

Michael Heinlein  
Norbert Huchler *Hrsg.*

# Künstliche Intelligenz, Mensch und Gesellschaft

Soziale Dynamiken und  
gesellschaftliche Folgen einer  
technologischen Innovation

 Springer VS

---

# Künstliche Intelligenz, Mensch und Gesellschaft

---

Michael Heinlein · Norbert Huchler  
(Hrsg.)

# Künstliche Intelligenz, Mensch und Gesellschaft

Soziale Dynamiken und  
gesellschaftliche Folgen einer  
technologischen Innovation

 Springer VS

*Hrsg.*

Michael Heinlein  
Institut für Sozialwissenschaftliche  
Forschung e. V. – ISF München  
München, Deutschland

Norbert Huchler  
Institut für Sozialwissenschaftliche  
Forschung e. V. – ISF München  
München, Deutschland

ISBN 978-3-658-43520-2      ISBN 978-3-658-43521-9 (eBook)  
<https://doi.org/10.1007/978-3-658-43521-9>

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

© Der/die Herausgeber bzw. der/die Autor(en), exklusiv lizenziert an Springer Fachmedien Wiesbaden GmbH, ein Teil von Springer Nature 2024

Das diesem Band zugrunde liegende Vorhaben wurde mit Mitteln des Bundesministeriums für Bildung und Forschung unter dem Förderkennzeichen 01IS20095 gefördert. Die Verantwortung für den Inhalt dieser Veröffentlichung liegt bei den Autorinnen und Autoren.

Das Werk einschließlich aller seiner Teile ist urheberrechtlich geschützt. Jede Verwertung, die nicht ausdrücklich vom Urheberrechtsgesetz zugelassen ist, bedarf der vorherigen Zustimmung des Verlags. Das gilt insbesondere für Vervielfältigungen, Bearbeitungen, Übersetzungen, Mikroverfilmungen und die Einspeicherung und Verarbeitung in elektronischen Systemen.

Die Wiedergabe von allgemein beschreibenden Bezeichnungen, Marken, Unternehmensnamen etc. in diesem Werk bedeutet nicht, dass diese frei durch jedermann benutzt werden dürfen. Die Berechtigung zur Benutzung unterliegt, auch ohne gesonderten Hinweis hierzu, den Regeln des Markenrechts. Die Rechte des jeweiligen Zeicheninhabers sind zu beachten.

Der Verlag, die Autoren und die Herausgeber gehen davon aus, dass die Angaben und Informationen in diesem Werk zum Zeitpunkt der Veröffentlichung vollständig und korrekt sind. Weder der Verlag noch die Autoren oder die Herausgeber übernehmen, ausdrücklich oder implizit, Gewähr für den Inhalt des Werkes, etwaige Fehler oder Äußerungen. Der Verlag bleibt im Hinblick auf geographische Zuordnungen und Gebietsbezeichnungen in veröffentlichten Karten und Institutionsadressen neutral.

Planung/Lektorat: Cori Antonia Mackrodt

Springer VS ist ein Imprint der eingetragenen Gesellschaft Springer Fachmedien Wiesbaden GmbH und ist ein Teil von Springer Nature.

Die Anschrift der Gesellschaft ist: Abraham-Lincoln-Str. 46, 65189 Wiesbaden, Germany

Das Papier dieses Produkts ist recycelbar.

---

# Förderhinweis



Das diesem Band zugrunde liegende Vorhaben wurde mit Mitteln des Bundesministeriums für Bildung und Forschung unter dem Förderkennzeichen 01IS20095 gefördert. Die Verantwortung für den Inhalt dieser Veröffentlichung liegt bei den Autorinnen und Autoren.

---

# Inhaltsverzeichnis

<b>Künstliche Intelligenz, Mensch und Gesellschaft: Fragestellungen und Perspektiven des Sammelbands</b> .....	1
Michael Heinlein und Norbert Huchler	
<b>Einblicke in die Nutzung Künstlicher Intelligenz</b>	
<b>KI als Kollegin (KIK) – Repräsentative Beschäftigtenbefragung zu Künstlicher Intelligenz am Arbeitsplatz</b> .....	15
Sabine Pfeiffer	
<b>Veränderungsdynamiken durch Technisierung von Arbeit am Beispiel von da Vinci als robotische Chirurgie-Assistenz – Erkenntnisse aus der Empirie</b> .....	41
Regina Wittal und Carolyn Hettinger	
<b>Künstliche Intelligenz als Bild und Mythos</b>	
<b>Das bewegliche Heer der Künstlichen Intelligenz. Ein Technomythos als Summe menschlicher Relationen</b> .....	75
Bruno Gransche und Arne Manzeschke	
<b>Bilder machen Menschen. Zur Bildermacht der Künstlichen Intelligenz</b> .....	109
Arne Manzeschke und Bruno Gransche	
<b>Maschinen lernen nicht! – „Machine learning“-Algorithmen entzaubert</b> .....	131
Irmhild Rogalla	

## **Normierung und Regulierung von Künstlicher Intelligenz**

<b>Die digitale Verantwortungslücke: Vorschläge zur Haftung für algorithmisches Fehlverhalten</b> .....	153
---	-----

Anna Beckers und Gunther Teubner

<b>Normierung, Regulierung, Governance: Wie, von wem und mit welchen Mitteln kann der Einsatz Künstlicher Intelligenz gesellschaftlich gestaltet werden?</b> .....	179
--	-----

Karsten Weber und Nadine Kleine

## **Gesellschaftliche Risiken und soziale Dynamiken Künstlicher Intelligenz**

<b>Soziale Dynamik der Künstlichen Intelligenz</b> .....	201
--	-----

Hartmut Hirsch-Kreinsen

<b>Risiken und Gefahren der ‚Künstlichen‘ ‚Intelligenz‘</b> .....	225
---	-----

Reinhard Kreissl und Roger von Laufenberg

<b>Toys are us: KI, Diversität und soziale Ungleichheit. Oder: KI für Alle?</b> .....	263
---	-----

Heike Raab

<b>Zukunftseuphorie als Trost. Verheißungserzählungen über Künstliche Intelligenz im Kontext gesellschaftlicher Erschöpfungsdiagnosen</b> .....	289
---	-----

Stefan Selke

## **Künstliche Intelligenz im Kontext von Macht, Herrschaft und Demokratie**

<b>Machine learning, political participation and the transformations of democratic self-determination</b> .....	321
---	-----

Jeanette Hofmann and Clara Iglesias Keller

<b>Hegemoniale Machtstruktur? Eine Kartierung der Akteure im aktuellen KI-Diskurs</b> .....	345
---	-----

Karin Hutflötz

---

<b>Künstliche Intelligenz und gesellschaftlicher Wandel – eine Herausforderung für demokratische Macht- und Herrschaftsverhältnisse</b> .....	363
Peter Imbusch und Joris Steg	
<b>Theoretische und methodische Zugänge zu Künstlicher Intelligenz</b>	
<b>Künstliche Intelligenz als kontingenzerzeugende Technik: Eine praxistheoretische Perspektive</b> .....	391
Michael Heinlein	
<b>Selektivitäten (subsymbolischer) Künstlicher Intelligenz</b> .....	431
Norbert Huchler	
<b>Von Interaktion zur Transformaktion: Die Folgen von Künstlicher Intelligenz für Theorien sozialen Handelns</b> .....	465
Valentin Rauer	
<b>Künstliche Intelligenz: Eine Methode für alles? Sozialwissenschaftliche Methodologie der KI-Forschung, ihre Herausforderungen und Möglichkeiten</b> .....	501
Peter Kahlert, Maryam Tatari, Suzette Kahlert, Silvan Pollozek, Johan Buchholz, Benedict Lang und Jan-Hendrik Passoth	



---

# Herausgeber- und Autorenverzeichnis

---

## Über die Herausgeber

**Michael Heinlein**, Dr., Institut für Sozialwissenschaftliche Forschung e. V. (ISF München). Forschungsschwerpunkte: Digitalisierung von Arbeit und Organisationen, Mensch-Maschine-Verhältnisse, Arbeit als Praxis und Handeln, Technik und Materialität, soziale Bedingungen und Folgen Künstlicher Intelligenz, Wissen und Gedächtnis.

**Norbert Huchler**, Dr., Institut für Sozialwissenschaftliche Forschung e. V. (ISF München). Forschungsschwerpunkte: Soziale Mechanismen und Koordination von Arbeit, Umgang mit Komplexität und Unsicherheit, Handeln und Strukturen, explizites/implizites Wissen, Mensch-Technik-Verhältnisse, Arbeits- und Technikgestaltung, Wertschöpfungsdynamiken, Künstliche Intelligenz, Industrie 4.0, Arbeiten 4.0.

---

## Autorenverzeichnis

**Anna Beckers**, Prof. Dr., Universität Maastricht. Forschungsschwerpunkte: Soziale Verantwortung von Unternehmen, globale Wertschöpfungsketten, Wirtschaft und Menschenrechte, unternehmerische Nachhaltigkeit, Verhältnis zwischen digitalen Technologien und Privat- bzw. Haftrecht.

**Johan Buchholz**, European New School of Digital Studies – Europa Universität Viadrina Frankfurt (Oder) und TUM School of Social Sciences and

Technology – Technische Universität München. Forschungsschwerpunkte: Organisationssoziologie, Science and Technology Studies, Digitalisierung der Arbeit, Digital Methods.

**Bruno Gransche**, PD Dr., Karlsruher Institut für Technologie (KIT) – Institut für Technikzukünfte ITZ. Forschungsschwerpunkte: Philosophie neuer Mensch-Technik-Relationen, gesellschaftliche und ethische Aspekte der Digitalisierung, Technikbilder und Menschenbilder, Zeitphilosophie, Zukunftsforschung, integrierte Forschung (Post-ELSI).

**Carolyn Hettinger**, B.A., Institut für Sozialwissenschaftliche Forschung e. V. (ISF München). Forschungsschwerpunkte: Gesellschaftliche Folgen Künstlicher Intelligenz, Arbeit als Praxis, Science and Technology Studies, Techniksoziologie.

**Hartmut Hirsch-Kreinsen**, Prof. Dr. i.R., TU Dortmund – Sozialforschungsstelle. Forschungsschwerpunkte: Innovations- und Techniksoziologie, Arbeits- und Industriesoziologie, digitale Transformation von Arbeit und Gesellschaft, gesellschaftliche Bedingungen und Perspektiven der Entwicklung von Künstlicher Intelligenz.

**Jeanette Hofmann**, Prof. Dr., Alexander von Humboldt Institut für Internet und Gesellschaft (HIIG), Wissenschaftszentrum Berlin für Sozialforschung (WZB) und Freie Universität (FU) Berlin. Forschungsschwerpunkte: Digitalisierung und Demokratie, Künstliche Intelligenz und Gesellschaft, digitale Regulierung, Internet Governance, Governance-Theorie, Internetpolitik, Science and Technology Studies.

**Karin Hutflöt**, Dr., Institut für philosophische Bildungsforschung und Beratung (philBB) und Katholische Universität (KU) Eichstätt-Ingolstadt. Forschungsschwerpunkte: Existenzphilosophie, Wertekonflikte als Bildungsort, Menschenbildung im Dispositiv des Digitalen, gesellschaftlicher Wandel durch die Verheißungen Künstlicher Intelligenz.

**Clara Iglesias Keller**, Dr., Wissenschaftszentrum Berlin für Sozialforschung (WZB). Forschungsschwerpunkte: Plattform-Governance und staatliche Regulierung, Desinformation, digitale öffentliche Verwaltungen, regulatorische und demokratische Implikationen der Ausweitung von Technologien der Künstlichen Intelligenz.

**Peter Imbusch**, Prof. Dr., Bergische Universität Wuppertal. Forschungsschwerpunkte: Politische Soziologie, Sozialstrukturanalyse, Konflikt- und Gewaltforschung, soziologische Theorie, Entwicklungssoziologie.

**Peter Kahlert**, M.A., Lehrstuhl für Techniksoziologie, European New School of Digital Studies – Europa Universität Viadrina Frankfurt (Oder). Forschungsschwerpunkte: Hackathons, Soziologie der Kreativität und Problemlösung, digitale Methoden, Pragmatismus und Sozialtheorie.

**Suzette Kahlert**, M.A., Lehrstuhl für Techniksoziologie, European New School of Digital Studies – Europa Universität Viadrina Frankfurt (Oder). Forschungsschwerpunkte: Gender Studies und Feminist STS, Infrastrukturen und Dezentralisierung, Poststrukturalismus und Sozialtheorie, Qualitative Methoden der Sozialforschung.

**Nadine Kleine**, M.A., Universität Osnabrück. Forschungsschwerpunkte: Technikakzeptanz und Technikwahrnehmung, Technikfolgenabschätzung, Autonomie und Arbeit, assistive Technologien.

**Reinhard Kreissl**, PD Dr., Wiener Zentrum für Sozialwissenschaftliche Sicherheitsforschung (VICESSE). Forschungsschwerpunkte: Soziologie des Rechts, Soziologie der Devianz und der sozialen Kontrolle, Soziologie des Wissens und der Wissenschaft, soziologische Aspekte der Kognitions- und Neurowissenschaften.

**Benedict Lang**, M.A., Lehrstuhl für Techniksoziologie, European New School of Digital Studies – Europa Universität Viadrina Frankfurt (Oder). Forschungsschwerpunkte: Smart cities und digitale Souveränität, Science and Technology Studies, Verantwortungsvolle Technologien und Innovationen.

**Arne Manzeschke**, Prof. Dr., Evangelische Hochschule Nürnberg (EVHN), Institut für Pflegeforschung, Gerontologie und Ethik (IPGE) und Fachstelle für Ethik und Anthropologie im Gesundheitswesen (FEAG). Forschungsschwerpunkte: Technik-, Medizin- und Wirtschaftsethik sowie ihre anthropologischen Grundlagen.

**Jan-Hendrik Passoth**, Prof. Dr., Lehrstuhl für Techniksoziologie, European New School of Digital Studies – Europa Universität Viadrina Frankfurt (Oder).

Forschungsschwerpunkte: Techniksoziologie, Science and Technology Studies, Digitale Infrastrukturen, Responsible AI.

**Sabine Pfeiffer**, Prof. Dr., Friedrich-Alexander-Universität (FAU) Erlangen-Nürnberg. Arbeitssoziologie, Digitalisierung und Informatisierung, Industrie 4.0, Arbeitsvermögen und Erfahrung, Qualifizierung, soziale Innovation, Nachhaltigkeit, Formalisierung, Controlling, soziale Ungleichheit, Arbeitslosigkeit, (Ernährungs-)Armut.

**Silvan Pollozek**, Dr., Lehrstuhl für Techniksoziologie, European New School of Digital Studies – Europa Universität Viadrina Frankfurt (Oder). Forschungsschwerpunkte: Kultur- und Techniksoziologie, Science and Technology Studies, Migrations- und Grenzregime, Datafizierung und Digitalisierung staatlicher Behörden, Infrastrukturen.

**Heike Raab**, Dr., Independent Scholar. Forschungsschwerpunkte: Gender-/Queer-/Disability Studies, Cultural Studies, Soziale Arbeit, Feministische Science & Technology Studies (STS).

**Valentin Rauer**, PD Dr., Türkisch-Deutsche Universität Istanbul. Forschungsschwerpunkte: Soziologische Theorie, Politische Soziologie, Wissens- und Kultursoziologie, Actor-Network-Theory, Science and Technology Studies, algorithmisches Handeln.

**Irmhild Rogalla**, Dr., Hochschule Bremen und Institut für praktische Interdisziplinarität. Forschungsschwerpunkte: Partizipation von Menschen mit Beeinträchtigungen bei der Entwicklung digitaler Medien, Tools und Systeme, Technikfolgenabschätzung, Arbeitsprozesse, Berufe und Kompetenzen.

**Stefan Selke**, Prof. Dr., Hochschule Furtwangen. Forschungsschwerpunkte: Öffentliche Wissenschaft, Digitalisierung und metrische Kulturen, soziale Utopien und Weltraumexploration, technische Assistenzsysteme, nachhaltige Entwicklung und BNE, Armutsökonomie.

**Joris Steg**, Dr., Bergische Universität Wuppertal. Forschungsschwerpunkte: Soziologische Theorie, Politische Ökonomie, Kapitalismus- und Krisentheorie, soziale Ungleichheit und sozialer Wandel.

**Maryam Tatari**, M.A., Lehrstuhl für Techniksoziologie, European New School of Digital Studies – Europa Universität Viadrina Frankfurt (Oder). Forschungsschwerpunkte: Science and Technology Studies, Infrastructure Studies, Value-in-Design, Public Service Media Platforms, Digital Methods.

**Gunther Teubner**, Prof. Dr. Jur. Dr. h.c. mult., Goethe Universität Frankfurt am Main. Forschungsschwerpunkte: Rechtssoziologie, Vertragsrecht, Rechtsvergleichung.

**Roger von Laufenberg**, PhD, Wiener Zentrum für Sozialwissenschaftliche Sicherheitsforschung (VICESSE). Forschungsschwerpunkte: Überwachung und Technologie, Critical Data Studies und Algorithm Studies, Science and Technology Studies.

**Karsten Weber**, Prof. Dr., Ostbayerische Technische Hochschule (OTH) Regensburg, Institut für Sozialforschung und Technikfolgenabschätzung und Regensburg Center of Health Sciences and Technology. Forschungsschwerpunkte: Technikfolgenabschätzung, Innovations- und Technikfolgenforschung, ELSA-Forschung, Akzeptanzfragen und ethische Begleitforschung bei Informations- und Kommunikationstechnologie und im Bereich Gesundheit, Energie und Mobilität, Digitalisierung und Künstliche Intelligenz im Gesundheitswesen.

**Regina Wittal**, M.A., DLR Projektträger Berlin – Bereich Gesellschaft, Innovation, Technologie. Forschungsschwerpunkte: Gesellschaftliche Folgen von Digitalisierung und Künstlicher Intelligenz, Arbeit als Praxis, digitale Technologien und Entwicklungen.



---

# Künstliche Intelligenz, Mensch und Gesellschaft: Fragestellungen und Perspektiven des Sammelbands

Michael Heinlein und Norbert Huchler

---

## Zusammenfassung

Künstliche Intelligenz (KI) wird vielerorts die Rolle einer Schlüsseltechnologie des gesellschaftlichen Wandels im 21. Jahrhundert zugeschrieben. Dafür spricht, dass mittlerweile zahlreiche technologische Anwendungen genutzt werden, die auf maschinellem Lernen und den damit verbundenen Möglichkeiten der statistischen Datensammlung, -nutzung und -verwertung aufbauen. Offen ist jedoch nach wie vor die Frage, welche tiefgreifenden und teilweise latenten Folgen für den Menschen als soziales Wesen und das gesellschaftliche Zusammenleben mit dem Einsatz und der Entwicklung von KI verbunden sind.

---

## Schlüsselwörter

Gesellschaftliche Transformation • Künstliche Intelligenz • Maschinelles Lernen • Nebenfolgen • Technikfolgen

---

M. Heinlein (✉) · N. Huchler  
Institut für Sozialwissenschaftliche Forschung e.V. – ISF München, München, Deutschland  
E-Mail: [michael.heinlein@isf-muenchen.de](mailto:michael.heinlein@isf-muenchen.de)

N. Huchler  
E-Mail: [norbert.huchler@isf-muenchen.de](mailto:norbert.huchler@isf-muenchen.de)

## 1 Künstliche Intelligenz als gesellschaftliche Herausforderung

Künstliche Intelligenz (KI) wird vielerorts die Rolle einer Schlüsseltechnologie des gesellschaftlichen Wandels im 21. Jahrhundert zugeschrieben. Für diese Annahme spricht, dass mittlerweile zahlreiche technologische Anwendungen genutzt werden, die auf maschinellem Lernen und den damit verbundenen Möglichkeiten der statistischen Datensammlung, -nutzung und -verwertung aufbauen. Indem KI große Datenmengen bewältigbar und verborgene Muster und Zusammenhänge sichtbar macht, werden viele Prozesse schneller, einfacher und effizienter – sei es im Alltag, in der Arbeit oder in Organisationen. KI-Systeme erfüllen dabei eine Reihe von Funktionen: Sie ermöglichen die automatische Erkennung und Analyse von Texten, Sprache und Bildern, unterstützen menschliche Aktivitäten in unterschiedlichen Graden, verrichten Tätigkeiten vollautomatisch, können mit Menschen interagieren und kommunizieren sowie als generative KI (leuchtendes Beispiel dieser Tage: ChatGPT) eigenständig Inhalte produzieren, die anschlussfähig für menschliche Sinnprozesse sind.

Das Grundprinzip von KI beruht dabei auf der programmierten Fähigkeit, in großen, in der Regel unstrukturierten Datenmengen Korrelationen und Muster zu erkennen, die die Basis für das technische Wirken von KI – im Sinne einer technisch realisierten ‚Entscheidung‘ – darstellen. Die Pfade, die mit der höchsten (Erfolgs-)Wahrscheinlichkeit erscheinen, werden genutzt und dabei weiter verstärkt. Die Komplexität des technischen Wirkens von KI in sozialen Zusammenhängen kann dabei mit den folgenden vier Dimensionen beschrieben werden:

- KI-Systeme müssen nicht vollends auf einen spezifischen Nutzungskontext hin programmiert sein, da sie sich bis zu einem gewissen Grad durch interne Prozesse („Lernen“) an situative Bedingungen anpassen und ihre Funktionalität kontextspezifisch entfalten können (Adaptivität).
- Auf welcher Grundlage KI-Systeme ihre internen algorithmischen Prozesse ordnen und in einen Output überführen, ist für die Nutzenden aus Gründen der Architektur von KI-Systemen (Stichwort „hidden layers“) in der Regel nicht durchschaubar – und bisweilen ist das Agieren von KI selbst nicht immer als Wirken von Technik zu erkennen (Intransparenz).
- Aufgrund dieser durch statistische Methoden ermöglichten Anpassungsleistung können die Ergebnisse, die KI-Systeme produzieren, sowohl im Zeitverlauf als auch in Abhängigkeit der sachlichen und sozialen Bedingungen einer

Situation unterschiedlich ausfallen und von Nutzenden nicht per se antizipiert werden (Kontingenz).

- Zugleich sind KI-Systeme in ihrem Zugriff auf die soziale Welt notwendigerweise immer beschränkt, da sie nur anhand von Daten operieren und ihre statistischen Methoden systematisch zu Verzerrungen und Pfadabhängigkeiten führen (Selektivität).

Damit ist bereits angedeutet, dass KI einerseits eine Reihe neuer Chancen bietet, mit technischen Mitteln bestimmte Aufgaben zu lösen und dadurch neues Wissen und neue Handlungsmöglichkeiten zu erzeugen. Zugleich wird aber auch deutlich, dass KI-Systeme (insbesondere neuere Formen der subsymbolischen KI) mehr sind als herkömmliche Techniken, die auf einem linearen Verhältnis von Input und Output und einer identifizierbaren Trennung von technischem Wirken und menschlichem Handeln beruhen. KI vermischt sich demgegenüber stärker mit sozialen Prozessen, indem sie diesen Daten entnimmt, die Daten verarbeitet und wieder in die sozialen Prozesse einspielt – und zwar, wie eben beschrieben, auf adaptive, intransparente, kontingente und selektive Weise. Offen bleibt damit die Frage, welche tiefgreifenden und teilweise latenten Folgen für den Menschen als soziales Wesen und das gesellschaftliche Zusammenleben mit dem Einsatz und der Entwicklung von KI verbunden sind. Wie wandelt sich das Verhältnis von Mensch und Technik durch KI und wie ist dieser Wandel zu bewerten? Welche Chancen, aber auch Risiken eröffnen sich durch den Einsatz und die Entwicklung von KI für Mensch und Gesellschaft? Welchen Grenzen unterliegt der Wandel und welche Gestaltungsmöglichkeiten bieten sich? Und nicht zuletzt: Was und wer bestimmt die Entwicklungspfade, die KI nimmt – mit welchen Folgen für wen?

Damit wird der Blick auf die Entwicklung und die Gestaltung von KI-Systemen ebenso wichtig wie der Blick auf die gesehenen und ungesehenen, gewollten und ungewollten gesellschaftlichen (Neben-)Folgen von KI. Dies bedeutet zugleich aber auch, dass das gesellschaftliche Phänomen KI mehr erfordert als Debatten über Datenschutz, Ethik und Diskriminierung – auch wenn diese Diskurse unentbehrlich für den gesellschaftlichen Umgang mit KI sind. Zusätzlich ist jedoch auch ein gesellschaftlich orientierter und damit sozialwissenschaftlicher Blick nötig, der zumindest in Teilen eine ‚post-algorithmische‘ Wende vollzieht: Es geht nicht mehr nur darum, den Algorithmus in den Mittelpunkt zu stellen, sondern die Folgen und (Wechsel-)Wirkungen algorithmischer Prozesse in und mit gesellschaftlichen Zusammenhängen aufzudecken und Antworten darauf zu finden, wie KI gesellschaftlich erzeugt wird und sich auf Gesellschaft rückbezieht bzw. in welcher Gesellschaft wir mit KI leben möchten.



Damit hängt unmittelbar die Frage zusammen, wie eine gerechte und nachhaltige Entwicklung und Einbettung von KI in soziale Zusammenhänge aussehen kann.

Um solch eine gesellschaftliche und folgenorientierte Perspektive einzunehmen, helfen weder Überhöhungen von KI, wie sie immer wieder in post- und transhumanen Diskursen auszumachen sind, noch Relativierungen, die KI als bloßes weiteres Werkzeug einer fortschreitenden Digitalisierung begreifen. Der Blick auf die (Neben-)Folgen von KI und ihre transformativen gesellschaftlichen Effekte benötigt vielmehr eine ‚realistische‘, d. h. an spezifischen Fragestellungen gewonnene und kritische Vorstellung des Wirkens von KI. Dahinter steht allerdings kein Technikpessimismus, sondern das Interesse an der Aufklärung über manifeste und latente Technikfolgen sowie Gestaltung und Gestaltbarkeit von KI in der Gesellschaft.

Das BMBF-geförderte Projekt „Künstliche Intelligenz, Mensch und Gesellschaft“ (KIMeGe) hat sich den Herausforderungen, die KI an den Menschen und die Gesellschaft stellt, zwei Jahre lang gewidmet und ethische, kulturelle, philosophische, politische, rechtliche und soziale Aspekte beleuchtet. Der aus 26 Expertinnen und Experten bestehende Verbund hat dabei zu gesellschaftlichen und sozialen Fragestellungen geforscht, die (teils prospektiv) die Folgen von KI für Mensch und Gesellschaft ausloten. In der Gesamtschau ergibt sich zwar kein lückenloses, jedoch recht umfassendes Bild der gesellschaftlichen Potenziale, aber auch Grenzen und Risiken von KI, das zentrale und wichtige Themen der (im weitesten Sinne) Sozialwissenschaften umfasst: Wissen und Nicht-Wissen, Macht und Herrschaft, Demokratie und politische Teilhabe, soziale Ungleichheiten und Geschlechterdifferenzen, sozialer Wandel und Zukunftsvorstellungen, Praxen und Diskurse, Akteure und Institutionen.

Der vorliegende Sammelband synthetisiert die Ergebnisse dieses Forschungsverbunds. Zugleich erschließt und diskutiert er verschiedene theoretische und konzeptionelle Herangehensweisen an das Phänomen ‚Künstliche Intelligenz‘ aus unterschiedlichen Perspektiven, zeigt politische und zivilgesellschaftliche Gestaltungsoptionen und Einbettungsmöglichkeiten von KI auf und identifiziert weiterführende Fragestellungen, die sich auf das Verhältnis von Mensch, Gesellschaft und KI beziehen. Die Ergebnisse umfassen einerseits Erkenntnisse zu bestehenden und möglichen Folgen des Einsatzes und der Entwicklung von KI in und für spezifische gesellschaftliche und soziale Felder. Andererseits werden aber auch aus den wissenschaftlichen Erkenntnissen abgeleitete Gestaltungsmöglichkeiten, -maßstäbe und -ideen präsentiert. Der Sammelband wendet sich daher nicht nur an ein Fachpublikum und die interessierte Öffentlichkeit, sondern auch an Entscheidungsträger\*innen in Wirtschaft, Politik und (Zivil-)Gesellschaft. Das Ziel besteht darin, gesellschaftlich relevantes Wissen über KI und ihrer Folgen

jenseits einseitig technischer und massenmedial verkürzter Diskurse zu vermitteln und zu einer kritischen Reflexion und Diskussion von KI im Zusammenhang mit gesellschaftlichen und sozialen Fragen anzustoßen.

Das Thema ‚Künstliche Intelligenz‘ hat in den Geistes- und Sozialwissenschaften sowie der Philosophie in den letzten Jahren vermehrt zu Publikationen geführt. Typisch dafür sind jedoch Monographien, die Einzelphänomene behandeln (z. B. soziale Robotik: Bischof 2017; Expertensysteme: Degele 1994) oder Anthologien, die sich entweder dem Phänomen KI als solchem, d. h. seiner Geschichte und seinen Funktionsweisen widmen (Engemann und Sudmann 2018; Rammert 1995; Rosengrün 2021) oder es auf bestimmte Praxisfelder beziehen (Arbeit: Hirsch-Kreinsen und Karacic 2019; Bildung: Schmohl et al. 2022; Kultur: Seyfert und Roberge 2017; Kunst: Dotzler und Karpát 2021). KI wird demnach entweder als eine historisch gewordene Technologie ohne Rekurs auf aktuelle und wahrscheinliche Folgen für Mensch und Gesellschaft behandelt oder aber in ihren Folgen anhand ausgewählter, für sich stehender Bereiche diskutiert. Vielfach verbleiben diese Arbeiten in ihren fachdisziplinären Diskursen; Fragen der gesellschaftlichen Gestaltung und Einbettung von KI werden zugunsten von konzeptionellen Überlegungen, Begriffs- und Theoriearbeit ausgeblendet. Eine umfassende, inhaltlich stringente Auseinandersetzung mit dem sozialen Phänomen KI und dessen Folgen für Mensch und Gesellschaft im Sinne einer öffentlichkeitsorientierten sozialwissenschaftlichen Aufklärung (die gleichwohl auch Begriffs- und Theoriearbeit betreibt) fehlt hingegen bislang.

Dies gilt in einem gewissen Maße auch für den internationalen Bereich, in dem sich vorrangig Publikationen finden, die sich stark an einer Einzelthese orientieren und diese – auch in anthologischer Form – zu belegen suchen (jüngste Beispiele: Studien zum Überwachungskapitalismus [Zuboff 2018] und zu *fake AI* [Kaltheuner 2021]). Der vorliegende Sammelband hingegen wird dem komplexen Phänomen KI und seinen vielschichtigen, teils ambivalenten und widersprüchlichen Folgen dadurch gerecht, indem er *erstens* disziplinär verschiedene, dabei jedoch durch inhaltliche Klammern miteinander abgestimmte Teilaspekte von KI bearbeitet und auf diesem Wege KI als soziales und gesellschaftliches Phänomen breit erschließt. *Zweitens* orientiert er sich an einer Vielzahl von Thesen, die gemeinschaftlich im Forschungsverbund diskutiert und auf der Website des Projekts als Thesenpapier veröffentlicht wurden. Die inhaltlichen Schwerpunkte des Sammelbands werden nachfolgend beschrieben.

## **2 Gesellschaftliche Dimensionen und Perspektiven von Künstlicher Intelligenz**

Im Lichte der Geistes-, Kultur- und Sozialwissenschaften sowie der Philosophie stellt KI ein vielschichtiges und mehrdimensionales Phänomen dar. Dies macht es einerseits leicht, technizistische oder massenmediale Verkürzungen und Überhöhungen zu entlarven, andererseits entzieht sich KI als soziales und gesellschaftliches Phänomen aber auch vereindeutigenden Deutungen und Zuschreibungen, sodass das Denken und Schreiben über KI stets auch selektiv vorgeht und blinde Flecken aufweist. Im Forschungsverbund KIMeGe wurde dieser Selektivität durch eine Strategie der Multiperspektivierung und des dialogischen Austauschs begegnet: Ein gesellschafts- und sozialwissenschaftlich angelegtes Vorhaben, das KI, Mensch und Gesellschaft systematisch zusammendenkt und -bringt, muss einerseits die Differenzen unterschiedlicher Perspektiven anerkennen, andererseits aber eine Entscheidungs- und Handlungsgrundlage schaffen, die keinen Minimalkompromiss darstellt, sondern jede Perspektive hinreichend würdigt und einbezieht. Die letztlich unvermeidbare Perspektivendifferenz auf KI und ihre gesellschaftlichen Folgen sollte demnach möglichst produktiv aufgelöst werden und nicht in einem konkurrierenden, exklusiven Nebeneinander von methodischen Zugängen und inhaltlichen Perspektiven bestehen bleiben.

Dies hat unmittelbare Konsequenzen dafür, welches Wissen für die gesellschaftliche und politische Gestaltung von KI erzeugt wird. KI wirft nicht nur in der Entwicklung und Anwendung drängende Fragen auf, sondern auch hinsichtlich ihrer gesellschaftlichen Folgen, die es in einem multiperspektivischen, dialogischen Verfahren möglichst breit und umfassend zu antizipieren gilt und denen bereits in der Technikentwicklung vorausschauend begegnet werden kann. Unerwünschte Folgen von KI wie etwa daten- und algorithmenbasierte Verstärkung von Ungleichheiten, ethisch wie rechtlich problematische Überwachungsmöglichkeiten, unklare Verantwortlichkeiten, Qualitäts- und Leistungsverschlechterungen, Dequalifizierung, soziale Entfremdung und neue, ungewollte Abhängigkeiten und Risiken, manipulative und destruktive Nutzungsweisen oder ein gesellschaftlich dysfunktionales Zurückdrängen des Menschen aus (system-)relevanten Bereichen können durch einen in diesem Sinne ganzheitlichen, weil perspektivisch aufeinander abgestimmten Zugang erkannt und im Idealfall ‚by design‘, mittels einer abgestimmten organisationalen Einbettung oder durch rechtliche Regulierung ausgeschlossen werden.

Für den Projektverbund waren insbesondere die folgenden Aspekte und Themen leitend:

*Herrschaft, Macht und Ungleichheit:* Technik und Maschinen sind im Laufe der Jahrhunderte zu einem integralen Bestandteil unserer Gesellschaft geworden. Sie prägen heutzutage mehr denn je unsere Kultur und unsere Lebensweise und spiegeln diese wider. Doch obwohl technologische Innovationen weitreichenden sozialen Wandel auszulösen vermögen, steht der Zusammenhang von Technik, Herrschaft und sozialer Ungleichheit jenseits offensichtlicher militärtechnischer Kontexte selten im Fokus der öffentlichen Reflexion. Die Entwicklung und Nutzung von KI birgt das Potenzial, gesellschaftliche, praktische und diskursive Macht- und Ungleichheitsverhältnisse auf vielfältige Weise zu verändern, aber auch zu reproduzieren oder zu bestärken.

*Menschenbilder, Technikbilder und Weltbilder:* Die Entwicklung und Nutzung von KI geht einher mit und ist eingebettet in spezifische Bilder vom Menschen und der Welt. Diese Bilder sind mehr als reine Vorstellungen davon, was KI in der Welt bewirken soll und welche Rolle der Mensch im Verhältnis zu KI spielt. Sie geben vielmehr Orientierungen vor, die reflektiert oder unreflektiert, manifest oder latent sein können und an denen sich konkretes Handeln ausrichtet. Den Imaginationen der KI-Entwicklung und -Nutzung wohnt damit eine normative Kraft inne, die soziale Prozesse der Entwicklung, Nutzung und Einbettung von KI zu prägen vermag und gesellschaftliche Strukturen verändern kann.

*Grenzen von Künstlicher Intelligenz:* Wie jede andere Technik auch verschiebt und schafft KI Grenzen, indem sie menschliche Fähigkeiten erweitert und in bestimmten Bezügen auch übertrifft. Nicht übersehen werden darf dabei jedoch, dass in diesem Zuge wieder neue Grenzen entstehen – und zwar nicht nur der technischen Machbarkeit, sondern auch der Einbettung und Umsetzbarkeit in gesellschaftlichen und sozialen Zusammenhängen. Diese Zusammenhänge sind komplex und eigendynamisch, von Unsicherheiten und Unwägbarkeiten geprägt und dementsprechend kontingent. Es geht somit nicht so sehr darum, was KI aus technischer Perspektive noch nicht kann, sondern welche (auch dauerhaften) Grenzen sich mit Blick sowohl auf die materielle Beschaffenheit bzw. Eigenlogik der Technik selbst wie auf die Beschaffenheit der sozialen Strukturen, in denen sie entwickelt, eingesetzt und genutzt wird, zeigen.

*Soziale Entwicklungs- und Nutzungskontexte:* Als Technologie ist KI menschengemacht und von ihrer Entwicklung bis hin zu ihrer Nutzung in soziale Kontexte und Praktiken eingebettet. Zugleich werden diese Kontexte und Praktiken in der Auseinandersetzung mit KI verändert – als Technologie wirkt KI zurück auf die sozialen Prozesse und Strukturen, innerhalb derer sie entwickelt und genutzt wird. Dies geschieht auf spezifische Weise, indem KI Daten erzeugt, diese Daten menschlichem Wissen und Handeln zur Verfügung stellt und es auf diese Weise verändert. In diesem Sinne unterscheidet sich KI nicht grundlegend

von anderen Technologien, ihre Folgen sind jedoch (auch mit Blick auf den Grad ihrer Auswirkung) andere.

*Gesellschaftliche Risiken und Gefahren:* Gesellschaften sind mit ihren komplexen Dynamiken, Strukturen, Institutionen und Akteur\*innen resiliente und vulnerable Gebilde zugleich. KI entsteht und wirkt innerhalb gesellschaftlicher und sozialer Zusammenhänge und bietet auf vielfältige Weise sinnvolle und produktive Anwendungs- und Einsatzmöglichkeiten. Zugleich entstehen mit ihrer Entwicklung und Nutzung aber auch Risiken und Gefahren, die sich aus ihren technologischen Eigenheiten ergeben. Einige davon werden bereits unter den Stichworten Diskriminierung durch KI und Datenschutz diskutiert, jedoch gehen die Effekte von KI auf gesellschaftliche Strukturen und Prozesse weit über diese Themen hinaus.

*Gesellschaftliche Gestaltungserfordernisse und -möglichkeiten von KI:* Wie alle technischen Errungenschaften haben KI-Anwendungen bestimmte Gebrauchseigenschaften und in ihnen angelegte Verwirklichungstendenzen der intendierten Folgen und Wirkungen. Blickt man näher auf die Interaktion mit und Aneignung von KI, dann zeigen sich Aspekte, die wichtig sind für ihre Entwicklung, Implementierung, Nutzung und Analyse. Zugleich kann KI als eine Innovation gesehen werden, die stetig verändert und weiter entfaltet wird und mit deren Hilfe sich neue Wege, Ziele, Funktionen und Inhalte erschließen lassen. Mit den Freiheiten in diesem Geschehen entsteht gleichzeitig eine Verantwortung, die auch auf gesellschaftlicher Ebene wahrgenommen und in Ideen zu einer gesellschaftlich nachhaltigen Technikgestaltung umgesetzt werden muss.

*Technikreflexion:* Um ein umfassendes Bild von KI als sozialer Technologie zu zeichnen, braucht es eine tiefgehende Reflexion, die ihre gesellschaftlichen und sozialen Zusammenhänge, Verhältnisse, Folgen, Risiken und Notwendigkeiten aus unterschiedlichen Blickwinkeln bedenkt und auslotet. Insbesondere geistes-, sozial- und kulturwissenschaftliche sowie philosophische Perspektiven kommen im KI-Diskurs bisher wenig zum Zuge, bieten aber wertvolle Erkenntnisse und eigenständiges Wissen.

Die in diesem Band zusammengeführten Texte rekurren jeweils auf eine oder mehrere dieser inhaltlichen Aspekte, erweitern die Diskussion um KI jeweils aber auch auf systematische Weise.

