubig, C.; Kaminski, A.; Nordmann, A. (Hg.), *Jahrbuch Technikphilosophie*, 2018 [Arbeit und Spiel], Baden-Baden: Nomos, 321–352.

Jonas, H. (1979): Das Prinzip Verantwortung. Versuch einer Ethik für die technologische Zivilisation, Frankfurt a.M.: Suhrkamp.

Kamlah, W. (1973): Philosophische Anthropologie. Sprachkritische Grundlegung und Ethik, Mannheim: Bibliographisches Institut.

Kopfmüller, J.; Brandl, V.; Jörissen, J.; Paetau, M.; Banse, G.; Coenen, R. (2001): Nachhaltige Entwicklung integrativ betrachtet. Konstitutive Elemente, Regeln und Indikatoren, Berlin: Edition Sigma.

Latour, B. (2005): Reassembling the social. An introduction to actor-network-theory, Oxford: Oxford University Press.

Loh, J. (2018): Trans- und Posthumanismus, Hamburg: Junius.

Luhmann, N. (1990): Die Zukunft kann nicht beginnen. Temporalstrukturen der modernen Gesellschaft, in: Sloterdijk, P. (Hg.), Vor der Jahrtausendwende. Berichte zur Lage der Zukunft, Frankfurt a.M.: Suhrkamp.

Mainzer, K. (2016): Wann übernehmen die Maschinen?, Heidelberg: Springer.

Misselhorn, C. (2018): Grundfragen der Maschinenethik, Stuttgart: Reclam.

Neugebauer, R. (Hg.) (2018): Digitalisierung. Schlüsseltechnologien für Wirtschaft und Gesellschaft, Heidelberg: Springer.

Nida-Rümelin, J.; Weidenfeld, N. (2018): Digitaler Humanismus. Eine Ethik für das Zeitalter der Künstlichen Intelligenz, München: Piper.

Nordmann, A. (2007): If and Then. A Critique of Speculative NanoEthics, in: *NanoEthics*, 1(1), 31–46.

Orwat, C.; Raabe, O.; Buchmann, E.; et al. (2010): Software als Institution und ihre Gestaltbarkeit, in: *Informatik-Spektrum*, 33, 626–633.

Petermann, T.; Bradke, H.; Lüllmann, A.; Poetzsch, M.; Riehm, U. (2011): What happens during a blackout. Consequences of a prolonged and wide-ranging power outage, Norderstedt: BoD – Books on Demand.

Rammert, W.; Schulz-Schaeffer I. (2002): Technik und Handeln, in: Dies. (Hg.), Können Maschinen handeln? Soziologische Beiträge zum Verhältnis von Mensch und Technik, Frankfurt a.M.: Campus Verlag, 11–64.

Ropohl, G. (1982): Zur Kritik des technologischen Determinismus, in: Rapp, F.; Durbin, P. (Hg.), Technikphilosophie in der Diskussion, Wiesbaden: Vieweg & Teubner, 3–18.

Safdar, N.M.; Banja, J.D.; Meltzer, C.C. (2020): Ethical considerations in artificial intelligence, in: *European Journal of Radiology*, 122. [https://doi.org/10.1016/j.ejrad.2019.108768].

van de Poel, I. (9. Auflage 2009): Values in Engineering Design, in: Meijers, A. (Hg.), Philosophy of Technology and Engineering Sciences, Amsterdam: North Holland, 973–1006.

van den Hoven, J.; Vermaas, P.; van de Poel, I. (Hg.) (2015): Handbook of Ethics, Values, and Technological Design, Dordrecht: Springer, 67–86. DOI: 10.1007/978-94-007-6970-0.

Notizen zu Macht und Algorithmen

Matthias Kettner

> The message of the electric light is total change. It is pure information without any content to restrict its transforming and informing power.[1]

Abstract: *Algorithmically controlled applications are becoming normal elements of our technoculture in more and more areas of practice. In a sober technical understanding, algorithms are automatable solution programs for machine-calculable goals. In the popular understanding, however, algorithms have become an object of fantasy. The popular topos of the ›power of algorithms‹ functions today in the digital revolution in much the same way as the topos of the ›power of genes‹ did thirty years ago in the euphoria of the molecular genetic revolution. Based on a re-analysis of Max Weber's classical definition of power, this essay develops a new dynamic understanding of power that allows for the analysis of power relations with respect to persons and quasi-personal corporate actors as well as with respect to a-personal software agents and other machine actors. The power that actors have in a situation is conceptualized modally and relationally as the capacity of actors, through forces that they can control counterfactually and robustly, to govern other forces in such a way that the actors come closer to achieving their goals.*

Keywords: *philosophical power theory; machine actors; algorithms; intelligent software agents*

1. Zur Fragestellung

Lassen sich auffällige gesellschaftliche Veränderungen auf Veränderungen in bestehenden Machtverhältnissen zurückführen und bringen sie ihrerseits neue hervor? Diese Frage liegt einerseits nahe, wenn wir Macht zunächst als unsere

1 McLuhan 1994: 52.

ubiquitäre Gattungsfähigkeit verstehen wollen, »sich gegen fremde Kräfte durchzusetzen« (Popitz 1992: 23). Andererseits bleibt der ›dynamische Gesichtspunkt‹, wie man die Blickrichtung dieser Frage nennen kann, nebulös wie der Machtbegriff selbst, dem Max Weber (1964: 38) bescheinigt, er sei »soziologisch amorph«. Formwandler sind unheimlich. Im Feld der Machtphantasien faszinieren die dunklen Mächte mehr als die guten. Im Feld des Machtdenkens haben Ansätze, die die Macht als ominöse und allgegenwärtige Größe mehr beschwören als beschreiben, wie es scheint, einen Bonus.[2] Noch jede neue Basistechnologie, die über die Schwelle zur begierigen Verbreitung in der Gesellschaft getrieben wurde, hat kollektive Hoffnungen und Ängste erzeugt und sich mit ausdrucksstarken Narrativen verbunden. Geschichtlich neu am Kulturprozess der Verbreitung von Digitaltechnologie ist die überwältigende Tragweite ihres Einsatzes, die unfassbare Geschwindigkeit ihrer Fortschritte, die grenzenlos erscheinende Eingriffstiefe ihrer Anwendungen. Auch wer ihr nicht traut, traut der neuen Basistechnologie zu, alle Verhältnisse umzuwerfen, – hochmeinende Fortschrittsoptimisten sogar, alle Verhältnisse, »in denen der Mensch ein erniedrigtes, ein geknechtetes, ein verlassenes, ein verächtliches Wesen ist« (Marx 1976[1844]: 385). Heute nicht mehr von der Hand zu weisen ist der Eindruck einer ›von uns‹ zwar bewerkstelligten, aber losgelassenen Verselbständigung des Fortschritts der schon immer als eine umwälzende ›Macht‹ erlebten Technik. Macht der Technik, Macht des Fortschritts, die Geschichte dieser Topoi wird gerade um das Kapitel Digitalisierung verlängert.[3] Darauf reagiert auch

2 Damit spiele ich vor allem auf die populäre Rezeption der Machtanalytik Foucaults aus den 1970er Jahren an. Foucaults einflussreiches Machtdenken und seine erstaunlichen Wendungen rekonstruiert Mathias Richter (Richter 2011: bes. 269–498) in erhellendem Vergleich mit Sartre.

3 Die ausgefeilteste mir bekannte technikphilosophische Abhandlung zu allen Registern der Rede von der Macht der Technik ist Hubig 2015. Der Ansatz: »›Macht der Technik‹ lässt zwei Lesarten zu, die die vereinseitigte und polarisierte Diskussionslage zu diesem Thema spiegeln: Im ersten Sinne kann ›Macht der Technik‹ als Genitivus subjectivus gelesen werden. In diesem Sinne, als Macht, die von der Technik ausgeht, macht sie den Befund eines sogenannten Technikdeterminismus oder eines quasi naturalistisch gefassten Technikevolutionismus aus. Beide heben darauf ab, dass die individuellen und sozialen Subjekte den ›Gesetzmäßigkeiten‹ der Technik unterliegen bzw. sich ihnen bei Strafe ihrer Selbstaufgabe anzupassen haben. Als Genitivus objectivus (Beherrschung von Technik, Gestaltungsmacht über die Technik) charakterisiert der Ausdruck ›Macht der Technik‹ das Konzept eines Konstruktivismus/Sozialkonstruktivismus, welcher Technik (in Aktualisierung der aufklärerischen Tradition) der Macht gesellschaftlicher Aushandlungs- und Gestaltungsprozesse unterstellt. Beiden ist gemeinsam, dass Technikentwicklung im Wesentlichen in Kausalschemata reflektiert wird. Gegen die Polarisierung der ›humanistischen‹ und ›posthumanistischen‹ Auffassung vom Gestaltungssubjekt von oder für Technik wenden sich Ansätze, die auf eine ›Symmetrie‹ im weitesten Sinne abheben und unter dem Topos von Macht als ›Netz‹ die Relationen zwischen ›Aktanden‹ als Artefakten und als ›Akteuren‹ (um vorwegnehmend eine Formulierung von Bruno Latour aufzugreifen) neu zu begreifen suchen.« (Hubig 2015: 8f.)

die Machttheorie, z.B. mit der begrifflichen Konstruktion neuer ominöser und allgegenwärtiger Größen. Teils ist das raffiniert und durchdacht, wie bei Shoshana Zuboff, die für ihre Analyse der digitalkulturellen Metamorphose des Kapitalismus in *Surveillance Capitalism* (Zuboff 2018) die Konstruktion eines neuen Begriffs für die Macht, Verhalten für kommerzielle Verwertung dienstbar zu machen, hilfreich findet: »instrumentäre Macht«.[4] Teils bleibt es plakativ und schlagworthaft bei den »Datenkraken« (Schröder/Schwanebeck 2017), aus deren Fängen kein Entkommen sei.[5]

Neben dem auffällig Disruptiven[6] erscheinen mir zwei weitere, auf den ersten Blick unvereinbare Entwicklungen charakteristisch für die gegenwärtige Phase des Kulturprozesses der Digitalisierung: (1) Eine erstaunliche Banalisierung: Dass Algorithmen – kurz: automatisierbare Lösungsprogramme für berechenbare Zielsetzungen – durch die expansive Anwendung in algorithmisch gesteuerten und steuernden Maschinen und wiederum durch deren enorme Verbreitung in immer mehr Praxisbereichen zu *normalen*, d.h. wie selbstverständlich als alternativlos wirkenden Elementen unserer Technokultur werden. (2) Eine Wiederbelebung großer utopisch-dystopisch polarisierender Erzählungen: Dass die Rede von Algorithmen utopische, skeptische und natürlich auch dystopische Erwartungen ambivalent verdichtet, vor allem in inzwischen ebenso weitverbreiteten wie spektakulären Vorstellungen über ›Künstliche‹ Intelligenz, ›autonome‹ Systeme und ›lernende‹ Algorithmen.