### 3 Die Beiträge des Sammelbands

Die Beiträge des Sammelbands spiegeln die Erkenntnisse, Perspektiven und Thesen wider, die im Projekt KIMeGe in laufenden Diskussionen entwickelt und geschärft wurden. Auch wenn sich die einzelnen Texte auf spezifische Aspekte des gesellschaftlichen Wirkens von Künstlicher Intelligenz konzentrieren, liegen doch immer wieder Themen quer zu den Beiträgen. Eine trennscharfe Clusterung ist daher nicht immer möglich, weshalb sich die folgende inhaltliche Sortierung als Angebot versteht, die Texte unter einem gewissen thematischen Blickwinkel zu lesen – jeder Text geht zugleich über die enge thematische Klammer hinaus und lässt sich als originärer Beitrag zu einer allgemeinen Diskussion von KI lesen.

Der Band beginnt praxisorientiert mit empirischen *Einblicken in die Nutzung Künstlicher Intelligenz*. *Sabine Pfeiffer* stellt in ihrem Beitrag „KI als Kollegin (KIK) – Repräsentative Beschäftigtenbefragung zu Künstlicher Intelligenz am Arbeitsplatz“ die Ergebnisse einer quantitativen webbasierten Primärstudie vor. Auf Basis von multivariaten Analysen identifiziert sie nicht nur vier Einstellungstypen gegenüber KI am Arbeitsplatz, sondern zeigt unter anderem auch, dass die Ansichten zu KI als Kollegin stärker vom Arbeitskontext geprägt sind als von allgemeinen Vorstellungen zu Technik und KI. Gleichwohl ist KI nur bei wenigen Beschäftigten spürbar angekommen. *Regina Wittal* und *Carolyn Hettinger* gehen in ihrem Forschungsbericht „Veränderungsdynamiken durch Technisierung von Arbeit am Beispiel von da Vinci als robotische Chirurgie-Assistenz – Erkenntnisse aus der Empirie“ der Frage nach, welche Folgen mit dem Einsatz von Robotern in der Chirurgie verbunden sind, und zeichnen mithilfe von qualitativen Interviews und ethnographischen Methoden ein dichtes Bild der Veränderungen der Praktiken im Operationssaal und darüber hinaus.

Im Themenfeld *Künstliche Intelligenz als Bild und Mythos* versammeln sich drei Beiträge, die gewissermaßen hinter die Vorstellungen blicken, die wir uns von Künstlicher Intelligenz machen. *Bruno Gransche* und *Arne Manzeschke* eröffnen die Abteilung mit ihrem Beitrag „Das bewegliche Herr der Künstlichen Intelligenz: Ein Technomythos als Summe menschlicher Relationen“ und zeigen darin, dass und wie implizite Bedeutungen, die interessensgeleitet oder unbewusst mit Künstlicher Intelligenz verbunden sind, ausgelegt werden müssen. Im Anschluss daran führen *Arne Manzeschke* und *Bruno Gransche* in ihrem Text „Bilder machen Menschen: Zur Bildermacht der Künstlichen Intelligenz“ komplementär zu ihrem ersten Beitrag aus, wie Bilder gesellschaftliche KI-Diskurse wirkmächtig bestimmen. Der Fokus liegt dabei auf anthropologischen

Erwägungen des Bildermachens, auf dem Verhältnis von Mensch- und Technikbildern zu Weltmodellen und Weltbild sowie auf visuellen Bildern. *Irmhild Rogalla* schließt das Themenfeld mit ihrem Beitrag „Maschinen lernen nicht! ‚Machine learning‘-Algorithmen entzaubert“ ab, indem sie entlang der Funktionsweise von Algorithmen zeigt, dass Analogien zwischen menschlichem Lernen und maschinellem Lernen nicht nur irreführend, sondern unzulässig sind.

Das Themenfeld *Normierung und Regulierung von Künstlicher Intelligenz* befasst sich mit den Herausforderungen, die sich durch den Einsatz von KI für rechtliche und regulierende Fragestellungen ergeben. *Anna Beckers* und *Gunther Teubner* werfen in ihrem Beitrag „Die digitale Verantwortungslücke: Vorschläge zur Haftung für algorithmisches Fehlverhalten“ die Frage auf, welche Konsequenzen sich für die rechtliche Haftung ergeben, wenn autonome Algorithmen Fehlentscheidungen treffen und entwickeln nach unterschiedlichen soziologischen Institutionen differenzierte Haftungskonzepte. *Karsten Weber* und *Nadine Kleine* zeichnen in ihrem Beitrag „Normierung, Regulierung, Governance: Wie, von wem und mit welchen Mitteln kann der Einsatz Künstlicher Intelligenz gesellschaftlich gestaltet werden?“ die – nach wie vor fluide – nationale und internationale Normierungs- und Regulierungsdebatte in Hinblick auf Künstliche Intelligenz nach und hinterfragen diese kritisch.

Im Anschluss daran werden *gesellschaftliche Risiken und soziale Dynamiken Künstlicher Intelligenz* in den Vordergrund gerückt. *Hartmut Hirsch-Kreinsen* zeigt in seinem Beitrag „Soziale Dynamik der Künstlichen Intelligenz“, dass sich mit KI ein spezifischer Innovationsmodus durchsetzt, der mit den tradierten und industrieorientierten Regelungen und Praktiken des Innovationssystems kaum mehr kompatibel ist und die Innovationspolitik vor neue Herausforderungen stellt. *Reinhard Kreissl* und *Roger von Laufenberg* geben in ihrem Beitrag „Risiken und Gefahren der ‚Künstlichen‘ ‚Intelligenz‘“ einen Überblick über die Risikodebatte von KI und analysieren ihre unterschiedlichen Stränge anhand von vier konkreten Einsatzszenarien. *Heike Raab* nähert sich in ihrem Beitrag „Toys are us: KI, Diversität und soziale Ungleichheit. Oder: KI für Alle?“ KI aus einer Ungleichheitstheoretischen Perspektive, die vor allem Bezug auf die Disability Studies nimmt und die These entfaltet, dass sich diskriminierende Muster und Machtverhältnisse durch KI zwar modernisieren, aber nicht auflösen. *Stefan Selke* geht in seinem Beitrag „Zukunftseuphorie als Trost: Verheißungserzählungen über Künstliche Intelligenz im Kontext gesellschaftlicher Erschöpfungsdiagnosen“ empirisch der Frage nach, inwieweit Narrative über KI als gleichermaßen prognostische wie entlastende Erzählungen begriffen werden können und welche epistemologische Position und gesellschaftliche Funktion ihnen zukommt.

Das Themenfeld *Künstliche Intelligenz im Kontext von Macht, Herrschaft und Demokratie* versammelt drei Beiträge, die dieses Feld aus unterschiedlichen Perspektiven analysieren. *Jeanette Hofmann* und *Clara Iglesias Keller* entwickeln in ihrem englischsprachigen Beitrag „Machine learning, political participation and the transformations of democratic self-determination“ eine Perspektive, die auf die Formbarkeit der Wechselwirkungen von KI und Demokratie abhebt und für ein dynamisches Verständnis des Verhältnisses von KI und Demokratie wirbt. *Karin Hutflötz* analysiert in ihrem Beitrag „Hegemoniale Machtstruktur? Eine Kartierung der Akteure im aktuellen KI-Diskurs“ die maßgeblichen Akteure im aktuellen Diskurs um KI und argumentiert, dass dieser Diskurs von einseitig hegemonialen Machtverhältnissen geprägt ist. *Peter Imbusch* und *Joris Steg* befassen sich in ihrem Beitrag „Künstliche Intelligenz und gesellschaftlicher Wandel – eine Herausforderung für demokratische Macht- und Herrschaftsverhältnisse“ mit dem Einfluss von KI auf gesellschaftliche Machtverhältnisse und Herrschaftsstrukturen und fragen danach, wie sich die Macht- und Herrschaftsverhältnisse in Wirtschaft, Politik und Gesellschaft durch den Einsatz von KI verändern und welche Folgen dies für die Demokratie hat.

Das letzte Themenfeld des Bandes diskutiert und entwickelt *theoretische und methodische Zugänge zu Künstlicher Intelligenz*, wobei KI als Herausforderung für die Soziologie und die Sozialwissenschaften selbst begriffen wird. *Michael Heinlein* entwickelt in seinem Beitrag „Künstliche Intelligenz als kontingenzerzeugende Technik“ eine praxistheoretische Perspektive, um zu zeigen, dass das technische Wirken von KI nicht vollständig vorhersehbar ist und Folgen zeitigt, die auf die Emergenz neuer Wirklichkeits- und Möglichkeitsräume des Wissens und Handelns verweisen. *Norbert Huchler* sondiert in seinem Beitrag „Selektivitäten (subsymbolischer) Künstlicher Intelligenz“ wesentliche soziale strukturierende Wirkungen von KI, die von der zielgerichteten Entwicklung über die technischen Funktionsweisen bis hin zu latenten sozialen und gesellschaftlichen Anpassungen reichen, und leitet hieraus Folgen für das Verhältnis KI-Mensch-Gesellschaft ab. *Valentin Rauer* geht in seinem Beitrag „Von Interaktion zur Transformaktion: Die Folgen von Künstlicher Intelligenz für Theorien sozialen Handelns“ der Frage nach, was es eigentlich heißt, mit KI zu handeln und entwickelt eine prozessorientierte Perspektive, die die wechselseitigen transformativen Dynamiken von menschlichem Handeln und handelnder KI in den Vordergrund rückt. Das Autorenkollektiv *Peter Kahlert*, *Maryam Tatari*, *Suzette Kahlert*, *Silvan Pollozek*, *Johan Buchholz*, *Benedict Lang* und *Jan-Hendrik Passoth* geht mit seinem Beitrag „Künstliche Intelligenz: Eine Methode für alles? Sozialwissenschaftliche Methodologie der KI-Forschung, ihre Herausforderungen und Möglichkeiten“ der Frage nach, wie die sozialwissenschaftliche Forschung und



ihre Methoden durch die fortschreitende Entwicklung und Implementierung von KI-Systemen und -Anwendungen herausgefordert werden und was die Methoden der KI-Forschung für die Sozialwissenschaften beitragen können.

Das diesem Band zugrunde liegende Vorhaben wurde mit Mitteln des Bundesministeriums für Bildung und Forschung unter dem Förderkennzeichen 01IS20095 gefördert. Die Verantwortung für den Inhalt dieser Veröffentlichung liegt bei den Autorinnen und Autoren. Wir bedanken uns an dieser Stelle für die Förderung sowie insbesondere bei Frau Ute Bernhardt, Leiterin des Referats Künstliche Intelligenz im BMBF, und Frau Dr. Ulrike Völlinger vom DLR Projektträger.

---

## Literatur

- Bischof, Andreas. 2017. *Soziale Maschinen bauen. Epistemische Praktiken der Sozialrobotik*. transcript.
- Degele, Nina. 1994. *Der überforderte Computer. Zur Soziologie menschlicher und künstlicher Intelligenz*. Campus
- Dotzler, Bernhard J., und Berkan Karpat (Hrsg.). 2021. *Götzendämmerung – Kunst und Künstliche Intelligenz*. transcript.
- Engemann, Christoph, und Andreas Sudmann (Hrsg.). 2018. *Machine Learning – Medien, Infrastrukturen und Technologien der Künstlichen Intelligenz*. transcript.
- Hirsch-Kreinsen, Hartmut, und Anemari Karacic (Hrsg.). 2019. *Autonome Systeme und Arbeit: Perspektiven, Herausforderungen und Grenzen der Künstlichen Intelligenz in der Arbeitswelt*. transcript.
- Kaltheuner, Frederike (Hrsg.). 2021. *Fake AI*. Meatspace Press.
- Rammert, Werner (Hrsg.). 1995. *Soziologie und künstliche Intelligenz – Produkte und Probleme einer Hochtechnologie*. Campus.
- Rosengrün, Sebastian. 2021. *Künstliche Intelligenz zur Einführung*. Junius.
- Schmohl, Tobias, Alice Watanabe, und Kathrin Schelling (Hrsg.). 2022. *Künstliche Intelligenz in der Hochschulbildung: Chancen und Grenzen des KI-gestützten Lernens und Lehrens*. transcript.
- Seyfert, Robert, und Jonathan Roberge (Hrsg.). 2017. *Algorithmenkulturen: Über die rechnerische Konstruktion der Wirklichkeit*. transcript.
- Zuboff, Shoshana. 2018. *Das Zeitalter des Überwachungskapitalismus*. Campus.

---

# **Einblicke in die Nutzung Künstlicher Intelligenz**



# KI als Kollegin (KIK) – Repräsentative Beschäftigtenbefragung zu Künstlicher Intelligenz am Arbeitsplatz

Sabine Pfeiffer

## Zusammenfassung

Über die Folgen von Künstlicher Intelligenz (KI) für Arbeit und Gesellschaft gibt es einen breiten medialen und interdisziplinären wissenschaftlichen Diskurs. Was aber Beschäftigte über KI in ihrem direkten Arbeitsumfeld denken, ist in der Forschung bislang nur rudimentär erforscht. Der Beitrag stellt dazu Ergebnisse einer quantitativen webbasierten Primärstudie mit  $N = 2.018$  Erwerbstätigen im Alter zwischen 20 und 65 Jahren vor. Das Sample wurde aktiv quotengemanagt und ist nach Geschlecht, Alter, Branchen und Qualifikation repräsentativ. Ein Teil der Befragung ging mithilfe eines selbst entwickelten „Denkzeug“ in Anlehnung an die Q-Methode einen eigenen interaktiv-ipsativen Befragungsweg, der komplexere Einstellungsbilder als üblich zu erfassen erlaubt. Dabei konnten auf Basis von multivariaten Analysen vier Einstellungstypen identifiziert: Die Störungsjonglierenden, die Arbeitsfokussierten, die Verantwortungstragenden und die Entlastungsuchenden. Für diese vier Typen zeigt sich trotz aller Unterschiede: die Ansichten zu KI als Kollegin sind stärker vom Arbeitskontext geprägt als von allgemeinen Vorstellungen zu Technik und KI. Noch allerdings ist KI bei den wenigsten Beschäftigten spürbar angekommen, sie fühlen sich recht gut zum Thema informiert – nicht aber durch ihre Arbeitgeber. Ängste und Sorgen sind geringer ausgeprägt als oft unterstellt und beziehen sich mehr auf andere als auf die eigene Arbeit. Während man den zuständigen Akteuren im Unternehmen etwas mehr als der Politik vertraut, ist der Anspruch an mehr Partizipation und

---

S. Pfeiffer (✉)

Institut für Soziologie, FAU Erlangen-Nürnberg, Nürnberg, Deutschland

E-Mail: [sabine.pfeiffer@fau.de](mailto:sabine.pfeiffer@fau.de)

© Der/die Autor(en), exklusiv lizenziert an Springer Fachmedien Wiesbaden GmbH, ein Teil von Springer Nature 2024

M. Heinlein und N. Huchler (Hrsg.), *Künstliche Intelligenz, Mensch und Gesellschaft*, [https://doi.org/10.1007/978-3-658-43521-9\\_2](https://doi.org/10.1007/978-3-658-43521-9_2)

Mitbestimmung beim Thema klar ausgeprägt. Die Erhebung ist eine Momentaufnahme und macht weitergehende Erhebungen notwendig. There is a broad media and interdisciplinary scientific discourse on the consequences of artificial intelligence (AI) for work and society. However, what employees think about AI in their direct work environment has so far only been rudimentarily explored in research. This article presents the results of a quantitative web-based primary study with  $N = 2,018$  employees between the ages of 20 and 65. The sample was actively quota-managed and is representative by gender, age, industries, and qualifications. One part of the survey used a self-developed „Denkzeug“ based on the Q-method, an interactive-ipsative survey method that allows to capture more complex attitudes than usual. Based on multivariate analyses, four types of attitudes were identified: The disruptive juggler, the work-focused, the responsibility-bearing, and the relief-seeking. For these four types, despite all the differences, it appears that views on AI as a colleague are more strongly influenced by the work context than by general ideas about technology and AI. However, AI has not yet made a tangible impact on very few employees, who feel quite well informed on the topic – but not by their employers. Fears and concerns are less pronounced than often assumed and relate more to others than to their own work. While the responsible actors in the company are trusted somewhat more than politicians, the demand for more participation and co-determination in the topic is clearly pronounced. The survey is a snapshot and requires further research.

---

### Schlüsselwörter

Verbreitung der generellen Einstellung • Bestehendes Wissen und gefühlte Informiertheit • Sorgen und Befürchtungen • Gewünschte Rolle • Vier Einstellungstypen • Akteure und Zuständigkeiten • Partizipation und Mitbestimmung • Gestaltung • Sorgen • KI als Kollegin • Komplexe Einstellungsmuster • KI am Arbeitsplatz

---

## 1 KI und Arbeit – zur Einleitung

Der Bezug auf den Einsatz von Künstlicher Intelligenz oder Maschinellen Lernen (nachfolgend KI/ML) am Arbeitsplatz wird in dieser Studie in den Mittelpunkt gestellt, weil sich in anderen Studien nur unerhebliche Zusammenhänge zwischen den Einstellungen zu KI und üblichen sozio-demografischen Daten wie Geschlecht, Alter, Einkommen und Qualifikation zeigten (vgl. Liang und Lee

2017). Möglicher- und naheliegenderweise sind dafür tätigkeitsbezogene Variablen relevanter – zumindest sollte dies der Fall sein, wenn es um den Einsatz von KI/ML am eigenen Arbeitsplatz geht. Wissenschaftlich dominieren in diesem Feld bislang jedoch weitgehend Prognosen zu möglichen Arbeitsmarkteffekten durch KI.

Die erstmalige Frage danach, ob Maschinen denken können, wird Alan Turing zugeschrieben, der sie 1947 auf einem Symposium formulierte; als Geburtsstunde der KI aber wird meist eine Konferenz gesehen, die 1956 am amerikanischen Dartmouth College stattfand (vgl. Konrad 1998; Nilsson 2010, S. 52–56). Erst seit den späteren 1980er Jahren kommt es zu immer neuen Sprüngen in Prozessorgeschwindigkeiten, mit denen sich die neueren Ansätze der KI wie etwa Unsupervised oder Reinforcement ML (vgl. ebd., S. 413–421) oder der Bayesschen Netzwerke (vgl. ebd., S. 381–397) realisieren lassen, die überwiegend gemeint sind, wenn aktuell von KI gesprochen wird. Von Anfang an wird KI wissenschaftlich wie in der öffentlichen Debatte das Potenzial zu weitgehender Automatisierung – vor allem nicht-stofflicher – Arbeitsprozesse zugesprochen.

Dabei ist die Betrachtung von KI/ML nicht losgelöst zu sehen von der allgemeinen Debatte um die digitale Transformation, die – auch wenn sie in manchen Sektoren lange nur graduelle Veränderungen bewirkt – ganze Wertschöpfungskontexte radikal verändern kann und teils bereits verändert hat (Schrape 2021) sowie zunehmend nicht mehr rein auf Rationalisierung im engeren Sinne, sondern auf neue betriebliche Rationalisierungsstrategien der Vermarktlichung (Menz et al. 2019), auf neue digitale Steuerungsformen von Marktbeziehungen und Wettbewerbsprozessen (Buss et al. 2022) zielt und sich damit zunehmend auf die Distributivseite der Produktivkräfte richtet (Pfeiffer 2021b). So erleichtert die Digitalisierung seit langem die Verlagerung von geistiger Arbeit an andere Standorte im globalen Kontext (Boes und Kämpf 2011) und kann und wird auch als Rationalisierungstechnologie für indirekte Tätigkeiten eingesetzt (Horton 2015; Seibold und Stieler 2016; Balu et al. 2020).

Seit der Begriff „Industrie 4.0“ Karriere macht(e), hat der aktuelle Digitalisierungsdiskurs in der internationalen Forschung und anknüpfend an frühere Arbeiten (Autor et al. 2003) zunächst für neue Impulse zur Messung und Prognose von Beschäftigungseffekten durch zunehmend autonom agierende Technik – insbesondere Robotik und Algorithmen bzw. Künstliche Intelligenz – geführt. Dabei kommt es einerseits zu drastischen Prognosen eines dramatischen Arbeitsplatzabbaus (Brynjolfsson und McAfee 2014; Frey und Osborne 2017). Der vielfach kritisierte (Fernández-Macías und Hurley 2014; Pfeiffer 2018) methodische Ansatz von Frey und Osborne, der für die USA einen Abbau von bis zu

47 % aller Arbeitsplätze prognostiziert, kommt angewandt auf Beschäftigungsdaten in Deutschland zu wesentlich geringeren Zahlen (12 % nach Bonin et al. 2015).

Andere Studien zeigen differenziert auf, warum die eng auf Technologien orientierten Studien zu kurz greifen (Krzywdzinski 2020). Zudem kommen andere methodische Vorgangsweisen ebenfalls für Deutschland zu geringeren Substituierbarkeitspotenzialen (Dengler und Matthes 2015). Nur wenige Ansätze berücksichtigen in der Gesamtbilanz auch neu entstehende Tätigkeiten (Acemoglu und Restrepo 2019) oder kommen deshalb sogar zu leicht positiven Effekten (Arntz et al. 2018). In der sozialwissenschaftlichen, technik- und arbeitssoziologischen Debatte dominiert einerseits die Sichtweise einer verteilten Handlungsträgerschaft zwischen Mensch und Technik auf der Mikroebene (Rammert 2016; Weyer et al. 2018) und andererseits die Betonung der Problematik von Diskriminierung z. B. bei algorithmengestützter Personalauswahl (vgl. Kadiresan et al. 2022).

Dabei sind gerade in der Verwendung von KI/ML im Arbeitskontext sehr viel mehr Themen für einen sinnvollen wie robusten Einsatz autonomer Technologien relevant. So verweist etwa Huchler (2019) auf drei „unhintergehbare“, aber gleichwohl in Gestaltungsprozessen oft „übergangene“ Grenzen der Adaptivität (teil-)autonomer Systeme: Immer wieder neu entstehende, nicht-digitalisierte Tätigkeiten sowie Probleme der Komplexität der physischen und sozialen Umwelt und schließlich Probleme der Nicht-Formalisierbarkeit (ebd., S. 150–159). Zudem negiert der in Wissenschaft wie Praxis weitgehend geteilte Konsens des „Keep the Human in the Loop“ zu sehr die Notwendigkeit, einzelne KI/ML-Anwendungen und deren Entscheidungen einzubinden in die diese umgebenden organisationalen Kontextbedingungen – „Keep the Organization in the Loop“ ist damit oft genug noch keine systematisch berücksichtigte Gestaltungsebene, wenn es zum Einsatz von KI/ML am Arbeitsplatz kommt (Herrmann und Pfeiffer 2022).

An anderer Stelle konnte schon auf Folgen eines „unbewusst destruktive[n] Einsatz“ (Pfeiffer 2021b, S. 284) von KI/ML hingewiesen werden, der gerade im Arbeitskontext aktuell oft zu beobachten ist; vor allem in Form einer uninformierten Anwendung von KI/ML. Ein unbewusst destruktiver Einsatz von KI/ML lässt sich beobachten, wenn etwa die Schlichtheit des statistischen Modells völlig über- und der gewählte Anwendungskontext in seiner Komplexität gleichzeitig unterschätzt wird; Kausalitäten unterstellt werden, wo es sich nur um Datenrauschen handelt oder Wissen um Skalenniveaus der Daten und um deren Notwendigkeit zur Passung sowohl im Hinblick auf den Algorithmus wie den sachlich-inhaltlichen Gegebenheiten des Anwendungskontexts fehlt. Meist werden auch algorithmenspezifische Dynamiken unterschätzt, die erst über längere

Zeit sichtbar werden (z. B., wenn zunehmend in die Mitte tendierende Algorithmen immer mehr Daten zu Ausreißern erklären). Auch, wie lange weitergelernt werden soll bzw. wann ein Lern-„freeze“ sinnvoll ist und nach welchen Kriterien dies von wem jeweils entschieden wird – all das und mehr (Pfeiffer 2021b, S. 283–285) sind Fragen, die zu den mittlerweile breit diskutierten Problemen von Bias und Diskriminierung noch hin zukommen – und möglicherweise sehr viel mehr Aufmerksamkeit bedürfen als ihnen derzeit zu teil kommt. Egal ob eine Ärztin eine KI-gestützte Therapie-Empfehlung gibt oder ob eine KI auf der Basis bildgebender Verfahren in der Qualitätssicherung ein Getriebegussteil als in Ordnung klassifiziert und dieses in ein Fahrzeug verbaut wird: Beim Einsatz von KI am Arbeitsplatz können solche unterschätzten Herausforderungen und ein falsch positives oder falsch negatives Ergebnis der KI zu dramatischen Folgen führen. Das Umgehen mit diesen Herausforderungen im Rahmen betrieblicher Gestaltungsprozesse und/oder ihren Folgen ist selbst Arbeit und muss von Beschäftigten geleistet werden. Wie sie über KI an ihrem Arbeitsplatz denken und ob sie sich KI als „Kollegin“ wünschen oder nicht, ist Gegenstand dieser Studie.

Bisherige Studien zur Einstellung von Beschäftigten sind noch relativ dünn gesät oder beziehen sich im engeren Sinne auf einen einzigen Anwendungskontext (z. B. Medizin). Relevante Studien, die zumindest zum Teil den Bezug zu Arbeit herstellen und deren Fragebatterien deswegen in dieser Studie teils mit aufgenommen wurden, sind etwa:

- Der Meinungsmonitor Künstliche Intelligenz (Dosenovic et al. 2020) ist angetreten, die „Bevölkerungsmeinung“ zu KI zu erheben und laufend zu monitoren, befragt wird ein repräsentatives Online-Panel von 1000 Personen in 16 Wellen im vierzehntägigen Abstand. Die Befragung enthält neben sich wiederholenden Themenblocks (z. B. persönliches Interesse an KI, Meinungen zu KI sowie Verhaltensintentionen im Zusammenhang mit KI) auch Fragen zu „psychologische[n] Wahrnehmungsphänomene[n]“, die „als erklärende Konstrukte“ verstanden werden wie etwa Chancen- und Bedrohungswahrnehmungen, Vertrauen, Legitimität (ebd., S. 9). Erwerbstätige sind in der Stichprobe nur mit  $n = 602$  vertreten und nur wenige Fragen beziehen sich auf den Einsatz von KI am Arbeitsplatz.
- Die Studie „Fear of autonomous robots and artificial intelligence (FARAI)“ (Liang und Lee 2017) basiert auf einem in USA erhobenen repräsentativen Samples ( $N = 1.41$ ). Dabei zeigte eine Latente Klassenanalyse bei 26 % der Befragten eine erhöhte Angst mit nur geringen Zusammenhängen zu soziodemografischen Variablen, aber klareren Zusammenhängen zum Konsum von Science-Fiction und zu anderen sozialen Ängsten (etwa vor Arbeitslosigkeit,

Obdachlosigkeit oder Einsamkeit). Hier spielt Arbeit kaum eine Rolle, weil aber deutlich wird, dass die Einstellungen sich oft aus ganz anderen Quellen (eben z. B. Science-Fiction-Konsum) speisen, wird aus dieser Studie die Skala mit leichten Veränderungen aufgenommen.

- Bei einer Umfrage der D21 Initiative erhebt den Wissensstand und allgemeine Einstellungen zu KI (Initiative D21 2020, S. 30). Auch in dieser Studie machen Erwerbstätige nur einen Teil des Samples aus und viele Fragen beziehen sich zu Digitalisierung im Allgemeinen (nicht nur KI/ML) und sehr stark auf die Nutzung im Privaten statt am Arbeitsplatz. Da das Gefühl der Informiertheit zum Thema auch am Arbeitsplatz eine große Rolle spielt und die digitale Nutzung bei der Einschätzung einen Unterschied machen kann, wird an diese Erhebung teils angeknüpft.

Basis der hier vorgestellten Ergebnisse ist eine repräsentative<sup>1</sup> Online-Befragung mit  $N = 2.18$ , die im Februar 2022 durchgeführt wurde. Angesprochen wurden weit über 4.00 Personen, deren Teilnahme über aktives Quoten-Management<sup>2</sup> zu dem gewünschten Sample führte. Die Erhebung erfolgte über ein professionelles Befragungsinstitut.<sup>3</sup> Für die Erhebung wurde das Online-Tool SoSci-Survey verwendet (Leiner 2019), das DSGVO-konform auf einem lehrstuhleigenen Server des Rechenzentrums der FAU Erlangen-Nürnberg installiert ist. Bei Erhebung und Analyse wurden nicht nur Datenschutzfragen sondern auch forschungsethische Kriterien angelegt (Pfeiffer 2021a).

---

<sup>1</sup> Im mathematischen Sinne ist Repräsentativität bei einem Online-Panel im engen Sinne nicht erreichbar, da dies eine echte Zufallsstichprobe erfordern würde, an der jede Person der Grundgesamtheit die gleiche Chance hat, ausgewählt zu werden – was auch für viele andere Befragungswege nicht in Gänze garantiert werden kann (vgl. dazu kritisch Schnell 2018).

<sup>2</sup> Teilnehmen konnten nur Erwerbstätige mit einer Wochenarbeitszeit von mindestens 20 h im Alter von 20 bis 65. Zusätzlich erfolgte eine ungekreuzte Quotierung nach Geschlecht und Berufsbildung (niedrig/mittel/hoch) sowie nach Branchen (10-Kodierung Wirtschaftszweige WZ2008). Als Zielgrößen dienten dabei die mit Mikrozensus gewichteten Daten aus der BIBB/BAuA-Erwerbstätigenbefragung 2018. Die Stichprobe ist nach Geschlecht, Alter, Branchen und Qualifikation repräsentativ. Die Befragten sind im Durchschnitt 45,2 Jahre alt (SD 11,763), 46 % sind weiblich, 54 % männlich und 0,2 % divers. Mit 52 % hat der Großteil eine Berufsausbildung als höchstem Abschluss und 15 % haben höhere berufliche Fortbildungen absolviert. 29 % haben eine rein akademische Ausbildung und 5 % gar keine formale Qualifikation (dieser letzte Wert erreicht als einziger nicht ganz die Zielgröße von 8,4 %).

<sup>3</sup> Dieses entwickelt Sampling und Befragungsmethoden in Zusammenarbeit mit der GESIS, beachtet die Vorgaben der Interactive Marketing Research Organization (IMRO 2015) und den Kodex der International Chamber of Commerce und ESOMAR (ICC/ESOMAR 2016).



Einstellungen zu KI sind komplexer und widersprüchlicher als es klassische Erhebungsmethoden abbilden können. Daher wurde in den Fragebogen als besondere Erhebungsmethode des am Lehrstuhl entwickelte „Denkzeug“<sup>4</sup> integriert<sup>5</sup>, eine Abwandlung der Q-Methode (Brown 1980). Dieses Erhebungs- und Analyseverfahren zur Untersuchung menschlicher Subjektivität erlaubt die „quantification of unique subjective perspectives on a topic, in which the subject has more control of data representation than usual“ (Johnson und Waishwell 2014, S. 1125).

Dabei wurden den Teilnehmenden 16 Aussagenkarten (Q-set; Müller und Kals 2004) zu KI am Arbeitsplatz präsentiert (vgl. Abb. 1), die dann von diesen in ein vorgegebenes Grid sortiert werden (Q-sort; vgl. Abb. 2), wobei an den äußeren Enden jeweils nur eine positivste bzw. negativste Karte und nach innen mehr Karten mit abnehmender Wertung sortiert werden. Die Einstiegsfrage dazu lautete:

„Künstliche Intelligenz (KI) ist in aller Munde. Vieles ist noch Zukunftsmusik. Manches aber kommt schon in unserem Leben an. Stellen Sie sich bitte nun einmal vor, KI kommt an Ihrem Arbeitsplatz an und spielt bei Ihrer alltäglichen Arbeit eine Rolle. Z.B. könnte die KI Ihnen Vorschläge für Entscheidungen machen, die bisher Sie alleine getroffen haben. Wie wünschen Sie sich eine solche Zusammenarbeit? Wie sollte es sein? Und was sollte nicht oder auf keinen Fall passieren?“

Die Q-Methode eignet sich besonders für die Erfassung von komplexen Meinungen oder Einstellungen, weil die subjektive Struktur der einzelnen Befragten

---

<sup>4</sup> Denkzeug ist ein Portmanteau. Die Analogie von Werkzeug und Denkzeug hat eine lange Tradition in der Reflexion der Bedeutung des Computers im Kontext der menschlichen Entwicklungsgeschichte, ähnliche Begriffe wie etwa die Denkmaschine finden sich schon im 17. Jahrhundert (vgl. Uhlig 2014). Haefner et al. plädieren dafür, den Computer als „persönliches Denk- und Lernzeug“ zu sehen (1987, S. 12), Sybille Krämer (1997) erweitert den Reigen vom Werkzeug über das Denkzeug zum Spielzeug und betont den körperlichen und nie rein kognitiven Zugang des Menschen zur Welt. Wir wählen den Begriff „Denkzeug“ für unsere Abwandlung der Q-Methode zur Erhebung des digitalen Wandels von Arbeit sowohl wegen des thematischen Gegenstands unserer Befragung als auch wegen der fast haptischen Nutzung beim Legen der Aussagen.

<sup>5</sup> SoSci hat zudem in Auftrag und in enger Zusammenarbeit mit dem Lehrstuhl die ältere HTMLQ-Applikation (Banasick 2022) technisch aktualisiert und durchgängig responsiv auf der Basis von Angular neu gestaltet sowie die vom Lehrstuhl entwickelten „Denkzeug“-Spezifika (z. B. veränderte Grid-Geometrie, anders als bei Q üblich wenige Items) integriert. Die methodische und konzeptionelle Entwicklung des Denkzeug erfolgte im Rahmen des DFG-Schwerpunktprogramms 2267 „Digitalisierung der Arbeitswelten“, gefördert durch die Deutsche Forschungsgemeinschaft (DFG) – Projektnummer 442171541 und kam in dieser Befragung in der Form zum ersten Mal in Einsatz.

Entscheidungshoheit	beim Mensch EH-M	Die letzte Entscheidung liegt immer bei mir, nie bei der KI. Das wäre ja noch schöner!	← Die meisten Entscheidungen nimmt mir die KI ab. Sie fragt mich nur dann, wenn die Situation unklar oder kritisch ist. →	bei der KI EH-T
Entscheidungskriterien	nachvollziehbar EK-N	Warum die KI genau zu einer Entscheidung kommt, das muss nachvollziehbar bleiben. Das ist bei meiner Arbeit ja auch nicht anders.	← Wie und warum die KI so oder so entscheidet, das muss ich gar nicht wissen. Dann bin ich auch nicht dafür verantwortlich. →	Blackbox EK-B
Lernfähigkeit	des Menschen LF-M	Wenn doch mal was schief läuft, muss ja doch ich ran. Deshalb bleibt wichtig, dass ich immer wieder Neues lernen kann – trotz KI.	← In der Arbeit ändert sich ja ständig etwas, deswegen bleibt die KI in der Lage, immer weiter zu lernen. →	des Algorithmus LF-T
Entscheidungsprozess	beeinflussbar EP-B	Bei der Arbeit geht es schließlich um etwas. Daher kann ich die KI jederzeit stoppen oder ihre Entscheidungen wieder zurücknehmen.	← Damit sich das Ganze auch lohnt, arbeitet die KI möglichst viel bis zum Ende völlig automatisiert ab, ohne dass ich eingreifen muss. →	automatisiert EP-A
Aufgabenlast	beim Mensch AL-M	Weil die KI mir die Routinesachen abnimmt, kann ich mich darauf konzentrieren, wo es hakt, ganz dringend ist oder „brennt“.	← Meine Arbeit ist viel weniger anstrengend, weil mir die KI sehr viel abnimmt und so den Rücken freihält. →	bei der KI AL-T
Arbeitsanforderung	wird anspruchsvoller AF-1	Die langweiligen und lästigen Aufgaben übernimmt die KI. So habe ich Zeit für meine anspruchsvollen und wichtigen Aufgaben.	← Die KI nimmt mir die komplexen und unübersichtlichen Aufgaben ab. Dafür ist sie da. Damit bei mir nicht nur das Schwierige übrig bleibt. →	wird einfacher AF-0
Kontextwissen	relevant für Einführung KW-R	So eine KI ist auch nur Software. Damit da was Ordentliches bei rauskommt, bin ich bei der Einführung von Anfang an dabei.	← KI ist was für Experten. Da halte ich mich raus. → Die Technik soll sauber funktionieren, wenn sie bei mir am Arbeitsplatz eingesetzt wird.	nicht relevant für Einführung KW-I
Verantwortung	bei Mensch/Anwender*in VW-M	Bei KI ist das auch nicht anders als wenn ich andere Technik einsetze: am Ende stehe nur ich gerade für meine Arbeit und das ist auch gut so.	← Wenn meine KI Quatsch macht, dann bin nicht ich schuld. Da muss dann der Programmierer ran oder die Softwarefirma. →	bei KI/Hersteller VW-T

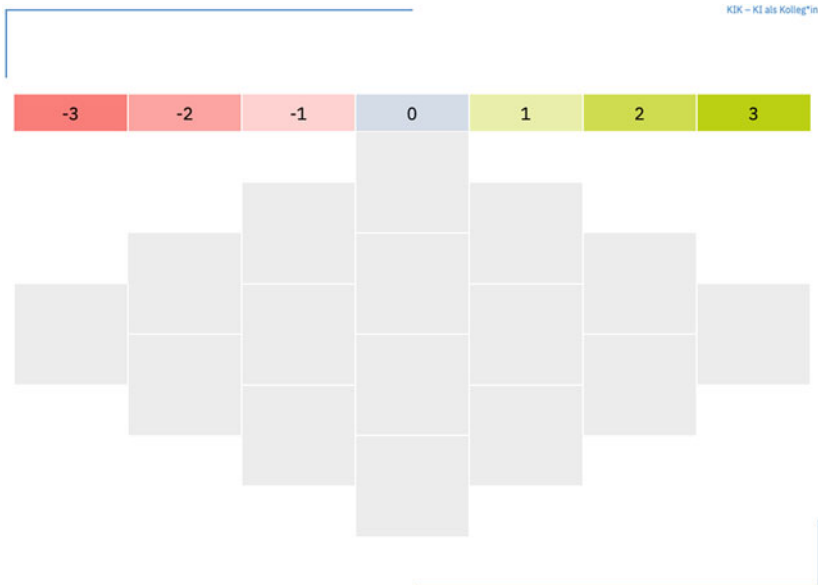
Abb. 1 Q-set – Aussagen zu KI des Denkzeugs. (Quelle: Eigene Darstellung)

ohne zu starke kategoriale Vorgaben abgebildet wird.<sup>6</sup> Damit werden die Items nicht wie in der Umfrageforschung üblich anhand einer Likert-Skala *normativ*, sondern *ipsativ* gemessen, d. h. alle Items werden *relativ* bewertet. Die verwendete interaktive Online-Umsetzung beinhaltet wegen der quasi-haptischen Legung der Karten Elemente der „aktiven Visualisierung“ (Pfeiffer et al. C) wie der Szenario-Vignetten (Aguinis und Bradley 2014).

Die vorliegende Studie ist eine der ersten, die in der Breite Einstellungen der Erwerbstätigen zum Einsatz von KI am Arbeitsplatz systematisch in den Mittelpunkt stellt. In dieser ersten Publikation werden Auszüge der vielfältigen

<sup>6</sup> Die oft deliberativ eingesetzte Methode verbessert die interne Validität und wird auch als Rationalisierung der Kodierung qualitativer Daten gesehen, die statistische Analyse wägt dabei widersprüchliche Werte ab und hilft somit dabei, polarisierende und konsensuale Ansichten besser zu verstehen (vgl. Khirfan und Peck 2021), sie liefert in diesem Sinne eine „beyond average information“ gegenüber Messverfahren mit Likert-Skalen (vgl. Brewer-Deluce und Sharma 2020).