---

4 »Der Überwachungskapitalismus ist der Puppenspieler, der uns durch das Medium des allgegenwärtigen digitalen Apparats seinen Willen aufzwingt. Ich bezeichne diesen Apparat als Big Other – das Große Andere. Ich verstehe darunter die wahrnehmungsfähige, rechnergestützte und vernetzte Marionette, die das menschliche Verhalten rendert, überwacht, berechnet und modifiziert. Big Other kombiniert diese Funktionen des Wissens und Tuns zu einem ebenso umfassenden wie beispiellosen Mittel zur Verhaltensmodifikation. Dirigiert wird die ökonomische Logik des Überwachungskapitalismus durch die immensen Fähigkeiten von Big Other zur Schaffung von *instrumentärer Macht*, die die Manipulation der Seele durch die Verhaltensmodifikation ersetzt.« (Zuboff 2018: 437)

5 Die tatsächliche Machtanalyse beschränkt sich hier auf »politische Macht, die durch technische Überlegenheit entsteht« (Schmidt 2015: 63); auf die vor allem finanzielle »Machtfülle der fünf Oligarchen des Westens: Amazon, Apple, Facebook, Google und Microsoft«; auf die »kommunikative Macht der Algorithmen« (ebd.: 141) alias den Einsatz von Social Bots. Von der »Macht der Algorithmen« gelte: »Bei aller Übermächtigkeit der Algorithmen gegenüber dem Menschen bleiben sie doch stets zugleich auf das menschliche Handeln bezogen und durch menschliches Handeln mitkonstituiert. Im Blick auf die Frage nach der Macht der Algorithmen gilt: Sie sind weder autonomer Automatismus noch gänzlich unselbständiges Werkzeug in der Hand des Menschen« (ebd.: 146). Krabbe et al. 2022 verspricht im Titel Machtanalyse, enthält aber keine. In Hobe et al. 2023 ist die besagte »Macht der Algorithmen« nur eine Chiffre für die Wirtschaftsmacht der Digitaltechnik-Konzerne.

6 Siehe hierzu den Beitrag von Armin Grunwald im vorliegenden Band.

Die ›Macht der Algorithmen‹, dieses Thema durchzieht viele popkulturelle Plots und findet auch entsprechende journalistische und kulturwissenschaftliche Aufmerksamkeit. Pars pro toto:

> »Algorithms are everywhere, supposedly. We are living in an ›algorithmic culture‹, to use the author and communication scholar Ted Striphas's name for it. Google's search algorithms determine how we access information. Facebook's News Feed algorithms determine how we socialize. Netflix's and Amazon's collaborative filtering algorithms choose products and media for us.« (Bogost 2015)
> »Algorithms are used in various ways and everyone who uses the internet will probably be in touch with algorithms at least once a day, likely over popular websites such as Google, Facebook, Twitter, Instagram, YouTube or Netflix. The impact they have on the information the user receives is tremendous.« (Schmidt 2015: 18)

Vor diesem Hintergrund interessiert mich folgende Fragestellung: Welche begrifflichen Unterscheidungen werden nötig, um machttheoretisch zu betrachten, was es heißt, dass Menschen algorithmisch gesteuerte und steuernde Maschinen zu der Leistung gebracht haben, einen inzwischen breiten Teil des Spektrums von tier- und personentypischen Intelligenzleistungen (Kettner 2022) zu virtualisieren?[7] Diese Frage erscheint mir dringlich, aber zu weit, um sie hier befriedigend zu behandeln. Zur Vorbereitung stelle ich im Folgenden einige Überlegungen an, die der Beantwortung einer engeren Frage dienen: Wovon kann die Rede sein, wenn von der Macht von Algorithmen die Rede ist? Es geht darum, geläufige Vorstellungen von Macht durch philosophische Analyse und Synthese gedanklich zu klären und den rhetorisch übermäßig verdichteten intuitiven Sinngehalt der Redewendung von der ›Macht der Algorithmen‹ in einem geeigneten Theorierahmen auseinanderzulegen.

2. Amorphe Macht, mit Max Weber weitergedacht

Das in der europäischen Begriffs- und Ideengeschichte gebildete Machtvokabular ist uralt (bes. *dynamis*, *energeia*, *potentia*, *potestas*, *dominium*, *auctoritas*) und beeindruckend vieldeutig. Theoretisch sinnvoller als *die* Klärung *des* Machtbegriffs erscheinen deshalb Versuche, klare von unklaren Fällen abzugrenzen und an ersteren die wesentlichen Begriffsmomente deutlich zu unterscheiden, die dann bestenfalls

7 ›Virtuell‹ sensu Peirce: »A virtual X (where X is a common noun) is something, not an X, which has the efficiency (virtus) of an X. – This is the proper meaning of the word; but (2) it has been seriously confounded with ›potential‹, which is almost its contrary. For the potential X is of the nature of X, but is without actual efficiency.« (Peirce 1902: 763)

eine Systematik unterschiedlicher Machtbegriffe ergeben.[8] Wird angesichts der Schwierigkeiten, die notorisch strittige Begriffe bereiten, fraglich, wozu philosophische Forschung über Macht gut sein soll, so wäre darauf zu verweisen, dass wir, vom theoretischen *l'art pour l'art* einmal abgesehen, in praktisch-politischen, evaluativen und moralisch-normativen Rechtfertigungskontexten um Machtfragen gar nicht herumkommen. Ich meine, von philosophischer Forschung über Macht sollten wir erwarten dürfen, dass sie die Problematisierung von Machtverhältnissen überzeugender macht, indem sie die Diskursivierungschancen für Identifikation, Kritik und Begründung von Machtverhältnissen verbessert.

Max Weber hat versucht, für klare Fälle von Macht deren allgemeine Form zu denken. Webers begrifflicher Vorschlag ist so einleuchtend, dass m.E. keine Machttheorie, gleich in welcher Disziplin sie ausgearbeitet wird, die wissenschaftliche Minimalanforderung der Erfahrungstreue erfüllt,[9] wenn sie nicht *wenigstens auch* Max Webers Realdefinition abdeckt.

> »*Macht* bedeutet jede Chance, innerhalb einer sozialen Beziehung den eigenen Willen auch gegen Widerstreben durchzusetzen, gleichviel worauf diese Chance beruht. *Herrschaft* soll heißen die Chance, für einen Befehl bestimmten Inhalts bei angebbaren Personen Gehorsam zu finden; *Disziplin* soll heißen die Chance, kraft eingeübter Einstellung für einen Befehl prompten, automatischen und schematischen Gehorsam bei einer angebbaren Vielheit von Menschen zu finden. […] Der Begriff ›Macht‹ ist soziologisch amorph. Alle denkbaren Qualitäten eines Menschen und alle denkbaren Konstellationen können jemand in die Lage versetzen, seinen Willen in einer gegebenen Situation durchzusetzen. Der soziologische Begriff der ›Herrschaft‹ muß daher ein präziserer sein und kann nur die Chance bedeuten: für einen *Befehl* Fügsamkeit zu finden.« (Weber 1964: 38)

Für vorzüglich halte ich diese Stammformel aus mindestens acht Gründen:

(#1) Sie ist auf Handeln, und zwar soziales Handeln unter Menschen gemünzt,[10] ist aber nicht strikt anthropozentrisch. Auch Machtverhältnisse jenseits der Menschen sind denkbar.

8 Für Versuche der ersten Art siehe z.B. Bertrand Russell (1938), der Macht kurzerhand definiert als »the production of intended effects. It is thus a quantitative concept: given two men with similar desires, if one achieves all the desires that the other achieves, and also others, he has more power than the other« (ebd.: 23). Zum unhaltbaren Reduktionismus dieser Definition siehe Knight 1939. Beispiele für die sinnvollere zweite Strategie: Morriss 1990; Röttgers 1990: bes. 241–346; Popitz 1992; Lukes 2005, der eine dreidimensionale Systematisierung versucht.

9 »Empirical adequacy«, siehe van Fraasen 1980: bes. 80–86. Man muss wissenschaftstheoretisch kein Empirist sein, um diese Adäquatheitsbedingung anerkennen zu können.

10 Soziales Handeln begreift Weber so: Akteure, die innerhalb irgendeiner sozialen Beziehung, d.h. ›sozial‹ handeln, orientieren ihr eigenes Verhalten *sinnhaft* am (1) Verhalten, an Erwartun-

(#2) Sie vermeidet die Konfundierung von Macht als Vermögen einer Machtinstanz[11] und der Ausübung dieses Vermögens.

(#3) Durch Verweise auf den Willen bezieht sie ein intentionales Begriffsmoment in den Machtbegriff ein, ohne die machtrelevante Intentionalität auf Handlungsabsichten zu reduzieren, denn

(#4) als »Chance« ist Macht als gegebene günstige Gelegenheit dispositional und kontextuell gedacht, da von den Umständen der jeweiligen Situation abhängt, wodurch welche wie und für wen günstige Gelegenheit gegeben ist.

(#5) Sie macht die Frage der *Machtmittel*, auf die viele Machttheorien solchen Wert legen, die Nachfrage nach dem Machtbegriff selber aber vernachlässigen (z.B. DeVrio et al. 2024), zu einer ungebundenen Variablen (»gleichviel worauf diese Chance beruht«). Machtausübung muss nicht per se gewaltförmig sein bzw. mit Hilfe gewalttätiger Mittel erfolgen.

(#6) Sie präjudiziert mit Bezug auf Machtverhältnisse keine notwendige Asymmetrie zwischen Macht und Gegenmacht (Unterwerfung, Überwältigung etc.), ja nicht einmal notwendigerweise Gegenmacht, sondern nur ein Widerstrebendes, ein machtbetroffener Widerpart von ausgeübter Macht. Darunter kann auch Widerständiges im Akteur selbst, das überwunden werden muss, fallen, z.B. Willensschwäche.

(#7) Sie vermeidet den in Machttheorien häufigen Fehler, die Relation des Beeinflussens bzw. des Einflussnehmens-auf zum Wesen von Macht zu erklären und damit das Machtverhältnis zu trivialisieren.[12]

---

gen von Verhalten und an Erwartungserwartungen von anderen ihresgleichen (= aktorialen Peers), und natürlich auch an (2) den Erwartungen des Verhaltens sachlicher Objekte. Akteure, die nicht sozial, also nicht innerhalb irgendeiner sozialen Beziehung handeln, orientieren ihr eigenes Verhalten *sinnhaft* mindestens an (2) und jedenfalls nicht an (1). Lediglich kausal, nicht auch sinnhaft, massenbedingtes, oder bloß reaktiv nachahmendes, oder eingelebtes ›traditionales‹ Handeln gelten Weber als »Grenzfälle sozialen Handelns« in dem Maße, wie sie »lediglich reaktiv, ohne sinnhafte Orientierung des eigenen an dem fremden Handeln« erfolgen (Weber 1964: 17).

11 Die Eigenschaft, Macht zu besitzen, verstehe ich (mit Max Weber) als das *Vermögen* – eine dispositionale Fähigkeitseigenschaft – eines geeigneten Akteurs, sie zu aktualisieren, also Macht *auszuüben*.