Ergebnisse – abgesehen von der Typenbildung auf Basis des Denkzeug – weitgehend deskriptiv dargestellt. Weitere Publikationen mit tiefergehenden und stärker erklärenden Analysen werden folgen. Im Folgenden finden sich, nach ersten Zahlen zur Verbreitung und der generellen Einstellung gegenüber KI (→ Kapitel 2), Ergebnisse zum bestehenden Wissen und der gefühlten Informiertheit zum Thema (→ Kap. 3) und Zahlen dazu, ob und mit welchen Sorgen und Befürchtungen Beschäftigte auf den Einsatz von KI in der Arbeit blicken (→ Kap. 4). Schließlich zeigen Einstellungen zur gewünschten Rolle von KI am Arbeitsplatz (→ Kap. 5) und vier Einstellungstypen (→ Kap. 6), wie sehr die Einstellungen zu KI in der Arbeit von den Beschäftigten aus der Perspektive ihrer Arbeit betrachtet werden. Die Ergebnisse zur Frage der Akteure und Zuständigkeiten (→ Kap. 7) bei der Gestaltung von KI zeigt, dass die Beschäftigten hier nur bedingt auf andere vertrauen, sich aber sehr viel mehr Partizipation und Mitbestimmung wünschen (→ Kap. 8). Ein kurzes Fazit (→ Kap. 9) schließt den Beitrag ab.



**Abb. 2** Grid des Denkzeugs. (Quelle: Eigene Darstellung)

## 2 KI am Arbeitsplatz: Verbreitung und generelle Einstellung

Künstliche Intelligenz (KI) und Maschinelles Lernen (ML) sind technisch gesehen keine völlig neuen Technologien, aber erst seit wenigen Jahren explodiert die Nutzung und auch in den Medien spielt KI eine zunehmende Rolle. Viele KI/ML-unterstützte Funktionalitäten, die in der Alltagsnutzung sowohl privat wie am Arbeitsplatz faktisch angekommen sind, sind für die Nutzenden teils nicht als KI/ML erkennbar. Noch gibt nur eine Minderheit von 11 % an, dass KI schon am Arbeitsplatz eingesetzt wird, bei weiteren 12 % sei ein Einsatz schon konkret geplant und in 17 % der Fälle wird zumindest im Betrieb bereits darüber diskutiert. Trotzdem bleibt KI für die Mehrheit von 60 % der Beschäftigten noch ein unbekanntes Terrain am Arbeitsplatz (Abb. 3).

Dies scheint zumindest nicht an der negativen Einstellung der Beschäftigten zu liegen; das untere Kreisdiagramm der gleichen Abbildung zeigt: Nur 14 % lehnen einen Einsatz von KI am Arbeitsplatz grundsätzlich ab, 37 % können

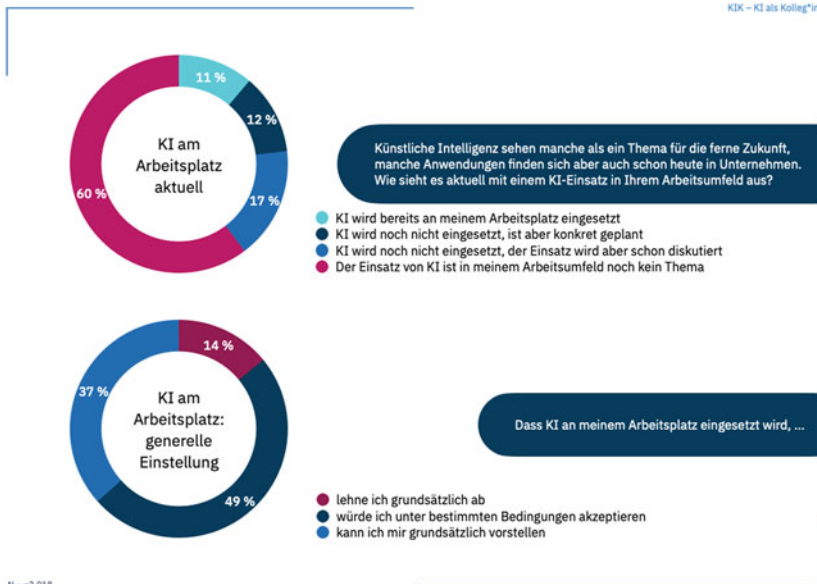


Abb. 3 Allgemeine Einstellung. (Quelle: Eigene Darstellung)

sich das grundsätzlich vorstellen und mit 49 % und damit nahezu der Hälfte der Befragten ist dies ein Thema der Gestaltung – es kommt sozusagen darauf an, die Akzeptanz knüpft sich an Bedingungen.

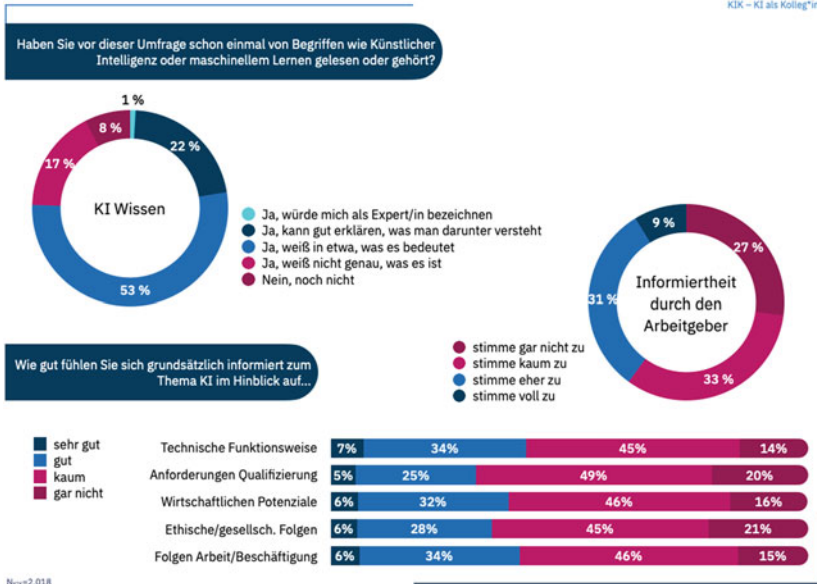
---

### **3 Wissen und Informiertheit**

Um den Stand des Wissens der Befragten zu erheben, wurde einerseits eine – notwendigerweise subjektive – Selbsteinschätzung zu den zentralen Begriffen abgefragt und andererseits erhoben, inwiefern sich die Befragten allgemein gut informiert fühlen zu verschiedenen Aspekten rund um Funktion bis zu Folgen von KI. Die Daten zeigen (Abb. 4) wenig erstaunlich, dass sich ein verschwindend kleiner Prozentsatz eine tiefgehende Expertise zum Thema zuschreibt. Gleichzeitig – und das dürfte eher überraschen – geben 22 % an, KI gut erklären zu können und mit 53 % ist die Mehrheit der Meinung, in etwa zu wissen, was KI bedeutet. 17 % sind sich dabei eher unsicher und 8 % hatten vor der Studie mit dem Begriff noch keine Berührung. Damit gibt die deutliche Mehrheit an, KI erklären zu können oder in etwa zu wissen, was KI bedeutet. Zum Vergleich: in einer ähnlichen Abfrage der Initiative D21 in 2020 waren dies 61 %. Der in unserer Abfrage höhere Wert spiegelt zum einen wider, dass hier alle Befragte im Arbeitsleben stehen und zudem das Thema eine deutliche Dynamik aufweist: auch in der D21 Studie stieg der Wert dazu zwischen der Befragung 2018 und 2020 um 9 % (Initiative D21 2020, S. 30).

Trotz dieser in der Summe recht breiten Einschätzung, mehr oder weniger zum Thema Bescheid zu wissen, ist das Gefühl, dazu gut informiert zu werden geringer ausgeprägt. Am stärksten erreicht die Befragten die Aspekte der technischen Funktionsweise (41 fühlen sich gut bis sehr gut informiert) und zu Folgen für Arbeit und Beschäftigung (40 % gut/sehr gut). Am wenigsten Informiertheit zeigt sich beim Thema Qualifizierung: 30 % sehen sich hierbei gut bis sehr gut informiert und damit die große Mehrheit von rd. 70 % als kaum bis gar nicht informiert. Definitiv ein Signal, das bildungspolitische wie unternehmensinterne Antworten als erforderlich kenntlich macht.

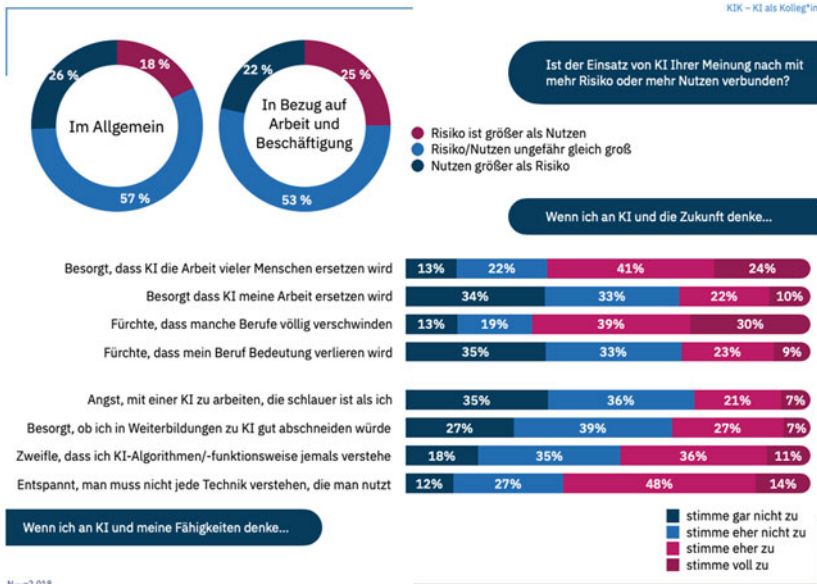
Vergleichsweise ist das Informiertheitsgefühl durch den Arbeitgeber deutlich schlechter: Die Mehrheit von 60 % stimmt der Aussage nicht zu, durch den Arbeitgeber gut zu KI informiert zu sein. Hier gibt es deutlichen Nachholbedarf.



**Abb. 4** Wissen und Informiertheit. (Quelle: Eigene Darstellung)

## 4 Sorgen und Befürchtungen

Während bspw. in einer US-amerikanischen Studie sich 19 % als sehr ängstlich in Bezug auf KI zeigten (Liang und Lee 2017), ist die Sichtweise von Beschäftigten in Deutschland differenzierter und weniger von Ängsten geprägt: Die oberen Kreisdiagramme (siehe Abb. 5) zeigen, dass die grundsätzliche Einschätzung zu KI im Allgemeinen wohlwollender ausfällt als konkreter in Bezug auf Arbeit und Beschäftigung: Im Allgemeinen sehen 18 % deutlich weniger Risiko als die Gruppe, die zu 26 % einen größeren Nutzen erwartet. In Bezug auf Beschäftigung und Arbeit stehen einer mit 25 % höheren Risikoeinschätzung leicht geringere 22 % an Nutzenerwartenden gegenüber. In beiden Fällen aber ist die Mehrheit (57 % bei KI allgemein und 53 % bei Arbeit und Beschäftigung) der Meinung, dass Chancen und Risiken ungefähr gleich groß seien. Die ausgewogene Mitte kann sowohl Ausdruck von Unsicherheit beim Thema oder Ausdruck der Hoffnung in gelingende und partizipativer Gestaltungsprozesse (vgl. Kap. 8) geschuldet sein kann.



**Abb. 5** Sorgen und Befürchtungen rund um KI. (Quelle: Eigene Darstellung)

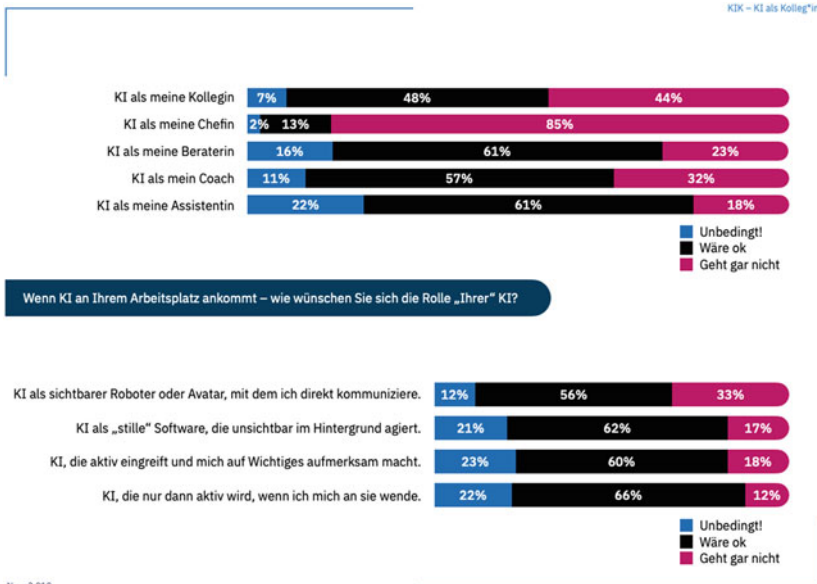
Bei Zukunftseinschätzungen und/oder Sorgen in Bezug auf KI, wenn diese stärker auf die eigene konkrete Betroffenheit zielen, zeigen die Balkendiagramme. Zwar wird in hohem Maße davon ausgegangen, dass die Arbeit anderer durch KI ersetzt werden wird (65 % stimmen eher oder voll zu), dagegen stimmen mit 32 % deutlich weniger der Aussage zu, dass ihre eigene Arbeit durch KI ersetzungsgefährdet sei. Ähnlich verhält es sich beim Beruf: 69 % sehen in diesem Zusammenhang auch ein Verschwinden von Berufen, mit ebenfalls 32 % ist diese Sorge aber für den eigenen Beruf deutlich geringer.

Auch beim in den Medien sehr beliebten Vergleich von Fähigkeiten der KI gegenüber den Fähigkeiten des Menschen, sind die Befragten durchgängig eher entspannt: 71 % haben keine Angst, dass sie mal mit einer KI zu arbeiten müssten, die schlauer ist als sie selbst und die deutliche Mehrheit von 66 % macht sich auch keine Sorgen, in einer Ausbildung zu KI nicht gut abzuschneiden. Zwar zweifeln 47 %, ob sie die Funktionsweise von KI je verstehen werden – 62 % aber finden, dass man das im Detail auch gar nicht muss.

## 5 KI: proaktive Kollegin oder stille Assistentin?

Die Erhebung hat auch abgefragt, in welcher Rolle sich Beschäftigte KI wünschen. Dabei ist die Einschätzung am deutlichsten, dass KI keine Chefin sein soll (siehe Abb. 6). Nur 3 % wünschen sich das, 85 % aber lehnen das klar ab. Bei der KI als Kollegin ist das Bild durchwachsen: Zwischen „wäre ok“ und „geht gar nicht“ hält sich die Einschätzung fast in der Waage und nur 7 % plädieren eindeutig für ein kollegiales Miteinander. Man wünscht sich KI eher in den unterstützenden Rollen, noch am stärksten wird mit 22 % KI als Assistentin gesehen.

Abgefragt wurde auch, wie sichtbar bzw. aktiv die KI in ihrer Rolle am Arbeitsplatz werden soll. Es dominiert auch hier jeweils das etwas uneindeutige und verhaltene „wäre ok“. Am eindeutigsten ist die Ablehnung von KI in Form eines sichtbaren Roboters oder Avatars: ein Drittel kann sich das nicht vorstellen. Bei den drei Varianten der Aktivität kristallisiert sich keine eindeutige Präferenz heraus. Insgesamt zeigen diese Werte vor allem eines: KI sollte so gestaltet



**Abb. 6** KI als Kollegin? Rollenerwartungen. (Quelle: Eigene Darstellung)



werden, dass die mit ihr umgehenden Beschäftigten die Rollenverteilung individuell bestimmen und beeinflussen können – am besten nicht nur während der Implementierungs- und Lernphase, sondern auch im späteren Prozess. Gestaltung ist auch hier „key“ (vgl. hierzu auch 8).

---

## 6 Vier Einstellungstypen

Mit dem besonders interaktiven Befragungstool „Denkzeug“ und aufwendigeren statistischen Analysen (PCA und Varimax Rotation) konnten vier Einstellungstypen identifiziert werden: Die Störungsjonglierenden, die Arbeitsfokussierten, die Verantwortungstragenden und die Entlastungsuchenden. Für diese vier Typen zeigt sich trotz aller Unterschiede: die Ansichten zu KI als Kollegin sind stärker vom Arbeitskontext geprägt als von allgemeinen Vorstellungen zu Technik und KI. Zur inhaltlichen Unterscheidung der vier Typen dienen zum einen die jeweils auf den extremeren Positionen liegenden Aussagen ( $\pm 2-3$ ) und zum anderen, die diesen von den Befragten zugeordneten offenen Kommentare (siehe Abb. 7).

Nachfolgend werden diese Typen kurz inhaltlich skizziert, dabei werden sowohl die über die Rotationsverfahren extrahierten differenzierenden Denkzeug-Aussagen einbezogen, die sozio-demografischen Merkmale der vier Einstellungstypen und die Schwerpunkte, die sich in den offenen Fragen als typisch für die jeweilige Gruppe zeigen.

*Störungsjonglierende* (30 %) – das zeigt die Legung der Aussagen (siehe Abb. 7) lehnen Automatismen der KI ab, wollen immer handlungsfähig bleiben – gerade bei Unwägbarkeiten und Störungen. Sie wollen wissen, wie die KI entscheidet, aber nicht erst dann einspringen, wenn die KI nicht mehr weiterweiß oder gar für deren Fehler einstehen. Zu diesem Faktor zählen 29,5 % der Befragten, er erklärt 13 % der Varianz.<sup>7</sup> Hierbei fällt auf, dass die Aussagen auf den positivsten Feldern, die mit  $p < 0,01$  hoch differenzieren und höchste Z-Scores aufweisen, die inhaltliche Handlungsfähigkeit des Menschen adressieren: zum einen im Hinblick auf die Nachvollziehbarkeit der Entscheidungskriterien (EK-N) und zum anderen im Hinblick auf den Erhalt der eigenen Lernfähigkeit

---

<sup>7</sup> Varianz ist ein Streuungsmaß; vereinfacht gesagt kann damit angezeigt werden, ob die errechneten Gruppen (statistisch ausgedrückt: Faktoren) in ihrer durch die Rotation vorgenommene Verdichtung mehr erklären als die einzelnen zugrunde liegenden Fälle (anders als bei üblichen Rotationsverfahren sind in der Q-Methode die Fälle und nicht die Variablen entscheidend). Insgesamt können mit den vier Gruppen rd. 66 % der Varianz erklärt werden, anders formuliert: Die Ergebnisse des Denkzeugs machen die Einstellung von Zweidrittel der Befragten zu KI am Arbeitsplatz pointiert sichtbar.

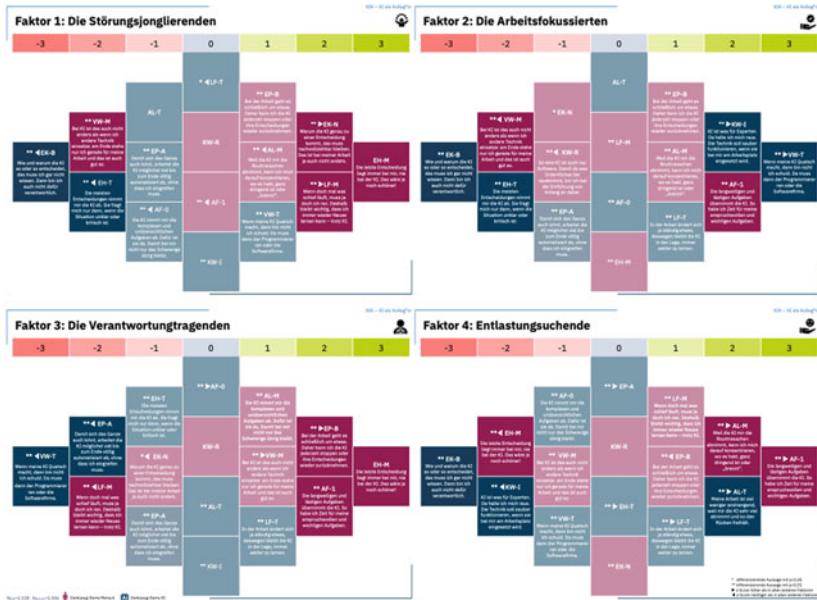
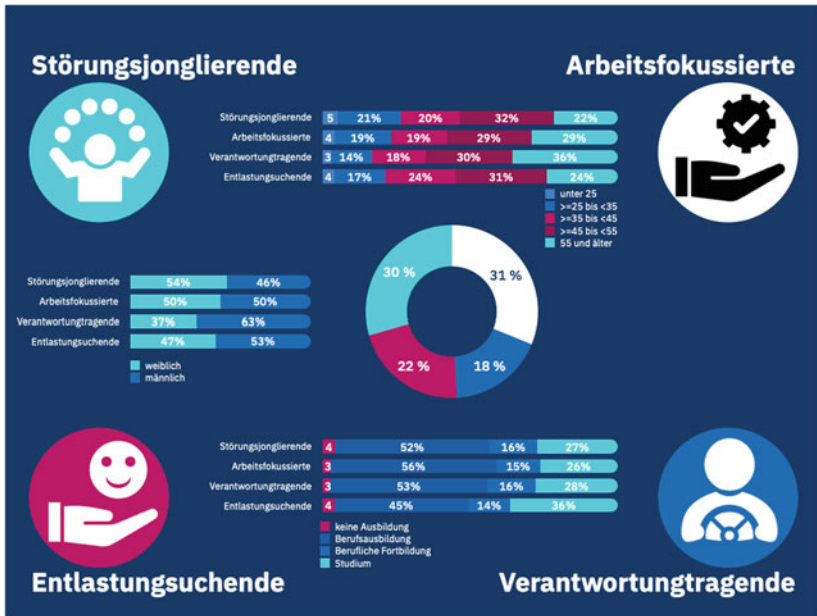


Abb. 7 KI-Einstellungstypen – Denzzeugresultate. (Quelle: Eigene Darstellung)

(LF-M), jeweils auf + 2. Im Negativen finden sich drei Aussagen mit hoher Differenzierung und teils höchsten Z-Scores: Dieser Typ will nicht nur wissen, wie die KI entscheidet (EK-B, -3), sondern auch die Entscheidungen nicht nur dann übernehmen, wenn die KI nicht mehr weiterweiß (EH-T, -2). Für die KI gerade stehen – oder deren Fehler – will dieser Typ ebenfalls nicht (EH-M, -2). Der Typ ist minimal weiblicher, deutlich jünger als die anderen und mehrheitlich beruflich qualifiziert (siehe Abb. 8). Typische Aussagen in den offenen Textfeldern dieser Gruppe fokussieren besonders auf die Komplexität und Dynamik von Arbeitssituationen („besondere Situationen“, „je nach Anwendungsfall“ seien vielfältige Faktoren zu berücksichtigen, „komplexe Arbeitssituationen“, „individuelle Fälle“, Bedeutung „zahlreicher Kleinigkeiten“ oder „neue Probleme in kurzer Zeit“) und dabei auf die Rolle menschlichen Handelns und nicht objektivierbarer Fähigkeiten („Erfahrung“, „gesunder Menschenverstand“, „Bauchgefühl“, „schnell reagieren können“).

Der Typ der *Arbeitsfokussierten* (31 %) ist vor allem auf die eigene Arbeit und der Erledigung der Aufgaben konzentriert, KI soll dabei vor allem nicht stören.



**Abb. 8** KI-Einstellungstypen – sozio-demografische Merkmale. (Quelle: Eigene Darstellung)

Zu dieser Gruppe zählen mit 31 % die meisten Befragten, er erklärt ebenfalls 13 % der Varianz. Bei diesem Typ steht ganz klar (alle sechs außen liegenden Aussagen sind hoch differenzierend für diesen Typ) die eigene Tätigkeit im Vordergrund, KI soll wie andere Technik auch, sozusagen ihren Job machen und vor allem nicht auffallen oder sich in den Vordergrund spielen. Wenn die KI nicht funktioniert (VW-T, + 3), sollen sich andere darum kümmern (deren Job das schließlich ist). KI soll still funktionieren (KW-I, + 2) und ist allenfalls für untergeordnete Tätigkeiten da (AF-1). Im Negativen finden sich hier die gleichen drei Aussagen wie bei den Störungsjonglierenden. Diese beiden Typen unterscheiden sich inhaltlich also im Positiven. Während die Gruppe der Störungsjonglierenden sozusagen von der KI mehr „will“ und gleichzeitig mehr bedacht ist, die eigene Kompetenz zu erhalten, sieht die Gruppe der Arbeitsfokussierten die KI weniger auf Augenhöhe, sondern eher als stumme Dienerin, mit der man sich auch kaum beschäftigen will. Sie soll einfach funktionieren. Dieser Typ ist vom Geschlecht her sehr ausgewogen, es dominiert die Altersgruppe ab 45 Jahren und bei der

Qualifikation finden sich ebenfalls mehrheitlich beruflich Qualifizierte. Auch in den offenen Angaben zu diesem Typ wird zudem noch mal besonders deutlich, dass den Befragten vor allem wichtig ist, dass die KI reibungslos funktioniert und nicht bei der eigentlichen Arbeit stört. Dafür wird die fachliche Verantwortung bei „dem Programmierer“ oder „der IT“ und „den Experten“ gesehen und die Verantwortung für eventuelle Folgen beim „Arbeitgeber“ oder „beim Chef“, der den Einsatz von KI entschieden hat.

Die *Verantwortungstragenden* (18 %) wollen vor allem am Ruder bleiben. Dieser Typ will weder Verantwortung noch Entscheidungshoheit an andere oder die KI abgeben und immer aktiv in die KI eingreifen können. Die anspruchsvollen Aufgaben sollen beim Menschen verbleiben, Verantwortung soll auch für die KI nicht an andere delegiert werden. Zum Typ der Verantwortungstragenden zählen 17,9 % der Befragten, er erklärt 18 % der Varianz. Auch bei den Verantwortungstragenden steht die eigene Arbeit im Vordergrund. Dieser Typ unterscheidet sich vor allem darin, dass das Thema der eigenen aktiven Rolle sich koppelt mit dem Übernehmen von Verantwortung. Dieser Typ will wie die Störungsjonglierenden nicht die Entscheidungshoheit abgeben (EH-M, + 3), hoch differenzierend aber sind hier die Aussagen zum aktiv Eingreifen-Können in die KI (EP-B, + 2) und dass die anspruchsvollen Aufgaben beim Menschen verbleiben sollen (AF-1, + 2). Die drei negativsten Aussagen sind jeweils hoch differenzierend und bestätigen das Bild: Die Verantwortung soll auch für die KI nicht an Andere delegiert werden (VW-T, -3), man will keinen Automatismus der KI, sondern weiterhin Eingriffsmöglichkeiten erhalten (EP-A, -2) aber man will nicht erst Eingreifen müssen, wenn bei der KI etwas schief läuft und sich dafür ständig weiterqualifizieren müssen (LF-M, -2). Die Verantwortung, die sich hier als roter Faden durchzieht, bezieht sich also immer aus der Arbeitsperspektive auf die KI. Dieser Typ ist überwiegend männlich und hat den höchsten Anteil bei den mindestens 55-Jährigen und ist ebenfalls überwiegend beruflich qualifiziert. Auch bei den offenen Aussagen bestätigt sich, dass bei dieser Gruppe die eigene Verantwortung für die Arbeit im Zentrum steht und nicht abgegeben werden soll an die KI. So finden sich hier vielfach Statements wie: „ich hafte am Ende für die Entscheidungen“ und „ich trage die Verantwortung“, deswegen solle die „Kontrolle bzw. endgültige Entscheidung [...] nach wie vor bei mir liegen“, man will „die Zügel niemals komplett aus der Hand geben“ und „keine Technik für [die eigenen] Aufgaben bzw. Entscheidungen verantwortlich machen“.

Die *Entlastungsuchenden* (22 %) erwarten sich von der KI die eigene und vielfältige Entlastung. Sie wollen entlastet werden von Routineaufgaben sowie von langweiligen und lästigen Aufgaben, um den Rücken für die eigentliche Arbeit freizuhaben. Zu dieser Gruppe zählen 21,5 % der Befragten, sie erklärt

10 % der Varianz. Entlastungssuche hat dabei viele Gesichter. Die positiv differenzierenden Aussagen zeigen, dass sowohl Routineaufgaben (AL-M, + 2) wie langweilige und lästige Aufgaben (AF-1, + 3) an die KI delegiert werden sollen. Nicht, weil dieser Typ nichts tun will, sondern in beiden Fällen, um sich auf die schwierigen und anspruchsvollen Arbeitsaufgaben konzentrieren zu können. Die KI soll hier ganz eindeutig den Rücken freihalten (AL-T, + 2), für das eigentlich Relevante: die eigene Arbeit. Denn im Hinblick auf die KI will sich dieser Typ nicht entlasten, sondern durchaus einmischen: man will schon wissen, warum die KI so oder so entscheidet (EK-B, -3) und die Verantwortung für die KI nicht an Andere abgeben (KW-I). Wenn beides gesichert ist, dann aber muss auch nicht jede Entscheidung beim Menschen liegen, da kann die KI gerne auch mal ran (EH-M) Das kann man ihr zutrauen, weil nach den Aussagen im Positiven ihre Rolle eine zuarbeitende und damit eine im kleinen Maße entlastende bleibt. Dieser Typ hat beim Geschlecht nur einen sehr kleinen Überhang an Männern. Bei der Altersverteilung zeigt sich nichts Auffälliges, bei der Qualifikation fällt auf, dass sich hier der im Vergleich höchste Anteil an rein akademisch Qualifizierten findet. Auch die offenen Aussagen zeigen, dass diese Gruppe sich durch den Einsatz von KI vor allem Entlastung bei „Routine-“ oder „ermüdenden“ oder „langweiligen“ Aufgaben erhofft, um sich den wirklich wichtigen Aufgaben zu widmen. Auch mehr „Effizienz“, „Prozessoptimierung“ werden oft genannt und immer wieder die Hoffnung und Erwartung, dass die KI eine bei der Arbeit unterstützende Rolle einnimmt.

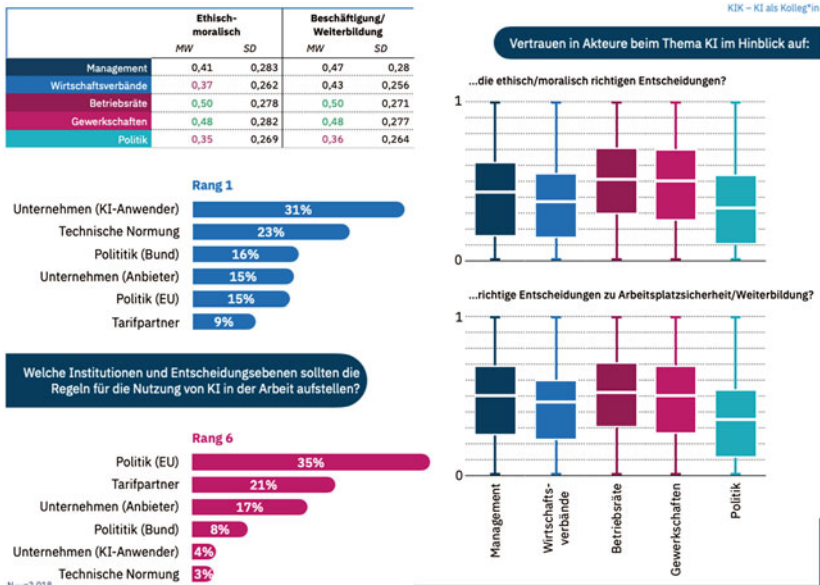
Die vier Typen zeigen zweierlei: zum einen lohnt eine komplexere Form der Befragung, wie sie mit dem Denkzeug ermöglicht wurde, um die komplexen und durchaus auch widersprüchlichen Sichtweisen zu erheben. Zum anderen zeigt sich, dass Beschäftigte den Einsatz von KI deutlich aus der Perspektive ihrer Arbeit bewerten. Fragen der Kommunikation im Unternehmen bei KI-Einführungsprozessen ebenso wie deren Gestaltung sollten daher die Unterschiedlichkeit dieser vier Einstellungstypen systematisch berücksichtigen.

---

## 7 Akteure und Zuständigkeiten

Im gesellschaftlichen Diskurs zu KI herrscht weitgehend Konsens, dass sich beim Einsatz von KI viele grundsätzliche Gestaltungsfragen stellen, deren Entscheidung nicht allein der je konkreten Technikentwicklung überlassen werden kann. Damit ist aber noch nicht beantwortet, welche Akteure was auf welcher Ebene zu klären oder zu regulieren haben. Auch dazu gibt es in unserer Erhebung eine Reihe von Fragen, von denen hier nur ein Teil dargestellt wird. Zu

den Balkendiagrammen links in Abb. 9 wurde gefragt, welche Institutionen und Entscheidungsebenen die Regeln für die Nutzung von KI in der Arbeit aufstellen sollten. Die Befragten konnten die angebotenen Akteure mit Kärtchen in ein Sechsraster einsortieren. Rang 1 (blau) zeigt, wo die Verantwortung am häufigsten verortet wird, nämlich mit 31 % in der konkreten Nutzung beim KI-Anwendungsunternehmen, weitere 23 % hoffen auf das Potenzial technischer Normung. Rang 6 (rot) zeigt, wer am wenigsten als „in charge“ gesehen wird und damit ein weitgehend konträres Bild: Tarifpartner und EU-Politik werden am wenigsten als relevante Akteure gesehen. Der Dreiklang aus technischer Normung, Bundespolitik und den KI-Anwendungsunternehmen ist für die Mehrheit der Befragten der bevorzugte Weg. KI-Anbieter-Unternehmen will man dabei auch nicht wirklich vertrauen, nur 15 % haben sie auf Rang 1 der zuständigen Akteure gesehen.



**Abb. 9** Akteure und Zuständigkeiten. (Quelle: Eigene Darstellung)

In den rechten Darstellungen<sup>8</sup> wird sichtbar, wie stark die Befragten den fünf genannten Akteuren zu trauen, die richtigen Entscheidungen zum KI-Einsatz zu treffen – im Hinblick auf ethisch-moralische Belange wie in Bezug auf Beschäftigungssicherung oder Weiterbildung. Die Skala steht für 0 gleich „kein Vertrauen“ und 1 für „volles Vertrauen“. Zu beiden Themen ist das Vertrauen in die Politik am geringsten ( $\bar{x}$  0,35 bei ethisch-moralischen Entscheidungen,  $\bar{x}$  0,36 bei Beschäftigungssicherung/Weiterbildung) und auch Wirtschaftsverbänden ( $\bar{x}$  0,37 bzw.  $\bar{x}$  0,43) traut man weniger als dem Management des eigenen Unternehmens. Die besten und jeweils gleich hohen Mittelwerte (bei leicht unterschiedlicher Streuung) erreichen bei beiden Themen die Gewerkschaften ( $\bar{x}$  0,48) und Betriebsräte ( $\bar{x}$  0,50). Das Vertrauen ist also in die Interessensvertretungen noch am höchsten, insgesamt aber zeigen die Mittelwerte, die nur in einem Fall die theoretische Mitte der Skala erreichen, ein doch eher gering ausgeprägtes Vertrauen in die Entscheidungsfähigkeit der Verantwortlichen.

---

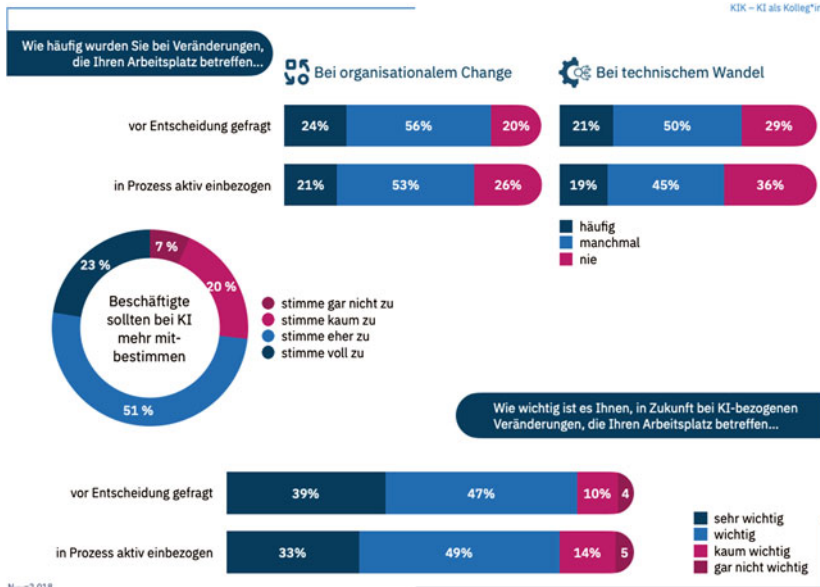
## 8 Partizipation und Mitbestimmung

Bisherige Erfahrungen mit Technik und deren Gestaltung am Arbeitsplatz prägen – so ist anzunehmen – auch stark, inwieweit auf die Gestaltungsfähigkeit des Neuen vertraut werden kann. In beiderlei Hinsicht aber sind die Erfahrung der Beschäftigten bisher – vorsichtig formuliert – durchwachsen. Bei der Abfrage zur Partizipation wurde nach der Partizipationsintensität unterschieden (nur gefragt oder aktiv einbezogen) und zum anderen, ob diese sich auf organisationalen oder technischen Wandel bezieht. Beschäftigte werden überwiegend vor der Entscheidung gefragt und etwas seltener aber ebenso überwiegend in einem aktiven Prozess einbezogen (siehe Abb. 10). Allerdings wird nur ein Fünftel bis ein Viertel der Beschäftigten häufig gefragt oder einbezogen. Und: bei organisationalem Change wird jede:r vierte Beschäftigte und mehr als jede:r dritte wird nie aktiv einbezogen, wenn es um technischen Wandel am Arbeitsplatz geht. Zudem ist der Einbezug beim Organisatorischen selbstverständlicher als der bei technischem Wandel.

KI wird in vielen Fällen am Arbeitsplatz mehr Partizipation erfordern als bislang. Und diese wünschen sich Beschäftigte auch vermehrt: 74 % stimmen der Aussage zu, dass Beschäftigte beim Thema KI mehr mitbestimmen sollten. Und wenn es um KI am eigenen Arbeitsplatz geht, übertreffen die Erwartungen an

---

<sup>8</sup> Die ‚Boxplots‘ sind zu verstehen als Draufsicht auf Verteilungskurven, wobei der Querstrich den Scheitelpunkt bzw. Median anzeigt.



**Abb. 10** Partizipationserfahrungen und -erwartungen. (Quelle: Eigene Darstellung)

den Einbezug deutlich, was bisher an Partizipation gewährt wird: 39 % ist es sehr wichtig und weiteren 47 % wichtig, vor der Entscheidung zu KI gefragt zu werden und 33 % ist es sehr wichtig und 49 % wichtig, aktiv in den Prozess einbezogen zu sein. Beschäftigte können und wollen Partizipation und wünschen sich bei KI mehr davon und mehr Mitbestimmung.

## 9 Fazit: Komplexe Einstellungen, wenig Angst, aber auch wenig Vertrauen in die Akteure

Diese komprimierte Darstellung einiger zentraler Ergebnisse unserer repräsentativen Beschäftigungsbefragung zu „KI als Kollegin“ (Kap. 1) zeigt, dass KI erst begonnen hat, spürbar in der Arbeitswelt anzukommen (Kap. 2) und Beschäftigte sich dazu mehrheitlich ganz gut informiert fühlen – allerdings zu wenig aus Richtung des eigenen Arbeitgebers (Kap. 3). Beschäftigte zeigen sich im Hinblick auf KI nicht so ängstlich und sorgenvoll, wie üblicherweise unterstellt wird und sie



sorgen sich beim Thema KI und Arbeit sehr viel mehr um andere als um sich selbst (Kap. 4). Beschäftigte wünschen sich die Rolle von KI am Arbeitsplatz nur bedingt KI als Kollegin, sondern eher als dienstbare und im Hintergrund still wirkende Assistentin (Kap. 5). Ein zentraler und besonders interaktiver Baustein der Erhebung – das so genannte „Denkzeug“ (Kap. 1) – zeigt auf der Basis multivariater Analysen, dass Einstellungsmuster zu KI am Arbeitsplatz nicht nur komplex sind, sondern dabei die Perspektive aus der eigenen Arbeit maßgeblicher ist als allgemeine Einstellungen zu Technik und KI. Es lassen sich zudem vier sehr komplexe Einstellungsmuster finden, die vor allem zeigen: KI am Arbeitsplatz wird aus der Perspektive von Arbeit bewertet und die ich in vier statistisch extrahierten Einstellungstypen (Kap. 6) – die Störungsjonglierenden, die Arbeitsfokussierten, die Verantwortungstragenden und die Entlastungsuchenden – abbilden.

Die Ergebnisse zeigen auch: bei Entscheidungen und Einführungsprozessen von KI ist das Vertrauen in die üblichen Akteure durchwachsen – oder andersherum: die Entscheidenden Akteure bei dem Thema sowohl in der Politik wie im Unternehmen müssen sich das Vertrauen der Beschäftigten beim Thema KI erst noch verdienen (Kap. 7). Übergeordneten potenziell zuständigen Akteuren der Gestaltung von KI wird dabei weniger vertraut als denen, die der betrieblichen Anwendung näher sind. Vielleicht auch deswegen finden sich klare Wünsche an eine verstärkte eigene Partizipation und Mitbestimmung bei der Gestaltung von KI am Arbeitsplatz (Kap. 8). Alle dargestellten Ergebnisse sind vor dem Hintergrund zu sehen: Die Mehrheit der Beschäftigten erlebt KI noch nicht an ihrem Arbeitsplatz und viele Beschäftigte fühlen sich zwar informiert zum Thema KI – allerdings nicht ausreichend.

Daher sind viele der erfragten Einstellungen und Erfahrungen zu KI als Kollegin heute noch nicht erfahrungsgesättigt. Je nachdem, welche Erfahrungen in der Nutzung von KI in den kommenden Jahren gemacht werden, werden sich Einstellungen dazu verändern, ausdifferenzieren oder auch verfestigen. Deswegen kann diese Befragung nur eine Momentaufnahme sein. Mit dem zunehmenden Einsatz und einer weiteren Verbreitung von KI wird sich auch die Meinung zu KI als Kollegin ändern. Daher sind weitere, regelmäßige Studien dieser Art notwendig, um den sich verändernden Blick von Beschäftigten auf diese noch neue Technologie sichtbar zu machen und um damit auch Beschäftigten in den gesellschaftlichen und betrieblichen Aushandlungsprozessen rund um KI immer wieder eine Stimme zu geben.

## Literatur

- Acemoglu, Daron, und Pascual Restrepo. 2019. *Automation and New Tasks: How Technology Displaces and Reinstates Labor*. Cambridge: National Bureau of Economic Research <http://www.nber.org/papers/w25684>.
- Aguinis, Herman, und Kyle J. Bradley. 2014. Best Practice Recommendations for Designing and Implementing Experimental Vignette Methodology Studies. *Organizational Research Methods* 17: 351–371.
- Arntz, Melanie, Terry Gregory, und Ulrich Zierahn. 2018. *Digitalisierung und die Zukunft der Arbeit: Makroökonomische Auswirkungen auf Beschäftigung, Arbeitslosigkeit und Löhne von morgen*. Mannheim: ZEW Zentrum für Europäische Wirtschaftsforschung.
- Autor, David H., Frank Levy, und Richard J. Murnane. 2003. The skill content of recent technological change: An empirical exploration. *The Quarterly Journal of Economics* 118: 1279–1333.
- Balu, V., Raga Lahari Tiruveedhula, und Akhilesh Panchapakesan Shastri. 2020. Paperles Office. *Mukt Shabd Journal* 9: 1697–1701.
- Banasick, Shawn. 2022. EQ Web Sort. <https://github.com/shawnbanasick/easy-htmlq>. Zugegriffen: 23. Juli 2022.
- Boes, Andreas, und Tobias Kämpf. 2011. *Global verteilte Kopfarbeit: Offshoring und der Wandel der Arbeitsbeziehungen*. Berlin: Edition Sigma.
- Bonin, Holger, Terry Gregory, und Ulrich Zierahn. 2015. *Übertragung der Studie von Frey/ Osborne (2013) auf Deutschland*. Berlin: Bundesministerium für Arbeit und Soziales.
- Brewer-Deluce, Danielle, und Bhanu Sharma. 2020. Beyond Average Information: How Q Methodology Enhances Course Evaluations in Anatomy. *Anatomical Sciences Education* 13: 137–148.
- Brown, Steven R. 1980. *Political Subjectivity – Applications of Q Methodology in Political Science*. New Haven, CT: Yale University Press.
- Brynjolfsson, Erik, und Andrew McAfee. 2014. *The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies*. New York, London: Norton.
- Buss, Klaus-Peter, Herbert Oberbeck, und Knut Tullius. 2022. Systemische Rationalisierung 0. Wie Wettbewerb und Geschäftsmodelle die Digitalisierung in Handel, Logistik und Finanzdienstleistungen prägen. *Berliner Journal für Soziologie*. 32: 35–68.
- Dengler, Katharina, und Britta Matthes. 2015. Folgen der Digitalisierung für die Arbeitswelt Substituierbarkeitspotenziale von Berufen in Deutschland. Nürnberg: IAB.
- Dosenovic, Pero, Kimon Kieslich, und Birte Keller. 2020. *Meinungsmonitor Künstlich Intelligenz – Methodensteckbrief*. Düsseldorf: Heinrich-Heine-Universität. [https://uniduesseldorf.datalion.net/custom/img/16/200629\\_Methodensteckbrief%20MeMoKI.pdf](https://uniduesseldorf.datalion.net/custom/img/16/200629_Methodensteckbrief%20MeMoKI.pdf).
- Fernández-Macias, Enrique, und John Hurley. 2014. *Drivers of recent job polarisation and upgrading in Europe: Eurofound Jobs Monitor 2014*. Luxembourg: Eurofound. [http://eurofound.europa.eu/sites/default/files/ef\\_files/pubdocs/2014/19/en/1/EF1419EN.pdf](http://eurofound.europa.eu/sites/default/files/ef_files/pubdocs/2014/19/en/1/EF1419EN.pdf). Zugegriffen: 16. Februar 2015.
- Frey, Carl Benedikt, und Michael A. Osborne. 2017. The future of employment: How susceptible are jobs to computerisation? *Technological Forecasting and Social Change* 114–280.
- Haefner, Klaus, Ernst H. Eichmann, und Claudia Hinze. 1987. *Denkzeuge. Was leistet der Computer? Was muß der Mensch selbst tun?* Basel: Birkhäuser.

- Herrmann, Thomas, und Sabine Pfeiffer. 2022. Keeping the Organization in the Loop: A Socio-Technical Extension of Human-Centered Artificial Intelligence. *AI & SocietOnline First*: online first. <https://doi.org/10.1007/s00146-022-01391-5>
- Horton, Richard. 2015. *The Robots are Coming*. London: Deloitte.
- Huchler, Norbert. 2019. Assimilierende versus komplementäre Adaptivität. Grenzen (teil)autonomer Systeme. In *Autonome Systeme und Arbeit. Perspektiven, Herausforderungen und Grenzen der Künstlichen Intelligenz in der Arbeitswelt*, Hrsg Hartmut Hirsch-Kreinsen und Anemari Karačić, 139–180. Bielefeld: Transcript.
- ICC/ESOMAR. 2016. *International Code on Market, Opinion and Social Research and Data Analytics*. Amsterdam.
- IMRO. 2015. IMRO Guidelines for Best Practices in Online Sample and Panel Management Insights Association. <https://www.insightsassociation.org/issues-policies/bestpractice/imro-guidelines-best-practices-online-sample-and-panel-management>. Zugegriffen: 5. Februar 2022.
- Initiative D21. 2020. *Wie digital ist Deutschland 19/20. Jährliches Lagebild zur Digitalen Gesellschaft*. Kantar.
- Johnson, Branden B., und Lynn Waishwell. 2014. Q method can identify diverse perspective on 'helpful' information on cancer clusters and inform risk communication generally *Journal of Risk Research* 17: 1125–1145.
- Kadiresan, Adheesh, Yuvraj Baweja, und Obi Ogbanufe. 2022. Bias in AI-Based Decision Making. In *Bridging Human Intelligence and Artificial Intelligence, Educational Communications and Technology: Issues and Innovations*, hrsg. Mark V. Albert, LinLin, Michael J. Spector und Lemoyne S. Dunn, 275–285. Cham: Springer International Publishing.
- Khirfan, Luna, und Megan Leigh Peck. 2021. Deliberative Q-method: A combined method for understanding the ecological value of urban ecosystem services and disservices. *MethodsX* 8: 101547.
- Konrad, Erhard. 1998. Zur Geschichte der Künstlichen Intelligenz in der Bundesrepublik Deutschland. In *Sozialgeschichte der Informatik: Kulturelle Praktiken und Orientierungen, Studien zur Wissenschafts- und Technikforschung*, hrsg. Dirk Siefkes, Peter Eulenhöfer, Heike Stach und Klaus Städtler, 287–296. Wiesbaden: Deutsche Universitätsverlag.
- Krämer, Sybille. 1997. Werkzeug — Denkzeug — Spielzeug. Zehn Thesen über unseren Umgang mit Computern. In *Informatik und Lernen in der Informationsgesellschaft*, hrsg Heinz Ulrich Hoppe und Wolfram Luther, 7–13. Berlin, Heidelberg: Springer.
- Krzywdzinski, Martin. 2020. *Automatisierung, Digitalisierung und Wandel der Beschäftigungsstrukturen in der Automobilindustrie. Eine kurze Geschichte vom Anfang der 1990er bis 2018*. Berlin: WZB Berlin Social Science Center. <https://ideas.repec.org/p/zbw/wzb/gwp/spiii2020302.html>.
- Leiner, Dominik Johannes. 2019. SoSci Survey (Version 3.1.06) [Computer software]. <https://www.soscisurvey.de>.
- Liang, Yuhua, und Seungcheol Austin Lee. 2017. Fear of Autonomous Robots and Artificial Intelligence: Evidence from National Representative Data with Probability Sampling *International Journal of Social Robotics* 9: 379–384.
- Menz, Wolfgang, Sarah Nies, und Dieter Sauer. 2019. Digitale Kontrolle und Vermarktlichung: Beschäftigtenautonomie im Kontext betrieblicher Strategien der Digitalisierung. *PROKLA. Zeitschrift für kritische Sozialwissenschaft* 49, 181–200.

- Müller, Florian H., und Elisabeth Kals. 2004. Die Q-Methode. Ein innovatives Verfahren zur Erhebung subjektiver Einstellungen und Meinungen. *Forum Qualitative Sozialforschung* 5: Art. 34.
- Nilsson, Nils J. 2010. *Quest for Artificial Intelligence: A History of Ideas and Achievements*. Cambridge, New York: Cambridge University Press.
- Pfeiffer, Sabine. 2018. The 'Future of Employment' on the Shop Floor: Why Production Job are Less Susceptible to Computerization than Assumed. *International Journal for Research in Vocational Education and Training (IJRVET)* 5: 208–215.
- Pfeiffer, Sabine. 2021a. *Datenmanagementkonzept zur Erhebung und Verarbeitung personenbezogener und/oder sensibler Daten für Forschungszwecke im Rahmen qualitativer und quantitativer Erhebungen*. v1.0 vom 6.11.2021. Nürnberg: Lehrstuhl für Soziologie (Technik – Arbeit – Gesellschaft) am NCT der FAU Erlangen-Nürnberg. <https://www.labouratory.de/files/downloads/WP-04-2021-Datenmanagementkonzept.pdf>.
- Pfeiffer, Sabine. 2021b. *Digitalisierung als Distributivkraft: Über das Neue am digitalen Kapitalismus*. Bielefeld: Transcript.
- Pfeiffer, Sabine, Petra Schütt, und Daniela Wühr. 2012. Smarte Innovation erfassen: Innovationsverlaufsanalyse und Visualisierung – Vorgehen und Samplebeschreibung. In *Smarte Innovation. Ergebnisse und neue Ansätze im Maschinen- und Anlagenbau*, Hrsg. Sabine Pfeiffer, Petra Schütt und Daniela Wühr, 49–74. Wiesbaden: VS Verlag für Sozialwissenschaften.
- Rammert, Werner. 2016. Die Zukunft der künstlichen Intelligenz: verkörpert – verteilt – hybrid. In *Technik – Handeln – Wissen. Zu einer pragmatistischen Technik- und Sozialtheorie* 227–242. Wiesbaden: Springer VS.
- Schnell, Rainer. 2018. *Warum ausschließlich 'online' durchgeführte Bevölkerungsumfragen nicht 'repräsentativ' sind*. Duisburg, Essen: Universität Duisburg-Essen. [https://www.uni-due.de/imperia/md/content/soziologie/schnell/repraesentativ\\_online.pdf](https://www.uni-due.de/imperia/md/content/soziologie/schnell/repraesentativ_online.pdf).
- Schrape, Jan-Felix. 2021. *Digitale Transformation*. 1. Aufl. Bielefeld: UTB GmbH.
- Seibold, Bettina, und Sylvia Stieler. 2016. *Digitalisierung der Bürotätigkeiten*. Stuttgart: IMU
- Uhlig, Franziska. 2014. Robinsons Pflug. Werkzeuge zwischen Nachbau und Erfindung. In *Werkzeug – Denkzeug*, Hrsg. Thomas H. Schmitz und Hannah Groninger, 165–190. Transcript.
- Weyer, Johannes et al. 2018. Big Data in soziologischer Perspektive. In *Big Data und Gesellschaft. Eine multidisziplinäre Annäherung*, hrsg. Barbara Kolany-Raiser, Reinhard Heil, Carsten Orwat und Thomas Hoeren, 69–149. Wiesbaden: Springer VS.



# Veränderungsdynamiken durch Technisierung von Arbeit am Beispiel von da Vinci als robotische Chirurgie-Assistenz – Erkenntnisse aus der Empirie

Regina Wittal und Carolyn Hettinger

## Zusammenfassung

Die fortschreitende Technisierung von Arbeit zeigt sich nicht nur in der Industrie und im Dienstleistungssektor, sondern auch in der Medizin. Von Röntgenapparaten über Computer- und Magnetresonanztomographie bis hin zu robotischen und auf Künstlicher Intelligenz (KI) basierenden Systemen zur Unterstützung medizinischer Tätigkeiten sind heutzutage zahlreiche technische Artefakte in den klinischen Alltag integriert. Neue Technologien bieten zwar neue Möglichkeiten, bergen aber ebenfalls Risiken und stoßen weitreichende und teilweise latente Wandlungsprozesse in Arbeitspraktiken an. Um diese Transformationsdynamiken zu verstehen, gilt es, stattfindende Veränderungen zu identifizieren und zu analysieren. Am Beispiel des Operationssystems ‚da Vinci‘ für die roboterassistierte Chirurgie (RAC) wird aufgezeigt, dass die Einführung des Roboters in sozialer, räumlicher, zeitlicher sowie sachlicher Hinsicht eine bedeutende Veränderungsdynamik auf der Mikroebene des Handelns im Operationssaal mit sich bringt. Dies gilt insbesondere für Tätigkeiten, in denen der Einsatz von KI-Tools vorbereitet und erprobt wird.

---

R. Wittal (✉)

Deutsches Zentrum für Luft- und Raumfahrt e.V., DLR Projektträger, Berlin, Deutschland  
E-Mail: [regina.wittal@dlr.de](mailto:regina.wittal@dlr.de)

C. Hettinger

Institut für Sozialwissenschaftliche Forschung e.V., ISF München, München, Deutschland  
E-Mail: [carolyn.hettinger@isf-muenchen.de](mailto:carolyn.hettinger@isf-muenchen.de)

© Der/die Autor(en), exklusiv lizenziert an Springer Fachmedien Wiesbaden GmbH, ein Teil von Springer Nature 2024

M. Heinlein und N. Huchler (Hrsg.), *Künstliche Intelligenz, Mensch und Gesellschaft*, [https://doi.org/10.1007/978-3-658-43521-9\\_3](https://doi.org/10.1007/978-3-658-43521-9_3)

---

**Schlüsselwörter**

Roboterassistierte Chirurgie • Da Vinci • Technisierung • Digitalisierung • Arbeitspraxis • Künstliche Intelligenz • Medizin

---

## 1 Einleitung

In den letzten Jahrzehnten schritt die technologische Entwicklung stetig voran. Im Zuge der Industrialisierung wurde Technik fester Bestandteil in Arbeitsprozessen zahlreicher Sparten, allen voran der Industrie und Produktion sowie der Dienstleistung (Butollo und Nuss 2019; Neuburger 2019; Kleemann et al. 2019). Dieser Prozess der Technisierung und Digitalisierung verursachte tiefgreifende Wandlungsdynamiken in Wertschöpfungsketten, Arbeitsorganisation und Arbeitsprozessen (Cascio und Montealegre 2016; Krzywdzinski et al. 2016; Krzywdzinski et al. 2022). Insbesondere die Weiterentwicklung algorithmischer Systeme, die in KI-Tools Anwendung finden, hat das Potenzial für weitere Veränderungsprozesse (Schuler et al. 2019). Die konkreten Auswirkungen von KI auf Arbeitswelten und Arbeitspraxen sind zum jetzigen Zeitpunkt schwer einzuschätzen, da es an empirischen Studien und genauen Zahlen mangelt und der Diskurs um KI in der Wissenschaft und der Öffentlichkeit sich unter anderem mit Hoffnungen, Erwartungen und Mythen mischt (Giering 2022; Lemke und Monett 2020). Zudem ist die KI-Entwicklung ein dynamisches Feld, das von auf- und abschwingenden Phasen gekennzeichnet ist (Hirsch-Kreinsen 2023). Die jüngsten Entwicklungssprünge lassen durch den weiten Fortschritt in den algorithmischen Methoden von KI-Systemen wie beispielsweise der Datenanalyse und Mustererkennung, der natürlichen Sprachverarbeitung, maschinellem Lernen oder ‚Computer Vision‘ auf Veränderungspotenziale schließen (Bundesministerium für Bildung und Forschung 2022; Neu et al. 2022; Hasenbein 2023; Kitzmann 2022; Schindler 2022).

Blickt man auf die Geschichte der Medizin, so lassen sich auch in diesem Feld Technisierungs- und Digitalisierungstendenzen erkennen (Kramme 2011). Wegweisende Verfahren der Medizintechnik wie der Röntgenapparat, der Computertomograph, die Kernspin- bzw. Magnetresonanztomographie und die Ultraschalltechnik ermöglichten den Blick in den Körper durch die Haut hindurch und bieten so neue Ansätze in der Diagnostik und Behandlung. Die neusten Entwicklungen im Feld der Nanotechnologie (Huss 2019), des 3D-Drucks (Kogelbauer und Krasser 2022) und der Telemedizin (Jedamzik 2019) sowie neuere Erkenntnisse in der Neuromedizin (Mainzer 2018), der regenerativen

Medizin und medizinischen Methoden, z. B. der Operationsrobotik (Brunner et al. 2023), machen deutlich, dass Technik heute standardmäßig als Teil medizinischer Arbeit im Einsatz, in ärztliches Handeln eingebunden und somit bedeutend für die medizinische Praxis ist. Auch KI-basierte Systeme bieten potenziell vielfältige Möglichkeiten für Therapie- und Diagnoseverfahren. So könnten umfangreiche Datenanalysen die personalisierte Medizin voranbringen, algorithmische Bildanalyse zur Entscheidungsunterstützung eingesetzt, die präoperative Planung sowie das postoperative Monitoring erleichtert oder mit KI-gestützten Instrumenten mehr Präzision und Genauigkeit bei Operationen erzielt und menschliche Fehler vermieden werden. Gleichzeitig stellen sich neue Fragen des Datenschutzes und der Datensicherheit, der Zuverlässigkeit und Verantwortung sowie weitere ethische, organisationale, ökonomische, soziale und professionsbezogene Fragen.

Im Jahr 2000 erhielt das robotische Assistenzsystem da Vinci seine Zulassung und ermöglichte damit flächendeckend roboterassistierte Chirurgie (Yates et al. 2011). Das System hat sich insbesondere für urologische, allgemein- und visceralchirurgische und gynäkologische Eingriffe sowie in der Thoraxchirurgie durchgesetzt und ist in zahlreichen Krankenhäusern standardmäßig im Einsatz. KI enthält das Robotersystem bisher nicht; allerdings ist da Vinci ein hochtechnisiertes chirurgisches Werkzeug, welches in ärztliches Handeln eingebunden wird und weitreichende Veränderungen bewirkt (Sergeeva et al. 2018). Zudem werden die Funktionen und Eigenschaften stetig weiterentwickelt. Deshalb eignet sich dieses Fallbeispiel, um das Arrangement von menschlichen und technischen Handlungsträgern am Operationstisch tiefergehend zu analysieren und so besser zu verstehen, welche Folgen und Anforderungen mit der Technisierung von medizinischer Arbeit einhergehen können.

In diesem Beitrag wird auf den komplexen sozio-technischen Praxiszusammenhang mit dem da Vinci System im Operationssaal eingegangen. Die Funktionen und die Art der Einbindung in die Handlungspraxis der Chirurg\*innen gibt Aufschluss darüber, wie diese Praxis funktioniert und in welcher Weise Technik und Mediziner\*innen aufeinander einwirken. Schlussendlich werden bereits vorhandene Potenziale des da Vinci-Systems zur weiteren Digitalisierung betrachtet, was Aufschluss über mögliche zukünftige Szenarien der roboterassistierten Chirurgie und die mögliche Rolle von KI in der Medizin gibt. Als Grundlage für diesen empirischen Bericht dienen mehrere qualitative Interviews mit Expert\*innen aus dem Bereich der ‚KI in der Medizin‘ sowie mit Chef- und Facharzt\*innen aus der Urologie, der Allgemein- und Viszeralchirurgie und der Thoraxchirurgie. Die ärztlichen Gesprächspartner\*innen arbeiten alle selbst mit dem da Vinci-System und sind in mehreren größeren Krankenhäusern innerhalb

Deutschlands beschäftigt. Zudem konnten im Rahmen einer dreitägigen Ethnografie mehrere Eingriffe mit dem da Vinci Roboter beobachtet sowie Anwesende befragt werden.

---

## 2 Minimalinvasive Chirurgie und das da Vinci-Operationssystem

Das da Vinci-Operationssystem wird bei minimalinvasiven Operationen (OP) eingesetzt und ist im Bereich der roboterassistierten Chirurgie (RAC) einzuordnen. Im Jahr 2000 erhielt das System die Zulassung der Food & Drug Administration in den USA (Yates et al. 2011). Seitdem wurden damit unzählige Eingriffe in Kliniken weltweit durchgeführt. Da Vinci ist außerdem ein telematisches System; die Nutzung ist also auch aus der Ferne über das Internet möglich. Die Grundlagen der Entwicklung gingen aus dem Militär hervor, um verwundete Soldat\*innen aus sicherer Entfernung von Expert\*innen weltweit operieren lassen zu können. Später setzte sich der Roboter in der zivilen Medizin durch (Lane 2018). Neben dem da Vinci-System entwickelten sich weitere medizinische Robotik-Typen und -anwendungen (z. B. Pugin et al. 2011; Yates et al. 2011; Kirchberg et al. 2016). Mit steigenden Einsatzzahlen setzte sich jedoch – nicht zuletzt aufgrund der Marketingstrategie des herstellenden Unternehmens Intuitive Surgical – vor allem das da Vinci durch. Um den Mehrwert des Systems zu verstehen, soll vorerst dessen Funktionsweise erläutert werden.

Minimalinvasive Chirurgie, metaphorisch auch ‚Schlüsselloch-Chirurgie‘ genannt, dient als Oberbegriff für Operationsmethoden, welche kleinstmögliche Verletzungen für notwendige Zugänge verursachen. Darunter fällt die laparoskopische Chirurgie, die Thorakoskopie und die Endoskopie. Es wird versucht, durch die natürlichen Körperöffnungen oder minimale Schnitte an das Operationsfeld zu gelangen, um das physische Trauma möglichst gering zu halten. Somit verringert sich die Massivität der Eingriffe, was zu weniger Schmerzen, schnellerer Genesung und damit geringerer körperlicher Belastung führt. Auch bleiben weniger (große) Narben zurück (Feußner und Wilhelm 2016; Zdichavsky et al. 2011).

Bei der laparoskopischen Chirurgie werden für die benötigten Zugänge kleine Schnitte gesetzt, durch die über Trokare (eine Art Metallröhre) mit einer Kamera (Endoskop) und weiteren Instrumenten oder Materialien das Operationsfeld erreicht werden kann. Das Kamerabild wird auf einen Bildschirm übertragen und lässt so den Blick in das Innere des Körpers zu. Bei der herkömmlichen Laparoskopie bekommt der/die Operateur\*in Unterstützung von OP-Pfleger\*innen und/



oder Assistenzärzt\*innen, die das Endoskop halten und führen, Instrumente anreichen, spülen, absaugen, Klammern setzen, Tupfer einführen oder entfernen und Ähnliches mehr. Die Bildgebung ist durch das händische Endoskopieren zumeist unruhig. Außerdem können Operateur\*innen das Kamerabild nicht selbstständig steuern. Die Instrumente sind starr und deswegen in der Beweglichkeit eingeschränkt, weswegen Handgriffe axial ausgeführt und Instrumente physisch gedreht werden müssen.

Bei der roboterassistierten Chirurgie werden ebenfalls Trokare gesetzt, über die das Endoskop, Instrumente sowie andere benötigte Materialien an das Operationsfeld gebracht und wieder entfernt werden können. Der Roboter besteht insgesamt aus drei Teilen: dem Robotercart, an dem vier Arme befestigt sind, der Konsole, an der die Operateur\*innen sitzen sowie einem Tower für den Computer und die Elektronik. Anders als bei der klassischen Laparoskopie sind die Trokare mit dem robotischen Assistenzsystem verbunden; so müssen diese nicht manuell gehalten und können über den Roboter bewegt werden. Der/die Operateur\*in steht nicht mehr am OP-Tisch, sondern sitzt in einem nicht-sterilen Bereich innerhalb des Operationssaals an der Konsole, über die die Kamera und die Instrumente gesteuert werden können (Hillebrandt et al. 2022). Das Bild des Endoskops ist für die Operateur\*innen durch die Konsole und für die anderen Anwesenden über Bildschirme im Operationssaal zu sehen.

In der empirischen Erhebung hat sich gezeigt, dass mit dem da Vinci-Operationssystem einige Veränderungen einhergehen. Um diese Erkenntnisse zu bündeln, bietet sich eine Auswertungssystematik in Tabellenform an (siehe Tab. 1). Zum einen werden verschiedene Ebenen des Wirkens betrachtet: Auf der Mikroebene liegt der Fokus auf den Folgen für die Prozesse im Operationssaal. Die Mesoebene spiegelt die Auswirkungen für die Organisation im Krankenhaus wider. Die Konsequenzen für die Profession der Medizin werden aus der Makroperspektive betrachtet. Mit der Einteilung in soziale, räumliche, zeitliche und sachliche Aspekte lassen sich die Effekte zum anderen inhaltlich bündeln. In diesem Praxisbericht soll es um die Mikroebene gehen, da sich hierbei am deutlichsten die Veränderungen der technisierten Arbeitspraxis aufzeigen. Die organisationale und professionelle Perspektive betrifft ebenfalls Anteile von Arbeit, die mit der tatsächlichen Tätigkeit in Wechselwirkung stehen, jedoch das Tun und Handeln der Akteure nicht ausreichend beschreiben können. Aufgrund dessen liegt hier der Fokus rein auf der Praxis im Operationssaal. In der untenstehenden Tabelle sind die Faktoren zusammengefasst, welche sich durch und mit da Vinci verändern. Im folgenden Kapitel wird auf die grau hinterlegten Faktoren konkreter eingegangen, wobei die adressierten Gesichtspunkte mit Auszügen aus

den Interviews belegt und verdeutlicht werden und so herausgearbeitet werden soll, wie da Vinci als technisches System beschaffen ist und wirkt.

---

### **3 Das da Vinci-System im Operationssaal**

Bei der robotischen Chirurgie steht bisher nicht die Automatisierung oder das Ersetzen von Arbeitstätigkeiten im Zentrum, sondern das Erleichtern der Arbeitspraxis und das Erweitern bzw. Verbessern der chirurgischen Fähigkeiten. Insbesondere das technisch vermittelte Sehen durch die Konsole, die Steuerung nach dem Master-Slave-Prinzip und die Kombination aus mechanischer Hardware und digitaler Software zeichnen das System aus. Im medizinischen Setting verändert der Einsatz von da Vinci als technisches System die Dynamik im OP-Saal und das Handeln der einzelnen Personen(gruppen); beispielsweise erfordert die RAC „new surgical and technology-related skills, changes in teamwork, improved utilization of available resources, and coordination of all critical elements to minimize risk and maximize performance“ (Catchpole et al. 2019, S. 11). Die technischen Anteile im System bestimmen die Nutzungsweise und den Interaktionsmodus, indem die unterschiedlichen Berufsgruppen da Vinci in ihr Handeln einbeziehen. Mit der Betrachtung der vier Dimensionen mit Fokus auf dem Geschehen im Operationssaal kann nachvollzogen werden, wie die Auswirkungen von da Vinci als technisches Artefakt auf die medizinische Arbeitspraxis verstanden werden können. Dazu werden die Eigenschaften des Systems und die einhergehenden Veränderungen genauer analysiert und eingeordnet. So werden z. B. das Wissen um die technischen Aspekte und Funktionen des Systems für die richtige Nutzung wichtig (Freschi et al. 2013, S. 396) und die Zukunft der roboterassistierten Medizin durch „enhanced manipulation capabilities, refined sensors, advanced vision, task-level automation, smart safety features, and data integration“ auf den Weg gebracht (Haidegger et al. 2022, S. 835).

#### **3.1 Die soziale Dimension im Operationssaal**

Ein bestimmender Faktor für den Operationsverlauf ist die Kommunikation und Zusammenarbeit im OP-Saal (Gill und Randell 2016; Randell et al. 2015; Cunningham et al. 2013). Die leitende Funktion hat der/die Operateur\*in inne. Ein Oberarzt der Hals-Nasen-Ohrenheilkunde schildert dies folgendermaßen:

**Tab. 1** Einhergehende Veränderungen mit dem da Vinci-System

	<i>sozial</i>	<i>räumlich</i>	<i>zeitlich</i>	<i>sachlich</i>
<i>Mikro (OP)</i>	<ul style="list-style-type: none"> <li>• Kommunikation, Zusammenarbeit</li> <li>• Verhältnis zwischen Arzt/Ärztin und Patient*in</li> <li>• Soziale Zuschreibungen an den OP-Roboter (Bilder, Akzeptanz/Vertrauen, eigene Identität)</li> </ul>	<ul style="list-style-type: none"> <li>• Anordnung im OP-Saal</li> <li>• Raumverteilung</li> <li>• verdoppelter Raum durch endoskopisches Bild</li> <li>• Virtualisierung und Immersion</li> <li>• Datenraum</li> </ul>	<ul style="list-style-type: none"> <li>• Phasen der Operation</li> <li>• Andere Zeit limitierende/erweiternde Faktoren</li> <li>• Verfügbarkeit über den Eingriff hinaus</li> <li>• Verändertes Zeitregime des Lernens</li> </ul>	<ul style="list-style-type: none"> <li>• Körperliche Voraussetzungen &amp; Anatomie</li> <li>• Sinnlichkeit, Wahrnehmung</li> <li>• Anforderungen während der Operation</li> <li>• Veränderung des Arbeitsprozesses</li> <li>• Datenerhebung</li> </ul>
<i>Meso (Org.)</i>	<ul style="list-style-type: none"> <li>• Aufgabenverteilung (Organisation unter den Beteiligten, Rollen)</li> <li>• Unentbehrlichkeit von Ärzt*innen</li> <li>• Unterstützung der Führungsebene</li> </ul>	<ul style="list-style-type: none"> <li>• (nicht) steriler Bereich im OP-Saal</li> <li>• Belegungsmanagement</li> </ul>	<ul style="list-style-type: none"> <li>• Effizienz der Eingriffe</li> <li>• Situativer Umgang mit Unvorhergesehenem</li> </ul>	<ul style="list-style-type: none"> <li>• Finanzieller Gewinn und Konkurrenzfähigkeit</li> <li>• Umgang mit Daten, Datenmanagement</li> <li>• Öffentliche Wahrnehmung</li> </ul>
<i>Makro (Prof.)</i>	<ul style="list-style-type: none"> <li>• Figur und Rolle der Ärzt*innen</li> <li>• Status in der Gesellschaft</li> <li>• Gefahren der Deprofessionalisierung</li> <li>• Involvierte Akteure</li> <li>• Technik- und Menschenbilder von Patient*innen</li> <li>• Akzeptanz und Widerstände in der Chirurg*innen-community</li> </ul>	<ul style="list-style-type: none"> <li>• Ausbildungs- und Trainingszentren</li> <li>• Standortdiversifizierung</li> </ul>	<ul style="list-style-type: none"> <li>• Weiterbildung</li> <li>• Ausbildungscurricula und Karrierewege</li> <li>• Beschleunigung medizinischen Wissens</li> </ul>	<ul style="list-style-type: none"> <li>• technivermittelter Gegenstand der Medizin</li> <li>• Anreicherung mit neuem Wissen und Fähigkeiten</li> <li>• (technische) Erweiterung ärztlicher Fähigkeiten und Fertigkeiten</li> </ul>

*„Der, der das Ruder in der Hand hat, der das letzte Wort hat, das bin dann schon immer ich [die Operateurin]. Und es ist immer ein Assistent natürlich dabei oder zwei, [...]. [Dann] gibt es natürlich den Narkosearzt, mit dem man sich dann noch verständigen muss, aber es gibt immer einen, der vorgibt, wie es läuft.“*

Für den reibungslosen Verlauf während der Operation ist es wichtig, dass jede\*r weiß, was er oder sie zu tun hat. Neben den Chirurg\*innen und den Assistenzärzt\*innen ist die Anästhesie, die Operations-technische/-medizinische Assistenz und teilweise die Krankenpflege anwesend. Ähnlich einem/r Dirigent\*in im Orchester gibt die Person an der Konsole die notwendige Orientierung und leitet die einzelnen Schritte und Tätigkeiten ein. Mit der Entfernung zwischen Konsole und Operationstisch verändert sich die Nähe zum weiteren Personal. Bei offenen oder laparoskopischen Eingriffen stehen alle Beteiligten direkt nebeneinander am Operationstisch. Bei der robotischen OP hingegen sitzen die Operateur\*innen separiert an der Konsole. Der/die Operateur\*in hat eine bessere Einsicht in das OP-Feld, da die Konsole das Bild des Endoskops vergrößert und dreidimensional überträgt. Alle anderen können das Bild auf Bildschirmen im Saal verfolgen. Ist der Blick in die Konsole gerichtet, kann jedoch das Geschehen im Saal nicht überblickt werden. Um zu erfahren, was im Saal geschieht, müssen die Operierenden also entweder ihre Arbeit unterbrechen und den Kopf aus der Konsole heben oder sie sind darauf angewiesen, dass ihnen Vorgänge durch das OP-Team verbal mitgeteilt werden:

*„das ist dann die Aufgabe des Assistenten, und der guckt natürlich die ganze Zeit, [...] dass im Prinzip der Roboter den Patienten natürlich nicht gefährdet. Das ist auch ein wirklich wichtiger Job, also dass man im Prinzip immer guckt, ist nicht so, dass jetzt der Roboterarm auf den Patienten drückt, das sieht ja der Chef nicht, der operiert. Da muss man sagen [...] passen Sie mal auf, [...] Sie drücken hier mit Ihrem [...] Einser-Arm, dann weiß der okay, das ist mein rechter Arm, Sie drücken gerade mit dem Einser-Arm auf das Becken des Patienten oder Sie gucken gerade mit dem Einser- und dem Zweier-Arm in die gleiche Richtung, und deswegen, wir nennen das clashen, also d.h., wenn die Roboterarme oben sich sozusagen außen, außerhalb des Körpers behindern, dann Achtung, Chef, Sie clashen zwischen Arm 2 und Arm 3, [...] und dann kann man dem Operateur den Roboter auch im Prinzip wieder wegnehmen, [...] kann man die Arme nehmen und verstellen sozusagen dahin, wo es dann hoffentlich wieder geht. [...] umso besser der Operateur ist, sage ich jetzt mal und umso standardisierter der Eingriff ist, umso weniger hat der Assistent zu tun, und umso neuer der Eingriff ist, umso komplexer der Eingriff ist, umso unerfahrener der Operateur ist, umso mehr hat der Assistent sicher auch zu tun.“*

Die befragte Oberärztin der Urologie schildert sehr genau, dass relevante Ereignisse wie ungewollter Druck auf Patient\*innen durch einen Roboterarm oder das

Kollidieren zweier Roboterarme außerhalb des Körpers negative Folgen haben, aber durch die Operateur\*innen nicht bemerkt werden können. Sie sind auf die Reaktion und den Hinweis durch die Assistenz angewiesen. Darüber hinaus wird Bezug genommen auf die Kompetenz der Operateur\*innen. Je mehr Erfahrung der/die Operateur\*in mitbringt, desto eher hat er/sie ein Gefühl dafür, was es bedeutet, die Instrumente zu bewegen. Werden die Instrumente angesteuert, bewegen sich gleichzeitig auch die Roboterarme außerhalb des Körpers, da beides miteinander verbunden ist. Mit zunehmender Erfahrung verbessert sich auch das Vorstellungsvermögen, das Wissen um die konkreten, zusammenhängenden und weiterreichenden Folgen des eigenen Handelns sowie das Orientierungsvermögen im Körper der Patient\*innen. Das Endoskop lässt sich in alle Richtungen drehen und steuern; die Chirurg\*innen müssen also selbstständig den Überblick behalten, wo im Körper sie gerade die Instrumente bewegen. Der Erfahrungswert bestimmt somit auch die Abhängigkeit zur Assistenz mit: ist der/die Operateur\*in in der Lage, beispielsweise die Kollision der Roboterarme außerhalb des Körpers durch entsprechende Steuerung zu vermeiden, ist er/sie nicht mehr auf die Hinweise der Assistenz zur Orientierung angewiesen.

Weiter beschreibt die Oberärztin, dass bei eingespielten Teams auch vereinfachte Verständigungslaute ausreichend sein können:

*„wenn wir mal in die Situation kommen, dass mein Chef und ich operieren, sage ich immer, wir kommunizieren in so Gutturallauten so. Hm. Hm. Hm. Hm. Also wir machen das so lange schon zusammen, dass wir eigentlich – der sagt nicht Nadel rein oder bitte hier saugen oder so, das sagt der einfach nicht mehr. Das weiß ich, dass er das will. [...] das dauert ein bisschen, bis man an den Punkt kommt. Ich muss jetzt, wenn ich mit den Assistenten operiere, immer sagen: Könntest du bitte mal [...] dein Endo Dissect nehmen und mal die Blase nach links halten. So. Dann macht der Assistent das netterweise. [...] Also ich habe ja ein Mikro in der Konsole, und das wird in den OP-Saal übertragen, d. h. selbst, wenn ich da in diesem komischen Kopfteil bin und da in meinen Bart nuschele sozusagen, dann hören die das trotzdem [...] Also wir versuchen relativ viel tatsächlich gerade in den Anfangs- und Lernphasen miteinander zu reden, damit halt natürlich der Patient nicht gefährdet wird.“*

Hier wird deutlich, dass zwischen neuen und erfahrenen Teams ein Unterschied besteht. Das Verhältnis zwischen der befragten Person und ihrem Chef ist von Vertrautheit und gemeinsamer Routine geprägt. Das führt dazu, dass die Kommunikation nicht mehr in Worten stattfinden muss, sondern Geräusche ausreichend sind, um zu verstehen. So verstärkt sich während der intensiven Operationsphase die Unabhängigkeit der Operierenden, die sich auf ihre Assistenz verlassen können. In der Ethnografie konnte beobachtet werden, wie zusätzlich auch nonverbale Kommunikationsformen, z. B. die Stellung der Instrumente, genutzt wurden.