12 Machtausübung an Einflussnehmen anzugleichen bedeutet, Macht mit der Erzeugung welcher sozialen Wirkung auch immer gleichzusetzen. Jede Einflussnahme von Akteuren aufeinander wird dann zur Machtausübung, und Gesellschaft, auf der Makroebene ihrer wichtigsten Institutionen als auch auf der Mikroebene persönlicher Beziehungen und sozialer Interaktionen, wird durchgängig zu einem Machtphänomen stilisiert, weil das Soziale die gegenseitige Beeinflussung von Individuen voraussetzt. Ein spezifischer Machtbegriff erübrigt sich dann aber. Der Begriff ›Einfluss‹ täte es und hätte zudem den Vorteil, dass er als Substantiv wie auch als Verb gebraucht werden kann.

(#8) Sie präjudiziert nicht die Beurteilung von Machtphänomenen unter moralischen und sonstigen normativen Standards. Keineswegs impliziert Machtausübung Freiheitseinschränkung, und schon gar nicht: unrechte Freiheitsberaubung.

Mit einigen Schritten von *conceptual engineering* möchte ich nun Webers Stammformel versuchsweise weiterdenken, um ihr Potential für die Aufklärung auch von Machtzuschreibungen, wie sie in Diskussionen über Digitalisierungsphänomene anfallen, zu heben.

In Webers Stammformel verweist ›Wille‹ auf das Moment der Intentionalität im Machtbegriff. Mit #1 ist klar, dass an Praktiken von Menschen gedacht ist, die etwas, das sich nicht von selbst einstellt, bewerkstelligen wollen, mit oder gegen andere Menschen, die etwas (anderes) bewerkstelligen wollen, oder in Auseinandersetzung mit etwas Widerständigem (Material, Sachen). Die Frage, ob und wieweit auch passend organisierten nichtmenschlichen Lebewesen die Fähigkeit zugeschrieben werden kann, etwas zu wollen und anderes nicht, führt in die philosophischen Labyrinthe der Willenstheorien. Könnten wir uns noch ungebrochen Kants praktischer Philosophie überlassen, wäre einfach zu sagen: Machtverhältnisse gibt es nur von und zwischen Wesen, die einen ›vernünftigen Willen‹ haben. Aber wir können die Bürde dieser Antwort m.E. dadurch etwas erleichtern, dass wir den Sinn der Stammformel durch den Zweckbegriff interpretieren, obwohl auch dieser seine metaphysischen Tücken hat: Macht (sensu Weber) meint so etwas wie ein situationsrelatives Vermögen von *zum Verfolgen von Zwecken befähigten Akteuren*, Beabsichtigtes durch Ergreifen von Verwirklichungschancen zu erreichen.

Diese Umformulierung macht deutlich, dass eine Menge Kognition und Evaluation und Motivation, kurz: Rationalität, in den Machtbegriff investiert wird. Gegebenheiten müssen *erkannt* und *eingeschätzt* werden als etwas, das möglicherweise (d.h. potentiell) zur Verwirklichung von Zwecken so und so gut beitragen würde, ebenso wie die Umstände, unter denen es dies würde, die also geeignete Umstände sein müssen. Und die Chancen müssen *ergriffen* werden, d.h. die Zwecke verfolgenden Akteure müssen *sich entscheiden*, zielstrebig *tätig zu werden*.

Wie müssen Akteure beschaffen sein, die diese begrifflichen Bürden schultern? Die beschriebenen Beschränkungen, die einem machtgeeigneten Akteur-Begriff auferlegt sind, lassen über menschliche Personen hinaus sicher auch Sozialgebilde zu, die wir so normativ organisiert haben, dass wir ihnen genug Identität und Intentionalität zuschreiben, um sie für handlungsfähig zu halten und dafür auch nach normativen Standards (also nicht bloß kausal) verantwortlich zu machen (Moore 1999, Bratu 2017). Erlauben sie die Ausdehnung auch auf (einige) Tiere? Und womöglich auch auf (einige) Artefakte, etwa maschinell intelligente Systeme, die wir so konstruiert haben, dass sie Zwecke verfolgend und Entscheidungen treffend

aktiv sind, wenngleich sie weder lebendig wie Tiere noch, wie etwa Unternehmen, normativ organisiert sind?[13] Ich komme in Abschnitt 5 auf die Akteursfrage zurück.

3. Macht als Steuerung von Veränderungsverläufen durch Kontrolle von Kräften

Obwohl unser vortheoretisches Verständnis von Macht zweifellos die begriffliche Komponente der *Kraft* oder *Stärke* enthält, können wir Macht darauf nicht reduzieren. Es fehlt etwas, das wir für selbstverständlich halten, sobald wir erklären müssen, was normalerweise für uns die menschliche Handlungsfähigkeit ausmacht: Der Aspekt der Kontrolle durch Absichten, die einer hat und handelnd in die Tat umsetzen will. Für normal sozialisierte menschliche Personen ist das: eine durch den vernünftig bestimmbaren Willen der handelnden Person sinnhaft bestimmte und sinngemäß wirksame Handlungsfähigkeit. Die im vorigen Abschnitt kurz auf den Nenner von Rationalität gebrachten intentionalen Fähigkeiten möchte ich nun als das ›Kontrollmoment‹ im Machtbegriff interpretieren. Dazu setze ich bei einfachen alltagspraktischen Intuitionen an.

Angenommen, ich betrete einen dunklen Raum und weiß nicht, dass die Luft wegen eines defekten Rohrs mit explosivem Gas angereichert ist. Ich ertaste den Lichtschalter, will Licht machen, und es gibt eine Riesenexplosion (die ich glücklicherweise überlebe). Mit meinem Verhalten habe ich tatsächlich eine sehr starke Wirkung hervorgerufen. Aber wir würden nicht sagen, dass ich ›die Macht hatte‹, den Raum in die Luft zu jagen. Die Wirkung und die Art und Weise, wie sie erzeugt wurde, lag außerhalb meiner Kontrolle, obwohl niemand anders als ich die Wirkung durch mein Tun verursacht habe. Wenn man von der Möglichkeit der Böswilligkeit absieht, war die Explosion ein Unfall, wurde also durch das unglückliche Zusammentreffen von Umständen für mein Handeln durch mein Handeln verursacht. Wenn wirkungsvolle Ereignisse durch das intentionale Verhalten eines Akteurs verursacht, aber außerhalb seiner Kontrolle eintreten, dann ist das gewiss nicht das, was wir meinen, wenn wir einem Akteur diesbezüglich Macht zuschreiben. Es liegt auf der Hand, dass wir in einen sinnvollen Machtbegriff ein Moment von Kontrolle einbauen müssen.

Wir haben gesehen: Macht, die irgendwie mächtige Akteure haben, lässt sich begrifflich nicht auf ihre Fähigkeit reduzieren, Ereignisse zu verursachen bzw. Wir-

13 Anscheinend ist das Rechtsmedium so plastisch, dass wir bei Bedarf auch Rechtskonstruktionen entwerfen können, unter denen die Frage, ob Roboter i.S. des Rechts *schuldhaft* handeln können, zu einer sinnvollen Frage wird (Hilgendorf 2012). Im Prinzip mag das gehen, aber es bleibt m.E. doch ein abenteuerlicher Gedanke, der an Tierprozesse und mittelalterliches Strafrecht erinnert.

kungen hervorzurufen. Ich schlage vor, den Begriff der Macht so zu konstruieren, dass er ein Verhältnis von Kräften impliziert. Macht und Machtverhältnisse kann es nur dort geben, wo es Kontrolle, Kräfte und Verhältnisse von Kräften gibt. Was es heißt, eine bestimmte Macht zu haben oder auszuüben, lässt sich am besten als Beherrschung von Kräften erklären, ohne Webers Stammformel zu negieren.

Wenn wir Webers #1 lockern, können wir den springenden Punkt von Macht so denken: Kontrolle über Verläufe durch Herrschaft über Kräfte innerhalb von Kräfteverhältnissen. Ich versuche nun diesen Gedanken ein wenig zu präzisieren. Was es für einen machtvollen Akteur A bedeutet, Macht zu haben, können wir als sein Vermögen erklären, Veränderungsverläufe, die A wichtig sind, kontrafaktisch robust zu steuern. Um das in einer prägnanten Formel schematisch auszudrücken:

> *Die Macht von A im Verhältnis zur Macht anderer Akteure A' ist eine Funktion der Fähigkeit, die A besitzt, bestimmte Kräfte (nennen wir sie F+) mit Hilfe bestimmter anderer Kräfte (nennen wir sie F*) zu beherrschen, die A bereits so sicher kontrollieren kann, dass A zu Recht glaubt, dass A F* auch dann noch kontrollieren und F+ steuern könnte, wenn andere als die aktuell wirklichen Umstände bestehen würden, d.h. unter kontrafaktischen Umständen.*

So gefasst, ist Macht eine Beziehung zwischen Kräften, von denen Personen (oder andere passende Akteure) mit einer Zweckverfolgungsfähigkeit, die der von Personen in relevanten Hinsichten ähnelt, zutreffend glauben, dass sie, in einem gewissen Spielraum der gegebenen Umstände zumindest, mit eigenkontrollierten Kräften F* andere Kräfte F+ steuern können (und dies im Fall der Aktualisierung ihrer Macht wirklich auch tun).

Wenn wir das Machtvermögen als ein Vermögen fassen, Veränderungsverläufe kontrafaktisch robust zu steuern durch eigenkontrollierte Kräfte, dann hängt offenbar der Machtbesitz von Akteuren davon ab, ob und welche Kräfte F+ sie mit Hilfe von bestimmten Kräften F*, auf deren Kontrolle sie vertrauen können, steuern können. Zwar könnten wir *beide* Weisen, wie F* und wie F+ sich auf die Akteursinstanz A beziehen, als Beziehungen der ›Kontrolle‹ beschreiben. Da Macht zu haben jedoch stets einige Kräfte voraussetzt, die bereits unmittelbar in den Aktionsmöglichkeiten von A verankert sind, finde ich es sinnvoller, den Begriff der Kontrolle für den Bezug auf *diese* Kräfte zu reservieren und die vergleichsweise breitere Bedeutung des Begriffs ›Steuerung‹ mit Bezug auf solche Kräfte zu verwenden, die A *mittelbar* kontrolliert (F+), nämlich mittels der Kräfte, die A *unmittelbar* kontrolliert (F*). Was es heißt, dass A F* *unmittelbar* kontrolliert, lässt sich so präzisieren: Kontrolle, die A über etwas hat, ist unmittelbar, wenn sie weder dezentralisiert noch delegiert werden könnte, ohne dass A aufhören müsste, für uns und/oder für sich selbst als der alleinige Urheber etwaiger Veränderungen zu gelten, die in dem, worüber A Kontrolle hat, auftreten.