In der Art und Weise, wie die Operateur\*innen ihr OP-Besteck zum Stillstand brachten, konnten die Assistenzen ablesen, welcher Schritt von ihnen erwartet wurde. Im Verhältnis dazu wird im Zitat beschrieben, dass die Kommunikation mit lernenden Assistenzen mehr konkrete Anweisungen erfordert. Der verbale Austausch wird in den direkten Zusammenhang mit einer Gefährdung der Patient\*innen gesetzt, was verdeutlicht, wie wichtig der Kontakt zwischen den beiden Personen ist. Da kein Blickkontakt zu den Personen besteht, müssen die fehlenden Informationen verbal ausgetauscht werden; im Falle der RAC vermittelt durch ein Mikrofon. Es zeigt sich, dass der Roboter in die Kommunikation im OP-Saal integriert wird und für die Verständigung untereinander als Instrument funktionieren kann. Außerdem verhindert die bildgebende Funktion für die Operateur\*innen den direkten Blickkontakt, ist jedoch notwendig, um die technischen Vorzüge von da Vinci – ein ruhiges, vergrößertes Bild in 3D, auszuschöpfen. Alle Beteiligten müssen sich dieser Eigenschaften bewusst sein, um sie in ihr Handeln und den Austausch untereinander einbeziehen und so als OP-Team funktionieren zu können.

Ein weiterer sozialer Aspekt, der sich mit dem Einsatz von da Vinci im OP-Saal verändert, ist das Verhältnis zwischen Arzt/Ärztin und Patient\*in. Im Idealfall ist die Beziehung von Vertrauen geprägt (Baumgart 2010). Die zu Behandelnden kommen als hilfeschuchende, bedürftige Personen zu den Behandelnden, die als Expert\*innen die Kompetenz und das Wissen haben, zu helfen und im besten Fall zu heilen. Es tut sich somit ein hierarchisches Verhältnis auf, dass sich mit zusätzlichen Informationsquellen wie Fachforen im Internet, Zeitschriften, Beiträgen in den sozialen Medien u.Ä. allerdings bereits verändert (ebd.). Der Einsatz des Roboters beeinflusst das Vertrauensverhältnis ebenfalls, indem er die Wahrnehmung von und die sozialen Zuschreibungen der Patient\*innen verändert, z. B. die Wahrnehmung der Chirurg\*innen als besonders fortschrittlich und kompetent, da sie die modernste Operationsweise beherrschen, was das Vertrauen in die Fähigkeiten bestärken kann. Das Vertrauen oder Misstrauen in die Technologie ist somit ebenfalls entscheidend und kann in die Vertrauensdynamik mit den Chirurg\*innen hineinspielen. Beispielsweise können sich die Rollenerwartungen an die Chirurg\*innen hin zu einer die Technik überwachenden, passiven Rolle verändern. Technikmisstrauen könnte so zu einem Vertrauensverlust führen. Wichtig ist hierbei z. B. die präoperative Kommunikation: Möglichkeiten und Grenzen des robotischen Systems sowie Fragen und Bedenken sollten offen besprochen werden.

Darüber hinaus ist die soziale Zuschreibung der Operierenden gegenüber dem System ein entscheidender Aspekt. Im Feld stellte sich heraus, dass diese Attribuerungen häufig durch Metaphern ausgedrückt werden, welche den Bezug zu

einer weithin geteilten Lebenswelt herstellen und bildhaft verdeutlichen. Diese Zuschreibungen drücken eine Betrachtungsweise des Roboters aus, welche auch die Nutzungsweise mitbestimmt. Ein befragter Oberarzt aus der Urologie macht mit dem Vergleich mit der Formel 1 deutlich, dass das System für ihn ein Mittel darstellt, um möglichst schnell voranzukommen:

*„Man muss sich das so ein bisschen vorstellen wie bei der Formel 1, da muss einfach jeder Handgriff sitzen, dann geht es schnell.“*

Er sieht sich selbst als den Fahrer des Wagens, um den herum alles Notwendige rasch getan werden sollte, damit er die Bestzeiten einfahren kann. Gleichzeitig nimmt er Bezug auf die Erwartungen an das weitere Personal, ohne das er sein Ziel nicht erreichen könnte: jede einzelne Aktion muss präzise ausgeführt werden, alle Abläufe müssen perfekt ineinandergreifen. In seiner Vorstellung treten er und sein Team gegen die Zeit an. Es lässt sich vermuten, dass mit der Absicht, Eingriffe möglichst zeitnah zu beenden auch das quantitative Ziel einhergeht, möglichst viele Operationen in kurzer Zeit zu schaffen. Der Anspruch an die Qualität des Zusammenwirkens und des Operationsergebnisses ist trotzdem hoch.

Des Weiteren macht folgendes Zitat deutlich, dass das da Vinci-System für den befragten Oberarzt der Erweiterung und Verbesserung der eigenen Leistungsfähigkeit dient:

*„Ja, also da Vinci muss man sich vorstellen wie in diesem Manga-Film, die diese Cyborg-Anzüge haben, so kann man sich das vorstellen. Ich kann damit körperliche Schwächen ausgleichen und ein bisschen besser werden in dem Bereich.“*

Das System kompensiert mögliche Schwächen und stärkt gleichzeitig die eigenen Fähigkeiten. In Bezug auf popkulturelle Darstellungen sieht der Befragte seinen eigenen Körper bestückt und im Einklang mit dem Roboter. In seiner Bildsprache verschmelzen Körper und System zu einem und agieren harmonisch und verbunden. Sein Team kommt in diesem Bild nicht vor. Lediglich er selbst profitiert vom Roboter als Erweiterung seiner selbst und seiner Fähigkeiten.

Im Gegensatz zu dem befragten Oberarzt, bezeichnet eine Oberärztin der Urologie das System ganz klar als Werkzeug:

*„Also er [der Roboter] hat keinen Namen, er ist auch nicht mein Freund, er ist ein Werkzeug. Für mich ein Werkzeug, genauso wie ich auch nicht zu den Leuten gehöre, die sich in Siri verlieben, ist es so, also, das ist ein Werkzeug. Es ist das Gleiche wie eine bipolare Schere für mich oder ein Retraktionsgerät, das ist ein Werkzeug, es ist eine Bosch Bohrmaschine oder so. Also das ist ein Ding ohne Herz so, ein Gegenstand.“*

*[...] Wenn der eine Schramme hat, tut mir das nicht weh so. Das ist ein Werkzeug für mich, was mir hilft, meine Arbeit gut zu machen und mehr ist das nicht. Und, wie gesagt, wenn es mehr kann demnächst, freue ich mich, wenn es meine Arbeit noch angenehmer machen kann [...].“*

Da Vinci stellt für die Oberärztin ein Objekt dar, dessen sie sich bedienen kann, um ihre Aufgabe auszuführen. Der Vergleich mit anderen (medizinischen) Werkzeugen verdeutlicht, dass sie keinen emotionalen Bezug zum System hat. Sie spricht zudem explizit an, dass es für sie keine vermenschlichte Rolle einnimmt, sondern ein Artefakt darstellt. Zudem grenzt sie da Vinci als Objekt klar vom Menschen ab, der ein Herz besitzt. Den Mehrwert sieht sie in der Unterstützungsleistung beim Ausführen ihrer Tätigkeit und begrenzt den Vorteil klar. Mit einer zukünftigen Erweiterung von Funktionen würde sich diese Betrachtungsweise nicht ändern. Stattdessen würde sich die Vielfalt der Einsatzzwecke und damit die Nützlichkeit des Werkzeugs erweitern. In Kontrast dazu beschreibt die Oberärztin den Umgang ihrer Vorgesetzten mit da Vinci:

*„Aber das ist ein Ding, ja, und das Ding kann nichts ohne mich, [...] also ich weiß, mein erster Chef, der hat ihm dann irgendwie mit Edding [...] einen kleinen Namen auf die Seite geschrieben, und dann mein zweiter Chef hat alle berühmten Chirurgen, die dann vorbeigekommen sind, auf dem Roboter unterschreiben lassen so, wie auf einem Arm so. [...] Wenn dann irgendjemand da war, der dann robotisch voroperiert hat, der hat das unterschreiben lassen und so. Ist auch irgendwie ganz lustig, aber nein.“*

Der erste Chef gibt dem Gerät einen Namen und personifiziert es damit. Das erweckt den Eindruck, dass er da Vinci nicht nur als “Ding” ansieht. Ähnlich einem Wettbewerbsgewinn oder einer Trophäe stellt der zweite Chef da Vinci zur Schau und lässt andere Operateur\*innen darauf unterschreiben. Dies kann als Ausdruck von Stolz verstanden werden: Stolz über den Besitz des Systems oder Stolz auf die eigene Fähigkeit, mit dem System umzugehen oder Stolz, dass andere fähige Chirurg\*innen mit diesem Gerät ihre Arbeit getan haben. Die Befragte grenzt am Ende des Zitates ihre eigenen Zuschreibungen zu denen ihrer Vorgesetzten ab.

In einem weiteren Statement von einer befragten Urologin bringt diese zum Ausdruck, dass sie sich über ihre Erfahrungen mit dem System anfreunden musste:

*„Am Anfang war es für mich wie ein Behinderungsinstrument. Wie eine Behinderung. Als ob zwischen mir und dem Organ jemand stehen würde, der es mir echt erschweren möchte. [...] erst so ganz langsam durchs Üben, Üben, Üben habe ich gemerkt, ah ja,*



*[...] es ist eigentlich mein Freund, auch wenn ich noch keine langjährige Freundschaft geschlossen habe, aber ich empfinde das schon, das ist jetzt nicht ein Teil von mir oder irgendwie, wenn ich sage, Mensch, das ist meine Vergrößerung oder das ist jetzt meine Hand. Also so close sozusagen sind wir noch nicht, aber mir macht es Spaß, da dran zu operieren in jedem Fall, ich genieße die Vergrößerung und die feine Präparation, das ist super.“*

Anfangs empfindet sie die da Vinci als Störfaktor, der es ihr aktiv erschwert, ihre Arbeit ungehindert auszuführen und sich (als technisches Artefakt) zwischen sie und den/die Patient\*in schiebt. Mit wachsender Routine wird der Roboter zu einem „Freund“ und sie lernt die Vorzüge zu nutzen und zu schätzen; sie erkennt den Mehrwert an. Sie grenzt das System klar von sich selbst ab und sieht es nicht als Teil ihrer selbst, sondern als eigenständigen Gegenstand. Damit steht auch fest, dass sie den aktiven und das System den passiven Part hat, sie jedoch unterstützt und ihre Tätigkeit erleichtert.

Die unterschiedlichen Zitate zeigen auf, dass Chirurg\*innen, die mit dem da Vinci-System operieren, auch unterschiedliche Zuschreibungen an das Gerät richten. Das bedeutet: obwohl die Funktionen und die grundsätzlichen Voraussetzungen in der RAC mit da Vinci für alle gleich sind, gibt es in der Arbeits- und Interaktionsweise mit dem System teils starke Unterschiede.

Zusammenfassend lässt sich sagen, dass die Einführung des da Vinci-Systems in die Arbeitspraxis von Chirurg\*innen in sozialer Hinsicht einige Veränderungen bewirkt. So wandeln sich die Kommunikation, Zusammenarbeit und Hierarchie im OP-Saal, das Verhältnis zwischen Chirurg\*innen und Patient\*innen sowie die sozialen Zuschreibungen gegenüber dem Gerät.

### **3.2 Die räumliche Dimension im Operationssaal**

Eine weitere Dimension, welche die Praxis im Operationssaal beeinflusst, ist die räumliche Gegebenheit. Dabei geht es nicht nur um das euklidische Raumkonzept, das sich durch Dreidimensionalität auszeichnet, in der Distanzen und Winkel fest definiert sind. Vielmehr kann sich Räumlichkeit auch auf andere Perspektiven beziehen, die eine unterschiedliche Auffassung von Raum und Umgebung haben oder sich mit nicht-physischen Aspekten der Räumlichkeit befassen, z. B. das psychologische, mentale Erleben und Konstruieren von Raum.

Die offensichtliche Veränderung im Operationssaal ist, dass mit da Vinci drei große physische Artefakte hinzukommen, die flexibel beweglich sind (Kanji et al. 2021). Das Robotercart mit seinen vier Armen wird an einer Seite des OP-Tisches

platziert, hängt über den Patient\*innen und nimmt dort Raum ein. Der Computertower dient der Datenübertragung und benötigt einen Ort zwischen dem Roboter und der Konsole. In der Regel ist am Tower ein Bildschirm zur Beobachtung des endoskopischen Bildes befestigt, das den Körperraum, in dem die Chirurg\*innen arbeiten, aus dem Körper heraus nach außen sichtbar macht. Die Konsole, von der aus die Chirurg\*innen operieren, benötigt einen dauerhaften Platz im Saal, der gut zugänglich ist. Außerdem braucht der Roboter eigene Instrumente und muss steril abgedeckt werden. Somit verändert sich die Anordnung im OP-Saal im Vergleich zur konventionellen Operationsmethode, bei der die zusätzlichen Gerätschaften nicht gebraucht werden. Zumeist sind die Operationssäle nicht von vornherein für die RAC konzipiert worden, sondern werden im Nachgang angepasst. Das kann eine andere Stellung des OP-Tisches, neue Bestückung und Ausführungen der Instrumentenwägen, andere Bedarfe von Materialvorräten u.Ä. bedeuten. Die benötigten Tools und die Positionen der einzelnen Gegenstände verändern sich also und die Artefakte bleiben veränderbar, wobei teilweise Abhängigkeiten bestehen. Mit einer veränderten Lage des OP-Tisches muss z. B. auch die Anästhesie mit ihren Gerätschaften den Platz wechseln. Die hinzugekommenen Artefakte nehmen Raum ein, bestimmen damit die räumlichen Strukturen im Operationsgeschehen mit und beeinflussen somit auch die Abläufe des Prozesses und die Tätigkeiten der Beteiligten: Laufwege, Erreichbarkeiten, Kommunikationsmöglichkeiten und die Verteilung der Handlungsschritte ändern sich. Damit werden die vorherigen Standards durch neue Routinen abgelöst.

Gleichzeitig entstehen neue unsichtbare räumliche Grenzen innerhalb des OPSaals. Etwa einen Meter um den OP-Tisch herum besteht die sterile Zone, die nur von entsprechend gekleideten Personen betreten werden darf. Auch der Teil des Roboters, der sich direkt am und über dem OP-Tisch befindet, wird mit sterilen Überzügen abgedeckt. Dieser Bereich wird sensibel geschützt; es dürfen nur sterile Gegenstände eingesetzt werden und ein Lüftungssystem reinigt die Luft möglichst keimfrei. Der restliche Raum bleibt unsteril, was bedeutet, dass auch die Konsole im nicht-sterilen Bereich steht. Die Operateur\*innen können nicht einfach aufstehen und an den OP-Tisch herantreten, da sie nicht steril eingekleidet sind. Durch die verbesserte Sicht in der dreidimensionalen Bildübertragung kommen sie einerseits sehr nahe an die Patient\*innen heran, tauchen geradezu in sie hinein. Andererseits befinden sie sich bei der RAC physisch weiter weg von den Patient\*innen als bei der laparoskopischen oder offenen Operation. Die sterile Hürde kann mit einigem Aufwand überwunden werden: nach dem sterilen Einkleiden darf die Grenze überschritten werden. In Notfällen beispielsweise kann dieser Aufwand jedoch entscheidend für die Bewältigung sein.

Nebstdem findet durch den Einsatz eines Endoskops eine technische Verdopplung des Raumes statt. Einerseits existiert der physische Körper, der durch das Hinsehen erfasst werden kann. Das Operationsfeld kann allerdings mit bloßem Auge nicht betrachtet werden, der Blick reicht nicht bis durch die Bauchdecke hindurch. Andererseits gibt es die Aufnahme des Endoskops, welche diesen Anblick ermöglicht: das Instrument wird in den Körper hineingeführt und erstellt dort die Aufnahmen, welche für alle Anwesenden zugänglich sind. Es digitalisiert somit die physischen Strukturen und transportiert das Bild aus dem Körper heraus. Das Personal im OP-Saal kann das übertragene Bild über die Bildschirme verfolgen. An der Konsole können die Chirurg\*innen durch das 3D-Bild zusätzlich die Verhältnisse, Tiefen, Relationen und Bezüge zwischen den einzelnen Gewebestrukturen erkennen. Zudem wird das digitale Abbild mit einer Vergrößerungsmöglichkeit, augmentierten Informationen und der Möglichkeit der Verzögerung beeinflussbar. Diese zusätzlichen Informationen und Einstellungen sind auf den Bildschirmen nicht verfügbar. Die Schnittstelle zwischen dem analogen und den digitalen Operationsfeld bildet die endoskopische Kamera. Blutungen, welche die Kameralinse verschmutzen oder Organe, die im Blickwinkel der Kamera hängen, können beispielsweise diese Schnittstelle stören. Es braucht physische Handlungsschritte, z. B. das Reinigen der Linse oder das Beiseitehalten der Organe, um den Blick wieder frei zu machen und den digitalen Raum in gewünschter und notwendiger Weise aufrecht zu erhalten. Außerdem müssen die Chirurg\*innen die Führung des Endoskops beherrschen, um das digitale Abbild so zu erzeugen, dass es die Bedarfe erfüllt.

Mit der erweiterten digitalen Kopie des Operationsfeldes werden darüber hinaus neue Dimensionen und Raumerfahrungen geschaffen: der Körper der Patient\*innen wird für die Chirurg\*innen virtualisiert. Durch die Konstruktion der Konsole sowie die Perspektive und die Steuerungsmöglichkeit hat diese Virtualisierung einen stark immersiven Effekt. Ähnlich einer Virtual Reality Brille wird die Person in der Konsole von der Außenwelt und Umgebungseindrücken abgeschirmt. Das virtuelle Abbild verhält sich nach den gewohnten physikalischen Gesetzmäßigkeiten – handelt es sich doch um ein Echtzeitbild des behandelten Körpers, zeigt sich durch die Konsole jedoch anders auf und hat seine eigenen virtuellen Gesetzmäßigkeiten. In der Regel operieren die Chirurg\*innen so, dass die Instrumente im unteren Drittel des Bildschirmrandes zu sehen sind. Da immer nur zwei Arme gleichzeitig gesteuert werden können, erinnert die Perspektive an die „first-person“-Perspektive in Computerspielen. Insofern und mit den augmentierten Informationen, z. B. der Lebensdauer der Instrumente, enthält die digitale Gestaltung der Bildübertragung des da Vinci Elemente, die einer Gamification nahekommen. Ähnlich wie im PC-Spiel entsteht ein immersiver Effekt,

der die Operateur\*innen in den Körper und digitale Abbild dessen eintauchen lässt, eine tiefe Involviertheit erzeugt und ein Gefühl von Präsenz entstehen lässt. Effekte können eine erhöhte Konzentrationsfähigkeit, Aufmerksamkeit und auch ein Gefühl von Verbundenheit sein, in dem man sich in der Aktivität, Erfahrung oder Umgebung verliert. Diese Virtualisierung und Immersion hat einen Effekt auf die Steuerungsmöglichkeit der Instrumente und die Handlungsmöglichkeiten der Chirurg\*innen. Auch kann dies Auswirkungen auf die Arbeitsbelastung, das Erleben der eigenen Handlungsfähigkeit und Rolle sowie den Bezug zu Patient\*innen und deren Körpern mit sich bringen.

Eine weitere, neue Form des Raumes, ist der entstandene Datenraum. Während des Eingriffs werden zahlreiche Parameter wie die Dauer der Operation, die Bewegungszeiten oder die Anzahl der Einsätze der einzelnen Instrumente erfasst. Zudem werden die Operationen zu Zwecken der Schulung, Qualitätssicherung, Nachsorge und für Lernzwecke der Chirurg\*innen selbst aufgezeichnet. Die aufgezeichneten Daten können von den Operierenden in einer App eingesehen, ausgewertet und gezielt dokumentiert werden. Somit entsteht ein unsichtbarer Raum, in dem das Geschehen im OP-Saal digital in Form von Daten abgebildet wird. Intuitive, der Hersteller des da Vinci-Systems, hat nach Freigabe ebenfalls Zugang auf erhobene Daten, die den OP-Saal verlassen. Das System lässt sich jedoch ohne die Datenfreigabe nicht nutzen und es wird nicht transparent gemacht, welche Daten in welchem Zeitraum erhoben oder wofür sie verwendet werden. Intuitive schafft sich damit einen eigenen Datenraum, dessen Dimensionen undurchsichtig sind und der für die Operierenden und das Krankenhaus unverfügbar ist. Dies verunsichert die Befragten häufig. Es zeigt sich, dass die Chirurg\*innen keine genaue Kenntnis darüber haben, was mit den Daten geschieht:

*„[...] aber ob Intuitive die Daten auch hat und ob die nicht unabhängig davon sozusagen ohne dass wir´s auch wissen aufzeichnen, keine Ahnung. Ich persönlich [...] glaube ja. Ich persönlich glaube, dass die hunderte und aberhunderte oder hunderttausende von OP-Videos sozusagen haben und versuchen, da irgendwelche Algorithmen draus zu ziehen [...]“.*

Der befragte Chefarzt an einer Klinik für Allgemein- und Viszeralchirurgie kann die Frage nach der Datenaufzeichnung nicht aufgrund von Vorschriften, Vereinbarungen oder Regularien beantworten, sondern vermutet, dass der Hersteller Zugang zu den Daten erhält und diese auch weiterverarbeitet. Inwiefern und zu welchem Zweck kann er nicht sicher sagen. Ein weiterer befragter Oberarzt der

Urologie kommt ebenfalls zu dem Schluss, dass vermutlich Daten abfließen, kann aber nicht konkret sagen wohin oder wozu:

*„Die können ja auch die Daten auswerten ohne Ende. Die können natürlich ... also das weiß ich jetzt nicht so, aber natürlich haben die da Interesse dran. Ich glaube, wir sind offline, wir haben es nicht eingestellt, aber rein theoretisch können die die Daten abfragen. Die können gucken, wer operiert. Die können gucken, was ... natürlich sind das Metadaten ohne Ende. Die können wahrscheinlich ... ich weiß nicht, was sie damit anstellen können.“*

Er äußert die Überzeugung, dass Zugang zu den Daten besteht. Zudem betont er den Wert der Informationen, die sich so ermitteln lassen. Gleichzeitig weist er das Problem vom sich, in dem er feststellt, dass der Assistenzroboter in seinem Krankenhaus offline betrieben wird. Einhergehend scheint für ihn klar, dass sich dann keine Daten erheben lassen. Ob das wirklich so ist, weiß er jedoch nicht sicher.

Zusammenfassend lässt sich sagen, dass die Einführung des da Vinci-Systems in die Arbeitspraxis von Chirurg\*innen in räumlicher Hinsicht einige Veränderungen bewirkt. Jenseits der Veränderungen im euklidischen Raum – das Hinzukommen des physischen Roboters sowie zusätzlich benötigten Material und eine andere Raumaufteilung – entstehen neue, nicht sichtbare Grenzen im Saal, welche die Praxis beeinflussen. Außerdem verdoppelt das endoskopische Bild den Raum im Körper, was neue Arten und Möglichkeiten der Raumerfahrungen, vor allem für die Chirurg\*innen, bewirkt. Dabei wird der Körper der Patient\*innen in neuer Weise virtualisiert. Zudem entsteht mit da Vinci ein Datenraum, der neue Informationen verfügbar macht, welche über die physischen Raumgrenzen hinweg im Digitalen den OP-Saal verlassen, sowie weitere Akteure außerhalb einbindet und Zugang gewährt. Im Zuge dessen entstehen neue (In-)Transparenzen sowie (Un-)Verfügbarkeiten von Daten. Außerdem zeigt sich eine Tendenz der Virtualisierung durch die gesteigerte Datenerhebung sowie die vollständige Digitalisierung und eine virtuelle Erweiterung des Sehens. Somit wird deutlich, dass auch die räumliche Dimension in ihren Facetten die Praxis der RAC mitbestimmt und in die Gestaltung zukünftiger Systeme einbezogen werden sollte.

### **3.3 Die zeitliche Dimension im Operationssaal**

Die größte Veränderung in zeitlicher Hinsicht entsteht in Bezug auf die einzelnen Phasen der Operation. Der Einsatz des da Vinci-Systems erfordert zwei neue Schritte, die Zeit in Anspruch nehmen: das Andocken und Abdocken der

Roboterarme an die Patient\*innen. Außerdem werden die Vorbereitung und die Nachbereitung der Operation umfänglicher, da der Roboter ebenfalls eingerichtet werden und hinterher wieder einsatzbereit sein muss. In den verschiedenen Operationsphasen treten unterschiedliche Personen(gruppen) auf verschiedene Weise mit dem Roboter in Interaktion. Insbesondere in der Vorbereitungs- und Nachbereitungsphase sind die OP-Pfleger\*innen aktiv. Während des An- und Abdockens wird das medizinische Personal gebraucht. Die Konsolenphase ist von den Operateur\*innen und den Assistent\*innen bestimmt. Die einzelnen Phasen erfordern im zeitlichen Verlauf also unterschiedliche Aktivitäten der einzelnen Berufsgruppen. Je nachdem, wie eingespielt das Team, wie geübt der/die Operateur\*in und wie kompliziert der Eingriff ist, dauert die gesamte OP kürzer oder länger. Das folgende Zitat eines Chirurgen der Urologie verdeutlicht dies:

*„Ja, also das ist ein bisschen aufwendiger, die OP-Vorbereitung, weil einfach der Eingriff so ist, dass man den Patienten relativ steil lagern muss und das ist einfach relativ aufwendig. Und dann auch das Andocken der Arbeits-Trokare, das ist auch ein bisschen aufwendiger als beim offenen Operieren. Wobei wir da mittlerweile so routiniert sind, dass das Ganze auch nur noch wenige Minuten in Anspruch nimmt. [...] Aber das geht halt nur mit einem sehr geübten Team.“*

Bei einem geübten OP-Team muss die zeitliche Verschiebung oder der Mehraufwand demnach nicht unbedingt große zeitliche Verzögerung bedeuten. Jedoch strukturiert sich der Zeitverlauf anders und die Tätigkeiten finden in anderen Zeiträumen statt. Mit der Zeit und der Erfahrung des Personals können sich diese Zeitstrukturen neu finden.

Darüber hinaus entstehen mit dem Einsatz des da Vinci-Systems andere Zeit limitierende oder erweiternde Faktoren. Die Gesamtdauer des Eingriffs ist z. B. relevant für die Anästhesie. Sie muss zu jeder Zeit sicherstellen, dass die Patient\*innen entsprechende Narkosemittel erhalten und stabil bleiben. Wenn die Operation länger dauert als geplant, kann beispielsweise die Lagerung des/der Patient\*in zu einem Problem werden, da bei einigen Eingriffen eine Kopf-tief-Lagerung erforderlich ist. Außerdem kann die geplante Narkosezeit nicht beliebig verlängert werden, ohne die Gesundheit der Patient\*innen zu gefährden. Der zeitliche Operationsverlauf ist demnach für die Anästhesie bedeutsam. Des weiteren können Komplikationen das Zeitregime herausfordern. Entsteht beispielsweise eine Blutung, sind schnell Handlungsschritte erforderlich. Die Ursache muss gefunden und möglichst rasch behoben werden. Sobald die Blutung so stark ist, dass das Endoskop stetig verschmutzt und kein Bild mehr aufnehmen kann, stößt die RAC mit da Vinci an ihre Grenzen. Im schlimmsten Notfall muss der Roboter abgedockt und die Operation offen weitergeführt werden, was mehr Zeit

in Anspruch nimmt. Nicht nur deshalb ist es wichtig, dass die Chirurg\*innen den (robotischen) Eingriff sicher beherrschen. Die höhere Präzision und die Vergrößerungsmöglichkeit lässt das Risiko für Zwischenfälle jedoch sinken, da Gewebestrukturen viel genauer erkannt und Schnitte akkurater gesetzt werden können. Nebstdem sorgt die verbesserte Sicht und die ruhige Kameraführung dafür, dass die Chirurg\*innen länger konzentriert arbeiten können.

Mit der digitalen Aufnahme der Eingriffe werden die Ereignisse im OP-Saal auch über die Dauer der Operation hinaus verfügbar gemacht. Zu jedem beliebigen Zeitpunkt können die Daten abgerufen werden. So lässt sich beispielsweise in der Nachsorge oder bei nachträglichen Komplikationen später nachvollziehen, wo gegebenenfalls ein Fehler lag oder worin die Ursache bestehen könnte. Die Speicherung und Dokumentation der Daten kann Informationen auch dauerhaft sichern oder z. B. die Lernkurve über eine Zeitspanne hinweg sichtbar machen. Über die Zeit sammeln sich zudem verschiedene Fälle oder Ähnlichkeiten tauchen auf – so werden mit der Dauer weitere Erkenntnisse sichtbar und können z. B. zu Lehrzwecken verwendet werden. Direkt nach der Einführung des da Vinci-Systems können demnach bestimmte Fälle noch nie eingetreten oder neue Einsatzmethoden noch nicht erprobt worden sein.

Ebenfalls zeigt sich am da Vinci-System das veränderte Zeitregime des Lernens. Bevor die Ausbildung am Roboter selbst beginnt, können die Chirurg\*innen z. B. live oder im Nachgang eine Operation über den Bildschirm mitverfolgen. Während der Schulungszeit in einem Trainingscenter von Intuitive können sie Erfahrungen machen, die ihnen später in der Praxis weiterhelfen. Im OP-Saal können sie erst assistieren und anschließend einzelne Schritte und nach und nach auch längere Sequenzen übernehmen – zeitlich angepasst an ihre eigene Lernkurve. Diese zeitliche Flexibilität und damit auch Intensität des Lernens kann Entlastung bringen und ermöglicht einen individualisierten Lernprozess. Auch Feedback kann durch die Aufzeichnungsmöglichkeiten zeitverzögert gegeben werden. Zusätzlich bauen die RAC-spezifischen Kompetenzen auf vorher erlernte Fähigkeiten auf und erweitern dieses Wissen.

Zusammenfassend lässt sich sagen, dass die Einführung des da Vinci-Systems in die Arbeitspraxis von Chirurg\*innen in zeitlicher Hinsicht einige Veränderungen bewirkt. So gestaltet das Hinzukommen des Roboters die zeitliche Einteilung der unterschiedlichen OP-Phasen um und erfordert neue Handlungsschritte, was sich auf die Rollen und die Zusammenarbeit im OP-Saal auswirkt. Überdies kann der zeitliche Aspekt erweiternden oder limitierenden Effekt auf den Verlauf des Eingriffs bedeuten. Die Digitalisierung in der RAC sorgt dafür, dass die (Un-)Verfügbarkeit der datenbasierten Informationen zeitversetzt besteht. Daneben beeinflusst der individualisierte und digitalisierte Lernprozess das Zeitregime

des Lernens. Diese zeitlichen Aspekte wirken sich besonders auf die Aus- und Weiterbildung der Chirurg\*innen aus, während der sie die Eigenheiten der RAC mit all ihren Erfordernissen, aber auch neuen Möglichkeiten erlernen.

### 3.4 Die sachliche Dimension im Operationssaal

Neben der sozialen, räumlichen und zeitlichen Dimension verändert da Vinci in der RAC auch die sachliche Dimension auf der Mikroebene im OP-Saal. Bevor Patient\*innen überhaupt in einem robotischen Operationssetting behandelt werden, entscheiden bereits die körperlichen Voraussetzungen über den Behandlungsverlauf. Starkes Über- oder Untergewicht beispielsweise sind Faktoren, die einen robotischen Eingriff stark erschweren. Der Körper sehr leichtgewichtiger Menschen bietet teilweise nicht genügend Raum, um die Instrumente des Roboters frei zu bewegen. Zudem kann die notwendige Kopf-tief-Lagerung bei bestimmten Eingriffen das Herz-Kreislauf-System zu stark belasten. Selbiges gilt für Menschen mit starkem Übergewicht. Zusätzlich weist der Körper in diesen Fällen viel Fettgewebe im Bauchraum und mehr Verwachsungen des Gewebes auf, was den Zugang zum Operationsfeld erschwert und Eingriffe zeitlich stark verzögern kann. Es erfolgt also eine Selektion der Patient\*innen nach körperlichen Gesichtspunkten (Zdichavsky et al. 2011, S. 40). Außerdem entscheidend sind diagnostische Befunde. Beispielsweise können Tumore, die an einer anspruchsvollen Stelle im Körper sitzen, nicht immer robotisch operiert werden:

*„[...] deswegen gibt es auch Eingriffe, die wir nicht robotisch machen, weil wir wissen, das sind gefährliche Eingriffe oder Eingriffe, die ein hohes Blutungsrisiko haben. Ich habe irgendwie letzte Woche eine Patientin operiert, die relativ jung ist, die einen riesengroßen Nierentumor hatte, wo ich wusste, wir gehen an die großen Gefäße ran, und da habe ich gesagt, da kann es, das mache ich nicht robotisch, weil ich da nicht diese Sicherheit habe, die mir meine Hände und ein Bauchtuch vermitteln so.“*

Das Zitat der Oberärztin der Urologie macht klar, dass risikoreiche Eingriffe an sensiblen Stellen nicht unbedingt operiert werden können und manche Befunde das robotische Verfahren nicht zulassen. Die befragte Person spricht an, dass neben dem medizinischen Befund auch ihre eigenen Erfahrungswerte in die Entscheidung einbezogen werden. In diesem Fall wollte sie das Risiko möglicher Komplikationen nicht eingehen, da ihr das taktile Sinnempfinden fehlen würde, welches ihr ein anderes Gefühl der Sicherheit vermittelt als die RAC – dazu folgend mehr. Die individuelle Anatomie in Zusammenhang mit der Diagnose und



den Erfahrungswerten der Operierenden spielen demnach in die Entscheidung für eine bestimmte Behandlungsmethode hinein. Beides entscheidet auch über den Operationsverlauf. So kann der Eingriff z. B. viel komplexer oder langwieriger werden als angenommen, wenn eine weitere Streuung eines Tumors entdeckt wird als vorher angenommen. Außerdem sind alle Eingriffe trotz hoher Standardisierung der Teilschritte mit der persönlichen Anatomie der Patient\*innen stets individuell. Dies verlangt den Operateur\*innen situative Entscheidungs- und Handlungsfähigkeit ab (Böhle 2017; Böhle et al. 2004) und gilt bei offenen und klassischen laparoskopischen Eingriffen ebenso.

Das Setting hat ebenfalls Einfluss auf das sinnliche Erleben der Tätigkeit als Operateur\*in. Der Roboter führt keine Handgriffe selbstständig aus. Stattdessen wird er durch die Bewegungen der Chirurg\*innen nach dem 'Master-Slave-Prinzip' gesteuert; da Vinci ist also ein Telemanipulator (Ficuciello et al. 2019). Das bedeutet, dass die Handbewegungen der Operateur\*innen durch die Technik über Distanz hinweg übertragen wird. Dabei entsprechen die Bewegungen der Operationsinstrumente nicht 1:1 den Bewegungen der Finger und Hände der Chirurg\*innen. Durch einen Algorithmus wird das Zittern der Hände (Tremor) ausgeglichen. Ruckartige Bewegungen werden durch das System erkannt und unterbunden. Um die Instrumente zielgerichtet steuern zu können, müssen Chirurg\*innen robotische Chirurgie praktisch erlernen. Der Bewegungsinput der Chirurg\*innen wird durch eine Steuereinheit über eine Software in die Bewegungen der Instrumente übersetzt und folgt dabei einer eigenen Logik. Die zeitliche Verzögerung, also die Geschwindigkeit der Übersetzung, kann voreingestellt werden. Durch die Zeitverzögerung ist ein genaueres Operieren und ein feineres Handling der Instrumente möglich. Außerdem bieten die Instrumente durch die sieben Freiheitsgrade eine bessere Beweglichkeit als das menschliche Handgelenk; die Instrumente können um 360 Grad gedreht werden. Diese Möglichkeit ist eine Eigenheit des Roboters und muss adaptiert werden, damit sie bewusst eingesetzt werden kann. Insgesamt müssen die Operateur\*innen sich der Maschinenlogik unterwerfen. Das bedeutet, dass sie die Steuerungsfunktion von da Vinci, das technische Eingreifen des Roboters mit dem Tremor-Ausgleich, die Distanz der Konsole zum Operationstisch, die Ansicht der Aufnahme des Endoskops und alle weiteren Funktionen und beschriebenen Eigenheiten annehmen und erlernen sowie entsprechend der Voraussetzungen ausführen müssen. Da die Operierenden die Instrumente über die Steuereinheit bewegen, fallen haptische Eindrücke weg. Druck und Widerstand werden nicht rückübertragen. Der Sehsinn rückt in den Vordergrund. Dies kann bewusst als Methode eingesetzt werden und einen Ausgleich zur taktilen Wahrnehmung schaffen, z. B. wenn vorsichtig an einem Gewebe gezogen wird, um die Elastizität zu testen. So entsteht eine Art

viso-taktilen Feedback, das Aufschluss über Gewebeschaffenheiten gibt. In der Empirie beschreiben die Befragten, dass sich die visuellen Eindrücke mit der Zeit verfeinern. Je mehr robotische Operationen sie durchgeführt haben, desto besser lassen sich die visuellen Wahrnehmungen einordnen. Es wird deutlich, dass die Steuerung des Roboters ein entscheidender Lernprozess ist, was folgendes Zitat einer Oberärztin der Urologie zeigt:

*„Ich habe keine metallenen Hände, und die sind auch nicht so hart wie die vom Roboter, d. h. das muss man am Anfang lernen, dass man sich auf diesen Dingen ein bisschen anders verhält. Das geht aber erstaunlich schnell, finde ich, also dass man so versteht, dass, ja, man eben nicht seine eigenen Hände hat [...] und ich denke mittlerweile auch nicht mehr darüber nach, wie ich mich bewegen darf oder nicht bewegen darf. Ich habe das einfach gelernt, wie ich mich bewegen darf.“*

Das Bewegen der Instrumente funktioniert nicht so intuitiv wie das Bewegen der eigenen Hand und die Instrumente haben eine andere Beschaffenheit und somit andere Anforderungen, um eine gewünschte Handlung auszuführen. Die befragte Person sagt aus, diese Anforderungen der Steuerung relativ schnell erlernt und eingeübt werden können, bis sich eine ‚gefühlsmäßige‘ Handlungsfähigkeit und Intuition für die Instrumente einstellt.

Außerdem beschreibt die Oberärztin, wie sich die Sinneswahrnehmung – auch durch da Vinci bedingt – über die Zeit ins Visuelle verschiebt:

*„[W]o wir alle immer sagen, das braucht man nach einer Zeit nicht mehr und das würde ich mittlerweile auch bestätigen, das unmittelbare taktile Feedback. [...] Das, was man aber so bekommt mit der Zeit, und das bestätigen, glaube ich, alle Kollegen, die lange robotisch operieren, ist, dass sie so eine Art visotaktilen Feedback bekommen. Also meint, man nimmt seine Roboterarme und macht das Gleiche, was man sonst mit seinen Fingern machen würde und sieht, wie sich das Gewebe bewegt und kriegt dann auch ein Gefühl tatsächlich dafür, jetzt nicht im Sinne von ich merke das an meinen Fingern, so ist es nicht, aber man denkt, das ist irgendwie, das reagiert komisch, das Gewebe. Das sieht mir derb aus.“*

Anstatt sich auf die taktilen Eindrücke zu verlassen, die in der RAC nicht mehr möglich sind, wird versucht, dieses Gespür zu ersetzen. Fehlende sinnliche Informationen werden ausgleichen, indem visuelle Eindrücke erzeugt werden. Blicken die Operateur\*innen in die Konsole hinein, sehen sie dort das endoskopische Bild in hoher Qualität, bewegungsruhig und dreidimensional. Tiefen- und Größenverhältnisse werden sichtbar und eine Vergrößerung ist möglich. Das Sehen mit da

Vinci stellt somit eine technische Erweiterung und Verbesserung des menschlichen Sehsinns dar. Die Verengung der Sinneseindrücke auf das Sehen birgt allerdings auch ein Risiko, welches ein Oberarzt der Urologie schildert:

*„Man hat ja auch kein forced Feedback. Das heißt, ich merke nichts. Ich muss das Gewebe erkennen. Ich merke nicht, ob ich gerade einen Knochen durchschneide oder ob ich gerade ein Haar durchschneide, das merke ich nicht, das erkennt man nur durch jahrelange Übung, [...] Irgendwann hat man das Gefühl, man tastet selber. [...] wenn ich weggucke, kann ich mit einem Arm gerade die Aorta zerreißen, ohne dass ich es merke.“*

Dieser Interviewausschnitt zeigt auf, dass mit dem auf das Sehen verengten Sinnerleben auch das Wissen um anatomische Details und das Erfahrungswissen wichtiger werden. Wie belastbar oder empfindlich eine Gewebestruktur ist, erkennt das geschulte Auge sehr viel besser. Im Falle von da Vinci sind diese Kenntnisse stärker herausgefordert, als bei der offenen Operation, da das Gewebe nicht zusätzlich ertastet und somit taktil erforscht und erlebt werden kann.

Mit dem eingeschränkten Sichtfeld auf den Bildausschnitt des Endoskops können daneben Zusammenhänge und benachbarte Gebiete und Strukturen im Körper nicht eingesehen werden. Durch die Verengung des Bildes entsteht ein Immersionseffekt, was die Konzentration auf den relevanten Ausschnitt bündelt – Operateur\*innen tauchen in den Körper ein: „floating in the body of the patient“ (Sergeeva et al. 2015, S. 5). Weil der restliche, über den Roboter zugängliche Raum im Körperinneren ebenfalls nicht eingesehen werden kann, gilt die goldene Regel, dass alle Instrumente möglichst im Sichtfeld sein sollten. Ansonsten bestünde die Gefahr, dass z. B. ungewollt Druck auf eine Körperstelle ausgeübt wird. Dies gilt auch für die Roboterarme außerhalb des Körpers. Neben der endoskopischen Bildübertragung werden unten im Bildrand einige Informationen zur Instrumentenbestückung, der Lebensdauer der Instrumente und deren Stromversorgung eingeblendet. Augmentierte Informationen zum Operationsfeld, -verlauf oder zu den Patient\*innen sind jedoch (noch) nicht einsehbar. Die Möglichkeiten sinnlicher Wahrnehmung während der Operation verändern sich somit stark, insbesondere für Operierende. Die Verengung auf visuelle Eindrücke hat, wie beschrieben, einen großen Einfluss auf die Operationstätigkeit.

Mit dem da Vinci-System verändern sich außerdem die Anforderungen während der Operation. Die Operateur\*innen müssen das Handling des Roboters beherrschen, ohne haptische Eindrücke zurechtzukommen, den Eingriff des Roboters an anderen Kommunikationsbedingungen anleiten und situativ auf das Operationsgeschehen reagieren können. Das beinhaltet auch, bei unerwarteten Komplikationen, Systemausfällen oder Unwägbarkeiten – wie einem abweichenden Befund oder

anatomischen Besonderheiten, handlungsfähig zu bleiben. Das Zitat einer Chirurgin der Urologie macht deutlich, dass das Operationsgeschehen sehr dynamisch ist und viele verschiedene Reaktionen gleichzeitig erfordert:

*„Ich glaube, das ist Außenstehenden gar nicht so klar, wie viel Bewegung da ist. Die OP-Schwester wechseln sich aus, der Anästhesist wechselt, [...] dann kommt irgendjemand und will Material holen oder irgendjemand hat eine Frage. [...] Dann klingelt das Telefon. Also ist ja relativ viel Unruhe in so einem OP-Saal und ist noch schwieriger, weil man in dieser Roboterwelt sozusagen in diese Konsole schaut und eine ganz andere Wahrnehmung seiner Außenwelt hat.“*

Es wird deutlich, dass die Operateurin nicht nur mit ihrer chirurgischen Tätigkeit an sich beschäftigt ist, sondern auch als Managerin, Ratgeberin, Koordinatorin usw. tätig werden muss. Die immersive Konstruktion der Konsole macht das gleichzeitige Da-Sein im Umfeld und im OP-Feld unmöglich. Entsprechend braucht es Aufwand und eine gezielte Herangehensweise, um allen Anforderungen gerecht werden zu können. Dafür erforderlich ist nicht nur bereits erlangtes (Erfahrungs-)Wissen, sondern es braucht spezifisches Wissen und neue Fertigkeiten und Kompetenzen, um mit dem System erfolgreich zu operieren. Praktisches Können und medizinisches Wissen aus konventionellen oder laparoskopischen Eingriffen helfen bei der robotischen Operation weiter. Umgekehrt wird in der RAC Wissen erzeugt, das bei anderen Operationsmethoden herangezogen werden kann; z. B. der Verlauf kleinerer Blutgefäße, die mit der vergrößerten Endoskopie gut erkannt werden können. Während die Chirurg\*innen die RAC erlernen, kommen immer mehr dieser Anforderungen hinzu. Von der Übernahme einzelner Teilschritte bis hin zur Anleitung eines robotischen Eingriffs hinter der Konsole braucht es sehr viel Übung. Das Vermitteln der RAC stellt wiederum eine besondere Herausforderung dar: die Anleitung lernender Chirurg\*innen kommt für die Operierenden zusätzlich zur Koordination der Operation hinzu. Diese kollegiale Unterstützung ist insbesondere zu Beginn der Ausbildung robotischer Chirurg\*innen wichtig.

Zudem ist der Arbeitsprozess ein anderer. Es sind mehr bzw. abweichende Handlungsschritte notwendig, was neue Aufgaben und eine andere Aufgabenverteilung für das gesamte Personal bedeutet. Auch muss die Operationsmethode angepasst werden, beispielsweise die Reihenfolge der einzelnen OP-Schritte. Für die Operateur\*innen bedeutet die RAC nicht mehr, mit den Händen oder Instrumenten am OP-Tisch direkt einen Körper zu bearbeiten, sondern aus der Distanz heraus über die Konsole robotische Instrumente zu steuern. Der Arbeitsprozess wandelt sich also mit dem System, ist aber grundsätzlich schematisch (neu) standardisiert.

Des Weiteren spielt der Aspekt der Datenerhebung in der RAC auch sachlich eine große Rolle. Da Vinci als technisches Artefakt lässt das Aufzeichnen zahlreicher und detaillierter Datensätze zu. Darüber hinaus haben die Operateur\*innen Zugriff auf eine App, in der sie ihre eigenen Statistiken übersichtlich einsehen und auswerten können. Zudem werden die Operationen häufig aufgezeichnet, um zu Lehrzwecken zu veranschaulichen oder Nachsorge betreiben zu können. Während der klassischen Laparoskopie können Daten nicht in diesem Umfang erfasst werden. All das stellt eine inhaltliche Erweiterung und Bereicherung der chirurgischen Arbeit dar und schafft neue Möglichkeiten, z. B. die Analyse der eigenen Lernkurve. Auch entsteht so mehr Transparenz über medizinische Details und die chirurgische Tätigkeit hinter der Konsole, es werden also mehr Informationen verfügbar. Entsprechend kommen auch Fragen nach dem Datenschutz sowie Datensouveränität auf.

Zusammenfassend lässt sich sagen, dass die Einführung des da Vinci-Systems in die Arbeitspraxis von Chirurg\*innen auch in sachlicher Hinsicht einige Veränderungen bewirkt. Die RAC ist aufgrund von Anforderungen an körperliche Merkmale nur einem eingeschränkten Personenkreis zugänglich. Zudem verändert sich der Arbeitsprozess für die Chirurg\*innen erheblich. Die Anforderungen werden mehr, Tätigkeitsabläufe wandeln sich und das sinnliche Erleben wird auf visuelle Eindrücke hin reduziert, wobei der Effekt der Immersion diese Verkürzung noch verstärkt. Gleichzeitig erweitert die Datenerhebung die sachliche Informationslage, wobei mit diesen zusätzlichen Informationen umgegangen werden muss.

Die Betrachtung der einzelnen Dimensionen auf der Mikroebene im Operationssaal macht bereits deutlich, dass mit dem Einsatz des da Vinci-Systems einige Veränderungen angestoßen worden sind. Demnach bedeutet das Hinzukommen der robotischen Technik in das Operationsgeschehen einen Wandel im Arbeitsprozess unterschiedlicher Berufsgruppen, insbesondere für die Chirurg\*innen. In der Tabelle (Tab. 1) wird ersichtlich, dass auch die Organisation im Krankenhaus sowie die Profession der Medizin mit Folgen konfrontiert sind. Darauf kann hier aus Gründen des Umfangs nicht näher eingegangen werden. Was im Feld ebenfalls deutlich wurde ist die Tatsache, dass das da Vinci-System bereits weitere technische Potenziale angelegt hat. Im Ausblick wird deshalb ein Blick in die Zukunft der roboterassistierten Chirurgie gewagt.

## 4 Ausblick

Betrachtet man das da Vinci-System als Beispiel für die fortschreitende Digitalisierung und Technisierung innerhalb der Medizin wird deutlich, dass in der Praxis zahlreiche Veränderungen zu beobachten sind. In sozialer Hinsicht verändern sich die Kommunikationsstrukturen und die Zusammenarbeit im OP-Saal, die Beziehung zwischen Ärzten/Ärztinnen und deren Patient\*innen und die sozialen Zuschreibungen der Chirurg\*innen an den Roboter beeinflussen deren Umgang damit sowie die Art und Weise, in der sie das System in ihren Arbeitsalltag einbeziehen. In räumlicher Hinsicht beeinflusst da Vinci die euklidische Raumgestaltung und zieht unsichtbare Grenzen, verdoppelt darüber hinausgehend die räumlichen Strukturen durch das endoskopische Abbild, virtualisiert den Körper der Patient\*innen und eröffnet durch umfangreiche Datenerhebung einen neuen Datenraum. In zeitlicher Hinsicht führt da Vinci zu neuen Standards im Operationsablauf, koppelt die Zeitstruktur an andere Faktoren, macht die Geschehnisse zeitunabhängig verfügbar und strukturiert das Zeitregime des Lernens neu. In sachlicher Hinsicht existieren neue anatomische, körperliche und diagnostische Merkmale für die Selektion von Patient\*innen, die Wahrnehmungsmöglichkeiten werden auf den Sehsinn fokussiert und verdichtet, da Vinci stellt andere Anforderungen an die Chirurg\*innen, Arbeitsprozesse laufen anders ab und mit der umfangreichen Datenerhebung kommen Informationen hinzu, die den Arbeitsalltag der Chirurg\*innen mitbestimmen.

Diese einzelnen Aspekte aus der empirischen Praxis lassen sich verdichten und deuten als eine Virtualisierung, Technisierung und Datafizierung der chirurgischen Tätigkeit. Die (Un-)Verfügbarkeit und (In-)Transparenz von und durch Daten und physischen Körpern sind Folgen davon. Weitere Technologien, insbesondere KI-Anwendungen, werden zukünftig in Arbeitsprozesse und -organisation integriert werden, auch in der Medizin (Zhang und Boulos 2023; Koohi-Moghadam und Ty Bae 2023; Duffourc und Gerke 2023; Tukra et al. 2021). KI kommt häufig dann zum Einsatz, wenn Abläufe oder Prozesse optimiert oder kostengünstiger und effizienter umgesetzt werden sollen. In der Medizin dient KI meist dazu, schnellere und bessere Behandlungserfolge zu erzielen, präzisere und zuverlässigere Behandlungsmethoden und -möglichkeiten zu bieten oder medizinisches Personal in ihrer Tätigkeit zu unterstützen.

Das da Vinci-System beinhaltet momentan zwar noch keine KI, wird jedoch vom Hersteller Intuitive schon jetzt als „Ökosystem“ vermarktet und bietet einige Ansatzpunkte für KI, z. B. den algorithmischen Tremor-Ausgleich, die Erhebung, Verknüpfung und Auswertung großer und vielfältiger Datenmengen

oder die Möglichkeit der Interaktivität und Verknüpfung mit anderen Geräten bzw. Anwendungen wie dem „motion table“ – ein vernetzter OP-Tisch, der das Umlagern von Patient\*innen während der Operation erleichtert. Neben dem Robotersystem selbst bietet das Unternehmen Zubehör, Dienstleistungen, Analysetätigkeiten, Beratung sowie Aus- und Weiterbildungsmöglichkeiten an (Intuitive Surgical Operations, 2023). Zukünftig sind weitere umfangreiche Funktionen und Dienstleistungen rund um das System denkbar (Azizian et al. 2020). Erweiternde bildgebende Verfahren könnten hochpräzise Vorausplanung von Eingriffen auf Basis präoperativer 3D-Bilder, das Einblenden von Live-Daten und 3D-Bildern sowie die dynamische Einordnung von Gewebe durch intraoperative Bildgebung möglich machen. Die Kombination von virtueller Realität und KI könnte ein chirurgisches Metaverse bilden (van Leeuwen und van der Hage 2022). Darüber hinaus könnten Machine Learning Verfahren durch Datenanalyse erweitertes Feedback für den Erwerb chirurgischer Fähigkeiten, der Effizienz chirurgischer Prozesse, die chirurgische Anleitung und die Vorhersage postoperativer Ergebnisse liefern. Spannungssensoren an den Roboterarmen und die Integration von AR-Methoden könnten dazu beitragen, die chirurgische Erfahrung zu verbessern und Organbewegungen zu überwachen. Das Inkludieren von KI in die RAC nach dem Master–Slave-Prinzip kann außerdem eine schrittweise, sorgfältige Prüfung der autonomen Roboterchirurgie ermöglichen (Andras et al. 2020). Zukünftig könnte es robustere haptische Systeme, die kinästhetisches und taktiles Feedback bieten, sowie durch weitreichendere Miniaturisierung auch Mikrorobotik geben. Zudem sind autonome Roboter denkbar, die Chirurg\*innen unterstützen können (Alip et al. 2022) – ähnlich dem robotischen System MAKO von Stryker, welches Chirurg\*innen in der Knie Endoprothetik beim akkuraten Ausfräsen der Knochen für den passgenauen Einsatz der Prothese hilft. Auch eine datengestützte Entscheidungshilfe und KI-basierter Robotik-Support können Chirurg\*innen unterstützen, ohne sie zu ersetzen (Bodenstedt et al. 2020). Der Entwurf einer Visualisierungstechnik für Telerobotik durch die Integration von Haptik und KI könnte einen Mehrwert bieten in Form von einer detaillierteren Methode zur Visualisierung medizinischer Bilddaten, einer verringerten Arbeitsbelastung für medizinisches Personal und weitere Möglichkeiten der Teleoperation (Minopoulos et al. 2023). Überdies gibt es, neben da Vinci, viele andere robotische Systeme und Ansätze für die medizinische Robotik, die ebenfalls zugelassen und im Einsatz sind (Brodie und Vasdev 2018).

In der Diskussion um KI in der Medizin wird deutlich, dass in der Forschung rasche Fortschritte erzielt werden und das Forschungs- sowie Anwendungsfeld hoch dynamisch ist (Bundesministerium für Bildung und Forschung 2019; Fraunhofer IKS o. J.). Zudem kommt immer mehr KI-basierte Software in der

Medizintechnik zum Einsatz und zahlreiche Unternehmen beteiligen sich an der Produktentwicklung medizinischer Technik mit KI-Anteil oder bieten bereits Medizinprodukte mit KI-Anteilen an. Merative, hervorgegangen aus IBM Watson Health, entwickelt beispielsweise Angebote zur klinischen Entscheidungsunterstützung, Datenanalyse in der Gesundheitsversorgung oder der Klinikentwicklung (Merative o. J.). Auch Siemens hat sich mit Siemens Healthineers der Medizintechnik angenommen und bietet in verschiedenen Fachbereichen unter anderem KI-gestützte Lösungen an, z. B. zur Entscheidungsunterstützung für die multimodale Bildgebung (Siemens Healthineers o. J.). Daneben beforschen Unternehmen wie Medtronic und SAP digitale sowie KI-gestützte Verfahren und Technologien im Gesundheitswesen und stellen diese teilweise schon bereit (Medtronic o. J., SAP o. J.).

Diese vielseitigen Ansätze machen deutlich, dass schon jetzt in der existierenden medizinischen (Roboter-)Technik ein großes Potenzial für die technische Weiterentwicklung der Systeme angelegt ist und weitere Technologien denkbar sind. Fraglich bleibt, welche Auswirkungen diese Entwicklungen auf die berufliche Tätigkeit von Chirurg\*innen und anderem medizinischen Personal haben wird, welche Ziele verfolgt werden – Automatisierung oder Erweiterung zur Unterstützung oder als Ersatz, welche (neuen) Verantwortlichkeiten entstehen und ob bzw. welche Grenzen sich aufzeigen. Durch das Hinzuziehen fachfremder Disziplinen, z. B. aus den Sozial- und Geisteswissenschaften wie die Technik- oder Medizinethik, die Soziologie und die Technikfolgenabschätzung, kann die technische Entwicklung neuer Medizintechnik übergreifend fachlich begleitet, ein ganzheitlicheres Verständnis gewonnen, innovative Denkansätze hervorgebracht, komplexe Sachverhalte durchdrungen, Fähigkeiten und Können erweitert sowie neue Netzwerke der Zusammenarbeit gebildet werden. Um verantwortungsvoll weitere technische Entwicklungen voranzutreiben bleibt es wichtig, die einhergehenden Veränderungen im Blick zu behalten, zu verstehen und entsprechend zu gestalten.

So betrifft die Frage nach den gesellschaftlichen Folgen von KI auch die Medizin und Gesundheitsversorgung als wichtigen gesellschaftlichen Bereich. Im Zuge der institutionellen Einbettung von KI braucht es Haftungs- und Regulationskonzepte sowie Modelle gesellschaftlicher Normierung und Kontrolle des Einsatzes von KI, was ebenfalls für Medizintechnologien gilt. Insbesondere in der Medizin, einem hochsensiblen und risikoreichen Feld, sind die Anforderungen für den sicheren und zuverlässigen Einsatz von KI sorgfältig zu ermitteln und zu reflektieren. Zudem kann die Frage nach gesellschaftlichen Mensch-KI-Verhältnissen und dem KI-getriebenen Wandel auch auf das Feld der Medizin hin spezifiziert werden, z. B. in welchem Verhältnis Chirurg\*innen zu einem



Entscheidungsunterstützungs-System stehen und was das für die Versorgung der Patient\*innen bedeutet. Es zeigt sich demnach deutlich: der durchdachte und vorteilhafte Einsatz von KI in der Gesundheitsversorgung erfordert interdisziplinäre Ansätze und Zusammenarbeit. Nur so kann sichergestellt werden, dass KI-basierte Medizinprodukte und die Folgen und Auswirkungen deren Einsatzes ausreichend verstanden werden sowie den Patient\*innen zugutekommen.

---

## Literatur

- Alip, S. L., Kim, J., Rha, K. H., und Han, W. K. 2022. Future platforms of robotic surgery. *Urologic Clinics*, 49(1): 23–38.
- Andras, I., Mazzone, E., Leeuwen, F.W.B. van, De Naeyer, G., Oosterom, M.N. van, Beato, S., Buckle, T., O’Sullivan, S., Leeuwen, P.J. van, Beulens, A., Crisan, N., D’Hondt, F., Schatteman, P., Poel, H. van Der, Dell’Oglio, P. and Mottrie, A. 2020. Artificial intelligence and robotics: a combination that is changing the operating room. *World Journal of Urology*38(10): 2359–2366.
- Azizian, M., Liu, M., Khalaji, I., Sorger, J., Oh, D., und Daimios, S. 2020. The da Vinci surgical system. In *Handbook of robotic and image-guided surgery*. pp. 39–55. Elsevier.
- Baumgart, J. 2010. Ärzte und informierte Patienten: Ambivalentes Verhältnis. *Deutsches Ärzteblatt*, 107(51–52), A2554–6.
- Bodenstedt Sebastian, Martin Wagner, Beat Peter Müller-Stich, Jürgen Weitz und Stefanie Speidel. 2020. Artificial Intelligence-Assisted Surgery: Potential and Challenges. *Visc Med* 4 December, 36 (6): 450–455. <https://doi.org/10.1159/000511351>.
- Böhle, F. 2017. *Arbeit als subjektivierendes Handeln*. Springer VS.
- Böhle, F., Pfeiffer, S., und Sevsay-Tegethoff, N. 2004. *Die Bewältigung des Unplanbaren*. VS Verlag für Sozialwissenschaften.
- Brodie, A., & Vasdev, N. 2018. The future of robotic surgery. *The Annals of The Royal College of Surgeons of England*, 100(Supplement 7): 4–13.
- Brunner, Stefanie, Dolores T. Müller, Jennifer A. Eckhoff, Alissa Reisewitz, Lars M. Schiffmann, Wolfgang Schröder, Thomas Schmidt, Christiane J. Bruns und Hans F. Fuchs. 2023. Innovative Operationsroboter und Operationstechnik für den Einsatz am oberen Gastrointestinaltrakt. In *Die Onkologie*. 29, 506–514. [Innovative Operationsroboter und Operationstechnik für den Einsatz am oberen Gastrointestinaltrakt | SpringerLink](#)
- Butollo F, Nuss S. 2019. *Marx und die Roboter*. Vernetzte Produktion, Künstliche Intelligenz und lebendige Arbeit. Karl Dietz Verlag Berlin GmbH. ISBN 978-3-320-02362-1.
- Bundesministerium für Bildung und Forschung. 2019. Was KI für die Medizin bedeutet. <https://www.bmbf.de/bmbf/shareddocs/kurzmeldungen/de/was-ki-fuer-die-medizin-bedeutet.html> (Letzter Zugriff: 18.09.2023).
- Bundesministerium für Bildung und Forschung. 2022. Künstliche Intelligenz. [Künstliche Intelligenz – BMBF](#). (Letzter Zugriff: 18.09.2023).
- Cascio, W. F., Montealegre, R. 2016. How Technology Is Changing Work and Organizations. *The Annual Review of Organizational Psychology and Organizational Behavior*, 3, 349–375.

- Catchpole, K., Bisantz, A., Hallbeck, S., Weigl, M., Randell, R., Kossack, M., und Anger, J. T. 2019. Human Factors in Robotic Assisted Surgery: Lessons from Studies ‘in the Wild’. *Applied Ergonomics*, 78, 270–276.
- Cunningham, S., Chellali, A., Jaffre, I., Classe, J., und Cao, C. G. 2013. Effects of experience and workplace culture in human-robot team interaction in robotic surgery: a case study. *International Journal of Social Robotics*, 5, 75–88.
- Duffourc M, Gerke S. Generative AI in Health Care and Liability Risks for Physicians and Safety Concerns for Patients. *JAMA*. 2023. 330(4):313–314. doi:<https://doi.org/10.1001/jama.2023.9630>.
- Feußner, H., & Wilhelm, D. 2016. Minimalinvasive Chirurgie und ‘robotic surgery’: Chirurgie 4.0?. *Der Chirurg*, 87(3).
- Ficuciello, F., Tamburrini, G., Arezzo, A., Villani, L., & Siciliano, B. 2019. Autonomy in surgical robots and its meaningful human control. *Paladyn, Journal of Behavioral Robotics*, 10(1), 30–43.
- Fraunhofer IKS. o.J. Künstliche Intelligenz in der Medizin. <https://www.iks.fraunhofer.de/themen/kuenstliche-intelligenz/kuenstliche-intelligenz-medizin.html> (Letzter Zugriff: 18.09.2023).
- Freschi, C., Ferrari, V., Melfi, F., Ferrari, F., Mosca, F., und Cuschieri, A. 2013. Technical review of the da Vinci surgical telemanipulator. *The International Journal of Medical Robots and Computer Assisted Surgery*, 9(4), 396–406.
- Giering O. 2022. Künstliche Intelligenz und Arbeit: Betrachtungen zwischen Prognose und betrieblicher Realität. *Zeitschrift für Arbeitswissenschaften*, 76: 50–64. doi <https://doi.org/10.1007/s41449-021-00289-0>.
- Gill, A., und Randell, R. 2016. Robotic surgery and its impact on teamwork in the operating theatre. *Journal of perioperative practice*, 26(3), 42–45.
- Haidegger, T., Speidel, S., Stoyanov, D., und Satava, R. M. 2022. Robot-assisted minimally invasive surgery—Surgical robotics in the data age. *Proceedings of the IEEE*, 110(7), 835–846.
- Hasenbein, Melanie. 2023. Wandel der Arbeits- und Organisationswelt durch Künstliche Intelligenz. In *Mensch und KI in Organisationen. Einfluss und Umsetzung Künstlicher Intelligenz in wirtschaftspsychologischen Anwendungsfeldern*, hrsg. Melanie Hasenbein, 35–55. Berlin, Heidelberg: Springer.
- Hillebrandt, K. H., Schmelzle, M., Sauer, I. M., & Pratschke, J. 2022. Robotik – Was geht in der Chirurgie? *OP-Journal*, 38(2), 104–110.
- Hirsch-Kreinsen H.2023. Artificial intelligence: a „promising technology”. *AI & Society*, 1–12. <https://doi.org/10.1007/s00146-023-01629-w>
- Huss, Ralf. 2019. Die Zukunft ist heute. In *Künstliche Intelligenz, Robotik und Big Data in der Medizin*, hrsg. Ralf Huss, 95–98. Berlin, Heidelberg: Springer
- Intuitive Surgical Operations. 2023. Da Vinci von Intuitive. <https://www.intuitive.com/de-de/products-and-services/da-vinci> (Letzter Zugriff: 18.09.2023).
- Jedamzik, S. 2019. Digitale Gesundheit und Pflege. Die Zukunft ist jetzt. In *Der Pneumologe*. 16, 391–396. [Digitale Gesundheit und Pflege | SpringerLink](https://www.springer.com/de/de/book/9783708914444)
- Kanji, F., Cohen, T., Alfred, M., Caron, A., Lawton, S., Savage, S., Shouhed, D., Anger, J. T., und Catchpole, K. 2021. Room Size Influences Flow in Robotic-Assisted Surgery. *International Journal of Environmental Research and Public Health*, 18, Artikel 7984.

- Kirchberg, J., Mees, T., und Weitz, J. 2016. Robotik im Operationssaal: Aus der Nische in die breite Anwendung (Leitthema). *Der Chirurg*, 87(12), 1025–1032.
- Kitzmann, Arnold. 2022. Veränderung unserer Lebensbereiche durch Künstliche Intelligenz. In *Künstliche Intelligenz*, hrsg. Arnold Kitzmann, 47–64. Wiesbaden: Springer.
- Kleemann, Frank, Jule Westerheide, und Ingo Matuschek. 2019. Arbeit und Technik. In *Arbeit und Subjekt. Aktuelle Debatten der Arbeitssoziologie*, hrsg. Frank Kleemann, Jule Westerheide und Ingo Matuschek, 39–67. Wiesbaden: Springer VS.
- Kogelbauer, Harald, und Andreas Krasser. 2022. 3-D-Druck in der Medizin und die Rolle der Radiologietechnologie. In *Radiopraxis*, 15 (04): 187–199. DOI: <https://doi.org/10.1055/a-1445-4964>. Thieme E-Journals – Radiopraxis / Abstract ([thieme-connect.com](http://thieme-connect.com))
- Koohi-Moghadam, M., Bae, K.T. 2023. Generative AI in Medical Imaging: Applications, Challenges, and Ethics. *J Med Syst* 47, 94. <https://doi.org/10.1007/s10916-023-01987-4>.
- Kramme R. 2011. *Medizintechnik. Verfahren – Systeme – Informationsverarbeitung*, Springer-Verlag GmbH, Berlin Heidelberg. ISBN-13 978–3–642–16186–5.
- Krzywdzinski, M., Jürgens, U., und Pfeiffer, S. 2016. *The Fourth Revolution: The Transformation of Manufacturing Work in the Age of Digitalization*. WZB Report, 22–25. [https://disciplinas.usp.br/pluginfile.php/5387512/mod\\_resource/content/1/2016\\_industry40%20%281%29.pdf](https://disciplinas.usp.br/pluginfile.php/5387512/mod_resource/content/1/2016_industry40%20%281%29.pdf) (Letzter Zugriff: 18.09.2023).
- Krzywdzinski, M., Pfeiffer, S., Evers, M., und Gerber, C. 2022. *Measuring work and workers: Wearables and digital assistance systems in manufacturing and logistics*. Discussion Paper SP III 2022–301, WZB Berlin Social Science Center. <http://hdl.handle.net/10419/251912>.
- Lane, T. 2018. A short history of robotic surgery. *The Annals of The Royal College of Surgeons of England*, 100(6): 5–7. <https://doi.org/10.1308/rcsann.suppl1.5>.
- Lenke, C., und Monett, D. 2020. KI und Automatisierung: Wirkung auf die Arbeit der Zukunft. *VQP. VQP, Sonderband Zukunft der Arbeit, Human Resources Consulting Review*, 12, 400–403.
- Mainzer, Klaus. 2018. Wie sicher ist Künstliche Intelligenz? In *Künstliche Intelligenz – Wann übernehmen die Maschinen?*, hrsg. Klaus Mainzer, 245–265. Berlin, Heidelberg: Springer.
- Medtronic. o.J. Eine KI-Revolution im Gesundheitswesen. <https://www.medtronic.com/de-de/our-company/ai-healthcare-technology.html> (Letzter Zugriff: 18.09.2023).
- Merative. o.J. Solutions. <https://www.merative.com/> (Letzter Zugriff: 18.09.2023).
- Minopoulos, G.M.; Memos, V.A.; Stergiou, K.D.; Stergiou, C.L. und Psannis, K.E. 2023. A Medical Image Visualization Technique Assisted with AI-Based Haptic Feedback for Robotic Surgery and Healthcare. *Appl. Sci.*, 13, 3592. <https://doi.org/10.3390/app13063592>.
- Neuburger, Rahild. 2019. Der Wandel der Arbeitswelt in einer Industrie 4.0. In *Handbuch Industrie 4.0 und Digitale Transformation. Betriebswirtschaftliche, technische und rechtliche Herausforderungen*, hrsg. Robert Obermaier, 589–608. Wiesbaden: Springer Gabler Wiesbaden.
- Neu, Matthias, Melanie Müller, Biju Pothen und Moritz Zingel. 2022. *Anwendungsfelder und Herausforderungen der Künstlichen Intelligenz*. Wie KI die Gesellschaft und Marketing beeinflusst. Wiesbaden: Springer Gabler.
- Pugin, F., Bucher, P., & Morel, P. 2011. History of robotic surgery: From AESOP® and ZEUS® to da Vinci®. *Journal of Visceral Surgery*, 148, e3–e8.

- Randell, R., Alvarado, N., Honey, S., Greenhalgh, J., Gardner, P., Gill, A., Jayne, D., Kotze, A., Pearman, A., & Dowding, D. 2015. Impact of robotic surgery on decision making: perspectives of surgical teams. In *AMIA Annual Symposium Proceedings* (Vol. 2015, S. 1057). American Medical Informatics Association.
- SAP. o.J. Healthcare. <https://www.sap.com/industries/healthcare.html> (Letzter Zugriff: 18.09.2023).
- Schindler, Matthias. 2022. Demokratisierung Künstlicher Intelligenz bei der BMW Group: Selbstservice für KI-basierte Computer Vision. In *Künstliche Intelligenz erfolgreich umsetzen*, hrsg. Ulrich Lichtenthaler, 149–170. Wiesbaden: Springer Gabler.
- Schuler, S., Hämmerle, M., und Bauer, W. 2019. *Einfluss Künstlicher Intelligenz auf die Arbeitswelten der Zukunft*. Digitale Transformation–Gutes Arbeiten und Qualifizierung Aktiv Gestalten, GITO Verlag: Berlin, 255–272.
- Sergeeva, A., Huysman, M., und Faraj, S. 2015, 13.–16. Dezember. Transforming work practices of operating room teams: the case of the Da Vinci robot [Konferenzbeitrag]. *Thirty Sixth International Conference on Information Systems*, Fort Worth: Texas.
- Sergeeva, A., Huysman, M., & Faraj, S. 2018, July. Losing touch: How robots transform the practice of surgery. In *Academy of Management Proceedings* (Vol. 2018, No. 1, p. 11429). Briarcliff Manor, NY 10510: Academy of Management.
- Siemens Healthineers. o.J. AI-Rad Companion. [https://www.siemens-healthineers.com/de/digital-health-solutions/ai-rad-companion?ste\\_sid=8be2d3bfe24fe63633d48a4b56824edf](https://www.siemens-healthineers.com/de/digital-health-solutions/ai-rad-companion?ste_sid=8be2d3bfe24fe63633d48a4b56824edf) (Letzter Zugriff: 18.09.2023).
- Tukra, S., Lidströmer, N., Ashrafian, H. und Giannarou, S. 2021. AI in Surgical Robotics. In *Artificial Intelligence in Medicine*, hrsg. Lidströmer, N., Ashrafian, H. . Springer, Cham. [https://doi.org/10.1007/978-3-030-58080-3\\_323-1](https://doi.org/10.1007/978-3-030-58080-3_323-1).
- Van Leeuwen, F. W. B. und van der Hage, J. A. 2022. Where Robotic Surgery Meets the Metaverse. *Cancers* 14(24): 6161. <https://doi.org/10.3390/cancers14246161>.
- Yates, D. R., Vaessen, C., und Roupret, M. 2011. From Leonardo to da Vinci: the history of robotic-assisted surgery in urology. *BJU International*, 108, 1708•1714.
- Zdichavsky, M., Kirschniak, A., Glatzle, J., Meile, T., Küper, M., Gögele, H., ... und Königsrainer, A. 2011. Minimalinvasive Chirurgie–Immer sinnvoll?. *Lege artis–Das Magazin zur ärztlichen Weiterbildung*, 1(01), 38–44.
- Zhang P, Kamel Boulos MN. 2023. Generative AI in Medicine and Healthcare: Promises, Opportunities and Challenges. *Future Internet*, 15(9):286. <https://doi.org/10.3390/fi15090286>.

---

# Künstliche Intelligenz als Bild und Mythos



# Das bewegliche Heer der Künstlichen Intelligenz. Ein Technomythos als Summe menschlicher Relationen

Bruno Gransche und Arne Manzeschke

## Zusammenfassung

KI-Diskurse finden in Bildern statt, in Sprachbildern, in Technikbildern und in visuellen Bildern. KI muss daher als vielschichtiges Phänomen in den Blick genommen werden, das technische und nicht-technische Aspekte enthält, und es muss im weiteren geistesgeschichtlichen Bedeutungskontext seiner Elemente betrachtet werden. Dazu gehören auch bildhafte Vorstellungen: Bilder, Urbilder, Abbilder, Vorbilder, Simulakren, sowohl bildhaft Dargestelltes und Sprachbilder als auch Vorstellungsbündel, wie sie etwa in der Rede von Menschenbildern, Technikbildern und Weltbildern angesprochen werden. Welche Elemente aber machen KI eigentlich aus? Welcher geistesgeschichtliche Kontext orientiert die Bedeutungen und Verständnisse von KI? Die hier zentralen technomorphen Menschenbilder und anthropomorphen Technikbilder sind nur ein Bruchteil derjenigen Phänomene, Narrative, Mythen, Schemata, Vorstellungen etc., mit denen KI semantisch freudig reagiert. Auch – vertikal gesprochen – ‚über‘ und ‚unter‘ dem Menschlichen sind Topoi bei der Verstehens- und damit Entscheidungs- und Handlungsorientierung wirkmächtig, die es zu berücksichtigen gilt, wenn KI, KI-Diskurse und deren Wirkungen

---

B. Gransche (✉)

Institut für Technikzukünfte ITZ, Karlsruher Institut für Technologie KIT, Karlsruhe, Deutschland

E-Mail: [bruno.gransche@kit.edu](mailto:bruno.gransche@kit.edu)

A. Manzeschke

Institut für Pflegeforschung, Gerontologie und Ethik, Evangelische Hochschule Nürnberg, Nürnberg, Deutschland

E-Mail: [arne.manzeschke@evhn.de](mailto:arne.manzeschke@evhn.de)

analysiert werden sollen. Dabei ist KI als dergestaltige Menge kategorial inhomogener Elemente als ein Inbegriff zu fassen. Wer die Interessen nicht beachtet, die solche Inbegriffe ausmachen, kann KI und ihre Diskurse nur oberflächlich verstehen. Die Reaktionsfreudigkeit und Kopplungsbereitschaft ist schwindelerregend, jedoch fundiert und strukturiert durch das verfügbare geistesgeschichtliche Repertoire an Kopplungskandidaten. Aber: KI entscheidet, lernt, handelt, denkt etc. trotz entsprechend semantischer Suggestionen genauso wenig, wie ein autonomes Auto autonom ist, wie ein Roboter zu etwas gezwungen werden könnte oder wie Informatiker, KI-Forscher und Data Scientists mit dem Schaffen von verblüffenden IT-Systemen zu Schöpfergöttern würden. Deshalb ist in diesem Zuge an vergessene oder aus der Aufmerksamkeit geratene Metaphorik zu erinnern sowie an die fundamentale Bildhaftigkeit und den vielfältigen Bildbezug der KI. KI ist – mit Nietzsche formuliert – ein Heer von Metaphern und Anthropomorphisierungen, eine poetisch gesteigerte Summe menschlicher Relationen. Deshalb muss implizites Bedeutungsgepäck der KI aufmerksam expliziert werden, müssen interessensgeleitet oder unbewusst hineingelegte Bedeutungen wieder ausgelegt werden. Solche KI-Auslegung muss KI-Entwicklung und -Verbreitung zwingend begleiten.

---

### Schlüsselwörter

Künstliche Intelligenz • Metapher • Bild • Technikbild • Menschenbild • Weltbild

### Vorbemerkung

Dieser Beitrag „Das bewegliche Heer der Künstlichen Intelligenz. Ein Technomythos als Summe menschlicher Relationen“ analysiert KI zunächst als Inbegriff, dann als Metapher und situiert dies in einem erweiterten Spektrum der Bedeutungsübertragungen. Dies wird verdeutlicht an einem Beispiel aus diesem Spektrum, nämlich dem Topos des Prometheus und diversen Verhältnissen von Schöpfertum und Geschöpflichkeit zwischen Gott, Mensch und KI. Der Beitrag schließt mit einer Problembetrachtung der unhintergehbaren Bildhaftigkeit der Vorstellungsorientierung allgemein sowie von KI speziell.

Der Beitrag „Bilder machen Menschen. Zur Bildermacht der Künstlichen Intelligenz“ (Manzeschke & Gransche in diesem Band) geht von den gleichen Prämissen aus wie der folgende Beitrag, komplementiert den Fokus auf Sprachbilder aber mit einem Schwerpunkt auf anthropologische Erwägungen des Bildermachens, auf das Verhältnis von Mensch- und Technikbildern zu Weltmodellen und Weltbild sowie

auf visuelle Bilder. Dabei steht der eigentümliche Bildbezug von KI im Mittelpunkt, denn KI prozessiert und generiert wesentlich Bilder und sie wird in speziellen Bildern vorgestellt und kommuniziert.

Beide Beiträge können eigenständig oder als zwei Seiten einer Medaille gelesen werden.

---

## 1 Einleitung

Was aber ist Künstliche Intelligenz? Obwohl KI derzeit im Zuge der Digitalisierung erhöhte Aufmerksamkeit erfährt, ist keineswegs klar, welche Phänomene, Anwendungen, Systeme oder Methoden unter dem Titelwort ‚Künstliche Intelligenz‘ eigentlich vorgestellt und angesprochen werden.