Kontrollieren, i.S. von etwas unter Kontrolle zu haben, bedeutet, etwas *maßgeblich* beeinflussen oder beherrschen zu können, *entscheidenden* Einfluss in einer Angelegenheit zu haben, oder *effektive* Autorität über bestimmte Personen auszuüben, die an bestimmten Praktiken beteiligt sind.[14] Die Kontrolle z.B., über die die spezifisch politische Regierungstätigkeit verfügen sollte, lässt sich wohl am besten als eine Gestaltungsmacht charakterisieren.[15] Als Regieren oder Herrschen (*governance*) darf jede Tätigkeit zählen, die einen gestaltenden Einfluss auf den Verlauf von Handlungen und Geschehnissen ausübt, indem sie das Verhalten von signifikanten Akteuren (u.U. einschließlich des regierenden Akteurs selbst) lenkt, in Schach hält, jedenfalls im Ablauf entscheidend beeinflusst, z.B. durch die Ausübung von Autorität oder die Durchsetzung von Disziplin, und dadurch die Verhältnisse gestaltet. Die Semantik von *Steuerung*, *Kontrolle*, *Regieren* überschneidet sich sehr weit, wenngleich Komposita in Gebrauch sind, um Machtvarianten auszuzeichnen (Kontroll-, Regierungs-, Steuerungsmacht). Ich wähle den Steuerungsbegriff statt des Regierungs- oder Herrschaftsbegriffs, weil wir ihn ganz abstrakt nehmen können, so dass er auf alle denkbar diversen Phänomene, in denen Macht in Erscheinung treten kann, passt. Wir können dann sagen: Keine Macht ohne eine durch Kontrolle von etwas vermittelte Steuerung von etwas anderem.

Wir haben uns in mehreren Schritten von Webers Stammformel wegbewegt, ohne sie aus dem Blick zu verlieren, und sind bei einer abstrakteren Realdefinition angekommen: Bezogen auf Menschen und relevant ähnliche Akteursinstanzen ist Macht ihr Vermögen, Veränderungsverläufe kontrafaktisch robust zu steuern durch Kontrolle eigener Kräfte.

In diesem Begriff ist eine interne und eine externe Relation zu unterscheiden. Die interne ist mit dem Kräfte*verhältnis* gegeben. Die Macht eines Akteurs A ist *innerlich* relativ, insofern als sie eine Beziehung zwischen unterscheidbaren aber aufeinander bezogenen Kräften beinhaltet. Die betreffenden Kräfte, nach ihren Beziehungen zum Akteur A unterschieden, habe ich schematisch als F* und F+ bezeichnet. Durch Gebrauch von solchen Kräften, über die A verfügt, weil sie im unmittelbaren Bereich und Umfang dessen liegen, was A kontrollieren kann, kann A den Bereich und Umfang dessen, was A überhaupt steuern kann, verändern (z.B. erweitern oder verkleinern) und damit den Spielraum der für A qua Machthaber verfügbaren Möglichkeiten, beabsichtigte Veränderungen von Verläufen zu bewerkstelligen. Falls andere Machthaber A' im Spiel sind, kann A auch den Spielraum der für A' verfügbaren Möglichkeiten, beabsichtigte Veränderungen von Verläufen zu bewerkstelligen, zu steuern versuchen. Das Verhältnis *zwischen* Machthabern ist die *externe* Relation im

14 Vgl. Weber im Zitat in Abschnitt 2 zu Disziplin und Herrschaft.

15 Die Einengung der Bedeutung von Regieren auf Politik ist relativ neu, die ursprünglich weite Bedeutung »bezog sich auf die unterschiedlichsten Formen der Führung von Menschen« (Foucault 2006: 161f.).

Begriff der Macht: Vermittelt durch ihre unmittelbaren und mittelbaren Kräfte F* und F+ verhalten A und A' sich zueinander in einem *Machtverhältnis*.

Die Macht, die *mehrere* Akteure in einer gemeinsamen Situation haben, findet für jeden einzelnen Akteur ihre Grenze in dem, was die eigenen Kräfte, über die jeder Akteur in dieser Situation verfügt, in dieser Situation ausrichten bzw. verändern könnte, und findet für alle zusammen ihre Grenze in dem, was sie mit den Kräften, über die sie verfügen würden, wenn sie sich zu einem korporativen oder kooperativen Akteur vereinen würden, an Veränderungsverläufen steuern könnten. Die *Auswirkung* von ausgeübter Macht zeigt sich in allen durch sie bewirkten Veränderungen in allen Praktiken eines bestimmten Praxis*bereichs* (der klein sein kann, wie eine punktuelle Interaktion, oder groß, wie ein soziales Funktionssystem, z.B. Wirtschaft), nämlich desjenigen Bereichs, auf den wir unsere, nie anders als ausschnitthafte Beobachtung von Machtverhältnissen und Machtwirkungen einstellen.

Ein kleinformatiges Beispiel: Eine im Vergleich zu den Gegenspielern bessere Ballkontrolle (F*) gibt mehr Chancen, den Verlauf des Fußballspiels in Richtung des beabsichtigten Ziels zu steuern (F+), und das gilt für einzelne Spieler sowie für Mannschaften als kooperative Akteure: ein asymmetrisches Machtverhältnis unterschiedlich machtvoller Akteure, die sinnhaft orientiert (u.a.) im sozialen Kraftfeld von sportlichem Wettbewerbsdruck aufeinander wechselwirken. Ein Beispiel im Makroformat: Angenommen, wir betrachten die Menge aller Mitglieder einer politisch verbundenen Gesellschaft, z.B. die Staatsbürger Deutschlands. Wie viel Macht ist in dieser Gesellschaft? Die machttheoretische Antwort lautet in größter Abstraktion: Die Gesamtheit aller Veränderungen von Verhaltensverläufen, die alle Mitglieder mit allen Kräften, die sie kontrollieren können, kontrafaktisch robust zu steuern vermögen würden, falls sie bestimmte, ihnen wichtige Veränderungen bewerkstelligen wollten.

4. Rechenkräfte und Geisteskräfte

Betrachten wir die Fähigkeit der meisten Menschen, einfache Rechnungen, z.B. die Addition zweier Zahlen im Zahlenraum bis 100, im Kopf auszuführen. Wenn ich will, kann ich meine vorhandene Fähigkeit des Kopfrechnens steigern, z.B. indem ich eine Praxis des Übens ausbilde oder indem ich eine einfache algorithmische Kulturtechnik erlerne, das Rechnen mit Papier und Bleistift, oder mit einem Abakus, oder, wie es heute weltkulturell selbstverständlich geworden ist, das Rechnen mit elektronischen Rechenmaschinen in Form winziger Computer. Ich kann lernen wollen, solche Techniken zu beherrschen, z.B. weil ich besser (komplexer, schneller, zuverlässiger, in einem größeren Zahlenraum) rechnen können will. Ich kann es beabsichtigen und auch bewerkstelligen.

Wir könnten hier von der Macht des Erlernens sprechen. Die Kulturtechnik, rechnen zu können, ohne Agenten in Anspruch nehmen zu müssen, muss erlernt werden und ermächtigt dann in begrenztem Umfang zum korrekten Rechnen. Aber wenn ich gelernt habe (F*), leistungsfähige rechnende Agenten zu kontrollieren und ihre Rechenkraft (F+) so zu steuern, dass sie sich meinen Wünschen gemäß verhalten, habe ich mittelbar durch die Rechenleistung des Technofakts mein Vermögen vergrößert, alle Arten von Zielen zu erreichen, deren Erreichung Rechenkraft erfordert: ein Machtzuwachs.

Die potentielle Rechenleistung, die stärkere Rechenkraft, die mir *nun* zu Verfügung steht, erweitert den Spielraum der möglicherweise rational, als Erfolg, verfolgbarer Zwecke, die ich mir setzen kann, und insofern: meine Handlungsmacht. Verfügt A über mehr Leistungskraft beim Rechnen als A', hat A größere Chancen als A', gesetzte Ziele, für die korrekte und rasche Berechnungen wichtig sind, zu erreichen. Angenommen, A und A' sind unternehmerisch tätig, in Wettbewerbspraktiken involviert, als solche aneinander sinnhaft orientiert und insofern wechselwirkend im sozialen Kraftfeld von *Marktkräften* situiert. Dann hat A im Wettbewerb mit A' mehr Macht als A' z.B. dann, wenn es darauf ankommt, schneller als die Konkurrenz die Profitabilität einer Investition zu kalkulieren.

Wie steht es mit der Macht, die man der Vernunft zuschreiben möchte? Halten wir uns an das vorige Beispiel vom Rechnen, an Regeln, Aktivitäten und Leistungen. Man könnte »die Vernunft« zu einem Machthaber verklären, dem die Vernünftigen loyal im Denken und Handeln verbunden sein sollen. Aber die Vernunft herrscht nicht. Ob Kants Wendung trägt, Vernunft als *ursprüngliche Selbstgesetzgebung* des vernünftigen Willens und aller Wesen, die einen solchen haben, zu begreifen, sei dahingestellt. Wir Pragmatisten sprechen lieber von rationalen Aktivitäten in Praktiken und von den nötigen Fähigkeiten und Kompetenzen: Aktivitäten in Praktiken des Denkens, Nachdenkens, Vorausdenkens (wozu auch die Fähigkeit gehört, Regeln der Logik und vielen weiteren Systemen von Regeln denkend zu folgen, deren Befolgung etwas leistet für rationale Aktivitäten in Praktiken); Aktivitäten in Praktiken des Verstehens und Erklärens; und alles integrierend: die Kompetenz, im kulturell ausgelegten Raum der Gründe umsichtig so sich zu orientieren, wie es die meisten anderen Menschen, die man als Interaktionspartner ernst nehmen würde, gleicherweise tun (Haugaard/Kettner 2020).

Wenn wir mit den Fähigkeiten und Kompetenzen, deren Besitz wir summarisch als den Besitz von Geisteskräften oder als die Fähigkeit, sich vernünftig orientieren und verhalten zu können, beschreiben, wirklich *arbeiten* und etwas *ausrichten* – innerhalb von Praktiken in Situationen, in denen es auf die entsprechenden Leistungen ankommt, weil sie Chancen für Erfolg im Ergebnis verändern – haben wir durch Besitz und Kontrolle unserer Geisteskräfte eine gewisse Macht, viele Kräfte, deren Spiel wir andernfalls bloß ausgesetzt wären, mehr oder weniger zu steuern bzw. zu ›regieren‹ (u.a. durch Urteilskraft), sodass wir Übel, die wir aus gutem Grund ver-

meiden wollen, fernhalten und Zielen, die wir aus guten Gründen verfolgen wollen, näherkommen können. Mit der ›Macht unserer Vernunft‹ finden wir uns in einem kulturell immer schon ausgelegten Raum von Gründen vor, der gleichsam wie ein Feld intelligibler Kräfte wirkt, und wir vermögen diesen Raum dank der Macht unserer Vernunft auch tatkräftig zu verändern.

5. Arbeitende Kräfte, Agenten und Akteure

Ich komme zurück auf die am Ende von Abschnitt 2 liegengebliebene Frage nach der allgemeinen Beschaffenheit machtvoller Akteursinstanzen.