KI ist nicht allgemein gültig definiert. Es gibt keine eigentliche Bedeutung von KI und keine konsensfähige Menge an Elementen, die zu KI notwendig oder hinreichend gehörten. KI ist ein strategisches Hülsenwort – oder ein „wissenschaftstaktisches Manöver“ (Zimmerli 2021, S. 200) – das je interessengeleitet gefüllt wird. Die Bedeutung des Begriffs wird in weiten Teilen in bildhaften, übertragenen, uneigentlichen Hinsichten konstruiert. KI ist – so viel ist klar – nicht nur eine Technologie. KI ist ein schillerndes Phänomen mit Elementen aus Technologie, Imaginationen, Diskursakten, Narrativen, Mythoselementen, Verheißung, Bedrohung, Geschäftsinteressen, Machtmittel, Ermöglichungsstrukturen usw. Aktuelle Definitionen greifen zwangsläufig auf Bedeutungsquellen zurück, die nur vor dem Hintergrund dieser Vielfalt verstanden werden können. Ein Beispiel soll dies einleitend verdeutlichen:

„Künstliche Intelligenz (KI) bezeichnet Systeme mit einem ‚intelligenten‘ Verhalten, die ihre Umgebung analysieren und mit einem gewissen Grad an Autonomie handeln, um bestimmte Ziele zu erreichen. KI-basierte Systeme können rein softwaregestützt in einer virtuellen Umgebung arbeiten (z.B. Sprachassistenten, Bildanalysesoftware, Suchmaschinen, Sprach- und Gesichtserkennungssysteme), aber auch in Hardware-Systeme eingebettet sein (z.B. moderne Roboter, autonome Pkw, Drohnen oder Anwendungen des ‚Internet der Dinge‘).“ (HLEG on AI 2019)

Diese ‚Definition‘ von KI der *Hochrangigen Expertengruppe für künstliche Intelligenz* (HEG-KI) zeigt, stellvertretend für solche Definitionsversuche, die Schwierigkeit einer konkreten eigentlichen Definition dieses komplexen Phänomens namens KI. Die ‚Definition‘ beginnt mit einem metaphorischen Konstrukt – „Künstliche Intelligenz“ – und erklärt dieses – quasi tautologisch – mit einem



Gattungsverweis („Systeme“) und einem Artspezifikum („mit ‚intelligentem‘ Verhalten“). Dabei ist dieses Spezifikum nicht nur mit dem Begriffskompositum nahezu identisch (Intelligenz/intelligent), sondern selbst nur in Uneigentlichkeit markierenden Anführungszeichen verwendbar. Entsprechend abgeschwächt wird eine weitere Metapher („Autonomie“) mit dem Hinweis „ein gewisser Grad“, was auf das wiederum metaphorisch zu verstehende „Handeln“ solcher Systeme bezogen wird. Eigene Ziele sich zu setzen und diese zu verfolgen, ist ein Kennzeichen (normativer) Autonomie, weshalb auch hier abschwächend nur von „bestimmten Zielen“, nämlich durch Menschen fremdbestimmten Ziele die Rede sein kann. Dann folgen Hinweise zu IT-üblichen Metaphern *harter* und *weicher* Ware (aka Software, Hardware).

Intelligenz, Autonomie und Handeln werden zur Beschreibung solcher Systeme namens KI in bildhafter Redeweise herangezogen. Damit ist die vermeintliche ‚Definition‘ fast eher ein narratives Gleichnis, das bestimmte Verstehensweisen orientiert, aber nicht *erklärt*, was KI eigentlich ist. Wie wir KI verstehen und unser Entscheiden und Handeln auf KI hin demnach ausrichten, hängt entschieden von unseren Bedeutungsräumen für die Bedeutungsquellen wie z. B. Intelligenz und Autonomie ab. Da diese in ihren (vermeintlichen) Ursprungskontexten wiederum alles andere als klar sind, wird hier (immerhin in Inversion des *ignotium per ignotius*) höchstens etwas Unbekannteres (KI) durch ein nur wenig vertrauterer Unbekanntes (menschliche Intelligenz) ‚erklärt‘. Dieselbe Expertengruppe der obigen Definition räumt dies ein:

„Mit dem Ausdruck ‚KI‘ wird explizit auf den Begriff der Intelligenz Bezug genommen. Da jedoch das Konzept der Intelligenz (in Maschinen wie im Menschen) unbestimmt ist – obwohl von Psychologen, Biologen und Neurowissenschaftlern ausführlich untersucht –, verwenden KI-Forscher in der Regel den Begriff der Rationalität. Darunter wird die Fähigkeit verstanden, unter Berücksichtigung bestimmter zu optimierender Kriterien und der verfügbaren Mittel das bestmögliche Handeln zu wählen, um ein bestimmtes Ziel zu erreichen. Rationalität ist natürlich nicht der einzige, aber doch ein wesentlicher Bestandteil des Konzeptes der Intelligenz.“ (HLEG on AI 2019)

Warum reden wir dann nicht von ‚Künstlicher Rationalität‘? Das Überleiten zu Rationalität verschiebt die Definition auf ein weiteres Unbekanntes, da im Definitionsversuch wiederum unbestimmte Elemente enthalten sind wie „zu optimierend“ (woraufhin?), „bestmöglich“ (nach wessen Bestem und wessen Möglichkeit?), „bestimmtes Ziel“ (durch wen und wie bestimmt?). Dass dies auch nur für einen Teil „des Konzeptes Intelligenz“ (welches?) hilft, begrenzt den Nutzen zusätzlich.

KI muss daher als größeres Phänomen in den Blick genommen werden, das technische und nicht-technische Aspekte enthält und es muss im größeren Bedeutungskontext seiner Elemente betrachtet werden. Dazu gehören auch bildhafte Vorstellungen: Bilder, Urbilder, Abbilder, Vorbilder, Simulakren, sowohl bildhaft Dargestelltes, Sprachbilder als auch Vorstellungsbündel wie sie etwa in der Rede von Menschenbildern und Technikbildern angesprochen werden.

Es ist nötig, nicht an den offensichtlich scheiternden Versuchen einer einheitlichen *Begriffsdefinition* von KI stehenzubleiben, sondern KI in einem ersten Schritt als *Inbegriff* zu denken (Abschn. 1). Im Gegensatz zu Begriffen ist die Vorstellung eines Inbegriffs in der Lage, kategorial völlig verschiedene Elemente zusammen zu denken. Im Falle der KI bestehen solche Elemente aus Artefakten, Technologien und Infrastrukturen, aber vor allem auch aus Bezeichnungen, Metaphern und Begriffen sowie aus Imaginationen, utopischen und dystopischen Verheißungen bzw. Narrativen. Diese Elemente werden in Sprachbildern, besonders Metaphern, auf KI übertragen, was es nötig macht, diesen Vorgang mit der jeweiligen Übertragungshinsicht, den identischen (übertragbaren) und differenten (nicht-übertragbaren) Elementen der Quell- und Zielkonzepte zu analysieren (Abschn. 2). Bereits auf dieser Ebene gilt es, Verengungen und Einseitigkeiten zu vermeiden und die enorme Spannweite möglicher Elemente des KI-Inbegriffs zu reflektieren. Vorstellungen, die KI fundieren und prägen, sind prominenterweise *anthropomorphe Technikbilder* wie intelligente, autonome oder lernende Systeme sowie *technomorphe Menschenbilder* wie die des Menschen als Maschine. Neben diesen wichtigen und sich reziprok stützenden Vorstellungen sind aber auch Bezüge jenseits von Mensch und Technik z. B. auf ‚Übermenschliches‘ und Göttliches sowie Animalisches und Technisches in einem erweiterten Spektrum möglicher Bedeutungsquellen und geistesgeschichtlicher Kontexte und Kopplungskandidaten zu berücksichtigen (Abschn. 3). Es ergeben sich eine unüberschaubare Vielzahl von Übertragungs- und Kopplungsoptionen wie (KI-induzierte) theomorphe Menschkonzepte, zoomorphe Technikkonzepte, anthropomorphe Götterkonzepte usw. Ein solcher Nexus wird in Abschn. 4 exemplarisch diskutiert, nämlich die Vorstellungen von KI als Geschöpf (und nicht bloß Artefakt), vom Menschen als quasi-göttlicher Schöpfer intelligenter Entitäten und von diesen KI-Entitäten als Kopfgeburten, autonome Wesen, Nachfahren etc. Schließlich ist in Abschn. 5 auf die nicht zu vermeidende Bildhaftigkeit unserer Vorstellungsorientierung einzugehen und Probleme dieses Umstandes herauszuarbeiten, mit dem Ziel, einen aufgeklärten Umgang mit der unvermeidlichen Bildhaftigkeit zu ermöglichen. In einer Variation von Cassirers Metaphysik der symbolischen Formen ließe sich dieses Ziel für Bilder und Bildhaftigkeit folgendermaßen fassen:

Es geht um das Durchschauen des bildhaften Grundcharakters der Erkenntnis selbst. Lösen können wir uns von Bildern nicht, aber wir können und müssen sie in ihrer relativen Notwendigkeit begreifen und einsehen. Das ist die einzig mögliche ideelle Befreiung vom Zwang zur Bildhaftigkeit<sup>1</sup>.

---

## 2 Bemerkende Interessen – KI als Inbegriff

Es ist eine Herausforderung, das Phänomen KI mit seinen heterogenen, technischen wie nicht-technischen Elementen zusammenzudenken, ohne einseitig Aspekte zu isolieren. So kann nicht auf lernende Algorithmen fokussiert werden, ohne die soziale Prägung der Trainingsdaten zu berücksichtigen; so kann nicht auf KI-Funktionalitäten abgehoben werden, ohne die teils kontrafaktischen und narrativ geprägten Zuschreibungen der Nutzer aufzugreifen usw. Von Husserl stammt der Vorschlag, eine Menge an kategorial inhomogenen Elementen unter einem ihnen gemeinsamem „einheitlichen Interesse“ zusammenzufassen, d. h. als *Inbegriff*:

„Ein Inbegriff entsteht, indem ein einheitliches Interesse und in und mit ihm zugleich ein einheitliches Bemerkende verschiedene Inhalte für sich heraushebt und umfasst. [...] Fragen wir, worin die Verbindung bestehe, wenn wir z.B. eine Mehrheit so disparater Dinge wie die Röte, der Mond und Napoleon denken, so erhalten wir die Antwort, sie bestehe bloss darin, dass wir diese Inhalte zusammen denken, in einem Acte denken.“ (Husserl 1891, S. 79)

Was haben also „vollkommen willkürliche Einzelinhalte“<sup>2</sup> wie ein Algorithmus, eine Konferenz in den 1950ern (Dartmouth), ein Haushaltsroboter, Star Trek und HAL9000, die Google-Suche, ein Forschungsfeld und eine Forschungsstrategie gemeinsam? Antwort: Sie können unter dem Inbegriff KI zusammengedacht

---

<sup>1</sup> Vgl. Cassirer (1995, S. 265).

<sup>2</sup> „Die Concreta, auf welche sich die abstrahierende Thätigkeit bezieht, haben wir bereits bezeichnet, es sind Inbegriffe bestimmter Gegenstände; wir fügen jetzt auch hinzu: vollkommen willkürlicher. In der That, für die Bildung concreter Inbegriffe giebt es in Beziehung auf die zu befassenden Einzelinhalte keinerlei Schranken. Jedes Vorstellungsobject, ob physisch oder psychisch, abstract oder concret, ob durch Empfindung oder Phantasie gegeben, kann zusammen mit einem jeden und beliebig vielen anderen zu einem Inbegriffe vereinigt und demgemäss auch gezählt werden. Z. B. Einige bestimmte Bäume; Sonne, Mond, Erde und Mars; ein Gefühl, ein Engel, der Mond und Italien u. s. w. Immer können wir in diesen Beispielen von einem Inbegriffe, von einer \* Vielheit und von einer bestimmten Zahl sprechen. Auf die Natur der einzelnen Inhalte kommt es also in keiner Weise an.“ Husserl (1891), S. 10–11.

werden. Dabei sind Inbegriffe nicht bloß ein Set oder eine Sammelkiste von Einzelteilen, sondern sie zeichnet eine je spezifische Relation zu ihrem Ganzen aus.<sup>3</sup>

KI als Inbegriff zu fassen, löst also das Problem, Imaginationen und Code, Hoffnungen und IT-Strukturen, deren Vorstellungen und Diskursmanifestationen zusammenzudenken. Zudem hilft auch Husserls Inbegriffdefinition dabei, drei wichtige Aspekte des KI-Phänomens von Anfang an in den Blick zu bekommen und zu behalten: das ausschlaggebende *Interesse* der jeweiligen Inbegriffassung, die konstitutive und andere Begriffsdefinitionen verunmöglichte *Heterogenität* der Elemente sowie das die Vielheit zu einem Ganzen vereinende einheitliche *Bemerken* bzw. das in *einem* Akte denken. Würden nur kategorial homogene Elemente zu jeweils verschiedenen KI-Teilbegriffen gefasst, blieben die jeweils anderen Elemente unverbunden außen vor. Dann könnte ein KI-Diskurs höchstens innerhalb homogener (z. B. je informatischer, politischer, juristischer etc.) Teilbereiche stattfinden und würde ansonsten riskieren, zu einer Aneinanderreihung von nicht aufeinander bezogenen und beziehbaren Parallelstatements über völlig Differentes zu geraten; dies wäre einzig durch das Homonym ‚KI‘ zusammengehalten, was dann ebenso wenig *ein* Diskurs ergibt wie Aussagen zu Toren (je als törichte Menschen und als Fußballtore). Ohne ein zusammendenkendes Bemerken und Interesse haben Aussagen über KI als *superhuman general intelligence* eben nichts Wesentliches mit Aussagen über KI als Streamingdienst-Empfehlungen gemein.

Dabei orientieren Inbegriffe wiederum das Verstehen von Begriffen und beispielsweise metaphorische Bedeutungsübertragungen. Inbegriffe lassen sich beliebig viele bilden, je nach Bemerken, Interesse und Denk-Akt, womit *Menschenbilder* (z. B. Homo faber, Homo deus, Mängelwesen, rationaler Entscheider) und *Technikbilder* (z. B. autonome Systeme, robotische Sozialpartner) ebenso wie *Intelligenzbilder* (rechnende, emotionale, soziale etc. Intelligenz) als verschiedene Inbegriffe mit verschiedenen Elementen in Betracht kommen. Was wiederum auf Begriffsebene verstanden wird und daraufhin als Handlungs- und Entscheidungsorientierung wirkt, ist bedingt durch die spezifischen Inbegriffe. Was KI also sein kann und wie wir unser Handeln in Bezug auf KI orientieren, ist bedingt durch

---

<sup>3</sup> „Es ist missverständlich zu sagen, die Inbegriffe beständen bloss aus den Einzelinhalten. Wie leicht man es auch übersieht, so ist doch über die Einzelinhalte hinaus etwas da, was bemerkt werden kann und was in allen Fällen, wo wir von einem Inbegriff oder einer Vielheit sprechen, notwendig vorhanden ist: die Verbindung der einzelnen Elemente zu dem Ganzen. Und es verhält sich hier, wie bei manchen anderen Klassen von Relationen es kann bei der grössten Verschiedenartigkeit der bezogenen Inhalte, doch in Hinsicht auf die verbindenden Relationen Gleichartigkeit bestehen.“ Husserl (1891). S. 13–14.

Inbegriffe wie verschiedene Technik-, Menschen-, Intelligenzbilder aber auch KI-Bilder. Dabei können ‚KI-Bilder‘ einerseits als Teilmenge von Technikbildern in Betracht kommen, nämlich insofern lediglich technische Elemente und lediglich solche technische, die spezifisch für KI sind, hervorgehoben werden. Andererseits werden unter KI-Bildern gerade auch nicht-technische und nicht KI-technische Vorstellungen von Menschlichem, Intelligentem, Göttlichem etc. orientiert, sodass Technik-, Menschen-, Intelligenzbilder etc. als Teil von KI-Bildern in Betracht kommen.

Was leistet ein Rückgang auf KI als Inbegriff und weitere wie Menschen- und Technikbilder – auch wie später zu zeigen sein wird Tier- und Götterbilder – neben der Verbindung des sonst Unverbundenen? Sie orientieren die Bedeutungsübertragungen metaphorischen Begriffsgebrauchs, der v. a. im Bereich von KI als reaktionsfreudiges Phänomen zentral ist. Solche Übertragungsfreudigkeit oder semantische Reaktionsfreude meint – selbst eine Metapher aus der Chemie – die Fähigkeit eines Begriffes (hier KI) Reaktionen mit anderen Konzepten (wie Kreatur, Kind/Nachkomme etc.) einzugehen und in der Folge mit diesen mitverstanden zu werden bzw. deren Verstehen zu orientieren. KI ist nicht nur Inbegriff, sondern auch eine Metapher und als solche unterliegt sie spezifischen Übertragungsproblemen.

---

### 3 Übliche Uneigentlichkeit – KI als Metapher

Die Elemente des Inbegriffs KI sowie die Bezeichnung ‚KI‘ selbst sind irreduzibel bildlich und uneigentlich in Gebrauch. Intelligenz, Entscheiden, Lernen, Wahrnehmen, Handeln etc. sind alles Übertragungen von (vor allem, aber nicht nur) Menschen auf technische Systeme. Dabei sind solche bildhaften Verwendungen v. a. Metaphern oder Analogien, aber auch (nicht zufällige<sup>4</sup>) Äquivokationen etc. Aus Gründen des Umfangs wird hier nicht zwischen den einzelnen rhetorischen Figuren wie Metapher, Metonymie oder Synekdoche etc. unterschieden,

---

<sup>4</sup> Die Arten des Äquivoken „kommen ‚zufällig‘ (casu) oder durch willentliche Benennung (consilio) zustande; letztere haben ihren Grund in einer Ähnlichkeit (similitudo) oder einem Verhältnis der Entsprechung (proportio), der Abkunft (ab uno) bzw. der Hinordnung (ad unum).“ Kluxen (2010) In Unterscheidung zum Metaphorischen ließe sich auch von Äquivokationen bspw. bei Intelligenz im menschlichen oder technischen Sinne sprechen. Anders als bei Toren (Törichte und Fußballtore), wäre diese aber nicht zufällig (casu), sondern vermutlich durch ein näher zu analysierendes Verhältnis (proportio) bzw. Ähnlichkeit (similitudo) zustande gekommen.

was aber eine spannende Vertiefung darstellte. Der KI-Diskurs, so lässt sich festhalten, findet in Bildern statt; in Sprachbildern, in Technikbildern etc. und in visuellen Bildern (zu letzterem vgl. Manzeschke und Gransche 2023). Hier wird exemplarisch ein Fokus auf Metaphern gelegt,<sup>5</sup> da Metaphern unhintergebar und konstitutiv für unser Verstehen, unser Denken und Handeln sind:

“If we are right in suggesting that our conceptual system is largely metaphorical, then the way we think, what we experience, and what we do every day is very much a matter of metaphor. [...] metaphor is pervasive in everyday life, not just in language but in thought and action. Our ordinary conceptual system, in terms of which we both think and act, is fundamentally metaphorical in nature.” (Lakoff und Johnson 2003, 1.)

In Anlehnung an Nietzsches Antwort auf die Frage „Was also ist Wahrheit?“ ließe sich die gleiche Antwort auf die Frage „Was also ist KI?“ geben:

„Ein bewegliches Heer von Metaphern, Metonymien, Anthropomorphismen kurz eine Summe von menschlichen Relationen, die, poetisch und rhetorisch gesteigert, übertragen, geschmückt wurden“ (Nietzsche 2009, 1.).

Im Kern wird bei einer Metapher eine Bedeutung von A nach B übertragen. Dabei müssen A und B als *ähnlich* angenommen werden, also sowohl identische Aspekte haben als auch differente. Zunächst wäre an der Bezeichnung ‚KI‘ selbst anzusetzen, die als komprimierte bzw. verkapselte Metapher ausgelegt werden kann, die in einer explizierten Form in etwa lautet:

/Solche IT-Systeme sind intelligent./

Mit dem Partikel ‚wie‘ als Vergleich erweitert, hieße das etwa:

/Solche IT-Systeme sind wie natürlich intelligente Entitäten im Problemlösen o.ä., nur ohne Emotionen usw. und mit perfektem ‚Gedächtnis‘/Datenspeicher usw./

Die Elemente dieser Formulierung lassen sich noch formal benennen:

/[**Solche IT-Systeme** (Bedeutungsziel B)] [**sind wie** (Identitätsmarkierung)] [**natürlich intelligente Entitäten** (Bedeutungsquelle A)] [**im Problemlösen o.ä.** (Übertragungshinsicht H)], **nur** (Differenzmarkierung) [**ohne Emotionen usw.** (Mängeldifferenz x)] **und** [**mit perfektem ‚Gedächtnis‘/Datenspeicher** (Überschussdifferenz y)]./

Rein formal ergibt sich dann:

---

<sup>5</sup> Auch weil es dafür zahlreiche theoretische Fundamente gibt, weiteres hierzu: Lakoff und Johnson (2003); Blumenberg (2001), (1998); Friedrich (2013); Heidenreich (2020). Auch in Bezug auf KI finden sich Vorarbeiten, z. B.: Barnden und Lee(2001); West und Travis (1991); Carbonell et al. (2016).

/B ist A in H ohne x mit y./

In einem ersten Schritt könnte man also sagen, dass KI als Unbekanntes durch die Referenz auf bekannte menschliche Intelligenz verstanden werden kann, nämlich als etwas Intelligentes wie Menschen, aber nicht in identischer Weise, sondern eben künstlich. Wie in obiger Definition der HEG-KI bereits gesehen, ist der Referenzbegriff „menschliche Intelligenz“ alles andere als *bekannt* und in diesem Sinne auch nur sehr eingeschränkt eine Bedeutungsquelle. Diese Sicht von KI als /Solche IT-Systeme sind intelligent wie Menschen nur künstlich/ hat – wie jede bildhafte Übertragung – triftige Probleme:

1. Das Bedeutungsziel (B) ist nicht hinreichend bekannt zur Übertragungsorientierung – Was sind *solche Systeme* genau? Es kommen bei Metaphern immer viele Bedeutungsaspekte als Übertragungskandidaten in Betracht; manche davon helfen, ein Verständnis von B zu orientieren, andere führen in die Irre. Um also zu entscheiden, welche sinnvoll sind, muss das unbekannte B bereits in einem Mindestmaß bekannt sein. Bei völlig unbekanntem B käme schlechthin alles zur Übertragung in Betracht, womit B und A aber identisch und nicht bloß ähnlich wären. Im Falle von KI ist wegen der Unbestimmtheit des Begriffs mitunter (und relativ zu gerade gebrauchten KI-Konzepten) zu wenig bekannt, um eine Übertragung von anderen Bedeutungsquellen sinnvoll zu orientieren. Dies ist auch ein Grund für die ‚Promiskuität‘ bzw. hohe Reaktivität des KI-Begriffs mit anderen Konzepten wie Kindern, Tieren, Dienern oder Golems etc. Die Unbestimmtheit lässt es kaum zu, Kopplungen mit solchen Konzepten zu irritieren und zurückzuweisen. Auf KI lässt sich somit so gut wie alles übertragen.
2. Die Bedeutungsquelle (A) ist nicht hinreichend bekannt zur Orientierung der sinnvollen Übertragungsoptionen – Was ist eigentlich (nicht-künstlich) *intelligent*? Intelligenz ist ein ihrerseits unterbestimmtes Konzept, dem heterogene Alltagsbedeutungen zugeschrieben werden (können) und teils hochdiverse wissenschaftliche Phänomene. Der Strauß an Intelligenzhinsichten zeigt die Diversität: *Emotional/ Practical/ Social/ Cultural/ Mating* (etc.) *Intelligence* (vgl. Sternberg und Kaufman 2011). A muss als Bedeutungsquelle aber insofern bekannt sein, als – noch bevor das Problem der orientierten Auswahl der Übertragungsoptionen auftritt – das Set an Aspekten gekannt werden muss, die allererst zur Auswahl stehen.
3. Einhergehend mit der unzureichenden Bekanntheit von A und B sind damit die Identitätselemente und Differenzelemente in Art und Menge unklar. Wird eines der beiden 0 % oder 100 %, dann sind A und B nicht mehr ähnlich, sondern

eben identisch oder völlig different, dann ist eine Ähnlichkeitsoperation qua Metaphern auch nicht mehr möglich.

4. Bei beidem (A und B) fehlt mitunter die kulturelle Tradierung, also eine überlieferte Bedeutung von Phänomenen, die wir kennen und von denen wir über ein tradiertes Set an Bedeutungsaspekten verfügen; auch ohne eigene Erfahrungen mit dem jeweiligen Phänomen. Jeder versteht so z. B. „Fußballgott“ oder „Höhle des Löwen“, ohne jemals einen Gott oder einen Löwen getroffen zu haben. Speziell KI selbst kann nicht auf diese Verankerung in der Geistesgeschichte verweisen und wird daher gekoppelt mit Phänomenen, die diese Verbindung herstellen können.<sup>6</sup>
5. Bei den Übertragungen handelt es sich nicht um lineare A-nach-B Bewegungen, sondern um ein Wechselspiel von Bedeutungen zwischen bekannteren und weniger bekannten Elementen, weshalb sich A und B auch qua Bekanntheitsgrad invertieren lassen, womit dann u. U. das historisch spätere zur Bedeutungsquelle des historisch früheren werden kann. Dies könnte z. B. bei der Bezeichnung *Roboter* der Fall sein, dessen tschechisches Ursprungswort (*robota*) Fronarbeiter hieß.<sup>7</sup> Die Ursprungsmetapher war also /Solche Maschinen dienen dem Menschen wie Fronarbeiter ihrem Feudalherren./<sup>8</sup> Nun sind heute Roboter ein ggf. bekannteres Phänomen als Fronarbeiter, weshalb die Inversion der Übertragung u. U. höheren Orientierungswert hat, also: /Fronarbeiter dienten einem Feudalherren wie heute Roboter den Nutzern./ Besonders bei Letzterem wird deutlich, wie bei solchen Hin- und Rückübertragungen in problematischer Weise Bedeutungsaspekte und Hinsichten aus dem Blick geraten; wie bspw. der Zwangsaspekt, der bei Agenten mit eigenen genuinen Präferenzen (z. B. Menschen) eine sinnvolle Hinsicht ausmacht und ethische Erwägungen erfordert. Bei Agenten ohne eigene genuine Präferenzen (z. B. Roboter) ist die Annahme von Zwang hinfällig und höchstens metaphorisch anregend, aber irreführend. Roboter wollen und präferieren nichts (nicht einmal die eigene Fortexistenz) und können deshalb auch zu nichts gezwungen werden, das diesen Präferenzen entgegenstände und in dieser Opposition überhaupt erst als Zwang gelten könnte.

---

<sup>6</sup> Vgl. Zimmerli (2021); Mayor (2018); Musa Giuliano (2020).

<sup>7</sup> “The etymological locus classicus by Josef and Karel Čapek, who defined “robot” in their play “Rossum’s Universal Robots (R.U.R.)” in 1920/1923 as a forced labourer (from Czech “robota” = forced or compulsory labour, corvée).” Funk und Coeckelbergh (2019).

<sup>8</sup> Gemäß der vorgeschlagenen Formel wäre das: /Solche Maschinen (B) sind wie menschliche Fronarbeiter (A) im Dienen (H) nur nicht einem Feudalherren (x), sondern dem Nutzer/Betreiber etc. (y)./



6. Es besteht das Problem der reziproken Angleichung der Bedeutungselemente und damit einer Transformation aller Elemente hin auf die gebräuchlichsten (s. Abschn. 5)

„...die Wahrheiten sind Illusionen, von denen man vergessen hat, dass sie welche sind, Metaphern, die abgenutzt und sinnlich kraftlos geworden sind, Münzen, die ihr Bild verloren haben und nun als Metall, nicht mehr als Münzen in Betracht kommen.“ (Nietzsche 2009, 1.)

In diesem Sinne wäre nicht mehr eine eigentliche Verwendung die Bedeutung eines Konzeptes, die dann um entsprechende Übertragungshinsichten (H) und Differenzen (x, y) ergänzt auf andere Bereiche teilweise angewandt wird, sondern alle Konzepte (z. B. Systeme, Algorithmen, Menschen, Tiere, Pflanzen etc.), die mit dem gleichen Wort (z. B. intelligent) verwendet werden, machen die Extension der Bedeutungsdefinition von Intelligenz gemeinsam aus. Damit hätte das Intelligente eben spezifische Unterarten wie technisch, menschlich, natürlich, sonstige. Das Primärverständnis von intelligent wäre dann das je gebräuchlichste und bei den weniger gebräuchlichen müsste man ihre Spezifika gesondert erwähnen. Beispiel: Solange menschliche Intelligenz der dominante Gebrauch des Wortes ist, muss bei KI eben ‚künstlich‘ (oder mustererkennend, rechnend etc.) hinzugefügt werden, während bei menschlicher Intelligenz dieser Hinweis entfallen kann. Niemand würde sagen: ‚Diese Person ist menschlich intelligent.‘ Wenn allerdings technische Intelligenz zur dominanten Verwendung von Intelligenz würde (wie es im Chinesischen der Fall zu sein scheint<sup>9</sup>), dann müssten die entsprechend abweichenden Aspekte nicht-technischer Intelligenz (soziale, emotionale etc.) hinzugefügt werden. So kann es passieren, dass Maschinen nicht nur metaphorisch, sondern eigentlich *intelligent* werden; nicht, weil sie Leistungen menschlicher Intelligenz erreicht haben werden, sondern weil die Bedeutung von Intelligenz sich zu dem transformiert haben wird, was Maschinen bereits können.

7. Weiter verkompliziert wird die übertragene Rede durch Komprimierung bzw. Verkapselung von Metaphern in Begriffsgebilde. So erscheinen Roboter und KI als Begriffe und ihre metaphorische Ausfaltung /Solche Maschinen sind

---

<sup>9</sup> Ein besonderer Dank ist geschuldet den Beteiligten am Deutsch-Chinesischen Workshops Künstliche Intelligenz und Governance – Internationale Tagung vom 04.11.–05.11.2021 am KIT in Karlsruhe, mit denen die kulturellen Differenzen des KI-Begriffs diskutiert werden konnten. Nach der Einschätzung der chinesischen Beteiligten der Tongji University werde in China das chinesische Wort für Intelligenz nur noch in der Kombination als KI verwendet, in Bezug auf Menschen kämen andere Begriffe zum Einsatz.

Zwangsarbeiter./ bzw. /Solche IT-Systeme sind intelligent./ wird nicht mehr kommuniziert. Mit dieser Verkapselung werden die Metaphern implizit und relevante Aspekte wie Übertragungshinsicht und Differenzen werden nicht mehr expliziert. Es bleibt unklar, unterbestimmt, unbewusst und entsprechend irrtumsanfällig, in welcher Hinsicht die Übertragung ein Verstehen orientieren könnte.

---

## 4 Menschliches, Übermenschliches – KI im erweiterten Spektrum

Bei der Betrachtung der unterschiedlichen Bedeutungsquellen, die genutzt werden, um das Verstehen (oder Vermarkten etc.) von KI zu orientieren, liegt es nahe und sind weit verbreitet einerseits anthropomorphe Technikbilder und andererseits technomorphe Menschenbilder. Im ersteren Fall werden Konzepte, die ihren üblichen Sinn in Anwendung auf Menschen haben, auf Technik übertragen (wie z. B. intelligente Maschinen), im letzteren Fall werden Konzepte des Technischen auf Menschen übertragen (wie z. B. das Gehirn als Rechner). Exemplarisch dazu etwa Rudolf Seising in „Es denkt nicht!“<sup>10</sup>:

„Das Gehirn wurde als eine Art natürlicher Computer angesehen und der Computer als eine Art künstliches Gehirn. [...] Aber weder sie noch die darauf aufsetzende Metapher eines Computerprozesses für den Geist, die nahelegt, dass sich der menschliche Geist zum Gehirn verhält wie die Software eines Computers zu dessen Hardware, beruht auf wissenschaftlichen Grundlagen.“ (Seising 2021, S. 118)

Diese beiden wechselseitigen Übertragungsquellen und -ziele sind aber nicht die einzigen und besonders im verstehenden Zugriff auf KI muss das Spektrum der Bedeutungsquellen deutlich erweitert werden. Neben dem Menschlichen als üblichem Bedeutungskontext sollte auch Übermenschliches, Übernatürliches oder Göttliches einerseits und Tierisches, belebt Natürliches oder Organisches andererseits berücksichtigt werden. Zwischen diesen Bereichsmarkierungen eines als

---

<sup>10</sup> Rudolf Seising macht in seiner Untersuchung „Es denkt nicht“ sehr schön deutlich, wie in dem Zusammengehen zweier zuvor getrennter Forschungsgebiete (der elektronisch basierten Informations- und Kommunikationstheorie einerseits und der biologisch und dann über das elektrische Reizleitungspotenzial der Nervenzellen auch elektrisch basierten Psychologie und Kognitionswissenschaften andererseits) die Analogie von Gehirn und Computer systematisch genutzt und zu einem metaphorischen Vokabular (z. B. Erinnern, Erkennen, Denken, Entscheiden, Generalisieren, Lernen, Verstehen usw.) ausgebaut wird – vgl. Seising (2021).

kontinuierlich zu denkenden Spektrums finden in alle Richtungen Übertragungen statt, was an dieser Stelle nicht für jede Kombination ausgeführt werden kann. So werden anthropomorphe und technomorphe Konzepte nicht nur auf Menschen und Technik, sondern auch auf Tier- und Göttervorstellungen übertragen; es werden auch theomorphe oder zoomorphe Konzepte auf alle Instanzen des Spektrums übertragen etc. Als Tabelle mit Bedeutungszielen, also Bildern von Menschen, Technik etc. in der X-Achse und mit Bedeutungsquellen, also techno-, anthropomorphe etc. Vorstellungen in der Y-Achse gedacht, lassen sich kreuzklassifikatorisch die verschiedenen Übertragungskombinationen sortieren. Formal lassen sich alle Kombinationen als *n-morphe X-Bilder* fassen, wobei (in einer nicht vollständigen Annäherung) für *n* mindestens *theo-*, *anthropo-* (inkl. *andro-*, *gyno-*), *zoo-*, *therio-*, *phyto-* und *techno-morph*<sup>11</sup> zur Berücksichtigung anstehen. Entsprechende Optionen für *X* wären beispielsweise Götter, Heroen, Übermenschen, Menschen (Männer, Frauen, Kinder), Tiere, Pflanzen, Technik. Dabei lassen sich viele Hierarchisierungen und Binneneinteilungen vornehmen wie *belebte Natur* für Menschen, Tiere, Pflanzen, *Gemachtes* für Götter und Technik (teils auch Tiere und Pflanzen), *Gewordenes*, *Künstliches*, *Kultürliches* etc. Dabei stehen stets gewisse Verstehens- bzw. Übertragungsinteressen im Fokus, mit denen auf spezielle Eigenschaften und Aspekte abgehoben wird. Es wird klar, dass z. B. Menschen weder als Element der Menge „Natur“ aufgehen, noch nur als nicht-natürlich oder über-natürlich infrage kommen, ebenso sind sie je nach Hinsicht sowohl geworden, als auch gemacht, zeichnen sich z. B. durch „natürliche Künstlichkeit“ (Plessners erstes anthropologisches Gesetz) aus etc. Es kann also nicht um eine wie auch immer geartete ontologische Einteilung gehen, sondern um Hinsichten der Vorstellungsorientierung im Sinne von Reflexionsbegriffen. Wie der Reflexionsbegriff Raum es uns erlaubt, etwas Räumliches an etwas vorzustellen (durchaus auch im Sinn von Nach-vorne-stellen, Herausheben), so erlaubt es uns der Reflexionsbegriff Technik, etwas Technisches an etwas zu thematisieren; Kultur, Kultürliches an etwas; Natur, Natürliches an etwas etc. Ein theomorphes Menschenbild (z. B. ebenbildliches Geschöpf Gottes, Hobbes *De Cive*: „Homo homini Deus“) orientiert andere Vor-Stellungen des Menschen, als ein zoomorphes Menschenbild (z. B. Säugetier, Hominidenart, Hobbes *De Cive*: „Homo homini Lupus“) und wiederum andere als ein technomorphes Menschenbild (z. B. La Mettrie 2015: *L'homme machine*, Pneumatisches/Hydraulisches wie Herz als Pumpe, Elektrotechnisches wie Nerven

---

<sup>11</sup> Theomorph = gottförmig, anthropomorph = menschenförmig, andromorph = mannförmig, gynomorph = frauformig, zoomorph = tierförmig, theriomorph = tiergestaltige Götter, phytomorph = pflanzenförmig, technomorph = technikförmig.

als Leitungen, Gehirn als Schaltzentrale oder Nervenbahnen als Telegrafennetz<sup>12</sup> etc.). Das gleiche gilt für Technikbilder: Anthropomorphe Technikbilder – z. B. autonome, intelligente, sensitive Agenten – orientieren andere Verstehensprozesse und Vorstellungs- sowie Handlungsorientierungen als zoomorphe Technikbilder – von den Roboterwachhunden der Phaiaken aus Homers Odyssee<sup>13</sup> bis zur Pflege-Roboterrobbe PARO (2014) – und wiederum andere als theomorphe Technikbilder – z. B. die Matrix (Wachowski und Wachowski 1999), Samantha aus Her (Jonze 2013), Talos mit olympischem Ichor in seiner Ader (Mayor 2018, S. 7 ff.) oder speziell auch KI. Wenngleich alle Übertragungskombinationen interessant sind, so stehen hier – mit Blick auf Technik- und Menschenbilder und KI-Diskurse – v. a. jene Kombinationen mit Technik und speziell KI im Fokus. Es zeigt sich dabei aber, dass gerade KI ein hoch reaktionsfreudiges Element in diesem Spektrum darstellt, dass nicht nur häufig anthropomorph gefasst wird, sondern das wiederum das Menschenbild in seinem Verhältnis etwa v. a. zum Übermenschlichen (Trans-, Post- oder Superhumanen), Magisch-Mächtigen oder Göttlichen ändert.

---

## 5 Der neue Prometheus: KI als Geschöpf und Menschen als Schöpfer

*Am Ende hängen wir doch ab*

*Von Creatures die wir machten*

Goethe, Faust II, 2. Akt

Die Vorstellung vom Menschen als Geschöpf bspw. eines biblischen Gottes, anderer Götter oder des Titanen Prometheus wurde in Bezug auf Technik mehrfach

---

<sup>12</sup> Letzteres wurde von Ernst Kapp unter vielen weiteren als Beispiel der Organprojektion in seiner Technikphilosophie prominent gemacht. Auch Organprojektion ist hier ein Zugriff auf Ähnlichkeiten, der vor einem Urbild/Original-Abbild Verhältnis betrachtet werden muss. Vgl. Kapp (1877).

<sup>13</sup> “We encounter a familiar range of AI prototypes exhibiting increasing levels of ‘intelligent’ capabilities too: ranging from the mundane household gadget (the automated gates), the industrial tool (Hephaestus’ automatic bellows), the exclusive/expensive AI device designed to impress (Hephaestus’ automated tripods), the AI device with potential for criminal exploitation (Hephaestus’ automated tripods again), the security system (the Phaeacian dogs), the self-steering autonomous vehicle (the Phaeacian ships), and the anthropomorphic AI assistant (Hephaestus’ amphipoloi).” Livey und Thomas Sam (2020), S. 43.

invertiert und der Mensch selbst in die Schöpferrolle gebracht.<sup>14</sup> Frühe Narrative hierzu wären die Geschichten vom Golem von Prag oder das Monster von Frankenstein. Beide wurden von Menschen geschaffen jeweils mittels einer Methode, die beide dem Grundsatz ‚gleiche Ursache, gleiche Wirkung‘ verpflichtet sind, nämlich Magie und Wissenschaft/Technik:

„Die Magie unterscheidet sich von der Wissenschaft zwar im Resultat, aber nicht in ihrem Prinzip und in ihrem Problem. Denn der Grundsatz: „Gleiche Ursachen, gleiche Wirkungen“ beherrscht auch sie und gibt ihr das Gepräge, das sie durchgängig zeigt.“ (Cassirer 1985, S. 56)

Der Golem von Prag ist eine Figur aus der jüdischen Literatur, die bis ins Mittelalter zurückreicht; ein aus Lehm geschaffener Humanoide mit übermenschlicher Größe und Kraft, der als eine Art autonomer Wachroboter für die jüdische Gemeinde funktionierte. Einen antiken Vorläufer findet der Golem im von Hephaistos geschaffenen bronzenen anthropomorphen ‚Roboter‘ Talos, der die Insel Kreta beschützte und in dessen Ader olympisches Blut – sogenanntes Ichor – floss (Mayor 2018, S. 7 ff.).<sup>15</sup> In einer späteren Version der Geschichte des Golems von Prag kommen die Elemente des Kontrollverlustes und der Auflehnung des künstlichen Geschöpfes gegen seinen menschlichen Schöpfer zum Tragen, wenn der Golem – nicht mehr zu bändigen – statt der potenziellen Feinde die eigene Gemeinde verwüstet. Dieses Element hat seine Vorläufer im Hybrismotiv des Ikarus-Mythos, in dem das Technikergenie Dädalus als heroisches sterbliches Äquivalent des Technikergottes Hephaistos unter dem Tod seines von ihm prothesenbeflügelten Sohnes zu leiden hat, der durch die Technikermächtigung übermütig zu hoch flog – also der Sonne und damit dem Olymp und olympisch göttlich zu sein zu nahe kam. Dieses Hybrismotiv spiegelt sich auch in Goethes berühmtem Zauberlehrling-Gedicht mit dem sprichwörtlichen Ausruf „Herr, die Not ist groß! Die ich rief, die Geister werd ich nun nicht los.“ Es findet vor allem in technikpessimistischen Debatten über die sogenannte starke KI einen Widerhall; z. B. der prominent unterzeichnete AI Open Letter: „our AI systems must do what we want them to do“ (Future of Life Institute 2015). Theomorph ist dieses Menschenbild vom Schöpfer intelligenter, eigenständiger/autonom und möglicherweise eigenwilliger oder aufbegehrender Geschöpfe, insofern diese Art des Schöpfertums geistesgeschichtlich eine primär göttliche

<sup>14</sup> Vgl. Foerst (2008).