Wenn wir, wie vorgeschlagen, Macht als eine komplexe Relation des Steuerns, Kontrollierens, Beherrschens von *Kräften* denken, unterstreichen wir eine wichtige Pointe von Macht, nämlich die Verrichtung einer *Arbeit*. Wo Macht ausgeübt wird, wird etwas getan, wird etwas ausgerichtet und geschieht etwas; im Falle der bloßen Potenzialität von Macht eine Arbeit, die stattfinden würde, wenn machtvolle Akteure ihre Machtpotentiale verwirklichen würden. Natürlich ist es auch möglich, dass ein Akteur über eine bestimmte Macht verfügt, um etwas zu *verhindern*, was andernfalls geschehen, sich materialisieren oder getan werden würde. Auch in diesem Fall wird Arbeit verrichtet, eine Widerstands- oder Blockadearbeit. Es muss Energie (= aktualisierbare Arbeit) aufgewendet werden.

Versuchen wir für einen Augenblick in einer naturalistisch klingenden, aber kultural gedachten Redeweise über Kräfte, Arbeit und Energie gewissen Analogien nachzugehen, die zwischen dem naturalistischen Machtvokabular der Mechanik, das in der Physik präzisiert wird, und dem der Lebenswelt, das in Kulturwissenschaften präzisiert wird, bestehen. Soweit solche Analogien sachhaltig sind, sind sie es nicht durch den Machtbegriff selbst, der im Begriffsrahmen des Naturalismus als solcher ortlos bleibt.[16] Sachhaltig sind sie vielmehr dank des Begriffs der Kraft. Er dient in der Physik zur Beschreibung und Erklärung des Verhaltens von Objekten bestimmter Art (z.B. mechanischer Körper), wenn sie durch eine Kraft (z.B. Gravitation) aufeinander wechselwirken. Kategorial haben Kräfte Vektorform, also Größe und Richtung. Sie bewirken (oder zeigen sich uns als wirksam in) Veränderungen oder im Verändern von Veränderungen (z.B. wenn

16 Nur in der Lebenswelt kann Macht missbraucht werden. Auch ist physikalisch weder ein Machtausübender, ein Subjekt von Herrschaft (Herrscher) zu denken, noch ein machtunterworfenes Objekt (Untertan). Allenfalls die Totalität der Naturgesetze, die zumindest nach Ansicht von physikalistischen Deterministen alles, was sich im Universum ereignen kann, vollständig beherrscht, böte eine Analogie zum Herrscher, eine allerdings übermäßig angestrengte. Sinnfälliger für Theologen ist die Allmacht Gottes.

eine Gegenkraft eine Veränderung, die andernfalls bewirkt würde, blockiert). Typische Veränderungswirkungen von Kräften, wie sie in der klassischen Mechanik interessieren (Gravitation, Reibungs- Druck- und Zugkräfte), sind Geschwindigkeitsänderungen (Beschleunigen, Abbremsen, Ändern der Bewegungsrichtung) und Formveränderungen (Verformungen). Veränderungswirkungen von typischen Kräften, wie sie nur in kulturellen Welten wirken (z.B. Willenskraft, Gesetzeskraft, Kaufkraft, Eigentum[17]), lassen sich nicht auf mechanische Kräfte und deren Veränderungswirkungen reduzieren, sondern *involvieren* sie nur (z.B. in Körperkraft, Feuerkraft). Denn die Kraftfelder, die von wesentlich *kulturell* konstituierten Kräften im Hintergrund der Lebenswelten sozialisierter Menschen aufgespannt werden, manifestieren sich in gerichteten Veränderungen im Willen und der Motivation von Personen sowie in gerichteten Veränderungen im Netzwerk ihrer Überzeugungen, mit allen Folgen, die derartige gerichtete Veränderungen für alle möglichen Praktiken haben.[18] In soziale Praktiken involviert, sind Personen, die sich so oder anders verhalten können, aneinander sinnhaft orientiert und insofern wechselwirkend.

Die vieldeutige Semantik des englischen ›power‹ ist viel näher an physikalisch verstandener Arbeitsleistung, Kraft und Stärke, als die von ›Macht‹, die sinngemäß die Herrschaft von Menschen über Menschen in den Vordergrund rückt. Zwar wäre es unsinnig, zur empirischen Angemessenheit eines Machtbegriffs zu fordern, er müsse gleichsinnig zum Sinn beitragen in Feststellungen über die Energie (*horsepower*) von Dampfmaschinen und die von Sumo-Ringern (*Körperkraft*). In den beiden Sätzen ist zwar Ähnliches gemeint, und beides hat auch mit Macht zu tun, aber nicht im selben Sinne. Eine Sinngemeinsamkeit wird erst auf einer abstrakteren Ebene deutlich, wenn wir den Arbeitsleister oder -verrichter als eine ontologisch offene Variable behandeln. Wir können die Werte dieser Variablen ›Agenten‹ nennen. Dienste leistende menschliche Arbeitskräfte sind machttheoretisch betrachtet Agenten, leistungsfixierte Maschinen sind es ebenso. Die Dampfmaschine soll eine bestimmte Arbeit leisten, zu nichts sonst haben ›wir‹ sie konstruiert, und diese Arbeit konnte auch von anderen Maschinen, z.B. Windmühlen geleistet werden als auch von abgerichteten tierischen Arbeitskräften (z.B. von Pferden) und vertraglich festgelegten oder in anderer Form eingebundenen und angewiesenen menschlichen Arbeitskräften (Freiwillige, beauftragte Dienstleister, Untergebene, Sklaven). Um im Vergleich zu bleiben: Jemand kontrolliert (F*) die Dampfmaschine so, dass die Arbeit, die sie leistet (F+), ein angesteuertes Ziel näherbringt. In diesem machtvollen

17 Eigentum zu haben (nicht: Besitz) verleiht innerhalb des normativen Kraftfelds von Eigentumsrechten die Macht, zu vermögen, andere vom Gebrauch des Eigentums auszuschließen.

18 Alle etablierten Praktiken bilden den »Background« sensu Searle 2019: 145–173. Hier verlässt der späte Searle auf erhellende kulturreflexive Weise den Rahmen seiner ansonsten eng intentionalistisch interpretierten Sprechakttheorie, die Soziologen mit einem gewissen Recht als reduktionistisch kritisieren (Witte/Suntrup 2017).

Mensch-Maschine-Ensemble ist die Maschine ein Agent, u.U. innerhalb eines Gefüges von mehreren Agenten, und ihre Aktivität wird von Personen, die untereinander in diversen Machtverhältnissen stehen können (von Bedienern, Auftraggebern, Besitzern, Gewerbeaufsehern u.a.), eingestellt, beherrscht und gesteuert. *Mutatis mutandis* für Knechte,[19] auch Rechenknechte bzw. leistungsfestgelegte algorithmisch gesteuerte Computer.

Es erscheint mir sinnvoll, begrifflich deutlich zu unterscheiden zwischen handlungsfähigen Personen, Akteuren und Agenten. Die übliche englischsprachige Terminologie ist hier unzuverlässig. Bejahen wir methodologischen Anthropozentrismus, dann bleiben Personen, die etwas bewerkstelligen wollen und dies vermögen, der paradigmatische und in puncto Intentionalität anspruchsvollste Fall einer Akteursinstanz im Machtbegriff. Wenn wir, wie in Abschnitt 2 vorgeschlagen, die Unterstellung eines gründeresponsiven freien Willens abschattieren und durch die Fähigkeit ersetzen, die Erfüllung von Zwecken zu leisten, dann können wir hier den Schnitt zwischen Personen und Agenten legen. Wo es auf Unterschiede nicht ankommt, kann der abstrakte Begriff eines ›Akteurs‹ einspringen.

Als Agenten und Agentensysteme[20] können zum einen rechtlich verfasste Entitäten gelten, aber nach dem Sprachgebrauch in Informatik und KI-Forschung auch maschinentechnische Artefakte, z.B. Computerprogramme, die in Aktion ein Stück weit unabhängig von Benutzereingriffen ablaufen, in einem definierten Spielraum selbstständig Entscheidungen treffen, mit anderen Programmen Informationen austauschen (›kommunizieren‹) und aus eigenen Kräften (›autonom‹) mehr oder weniger adaptiv (›intelligent‹) auf veränderte Randbedingungen und Inputs reagieren können.[21]

Genügt es machttheoretisch vielleicht, Agenten als ›Mittel‹ oder ›Hilfsmittel‹ (für Personen, die etwas bewerkstelligen wollen) zu denken? In vielen Fällen wird diese eher banale Kategorisierung tatsächlich genügen. Etwa bei teilautonomen Systemen, die ihre Arbeitsaktivitäten zwar selbst steuern, dabei aber auch auf menschliche Eingaben angewiesen bleiben, sogenannte Human in the Loop (HITL)

19 Knechte erbringen Arbeitsleistungen, in der Regel ökonomische, für ihre Herren. Hier berühren wir das seit Hegels *Phänomenologie des Geistes* (Kapitel IV) berühmte Denkmodell eines Zusammenhangs von Selbstbewusstsein, Abhängigkeit, Unabhängigkeit, Herrschaft und Knechtschaft, auf das ich an dieser Stelle nicht eingehen kann, das aber für die immens gesteigerte Technikabhängigkeit, die wir im Kulturprozess der Digitalisierung eintreten lassen, erhellend ist.

20 Die Konstruktion von Multiple-Agenten-Systemen spielt für die Interaktion von Menschen und Robotern eine wichtige Rolle (Dahiya et al. 2023). Solche Systeme machttheoretisch zu analysieren wäre interessant.

21 Mit Blick auf das neue philosophische Forschungsfeld Maschinenethik siehe Cervantes et al 2020; für eine Übersicht über die seit den 1960er Jahren immens angewachsene Forschungsliteratur speziell zu Konversations-Agenten siehe Schöbel et al. 2024.

Systeme. Solche Systeme sind *machterweiternde* Hilfsmittel für Personen, oder ›Machtmittel‹, wenn man so will. Machtmittel ermöglichen Macht, haben aber keine. Es fehlt die Instanz eines rational wollenden Subjekts (Person) oder eines wenigstens Zwecke verfolgenden Akteurs. Algorithmen als Code bzw. Programm verfolgen keine Zwecke, sondern wir mit ihrer Hilfe. Auch als in der passenden maschinellen Umgebung laufendes Programm bzw. als Performanz verfolgen sie keine eigenen Zwecke, sondern verrichten nur die Arbeit, die wir ihnen zugedacht haben.

Wenn wir algorithmisch gesteuerte Maschinen allerdings so konstruieren wollen, dass sie funktionsäquivalent reproduzieren, was unter Menschen (und einigen Tieren) die Macht des Lernens ausmacht, *ermächtigen* wir sie in bestimmten Hinsichten. Dadurch entsteht innerhalb der Beziehung von Personen auf Maschinen ein Machtverhältnis. Man könnte das Ermächtigen der Maschine als ›Externalisierung‹ menschlicher Macht charakterisieren.[22] Sie tritt bei vollautonomen Systemen hervor, die nicht mehr auf von Menschen kontrollierte Eingaben angewiesen sind und insofern als Human-out-of-the-Loop (HOOL) bezeichnet werden. Denken wir z.B. an militärische HOOL-Waffensysteme, etwa autonom ihr Ziel verfolgende Flugabwehrraketen, dann ist diese Machtexternalisierung, einmal losgelassen, vielleicht nicht mehr einzufangen, bleibt aber doch in dem gewollten und technisch eingerichteten Spielraum unserer Zwecke für diese Art von Gerät. Weitergehende Externalisierung von Macht ist denkbar mit selbstlernenden Maschinen. Die Macht des Lernens, die wir übertragen haben, ermächtigt selbstlernenden Maschinen womöglich, den gewollten und technisch eingerichteten Spielraum unserer Zwecke für sie zu verlassen (Davidson 2023).

Wenn ich mir passende Software-Agenten dienstbar machen kann, kann ich dadurch meine Handlungsmacht für sehr vielfältige Aufgaben steigern.[23] Falls es mir wichtig ist z.B. für Zwecke der gezielten Recherche im Internet, zum Prüfen und Priorisieren meiner Mails, zum Ausfüllen von elektronischen Formularen, zum Synchronisieren von Profilen in sozialen Netzwerken und Kuratieren von Nachrichten-Feeds, und *last not least* zum Finden guter Angebote im Online-Handel. Digitalisierung als Kulturprozess zeigt sich nicht zuletzt durch die Aufnahme von immer mehr algorithmisch organisierten Agenten in die Routinen und Gewohnheiten unserer Alltagspraxis. Der durchschlagende Welterfolg von Sprache, Bild und Ton generierenden KI-Modellen, der nach dem Coup von *Open AI* Ende 2022 einsetzte –

22 ›Delegieren‹, wie es im Marketingjargon der Werbung für digitalkulturelle ›persönliche intelligente Assistenten‹ manchmal heißt, möchte ich es nicht nennen, wegen der fehlenden Reziprozität (Loer 2021) von Person zu Person. Wenn ich mein Smartphone anweise, mich um 5 Uhr zu wecken, delegiere ich das Aufwecken nicht an mein Smartphone.

23 Anschaulich zum Stand der Technik im Zeichen von KI: [https://www.personal.ai/] und [https://openai.com/index/gpt-4/].

eine vielleicht ähnlich bedeutsame historische Zäsur wie die Ankunft marktreifer Smartphones es 2007 war –, bezeugt ein ungeheuer starkes allgemeines Interesse an machterweiternden Technofakten jeder Art. Es wäre naiv, dieses Interesse für unvermittelt zu halten.

Wir stoßen hier auf einen neuen wichtigen Aspekt der ›Macht der Algorithmen‹, nämlich die ideologische Kraft digitalkultureller Innovationen, als jüngste die der ›künstlichen Intelligenz‹, die sich allesamt im Phantasma des Algorithmus resümieren. *Being digital* war die Zukunftsformel der 1990er Jahre, als das Internet zu einer globalen Wirklichkeit wurde, *becoming smart* ist heute das einzige libidinös besetzte Narrativ eines Fortschritts, der die Reste früherer Sozialutopien ablösen soll in einer Welt der künstlichen Intelligenz. Imaginär begrüßen die Fortschrittsfreudigen den Algorithmus als Symbol einer besseren Zukunft. Der Vergleich mit der Alchemie, auf die die Goldmacher hofften, drängt sich auf. Das Vermögen, Faszination und mimetische Konkurrenz anzukurbeln, das Blaue vom Himmel zu versprechen und damit durchzukommen, ist eine Macht, die Macht der Steuerung kollektiver Aufmerksamkeit zum Ziel von Verklärung, Hype und Affirmation. Bekanntlich sind die machtvollsten Akteure dieser sehr ungleich verteilten Macht kommerzielle.

Mir scheint, die Perspektivierung von KI und Algorithmen als verklärende Ideologie eröffnet ein weites Feld fruchtbare Forschungsfragen für Philosophie, Kultur- und Sozialwissenschaften.

Eine Formulierung von David Beer weist in die richtige Richtung:

> »The notion of the algorithm is part of a wider vocabulary, a vocabulary that we might see deployed to promote a certain rationality, a rationality based upon the virtues of calculation, competition, efficiency, objectivity and the need to be strategic. As such, the notion of the algorithm can be powerful in shaping decisions, influencing behaviour and ushering in certain approaches and ideals. The algorithm's power may then not just be in the code, but in that way that it becomes part of a discursive understanding of desirability and efficiency in which the mention of algorithms is part of ›a code of normalization‹.« (Beer 2016: 9)

Ein Gleiches gilt für ›KI‹ und das damit verbundene Vokabular, das inflationär verwendet wird, um eine gewisse Rationalität zu beschwören, die auf den Tugenden der Berechnung, des Wettbewerbs, der Effizienz, der Objektivität und der Strategie beruht. Bei der Formulierung der suggestivsten Schlagworte, Rhetoriken und Narrative gibt es viele Akteure mit sehr geringer und einige wenige mit enormer Kommunikationskraft, die die entsprechenden *Frames* zu formen und zu verstärken vermögen: ein stark asymmetrisches Machtverhältnis.

Wir sollten die Art und Weise, in der Algorithmen Teil umfassenderer Programme des sozialen Wandels und der Entwicklung sind, die als rational und fortschritt-

lich angepriesen werden, sorgfältig untersuchen. Ein weiteres lohnendes Anliegen auf der Agenda philosophischer Digitalisierungsforschung zu Machtfragen könnte darin bestehen, das massive unreflektierte Vertrauen aufzudecken, das in Systeme gesetzt wird, die als algorithmisch und als künstlich intelligent bezeichnet werden. Interessant erscheint mir vor allem die grundlos optimistische Vorstellung, dass diese Systeme, sobald einige unschöne Probleme digitaler kultureller Diskriminierung technisch gelöst sind, neutral und vertrauenswürdig sein und viel besser funktionieren werden als alles, was Menschen ohne sie vermögen. Diejenigen unter uns, die sich als kritische Humanisten verstehen, sollten untersuchen, wer durch die soziale Konstruktion und selektive Verteilung dieses Vertrauens gewinnt und wer verliert. Auch in dem Vermögen, die von Günther Anders auf den Begriff einer ›prometheischen Scham‹ gebrachte Selbstabwertung angesichts der Vollkommenheit unserer Maschinenwelt zu steuern, liegt Macht.[24]

6. Algorithmische Steuerung, normative Steuerung, Diskursivität

Wenn Steuerung viele Formen annehmen kann, worin unterscheiden sich die Steuerung durch Normen und die Steuerung durch Algorithmen? Aufgrund ihrer Verankerung in Sprechakten haben normative ›Programme‹, von den banalsten und schwächsten, die individuell für private Zwecke als Mittel gewählt werden, wie etwa ein Kochrezept, bis hin zu den kraftvollsten, politisch autorisierten allseits bindenden Gesetzen, von Natur aus die Eigenschaft, dass sämtliche normativen Kräfte, die durch Sprechakte ins Spiel gebracht werden, und alle Wirkungen, die durch sie ins Spiel kommen, von nahezu jedem, der zur einschlägigen Kommunikationsgemeinschaft gehört, in der das betreffende normative ›Programm‹ läuft, umfassend reflektiert werden können. Personen können Adressaten, aber auch Autoren oder Modifikatoren normativer Texturen sein, d.h. kulturell ausgebildeter Gefüge normativer Kräfte. Sie können die Normen auf ihre angebliche Sinnhaftigkeit befragen, gute Gründe für deren Anerkennungswürdigkeit einfordern und angeben, wirkliche von bloß vermeintlichen unterscheiden und schlechte verwerfen. Die Ausübung von Geltungsreflexion, die ›Macht der Reflexion‹ (F*) kann die verfügbaren Rechtfertigungen und damit die normative Kraft (F+) der fraglichen Normen verändern, stärken oder schwächen.

Die prinzipielle Offenheit für Geltungsreflexion – kurz: ›Diskursivierbarkeit‹ – ist für normative Texturen, die kulturell formiert und in einer natürlichen Sprache formuliert sind, eine angestammte Eigenschaft. Algorithmische Programme

24 Der von Anders geprägte sozialpsychologische Begriff meint eine »Scham vor der ›beschämend‹ hohen Qualität der selbstgemachten Dinge« (Anders 1985: 23) und hat m.E. im Kulturprozess der Digitalisierung einen beträchtlichen Erkenntniswert.

hingegen, d.h. die in Programmiersprachen formulierten Kompositionen von Befehlen, die zum Zweck der Steuerung der Veränderungsverläufe der Operationen von Rechenmaschinen entworfen werden, haben diese Eigenschaft nicht von Haus aus. Können algorithmische Steuerungsprogramme diese Eigenschaft überhaupt erhalten? In gewisser Weise ja, aber nur in dem Maße, wie ›wir‹ Diskursivierbarkeit zielgerichtet und bewusst aufbauen. In die Programme selbst wird sie sich umso weniger einschreiben lassen, je mehr ›wir‹ die Programme befähigen, sich selbst zu verändern, und zwar so, dass sich ihre Selbstveränderung nach einer Anfangsphase von der menschlichen Macht des Belehrens immer unabhängiger machen soll. Der maschinell lernende Apparat ist in einer Trainingsphase belehrungsbedürftig, aber wenn er funktioniert, wie er soll, lernt er weiter wie von selbst, und die Box wird schwarz. Die genauen Aktivitäten des Programms innerhalb seiner unmittelbaren operativen Umgebung werden für uns dann opak. ›Wir‹ haben diesen Kontrollverlust gewollt oder nehmen ihn zumindest in Kauf. Wollen ›wir‹ ihn rückgängig machen oder wenigstens in Grenzen halten, dann sind technische Anstrengungen erforderlich, die (derzeit noch) mit fast absurd anmutendem Aufwand verbunden sind.[25] Und selbst wenn die XAI-Forschung große Fortschritte machen sollte und die technische Seite des Problems intransparenter maschineller Lernprozesse löst (Ali et al. 2023), bleibt die normative, letztlich politische Problematik, dass unterschiedliche Gruppen an ganz unterschiedlichen Positionen im Gesamtzusammenhang der Entwicklungs- und Anwendungspraktiken eines (teil)autonomen ML-Systems in der Regel auch ganz unterschiedliche Kontrollinteressen und -auffassungen haben. ›Wir‹ sind viele und haben vielfältige Interessen.

Man denke etwa an die Schüler, deren Eltern, das Lehrpersonal, die Verwaltungsbeschäftigten, Systementwickler, Unternehmer, Datenschutzbeauftragten und Bildungspolitiker im Rahmen der Transformation einer herkömmlichen Schule in eine *Smart School* (McConvey/Guha 2024). Die utopische (dystopische?) Vision ist hier ein KI-gesteuerter personalisierter Unterricht und die automatisierte Überwachung aller pädagogisch relevanten psychosomatischen Parameter aller Schüler (Dimitriadou/Lanitis 2023). Was sie interpretiert haben wollen und was nicht, was sie unter besseren und schlechteren Erklärungen verstehen und worauf jeder ein Recht zu haben meint, all das wird ohne vernünftige politische Aushandlungsprozesse nicht auf einen gemeinsamen Nenner zu bringen sein. Solche politischen Aushandlungsprozesse können nicht ihrerseits an algorithmisch gesteuerte Agenten delegiert werden, es sei denn, die in politisch regierenden Akteuren organisierte Gestaltungsmacht wollte ihre Selbstauflösung gestalten.

Beim Problem der Diskursivierbarkeit von Praktiken, in die ›wir‹ algorithmisch gesteuerte maschinell lernende Systeme eingelassen haben, geht es um mehr als

25 Zum Forschungsfeld XAI siehe den Beitrag von Alpsancar im vorliegenden Band. Mit Bezug auf normative Regulationsbemühungen siehe Pangutti et al. 2023.

nur um ›erklärbare‹ und ›interpretierbare‹ KI.[26] Vielmehr wirft es wichtige wissenschaftliche und politische Fragen auf (Hedlund/Persson 2024; Walter 2024). Wie sollen die Befugnisse verteilt werden? Wer von ›uns‹ kann, wer sollte befähigt und ermächtigt werden, im Zweifelsfall Rechtfertigungsgründe einzufordern, zu geben, zu bewerten? Und welche Akteure füllen dieses idealisierte und potenzielle ›Wir‹ heute tatsächlich aus?

Diese Fragen verweisen auf die komplexe Problematik, ob im Kulturprozesses der Digitalisierung die Innovation, Dissemination und Applikation digitaltechnischer Produkte überhaupt noch gesteuert werden können durch die sozialen Kräfte, über die der politische Souverän verfügt. Angesichts heutiger Ernüchterung über die Steuerungskapazität politischer Macht in Demokratien erscheint die Regulierungsproblematik als eine Herausforderung, die so massiv ist, dass sie die politische Gestaltungsmacht zu überfordern droht.

Literatur

Ali, S; Abuhmed, T.; El-Sappagh, S. et al. (2023): Explainable Artificial Intelligence (XAI). What we know and what is left to attain Trustworthy Artificial Intelligence, in: *Information Fusion*, 99, 101805. [https://doi.org/10.1016/j.inffus.2023.101805].

Anders, G. (1985): Über prometheische Scham, in: Ders., Die Antiquiertheit des Menschen, Bd. I, Über die Seele im Zeitalter der zweiten industriellen Revolution, München: Beck, 21–98.

Beer, D. (2017): The social power of algorithms, in: *Information, Communication & Society*, 20(1), 1–13.

Bogost, I. (2015): The Cathedral of Computation, in: *The Atlantic*, 15.01.2015. [http://www.theatlantic.com/technology/archive/2015/01/the-cathedral-of-computation/384300/].

Bratu, C. (2017): Korporative und kooperative Verantwortung, in: Heidbrink, L.; Langbehn, C.; Loh, J. (Hg.), Handbuch Verantwortung, Wiesbaden: Springer, 477–499.

Cervantes, J.A.; López, S.; Rodríguez, L.F. et al. (2020): Artificial Moral Agents. A Survey of the Current Status, in: *Science and Engineering Ethics*, 26, 501–532.

Dahiya, A.; Aroyo, A.M.; Dautenhahn, K.; Smith, S.L. (2023): A survey of multiagent Human-Robot Interaction systems, in: *Robotics and Autonomous Systems*, 161, 104335. [DOI:10.48550/arXiv.2212.05286].

26 Wie können wir maschinell lernende KI-Systeme in dem Sinne explizierbar machen, dass ihre äußerlich undurchsichtigen Input-Output-Funktionen für Menschen mit natürlicher menschlicher Intelligenz nachvollziehbar werden?

Davidson T. (2023): The Danger of Runaway AI, in: *Journal of Democracy*, 34(4), 132–140.

DeVrio, A.; Eslami, M.; Holstein, K. (2024): Building, Shifting, & Employing Power. A Taxonomy of Responses From Below to Algorithmic Harm, in: *FAccT '24. Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, June 2024, 1093–1106.

Dimitriadou, E.; Lanitis, A. (2023): A critical evaluation, challenges, and future perspectives of using artificial intelligence and emerging technologies in smart classrooms, in: *Smart Learning Environments*, 10(12), 1–26.

Foucault, M. (2006): Sicherheit, Territorium, Bevölkerung. Geschichte der Gouvernementalität I. Vorlesungen am Collège de France 1977–1978, Frankfurt a.M.: Suhrkamp.

Haugaard, M.; Kettner, M. (2020) (Hg.): Theorising Noumenal Power. Rainer Forst and his Critics, London: Routledge.

Hedlund, M.; Persson, E. (2024): Expert responsibility in AI development, in: *AI & Society*, 39, 453–464.

Hilgendorf, E. (2012): Können Roboter schuldhaft handeln?, in: Beck, S. (Hg.), Jenseits von Mensch und Maschine. Ethische und rechtliche Fragen zum Umgang mit Robotern, Künstlicher Intelligenz und Cyborgs, Baden-Baden: Nomos, 119–132.

Hobe, S. (2023) (Hg.): Die Macht der Algorithmen, Baden-Baden: Nomos.

Hubig, C. (2015): Die Kunst des Möglichen III. Grundlinien einer dialektischen Philosophie der Technik. Macht der Technik, Bielefeld: transcript.

Kettner, M. (2018): The Forstian Bargain. Overrationalizing the Power of Reasons, in: *Journal of Political Power*, 11(2), 139–150.

Kettner, M. (2021): Die künstliche und die natürliche Intelligenz der Gesellschaft, in: Held, B.; Oorschot, F. (Hg.), Digitalisierung: Neue Technik – neue Ethik? Heidelberg: HEIbook, 182–209.

Knight, F.H. (1939): Bertrand Russell on Power, in: *Ethics*, 49(3), 253–285.

Krabbe, A.; Niemann, H.M.; von Woedtke, T. (2022): Künstliche Intelligenz. Macht der Maschinen und Algorithmen zwischen Utopie und Realität, Leipzig: Evangelische Verlagsanstalt.

Loer, T. (2021): Reziprozität. Annäherungen an eine Grundlegung der Kultur- und Sozialwissenschaften, Wiesbaden: Springer.

Lukes, S. (2005, 2. Aufl.): Power. A Radical View, New York: Macmillan.

Marx, K. (1976[1844]): Zur Kritik der Hegelschen Rechtsphilosophie. Einleitung, in: Marx, K.; Engels, F.: Werke, Band 1, Berlin: Dietz Verlag, 378–391.

McConvey, K.; Guha, S. (2024): »This is not a data problem«. Algorithms and Power in Public Higher Education in Canada, in: *CHI '24: Proceedings of the CHI Conference on Human Factors in Computing Systems*, Article No. 16, 1–4. [https://doi.org/10.1145/3613904.3642451].

McLuhan, Marshall (1994): Understanding Media. The Extensions of Man, Cambridge (MA): The MIT Press.

Moore, G. (1999): Corporate Moral Agency. Review and Implications, in: *Journal of Business Ethics*, 21, 329–343.

Morriss, P. (2002): Power. A Philosophical Analysis, Manchester: Manchester University Press.

Panigutti, C.; Hamon, R.; Hupont, I. et al. (2023): The role of explainable AI in the context of the AI Act [https://www.forbes.com/sites/glenngow/2021/10/10/the-eu-is-regulating-your-ai-five-ways-to-prepare-now/].

Pasquale, F. (2015): The black box society. The secret algorithms that control money and information, Cambridge (MA): Harvard University Press.

Peirce, C.S. (1902): Virtual, in: Baldwin, J.M. (Hg.): Dictionary of Philosophy and Psychology, Vol. II, London: Macmillan, 763–764.

Popitz, H. (1992, 2. Aufl.): Phänomene der Macht, Tübingen: Mohr.

Richter, M. (2011): Freiheit und Macht. Perspektiven kritischer Gesellschaftstheorie – der Humanismusstreit zwischen Sartre und Foucault, Bielefeld: transcript.

Röttgers, K. (1990): Spuren der Macht. Begriffsgeschichte und Systematik, Freiburg: Alber.

Russell, B. (1938): Power. A new social analysis London: George Allen & Unwin.

Schmidt, R. (2015): The Power of Algorithms. The Use of Algorithmic Logic and Human Curation at *The Guardian*. MA-Thesis, Universität Stockholm.

Schöbel, S.; Schmitt, A.; Benner, D. et al. (2024): Charting the Evolution and Future of Conversational Agents. A Research Agenda along Five Waves and New Frontiers, in: *Information Systems Frontiers*, 26, 729–754.

Schröder, M.; Schwanebeck, A. (2017) (Hg.): Big Data – In den Fängen der Datenkraken. Die (un-)heimliche Macht der Algorithmen, Baden-Baden: Nomos.

Searle, J. (2010): Making the Social World, Oxford: Oxford University Press.

van Fraasen, B.C. (1980): The Scientific Image, Oxford: Clarendon Press.

Walter, Y. (2024): Managing the race to the moon: Global policy and governance in Artificial Intelligence regulation. A contemporary overview and an analysis of socioeconomic consequences, in: *Discover Artificial Intelligence*, 4, Artikel 14. [https://doi.org/10.1007/s44163-024-00109-4].

Weber, M. (1964): Wirtschaft und Gesellschaft. Grundriß der verstehenden Soziologie. Erster Halbband. Studienausgabe, Köln/Berlin: Kiepenheuer & Witsch.

Witte, D.; Suntrup, J.C. (2017): John Searle on Power and Human Rights. Critical Reflections on Recent Developments in his Social Ontology, in: Gephart, W.; Suntrup, J.C. (Hg.), The Normative Structure of Human Civilization. Readings in John Searle's Social Ontology, Frankfurt: Klostermann, 89–118.

Zuboff, S. (2018): Das Zeitalter des Überwachungskapitalismus, Frankfurt/New York: Campus.

Zu den Autorinnen und Autoren

Rainer Adolphi ist Professor für Philosophie [i.R.], TU Berlin; Arbeitsbereiche: Sozialphilosophie, Theorie der Kultur, politische Philosophie, Anthropologie, History of Ideas.

Suzana Alpsancar ist Juniorprofessorin für Angewandte Ethik mit Schwerpunkt Technikethik in digitalen Welten an der Universität Paderborn. Sie untersucht philosophische und ethische Herausforderungen der Technisierung sowie die besondere Wissens- und Reflexionsform, die in diesen Problematisierungen Ausdruck findet. Aktuell untersucht sie die Ethik und Normativität erklärbarer KI (als Forschungsgruppenleiterin im Sonderforschungsbereich/Transregio »Co-Constructing Explainability«) sowie das Verhältnis von Nachhaltigkeit und Digitalisierung (als Forschungsgruppenleiterin des NRW-Forschungsnetzwerkes »Sustainable Life-Cycle of Intelligent Socio-Technical Systems« sowie des EU-geförderten Forschungsprojektes »Cultures of the Cryosphere«).

Kai Denker studierte Philosophie, Geschichte und Informatik an der TU Darmstadt. Nach der Promotion 2018 in Philosophie mit einer Arbeit zur Philosophie der Mathematik bei Gilles Deleuze war er PostDoc im BMBF-geförderten Verbundvorhaben »Parallelstrukturen, Aktivitätsformen, Nutzerverhalten im Darknet (PANDA)«. Seit 2021 ist er Verbundkoordinator des BMBF-geförderten Vorhabens »Meme, Ideen, Strategien rechtsextremistischer Internetkommunikation (MISRIK)« sowie seit 2024 des BMBF-geförderten »Kompetenznetzwerks Datentreuhandmodelle (K-Netz_DTM/DaTNet)«. Er ist Mitglied im DFG-Netzwerk »Philosophie der Digitalität: Phänomenologische und systematische Perspektiven«. Seine Forschungsinteressen reichen von den philosophischen Grundlagen der Informatik, insbes. der Informationstheorie und Algorithmik, bis zu Aneignungsstrategien der extremen Rechten auf Online-Kommunikationsplattformen.

Gabriele Gramelsberger ist Professorin für Wissenschaftstheorie und Technikphilosophie an der RWTH Aachen sowie Direktorin des Aachener Käte Hamburger Kol-

legs »Kulturen des Forschens«. Sie beschäftigt sich mit der Digitalisierung der Wissenschaften, insbesondere mit der Frage nach neuen Methoden der Erkenntnisproduktion. Zu diesen neuen Methoden gehörte von Beginn der Computerentwicklung an die Computersimulation und seit einigen Jahren das maschinelle Lernen basierend auf künstlichen neuronalen Netzen. Dadurch hat sich der Zugang zu Wissen über komplexe Systeme, über zukünftige Trends sowie über Zusammenhänge in großen Datenmengen eröffnet. – Aktuelle Veröffentlichungen: Software in science is ubiquitous yet overlooked (in: Nature Computational Science 2024); Philosophie des Digitalen zur Einführung (Junius 2023); Künstliche Intelligenz, Computerspiele und Sozialität (in: Acta Historica Leopoldina 2023); Operative Epistemologie (Meiner 2020); Natures of Data (diaphanes 2020).

Armin Grunwald ist seit 1999 Leiter des Instituts für Technikfolgenabschätzung und Systemanalyse (ITAS) am Karlsruher Institut für Technologie (KIT); seit 2002 leitet er ebenfalls das Büro für Technikfolgen-Abschätzung beim Deutschen Bundestag (TAB). Seit 2007 ist er Professor für Technikethik und Technikphilosophie am KIT. Seine Arbeitsgebiete sind die Theorie und Methodik der Technikfolgenabschätzung, Konzeptionen nachhaltiger Entwicklung, Ethik der Technik, insbesondere der Digitalisierung sowie wissenschaftliche Politikberatung. Armin Grunwald ist Mitglied der Deutschen Akademie der Technikwissenschaften (acatech) seit 2009, seit 2014 Mitglied im Präsidium. Er war Mitglied der Endlagerkommission des Deutschen Bundestages (2014–2016) und der Ethik-Kommission für autonomes und vernetztes Fahren des Bundesverkehrsministeriums (2016/2017); aktuell ist Armin Grunwald Ko-Vorsitzender des Nationalen Begleitgremiums Endlagersuche und Mitglied des Deutschen Ethikrates.

Susanne Hahn lehrt Philosophie an der Heinrich-Heine-Universität Düsseldorf. Ihre Arbeitsschwerpunkte sind: Philosophische Fragen der Digitalisierung, Rationalität, Normativität und Wirtschaftsethik. Für ihr Buch *Rationalität. Eine Kartierung* wurde sie 2017 mit dem Deutschen Preis für Philosophie und Sozialethik ausgezeichnet.

Kerrin Artemis Jacobs ist Assoziierte Professorin für Praktische Philosophie. Sie forschte und lehrte an verschiedenen Einrichtungen der Universitäten Osnabrück, Göttingen, Mainz, Witten/Herdecke und der Hokkaido-Universität in Japan. Ihre Forschungsschwerpunkte liegen in den Bereichen der Sozialphilosophie, der Persönlichkeitspsychologie und der Einsamkeitsforschung. Am Institut für Erste Person-Forschung (Department für Psychologie) der Universität Witten/Herdecke forscht sie gegenwärtig zur Thematik des veränderten Erlebens von sozialer Bezogenheit in psychischen Erkrankungen unter besonderer Berücksichtigung von Prozessen intuitiver Selbst- und Welterfahrung.

Natalia Juchniewicz ist Assistenzprofessorin an der Philosophischen Fakultät der Universität Warschau; Doktor der Philosophie (2015) und Doktor der Soziologie (2018). Sie forscht auf dem Gebiet der Technologiephilosophie, der Soziologie der neuen Medien und der künstlichen Intelligenz in Bezug auf die klassische Sozialphilosophie.

Matthias Kettner ist Professor für Philosophie und Diplompsychologe. Er wurde 2002 auf den Lehrstuhl für Praktische Philosophie der damaligen Fakultät für Kulturreflexion der Universität Witten/Herdecke berufen und ist seit 2020 Seniorprofessor an der Fakultät für Wirtschaft und Gesellschaft im Department für Philosophie, Politik und Ökonomik. Von 1994 bis 2002 forschte er am Kulturwissenschaftlichen Institut Essen (KWI) in Arbeitsgruppen zu Pragmatismus, Neuen Medien und Demokratie. Seine aktuellen Forschungsschwerpunkte sind: Bereichsethiken und Diskursethik, Wissenschaftstheorie der Psychoanalyse, Kritische Theorie institutioneller Pathologien, Digitalisierung als Kulturprozess. – Publikationen siehe [orcid.org/0000-0001-5896-7861].

Sybille Krämer, bis zum Ruhestand 2018 Professorin für Philosophie an der FU Berlin; jetzt Seniorprofessorin Leuphana Universität Lüneburg. Gastprofessuren an Universitäten in Tokyo, Yale, Santa Barbara, Santiago de Chile, Wien, Zürich. Diverse Fellowships. 2016 Ehrenpromotion der Universität Linköping/Schweden. Ehemals Mitglied des Wissenschaftsrates, des Scientific Panel des European Research Council (Brüssel) und des Senats der Deutschen Forschungsgemeinschaft und *permanent fellow* am Wissenschaftskolleg zu Berlin. Forschungsbereiche: Philosophischer Rationalismus (Leibniz, Descartes), Epistemologie und Theorie des Geistes, Sprach- und Medienphilosophie, Theorien des Performativen und der Kulturtechniken, Symbolische Maschinen, Konzepte und Geschichte des Digitalen, Künstliche Intelligenz. – Veröffentlichungen: Figuration, Anschauung, Erkenntnis. Grundlinien einer Diagrammatologie (Suhrkamp 2016); Was ist digitale Philosophie? Phänomene, Formen und Methoden (Hg., zus. mit J. Noller, Brill/mentis 2024); Der Stachel des Digitalen. Geisteswissenschaften und Digital Humanities (im Erscheinen, Suhrkamp 2025).

Tobias Matzner ist Professor für Kulturen der Digitalität/Digital Humanities an der Universität Paderborn. Studium der Informatik und Philosophie in Karlsruhe, Rom und Berlin; Promotion in Philosophie am Karlsruher Institut für Technologie (KIT). Danach war er am Internationalen Zentrum für Ethik in den Wissenschaften der Universität Tübingen sowie an der New School for Social Research in New York tätig. Seine Arbeit liegt an den Schnittstellen von politischer Philosophie, kritischen Theorien einerseits und Medientheorie sowie Theorien digitaler Technologie ande-

rerseits. – Letzte Veröffentlichungen u.A.: Algorithms. Technology, Culture, Politics (Routledge 2024).

Nicola Mößner ist Vertretungsprofessorin für Theoretische Philosophie am Institut für Philosophie der Leibniz Universität Hannover. Zuvor war sie als Vertretungsprofessorin am Institut für Philosophie der Universität Stuttgart sowie am Philosophischen Seminar der Westfälischen Wilhelms-Universität Münster tätig. Ihre Forschungsschwerpunkte umfassen die Wissenschaftsphilosophie (v.a. Fragen zur Digitalisierung und ihre epistemologischen Folgen, Ludwik Fleck, soziale Prozesse der Wissensgenese, visuelle Repräsentationen in epistemischen Prozessen), soziale Epistemologie sowie analytische Bildtheorie. – Wichtigste Publikationen: Wissen aus dem Zeugnis anderer – der Sonderfall medialer Berichterstattung (2010); Knowledge, Democracy, and the Internet (2017, zus. mit Philip Kitcher); Visual Representations in Science – Concept and Epistemology (2018); Kalibrierung der Wissenschaft. Auswirkungen der Digitalisierung auf die wissenschaftliche Erkenntnis (2022, zus. mit Klaus Erlach).

Vincent C. Müller ist Alexander von Humboldt Professor für Theory and Ethics of Artificial Intelligence und Direktor des Centre for Philosophy and AI Research (PAIR) an der Universität Erlangen-Nürnberg, ferner Gastprofessor an der TU Eindhoven; Präsident der European Society for Cognitive Systems, Vorsitzender der Society for the Philosophy of Artificial Intelligence und Vorsitzender der euRobotics-Themengruppe zu ›ethischen, rechtlichen und sozioökonomischen Fragen‹. Er beschäftigt sich vorwiegend mit philosophischen Problemen im Zusammenhang mit künstlicher Intelligenz, sowohl in der Ethik als auch in der theoretischen Philosophie. Organisator eine Konferenzreihe zur Philosophie der KI (PT-AI/PhAI); Mitherausgeber der Zeitschrift *Philosophy of AI*. – Einige Publikationen: Ethics of AI and Robotics (Artikel in der Stanford Encyclopedia of Philosophy); Oxford Handbook of the Philosophy of Artificial Intelligence (im Erscheinen, Oxford University Press); Can Machines Think? Fundamental Problems of Artificial Intelligence (im Erscheinen, Oxford University Press); Artificial Minds (mit G. Löhr; im Erscheinen, Cambridge University Press). Die Arbeiten von Vincent C. Müller werden mehr als einmal täglich zitiert. Er hat für seine Institutionen Drittmittel von ca. 9,8 Mio. € eingeworben.

Micha H. Werner studierte Philosophie, Soziologie und Literaturwissenschaften in München, Berlin und Tübingen und promovierte 2001 an der FU Berlin mit einer Arbeit zur Diskursethik als Maximenethik. Nach Stationen in Tübingen, Freiburg/Br. und Utrecht ist er seit 2012 Professor für Philosophie mit dem Schwerpunkt Praktische Philosophie an der Universität Greifswald. Zu seinen Interessen zählen Grundlagen- und Methodenfragen normativer Ethik (bes. neo-kantische Ansätze) sowie

Einzelthemen politischer und angewandter Ethik, insbesondere der Medizin- und der Kommunikationsethik.