<sup>15</sup> Talos, der Schiffe vor Kreta mit Steinwürfen versenkte, sich stark erhitzen und Feinde an der glühenden Bronzebrust zerquetschen und verbrennen konnte, ist Namensvorbild für ein seegestütztes Boden-Luft-Lenkwaffensystem der US-Navi, vgl. Garten und Dean (1982).

Zuschreibung ist, die erst später auch für übernatürlich unterstützte Schaffende (Genies, Genius; „große Menschen“<sup>16</sup>) in Anschlag gebracht wurde. Dies gilt nicht nur für den biblischen Gott, sondern beispielsweise auch für den titani-schen Menschenschöpfer Prometheus.<sup>17</sup> In dieser Figur ist auch das Motiv des Aufbegehrens zentral, da Prometheus selbst sich gegen den obersten Machthaber Zeus stellt, indem er den Menschen das von Zeus gestohlene olympische Feuer zur Verfügung stellt und damit Technik, Wissenschaft und Kultur (*techne*). In dieser Linie wäre auch KI eine der aktuellsten Manifestationen der prometheischen Gabe. Genau mit diesem Geschenk stellt Prometheus den Menschen die Mittel zur Verfügung, sich von der Gunst der Götter unabhängig zu machen und zu emanzipieren. Hier zeigt sich das Motiv ‚Technik statt Gott‘ bzw. ‚wer Technik hat, braucht nicht auf göttliche Hilfe zu hoffen, noch göttliche Strafe zu fürchten‘. Dieses Emanzipationsmotiv wird v. a. in der Prometheus-Rezeption ab der Renaissance wichtig und zeigt sich wiederum in Goethes Prometheus-Hymnus: „Ich dich ehren? Wofür? [...] Hier sitz ich, forme Menschen / Nach meinem Bilde / Ein Geschlecht das mir gleich sei: / Zu leiden, zu weinen / Zu genießen und zu freuen sich / Und dein nicht zu achten / Wie ich!“ (Goethe o. D.)<sup>18</sup> Hier zeigen sich ursprüngliche Schöpfervorstellungen – bereits vor der christlichen Genesis – verschränkt mit emanzipatorischen Autonomieerklärungen sowie poetische Bild-Bildner-Verhältnisse:

---

<sup>16</sup> So beginnt auch die bekannteste Golem von Prag Erzählung: „Große Menschen waren einst durchaus im Stande große Wunder zu vollbringen.“ (Jizchok-Leib Perez, 1890)“ Augustin (2016). Wunder zu vollbringen war ebenso eine dezidiert göttliche Domäne und es wundert nicht, dass die KI-Forscher sich aktiv narrativ selbst an die Golemerzählung koppeln (s.u.), denn wenn nicht direkt Selbstvergötterung, so liegt damit eine Selbstinszenierung als „große Menschen“ nahe.

<sup>17</sup> In der biblischen Tradition ist im Unterschied zur antiken die Schöpfung eine kontinuierliche im Sinne eines kontinuierlich erhaltenden Handelns Gottes. Was in einem Moment ex nihilo geschaffen wurde, bedarf demnach dann der Erhaltung durch einen außerhalb dieser Schöpfung stehenden Gott. Sofern man dieses Moment der biblischen Schöpfungstradition unterschlägt, lässt man schnell Innovation und Kreation zusammenfallen und verfehlt damit die theologische Pointe des Kreativen in diesem Ansatz.

<sup>18</sup> Wenn mit der KI einerseits ein emanzipatorischer Gedanke verbunden wird, dass diese menschliche Schöpfung sich von ihrem Schöpfer abwenden und ‚eigene Wege‘ gehen könnte, so wird doch andererseits auch das Motiv von der göttlichen Erhaltung seiner Schöpfung in biblischer Tradition aufgerufen: Wenn z. B. der KI-Forscher Schmidhuber davon spricht, dass wir Menschen vor einer möglichen Übermacht der KI keine Angst zu haben brauchten, weil diese sich uns gegenüber erhaltend-fürsorglich zeigen werde wie wir Menschen gegenüber unseren Hauskatzen (vgl. Wiedemann 2017).

„Was in diesem Prozeß geschaffen wird, ist nicht die Menschheit im biologischen Sinne, sondern das Menschen-„Geschlecht“ als emanzipiertes Subjekt seiner eigenen Geschichte, ein „Geschlecht“ eben „nach meinem Bilde“ – und das kann nur heißen: nach dem Bild, das Prometheus in dem Gedicht von sich selbst entwirft. [...] als Autonomieerklärung des ästhetischen „Genies“, ist die Schlußstrophe immer wieder als Selbstvergötterung des Künstlers gelesen worden, der sich an die Stelle des biblischen Schöpfers setzt. Ihre blasphemische Zuspitzung gewinnt diese Selbstvergötterung allerdings erst daraus, daß sie die Gottesebenbildlichkeit (Schmidt, S. 267), ja mehr: jede Abbildlichkeit des Menschen (Wellbery, S. 322ff.) negiert. Nicht „Gott schuf den Menschen ihm zum Bilde“ (1. Mose 1,27), vielmehr sind die Götter anthropomorphe Projektionen der menschlichen Phantasie. Vor allem aber gilt: das „Geschlecht“, das Prometheus ‚nach seinem Bilde‘ formt, wird ihm gerade darin „gleich“ sein, daß es sich seine eigenen, originären Züge verleiht.“ (Witte et al. 2004, S. 113f.)

Oder – ließe sich 250 Jahre nach Goethes Hymnus aktualisieren – die Menschen sehen sich gerade darin als *schöpferhaft* (göttlich, titanisch, qua Prometheusgabe), als sie nun ein ‚Maschinen-Geschlecht‘ nach ihrem Bilde formen, dass ihnen gerade darin gleicht, dass es seine eigenen originären ‚Züge‘/Strukturen ausformt; nämlich durch maschinelle Intelligenz, Lernen und technische Autonomie vom bevormundend steuernden Menschen.

In Gestalt der theomorphen Technikbilder findet die Vorstellung vom Menschen als technogener Gott (Prothesengott, Freud 1930, 3.1) eine weitere Wendung in der Vorstellung von Technik selbst als neuem Gott bzw. von Technik mit Gotteseigenschaften. Damit wäre auf problematische Weise die erreichte Emanzipation von übermenschlichen Instanzen wieder eingerissen und nur die ‚Leitungsposition‘ ausgetauscht; eine theomorphe Überhöhung von Technik stellte den Menschen erneut in eine orientierungslose nicht-emanzipierte Abhängigkeit von etwas ‚Höherem‘ – einem „übernatürlichen Herrn“ (Gadamer s. u.) –, was, mit Goethe gesprochen, bereits Prometheus überwunden hatte und mit ihm in der Spur Nietzsches und heutiger Säkularisierung aufgeklärte Menschen: „Da ich ein Kind war, / Nicht wußte wo aus noch ein, / Kehrt ich mein verirrtes Auge / Zur Sonne, als wenn drüber wär / Ein Ohr, zu hören meine Klage“ (Goethe o. D.). Die aufgeklärte Erkenntnis, dass *über den Wolken* kein Ohr eine Klage hört (etwa über Kindstode, Hungersnöte, Klimawandel, Kriegsverbrechen etc.), sollte nicht technoreligiös revidiert werden mit der Vorstellung, ein solches Ohr könne es *in der Cloud*, den Daten, Algorithmen und übermenschlichen IT-Potenzen geben.<sup>19</sup>

<sup>19</sup> Die Beispiele in denen KI ‚weltrettende‘ Potenziale zugeschrieben werden und die Fähigkeit, Probleme wie Hunger, Klimawandel, Krankheit und Kriminalität zu lösen sind unzählig;

Ein Narrativ des frühen 19. Jahrhunderts, das den Menschen (und genauer den nicht geschlechtlich, sondern technisch allein ‚(er)zeugungsfähigen‘ Mann – wie Rabbi Löw) als Schöpfer intelligenten (wenn auch unfertig grobschlächtigen<sup>20</sup>) Lebens thematisiert, ist das von Victor Frankenstein (Mary Shelley) und dessen namenlosem Monster, das nicht zufällig den Untertitel „The modern Prometheus“ trägt:

„Prometheus, der titanische Widersacher der olympischen Götter, wird zur Verkörperung der schöpferischen Kräfte des Menschen. Die antike Überlieferung späterer Zeit kennt ihn als den Menschen-Schöpfer. Jetzt sieht sich der Mensch selbst als einen solchen Schöpfergeist. Sein Selbstbewußtsein wird das des Künstlers, des zweiten Machers, des zweiten Gottes, der selbstherrlich und unabhängig seine eigene Welt erbaut.“ (Gadamer 2021, S. 260)

Es lassen sich viele Parallelen zwischen Frankenstein und heutigen KI-Vorstellungen ziehen (vgl. Yalouris 2019)<sup>21</sup>, die meist auch mit den Motiven der Hybris und des Kontrollverlustes von Ikarus bis Zauberlehrling versehen sind. Intelligentes Leben zu schaffen, stellt sich als theomorphe Menschvorstellung dar, insofern eigenständige singuläre Schöpfungsakte (wie dem von Adam und Eva oder eben Prometheus) die Bedeutungsquelle stellen. Als intelligentes Leben schaffend bzw. schenkend kommt neben den Schöpfergöttern/-titanen (der Vater) auch die Frau bzw. das Weibliche – insofern gebärfähig – in Betracht (die Mutter).

Dass gerade KI-Forscher nicht nur passiv in diese theomorphe Sicht gerückt werden, sondern sich aktiv in Bezug zu Golem, Prometheus, Schöpfer etc. selbst positionieren, zeigt beispielsweise Norbert Wieners *God and Golem, Inc.* (1963) – “The Machine, as I have already said, is the modern counterpart of the Golem of

---

exemplarisch vgl.: Strasser et al. (2021); Ekin (2021); DAIA (2018); World Economic Forum (2022).

In aufgeklärter Sicht würde kein solches ‚Ohr über den Wolken‘ angenommen – weder ein religiöses noch ein technoreligiöses. Aus gläubiger Sicht würde in Fundamentaldifferenz von KI und Gott zwar daran festgehalten, dass es ein solches Ohr gebe, das der ‚Menschenkinder Klage‘ höre. Solche Gläubige verstünden sich aber – im Gegensatz zu Apologeten vergöttlichter Technik – gerade darin als ‚aufgeklärt‘, dass sie ihre Hoffnung nicht auf ein selbst geschaffenes ‚KI-Ohr‘ setzten, sondern diese außerhalb ihrer selbst festmachten.

<sup>20</sup> Frankensteins Monster ist gemäß dieser Charakterisierung auch ein Golem, das so viel wie unfertig, unvollendet (etwa wie ein Embryo) bedeutet („hebr. (von galam, zusammenfalten) eig. das unentwickelte, der fötus,“ <https://www.dwds.de/wb/dwb/golem>, zuletzt überprüft: 01.07.2022).

<sup>21</sup> Vgl. auch die Rolle von Frankenstein um artificial life, z. B. Botting (2021).



the Rabbi of Prague.” (Wiener 1963, S. 95) – oder folgende Anekdote über die KI-Anfänge am MIT:

“an astonishing story involving three Jewish AI pioneers from MIT; Joel Moses, Gerry Sussman and their famed teacher, Marvin Minsky. Moses told that on the occasion of his bar mitzvah his grandfather called him apart to tell him that he was a descendent of the actual Rabbi of Prague who had created the original golem, and furthermore that the golem had not been destroyed, as the legend claimed, but was actually dormant in suspended animation. He then proceeded to bestow upon him the secret spell that could awaken the golem, entrusting him to transmit it in turn to future generations. After hearing this, Sussman was speechless. He had been told the exact same story by his own grandfather on his bar mitzvah. Supposedly, each of them then proceeded to go to a corner of the room and write down the spell independently. When they compared both spells, these turned out to be equal. Suddenly, Minsky came out of his office and seeing the students in such a state of shock he asked what was going on. After hearing the story, he said it was utter nonsense, for he too had heard that from his own grandfather on his own bar mitzvah, but had not believed it for a second” (Musa Giuliano 2020, S. 1011)

Die Idee, Leben schaffen zu können, ist einerseits ein theomorphes Menschenbild, andererseits aber auch ein gynomorphes (also frauenartiges) Männerbild. Die KI wird auch als *Kind* von Menschen verstanden und zwar im Gegensatz zu menschlichen Kindern als reines Geisteskind (eine Kopfgeburt und Ergebnis von Rationalität) bzw. *Mind Children* wie der Titel eines Buches von Hans Moravec lautet (Untertitel: *The Future of Robot and Human Intelligence*).<sup>22</sup> Robotische bzw. künstliche Intelligenz wird darin als ‚künstliche Nachkommenschaft‘ der Menschen aufgefasst und direkt mit dem Topos der *Usurpation* gekoppelt. Auch dies ist ein klassischer Topos der griechischen Mythologie, der vor allem um die Figur des Zeus entfaltet wurde, der seinen Vater Kronos<sup>23</sup> in der Titanomachie bekriegte und stürzte und der Prometheus später auch deshalb folterte, weil er

---

<sup>22</sup> Moravec schreibt darin über die kommende von KI bevölkerte postbiologische Welt: “It is a world in which the human race has been swept away by the tide of cultural change, usurped by its own artificial progeny. [...] Today, our machines are still simple creations, requiring the parental care and hovering attention of any newborn, hardly worthy of the word “intelligent.” But within the next century they will mature into entities as complex as ourselves, and eventually into something transcending everything we know – whom we can take in pride when they refer to themselves as our descendants.” (Moravec 1988, S. 1).

<sup>23</sup> Dieser Kronos wiederum stürzte nicht nur seinen Vater Uranos, sondern entmannte ihn auch mit einer Sichel. Die generationenübliche Usurpation durch die eigenen Nachfahren zieht sich wie eine rote Blutspur durch die Mythologie, weshalb sie für selbstvergötterte KI-Schöpfer geistesgeschichtlich unmittelbar nahe liegt und mit Blick auf die gewaltsame Entmachtung und Entmannung dystopisch besetzt ist.

von diesem erfahren wollte, wer ihn eines Tages usurpieren werde. In Bezug auf dieses anthropomorphe Technikbild ‚KI ist ein Kind/Nachfahre‘ hat Joseph Weizenbaum über Moravec gesagt:

„Es ist ein Mann, der so spricht. Ist das Zufall? Ich glaube nicht, wenn man sich klar-macht, dass das Feld der KI von Männern dominiert wird. Mir scheint hier nicht nur der Wahn, Gott zu spielen, sondern auch der Neid auf die Frauen und ihre Fähigkeit, Kinder zu gebären, als ein treibendes Motiv. Was hier zum Ausdruck kommt, würde ich als Uterusneid bezeichnen. [...] Man tut nun so, als könne man auch Kinder hervorbringen – nur sind diese eben, wie man verbreitet, besser und intelligenter als jedes menschliche Wesen.“ (Weizenbaum und Pörksen 2000, S. 7)

Und wer ‚bessere und intelligentere Kinder‘ hervorbringen kann, der meint sich im Umkehrschluss auch als ‚bessere Mutter‘ und ‚besseren Schöpfer‘ gerieren zu dürfen. Auch die Idee der nichtgeschlechtlichen Kopfgeburt geht auf die griechische Mythologie, in der Zeus seine Tochter Athene – die Göttin der Wissenschaft, Weisheit, Kunst und Technik (*techne*) – nicht als wehrlosen Säugling, sondern als Erwachsene inklusive Helm, Schild und Speer aus seiner Stirn gebiert (daher Kopfgeburt)<sup>24</sup>. Geburtshelfer und Hebamme hierbei ist die olympische Technikinkarnation Hephaistos, der Zeus’ Schädel mit einer Axt spaltete, um Athene zur Welt zu helfen. In dieser typischen Frauenrolle agierend wäre Hephaistos hier als gynomorphe Männer- oder Gottvorstellung im Spiel, ebenso der gebärende Zeus. KI als Geisteskinder zu betrachten und diese in ihrer Leistungsfähigkeit und Intelligenz über menschliche Kinder zu stellen, vergöttlicht vor diesem Hintergrund die so als ‚Väter‘ vorgestellten KI-Schöpfer im übertragenen Sinne. Wer metaphorisch sagt ‚Mein KI-Produkt (B) ist wie ein Kind (A) nur besser als menschliche Kinder (y)‘, der impliziert auch ‚Ich (B) bin wie eine Mutter (A) nur besser als menschliche Mütter/geradezu göttlich (y)‘<sup>25</sup>. Im Falle der KI als nicht nur menschen-ebenbürtig, sondern tendenziell ‚übermenschlich intelligent‘ kommt hier ein klassischer Topos des fiktiven Nachdenkens über Mensch-Technik-Verhältnisse ins Spiel. Die Logik ist folgende: Wenn der Mensch als ebenbildliches Geschöpf Gottes verstanden wird, dann wird er selbst göttlich,

<sup>24</sup> “... therefore did Zeus, who sits on high and dwells in the aether, swallow her [Metis, Anm.d.A.] down suddenly. But she straightway conceived Pallas Athene: and the father of men and gods gave her birth by way of his head on the banks of the river Triton. [...] And with it (Zeus) gave her birth, arrayed in arms of war.” Hesiod (1914), 929a.

<sup>25</sup> Vgl. Kanz (2009).

wenn er seinerseits ihm ebenbildliche Geschöpfe kreiert, und er wird zum besseren Gott<sup>26</sup>, wenn er bessere Geschöpfe als dieser schaffen kann (analog: zur besseren Frau, wenn er bessere Kinder schaffen kann). Wenn sich der Mensch (als Gottesgeschöpf) durch Intelligenzleistungen (qua Vernunftgebrauch, Wissenschaft und Technik) als Homo Faber (wie von Prometheus ermöglicht) von göttlicher Abhängigkeit emanzipiert, dann wird sich eine KI (als Menschengeschöpf) durch übermenschlich performante Intelligenzleistungen auch vom Menschen emanzipieren. Emanzipierte Wesen (also auch *Mind Children*) lassen ihre Abhängigkeit hinter sich und verfolgen eigene Pläne nach eigenen Präferenzen, die folglich auch denen der jeweiligen Schöpfer entgegenlaufen können. Damit kommt in Betracht, dass solche ‚Kinder‘ sich zum Homozid (als Patri- und Matrizid) aufmachen, wie der Mensch inzwischen zum Theozid:

„Gott ist tot! Und wir haben ihn getötet! Dies Gefühl, das Mächtigste und Heiligste, was die Welt bisher besaß, getötet zu haben, wird noch über die Menschen kommen, es ist ein ungeheures neues Gefühl!“ (Friedrich Nietzsche 1879, 14[26])

Analog müsste eine philosophierende KI eines Tages schreiben: *Der Mensch ist tot! Und wir haben ihn getötet!*<sup>27</sup> Diese Zeit nach dem Menschen, wird heute ebenfalls imaginiert und ist engstens mit KI-Entwicklungen verbunden, etwa im Posthumanismus. Dabei wären beide ‚Todesnachrichten‘ nicht im Ton nostalgischen Bedauerns geäußert, sondern in der Euphorie der Überwindung und des Neuen:

„Was aber ist für den Menschen, der selbstherrlicher Schöpfer einer Welt sein will, seinesgleichen? [...] So muß sich auch für den Jünger des Gottes Dionysos [Nietzsche, Anm.d.A.] das gemeinsame Menschenbild des Abendlandes auflösen. Nicht mehr sieht er in allem, was Menschenantlitz trägt, seines- und Gottes gleichen, sondern auch dies noch, was die Menschen zu sein haben, ist den großen Schaffenden

<sup>26</sup> Diese Vorstellung – wenngleich im KI-Diskurs wirkmächtig – entspricht allerdings einem profanierten Gottes- und Menschenbild. Es unterschlägt aus theologisch informierter Perspektive die kategoriale Differenz zwischen menschlicher und göttlicher Schöpfung; für die göttliche Schöpfung gibt es in der Bibel ein spezifisches Wort (*bará*), dessen Subjekt stets Gott ist. Für menschliches Schaffen oder Schöpfen werden andere Termini eingesetzt. Demnach wäre die Formulierung ‚menschliche Schöpfer‘ (im *bará*-Sinne) ein Oxymoron.

<sup>27</sup> Demgegenüber sowie entgegen des im KI-Diskurs prominenten Verdrängungstpos des Menschen durch die KI hat sich die Idee einer Ko-Existenz in umgekehrter Haustierrelation als Zukunft des Menschen in einer KI dominierten Welt verbreitet: “They’re going to be smarter than us and if they’re smarter than us then they’ll realise they need us [...] We want to be the family pet and be taken care of all the time.” Gibbs (2015). Vgl. auch Wiedemann (2017).

eine Frage ihres Wollens. [...] Die Idee des Übermenschen hat gerade darin ihre Auszeichnung, daß sie etwas ist, das nicht da ist. ‚Der höchste Gedanke: Der Mensch ist etwas, das überwunden werden soll.‘“ (Gadamer 2021, S. 264)

Mit der Tötung der Götter emanzipierte sich der Mensch (manche Menschen) von göttlicher Bevormundung und damit dem Zwang zur Rechenschaft (vor höchstem Gericht etc.), aber auch von Anleitung und Orientierung durch Götter (die immer durch Menschen vermittelt und deren Authentizität/ göttlicher Ursprung eine Glaubenssache war: z. B. Pythia von Delphi, Propheten, Priester, Exegeten etc.).

Nietzsche hat es als das Glück der Zeit empfunden, daß wir Menschen niemanden mehr haben, dem wir Rechenschaft schuldeten, als uns selbst: ‚Die Menschheit kann von nun an durchaus mit sich anfangen, was sie will.‘ Es ist der Fortfall eines übernatürlichen Herrn, des christlichen Gottes und der Moral, was dieses neue Selbstgefühl verleiht.“ (Gadamer 2021, S. 264)

Diesem Selbstgefühl muss man aber mit eigener Orientierungsleistung gerecht werden und den Implikationen sinnstabil und emotional standhalten können (endliche Existenz ohne Himmel/Hölle, kein kosmischer Sinn, kein gütiger Vater, usw.). Es mangelt daher nicht an Versuchen, die freigewordene Stelle mit Nachfolgern zu besetzen, die zwar nicht mehr göttlich sind, wohl aber doch wenigstens einige der Orientierungs- und gefühlten Sicherheitsfunktionen übernehmen könnten. Große Herrscher und Führer samt entsprechender Ideologien waren solche Kandidaten, die unsichtbare Hand des Marktes (Adam Smith) wäre hier nahe oder Intellektuelle, Eliten usw. Auch KI scheint (anthropomorphisierend formuliert) hier den Hut in den Ring zu werfen bzw. wird KI von KI-Apologeten als neue Ordnungsmacht qua übermenschlicher Einsicht in Stellung gebracht. Dass dies Proponenten von ‚Traue der KI, sie weiß es besser als du‘ (etwa bei KI-Expertensystemen oder Entscheidungsunterstützungssystemen) üblicherweise aus denjenigen Lagern kommen, die Zugang zu und Einfluss auf solche Systeme haben, sollte wenig verwundern.

KI hat neben starker Anthropomorphisierungstendenz speziell hohe Reaktionsfreudigkeit mit dem Bereich des Übermenschlichen, Magischen und Göttlichen. Wobei auch theomorphe Technikbilder sich indirekt als spezifisch gesteigerte (eminente) anthropomorphe Technikbilder auffassen lassen. Nämlich dann, wenn die Vorstellungen des Göttlichen wiederum (nur) Vorstellungen des Menschlichen in eminenter Potenz darstellen. Berühmter Vertreter für die Idee anthropomorpher Gottvorstellung ist Ludwig Feuerbachs Projektionsthese, wonach Göttliches nur

eine Projektion von Menschlichem, der eigenen Wünsche nach Unsterblichkeit, Allmacht, Allwissenheit etc., darstellt:

„So vergötterten die Römer und Griechen Akzidenzen als Substanzen, Tugenden, Gemütszustände, Affekte als selbständige Wesen. Der Mensch, insbesondere der religiöse, ist das Maß aller Dinge, aller Realität (Wirklichkeit *C*). Was nur immer dem Menschen imponiert, was nur immer einen besondern Eindruck auf sein Gemüt macht – es sei auch nur ein sonderbarer, unerklärlicher Schall oder Ton –, verselbständigt er als ein besonderes, als ein göttliches Wesen. [...] Selbst die Affekte der Furcht und des Schreckens hatten in Rom ihre Tempel. Auch die Christen machten Gemütserscheinungen zu Wesen, ihre Gefühle zu Qualitäten der Dinge, die sie beherrschenden Affekte zu weltbeherrschenden Mächten, kurz, Prädikate (Eigenschaften *C*) ihres eignen, sei es nun bekannten oder unbekannt, Wesens zu für sich selbst bestehenden Subjekten (Wesen *C*). Teufel, Kobolde, Hexen, Gespenster, Engel waren heilige Wahrheiten, solange das religiöse Gemüt ungebrochen, ungeteilt die Menschheit beherrschte.“ (Feuerbach 2006, S. 59f.)

Hier fragt sich nun, ob KI in diesem Sinne das neue Set an Kobolden und heutigen Engeln bereithält und so als Äquivalent von ‚heiligen Wahrheiten‘ zählt, solange Technik theomorph aufgefasst wird: Die Wege des Algorithmus sind unergründlich (Black Box)? „Plato hat es so gesagt: Die Philosophie strebt an, den Menschen, soweit es angeht, unsterblich zu machen, d. h. aber: dem Sein der unsterblichen Götter anzugleichen.“ (Gadamer 2021, S. 257) In der Folge der Feuerbachschen Projektionsthese gälte, dass Göttliches als Menschenprojektion also menschengespezifisch und menschlich vorgestellt wird. Auch hier gibt es quantitative Unterschiede: der Mensch sei wissend, mächtig/stark, sterblich etc., Gott sei allwissend, allmächtig, unsterblich etc. Damit wären aber theomorphe Vorstellungen letztlich gesteigerte oder spezifisch gefilterte anthropomorphe Vorstellungen und damit wären theomorphe KI-Bilder letztlich eine Unterart anthropomorpher KI-Bilder.

Die Übertragungen dazu hießen einerseits:

/Gott (B) ist wie ein Mensch (A) nur z. B. unsterblich (x) und z. B. allwissend (y)./

Andererseits:

/KI (B) ist (z. B. unsterblich, allwissend und z. B. menschengeschaffen etc.) wie ein Gott (A) nur ohne z. B. Emotionen (wie Neid und Zorn, x) und z. B. selbst Geschöpf/ technisch realisiert (y)./

Entsprechend ließe sich der Mittelbegriff ‚Gott‘ streichen und aus ‚Gott ist wie Mensch‘ und ‚KI ist wie Gott‘ folgt:

/KI (B) ist wie ein Mensch (A) nur z. B. nicht sterblich (x) und z. B. allwissend<sup>28</sup> (y)./

Auch hier würde die Angabe der Übertragungshinsicht zu Differenzierungen beitragen.

Ähnlich wie im Spektrum zwischen Menschlichem und Göttlichem Platz für Graduierungen sind, z. B. des Heroischen, Genialen, Halbgöttlichen etc. so wäre bei den Eigenschaften zu skalieren, etwa um ‚wissend‘ und ‚allwissend‘ noch ‚viel wissend‘ und ‚nichts/kaum/wenig wissend‘ etc. KI könnte dann auf übermenschliche Positionen bezogen sein (zwischen Menschen und Göttern), wenn etwa zwischen ‚wissend‘ und ‚allwissend‘, ‚digital viel wissend via Big Data, Mustererkennung, Bilderkennung etc.‘ träte.

---

## 6 Probleme der unhintergebar bildhaften Vorstellungsorientierung

Die bildhafte Vorstellungsorientierung bzw. die Übertragungen von einem Bereich des skizzierten Spektrums zu einem anderen bieten also einerseits wichtige Verstehensorientierungen beim Umgang mit neuen Phänomenen wie KI, zu dem (noch) keine oder kaum eigene Lebenserfahrung zur Erwartungsorientierung gemacht wurde; Metaphern leisten viel. Andererseits bieten diese Übertragungen erhebliche problematische Desorientierungspotenziale, die untauglichen bzw. irreführenden Vorstellungen wie von KI-Systemen als selbst-bewusste, intentionale Wesen (postbiotische Bewusstseine etc.) Vorschub leisten können.

Eine Forderung, angesichts dieser Problematik zu rein eigentlichen, nicht-übertragenen Konzepten überzugehen, ist hinfällig, wenn – wie exemplarisch mit Lakoff und Nietzsche erwähnt – es gar keine ‚reine Eigentlichkeit‘ und nur mehr oder weniger vergessene bzw. übliche Übertragungen gibt. Was sollten denn solche eigentlichen Konzepte sein? In Anlehnung an Stephen Hawking und Leonard Mlodinows *modellabhängigem Realismus* (Hawking und Mlodinow 2010) – d.i. solange sie richtige Vorhersagen zulassen, kann unter konkurrierenden Modellen des Universums (wie z. B. dem heliozentrischen oder geozentrischen Weltbild) nicht eines als *das* richtige beurteilt werden –, ließe sich hier von einem *üblichkeitsabhängigen Realismus* sprechen, der besagt: Solange sie erfolgreich

---

<sup>28</sup> „Es wird in der Theologie seit längerem diskutiert, ob ‚allwissend‘ als Attribut Gottes nicht aufgegeben werden sollte, weil es problematische Vorstellungen wecke und Fehlschlüsse provoziere. Der Topos ist zu komplex, um ihn hier darstellen. Bemerkenswert ist in jedem Fall, dass und wie über den KI-Diskurs fundamentale religions-affine Topoi eine Renaissance erfahren.“

Handlungen koordinieren und Vorstellungen orientieren, kann unter konkurrierenden Begriffen (wie z. B. Intelligenz oder Rationalität bzw. autonom oder automatisiert etc.) nicht einer als *der* richtige beurteilt werden.

Die Metapher der KI-Systeme (bzw.: ‚Solche IT-Systeme sind intelligent‘) lässt sich also nicht durch eine Rückführung auf (eigentlich) menschliche Intelligenz *klären*. Der Versuch führt vielmehr entweder zu einer strategischen Setzung einer gewünschten Bedeutung von Intelligenz (im gegenwärtigen Diskurs vor allem eine technizistische, um der besseren Passung zu KI-Produkten willen, und dies um des besseren Absatzes von KI-Produkten willen) oder aber zu einer begriffsarchäologischen Explikation der tatsächlichen Uneigentlichkeit des vermeintlich eigentlichen Aspektes ‚Intelligenz‘. Letzterer erweist sich dann als vergessene Metapher aus dem Lateinischen, wo es *dazwischen (inter-) auswählen/auslesen (-legere)* bedeutete. Das meinte den konkret physischen Vorgang zwischen gemischten Mengen (z. B. Linsen und Steinchen) die jeweils unerwünschten auszulesen/auszusortieren: „Die guten ins Töpfchen, die schlechten ins Kröpfchen“ (Grimms Aschenputtel) war also in dieser Hinsicht eine Intelligenzleistung. Schon in klassischer Zeit wurde diese Art eines konkret manuellen Vorgangs auf geistige Leistungen *übertragen*, so wurde bspw. aus dem *Auflesen* und *Auslesen* (vgl. die *Traubenlese* und die *Beerenauslese*) das Buchstaben-*Lesen* als *intellektuelle Fähigkeit* oder wurde die Fähigkeit zu Unterscheidungen zu einem Synonym von Vernunftgebrauch. Die Proto-Indo-Europäische (PIE)-Wurzel von *\*leg*, das dem lateinischen *legere* zugrunde liegt, bedeutet sammeln, auflesen mit abgeleiteten Bedeutungen wie *Sprechen* im Sinne von *Worte auswählen und sammeln*.<sup>29</sup> In gleicher Metaphorik ist die *Kritik* ein unterscheiden können, das griechische *krinein* bedeutet (aus)sieben, trennen und (zwischen Töpfchen und Kröpfchen) entscheiden. Die PIE-Wurzel *\*krei-* bedeutet sieben, also unterscheiden. Intelligenz ist also eine Metapher, die besagt:

/Solche (intelligente) Menschen (B) sind wie manuell auslesende/siebende Menschen (A) im gekonnt Unterschiede feststellen (H) nur nicht mit Händen (x), sondern im Geiste (y)./

‚Intelligente Menschen‘ ist also eine Auslese-/Sieb-Metapher, deren Uneigentlichkeit vergessen wurde. ‚X ist intelligent‘ ist also keine Metapher, die bei Menschen *eigentlich* und bei IT-Systemen *uneigentlich* gälte, sie gilt bei beiden *uneigentlich*, aber in unterschiedlich bewusstem Maße. Es gilt: „An sich herrscht hier nur der Unterschied zwischen Gewöhnung und Neuheit, Häufigkeit und Seltenheit.“ (Friedrich Nietzsche 2022, 19[228]).

---

<sup>29</sup> Siehe die Einträge *intelligence* und *critic* in <https://www.etymonline.com/>, zuletzt überprüft: 30.05.2022.

Dasselbe gilt bei der Übertragung von Autonomie von Menschen auf technische Systeme: ‚X ist autonom‘ ist nicht auf Menschen *eigentlich* und auf Autos/Roboter etc. *uneigentlich* bezogen, sondern beide sind unterschiedlich uneigentlich. Autonome Menschen sind bereits eine – wesentlich von Kant vorangebrachte – Übertragung, die Menschen in einer Staatsmetapher fasst. Bsp.: „Die Gesetzgebung der menschlichen Vernunft (Philosophie) hat nun zwei Gegenstände...“ (Kant 1998, S. 867). ‚Gesetzgebung‘ ist ein staatlicher, politischer, gesellschaftlicher Vorgang, der nur im Zwischenmenschlichen Sinn macht, also auf Menschengruppen bezogen ist. *Autonom*<sup>30</sup> waren Staaten in ihrer Gesetzgebung nach innen, wenn kein anderer Staat von außen ihnen (heteronom) andere Gesetze vorschreiben und aufzwingen konnte. Ein Mensch muss in sich keine Bevölkerung durch Gesetze koordinieren. Autonomie als Selbstgesetzgebung bei Menschen ist vergessene Staatsmetaphorik (vgl. Gransche 2022), genauso wie Intelligenz als Verstandesleistung bei Menschen vergessene Sortiermetaphorik ist. Dieser Nichtrückführbarkeit auf Eigentlichkeit zu entgehen zu versuchen, indem von *Intelligenz* auf *Rationalität* übergegangen wird (s. Einleitung) changiert stattdessen auch nur die Metaphernbereiche von *manueller Auslese* (Intelligenz von lat. *interlegere*) zu *berechnet* und *rechtskräftig gültig* (Rationalität von lat. *ratus*).

Statt eines vermeintlichen ‚Zurück zur Eigentlichkeit!‘ kann die Explikation der Orientierungsschwierigkeiten durch und der geistesgeschichtlichen Einbettung von metaphorischen Verständnisversuchen v. a. als Anlass dienen, sich dieser Dynamik von Sprachbildern und unterschiedlich vergessener Uneigentlichkeit bewusst zu werden. Eine solch geschärfte Aufmerksamkeit auf bildhafte Elemente und damit einhergehende Potenziale und Probleme der Verstehensorientierung würde wiederum die Auseinandersetzung und davon abgeleitet eine adäquate Regulierung, Nutzung, Verbreitung etc. von KI orientieren helfen. Dass nämlich KI nicht einfach (annähernd) menschlich intelligent ist, sondern eben nur mit Blick auf eine gewisse (und häufig interessen geleitete) Übertragungshinsicht, erhöht die Aufmerksamkeit und den Rechtfertigungsdruck zur Explikation der beabsichtigten Übertragungshinsicht. Die Einsicht, dass selbst unter möglichst

---

<sup>30</sup> Staatstheoretisch ebenfalls gebräuchlich ist hierfür die Souveränität, die aktuell ebenfalls aus dem Staatskontext auf Individuen übertragen wird, etwa in der Rede von digitaler Souveränität, Verbrauchersouveränität etc. z. B. Wittpahl (2017). In der griechischen Antike war Autonomie (αὐτονομία) eine dezidiert politische Kategorie, von Herodot etwa gefasst als innere und äußere politische Freiheit im Gegensatz zu einer äußeren Abhängigkeit durch Fremdherrschaft und zu der inneren Staatsform der Tyrannis; Vgl. Pohlmann 2010, S. 701–702. Aber bereits Sophokles gebrauchte Autonomie dichtend als innere Haltung: Von zehrendem Übel unversehrt, Um keines Schwertes bedungenen Lohn, Dir selber Gesetz [αὐτόνομος, Anm.d.A.], gehst einzige du Lebend hinunter zum Hades.“ Sophocles. 1982, S. 81, Verse 819–822, Hervorh. d.A.



geklärter Übertragungshinsicht (H) die Bekanntheitsgrade und die Verständnisse von A und B, von menschlicher und künstlicher Intelligenz, mitunter mangelhaft und teils inkompatibel heterogen sind, ist notwendige Bedingung dafür, vorgängige Informationsbedarfe zum Diskurseintritt zu identifizieren und unvermittelte Parallelkommunikation über gänzlich differente, aber äquivoke Begriffe zu vermeiden. Zu fragen wäre also: Was muss ich wissen, um in *diesem* Diskurs überhaupt mitreden, um dessen Übertragungen überhaupt verstehend dechiffrieren zu können? Dass jede heterogene Vielheit (wie KI) in inbegrifflicher Perspektive von einem ‚Interesse‘ zusammengehalten wird und ein (teils strategisches) Interesse auch bei der Übertragungswahl und Metaphernkomposition eine maßgebliche Rolle spielt, sensibilisiert dafür, dass ‚solche IT-Systeme‘ eben nicht einfach intelligent sind, sondern jemand ein Interesse daran hatte und hat, die mit Intelligenz üblicherweise assoziierten Bedeutungselemente nahezulegen. Ein Fokus auf die hintergründigen Interessen der Entscheidung über Sprachbilder hilft bspw. ein „wissenschaftstaktisches Manöver“ (Zimmerli 2021, S. 200) wie die Kombination von *artificial/künstlich* mit *intelligence/Intelligenz* aufzuklären und sich so nicht ‚ausmanövrieren‘ zu lassen. Diese Notwendigkeit für eine Sensibilisierung für das Bildhafte des Denkens und der Sprache gilt ganz allgemein. Sie ist aber besonders sinnvoll beim Phänomen der KI, da KI gewissermaßen als hochreaktiver Hülsenbegriff verstanden werden kann. Er ist Hülse, insofern er inhaltsleer ist und zur interessengeleiteten situativen Füllung qua Sprecherintention taugt. Die Sprecherinteressen der Hülsenschöpfer im Antrag zu Finanzierung des *Dartmouth Summer Research Project on Artificial Intelligence*<sup>31</sup> waren bekanntlich Finanzierungserfolg (etwas, was KI aktuell wieder hervorragend leistet) sowie Distinktionsgewinn, nämlich zur Abgrenzung der damals wirkmächtigen Kybernetik und deren populären Vertreter Norbert Wiener (vgl. Zimmerli 2021; Manzeschke und Gransche in diesem Band). Als besondere Dynamik des KI-Diskurses kann gelten, dass KI – um eine Chemiemetapher zu bemühen – eine hohe Reaktivität mit anderen Bedeutungsbereichen besitzt und vielfältige Kopplungen mit deren Elementen eingeht. So kann – selbst ohne jede weitere Idee, was KI genau auszeichnet – KI reaktionsfreudig als Werkzeug, Golem, (Geistes)Kind, Schöpfung, Nachfahre, Struktur, Verfahren, Roboter, Agent, Forschungszeitung, Entwicklungsziel, Techno-Göttliches etc. vorgestellt werden. Wenn etwa Walther Zimmerli KI selbst als „epochalen Mythos“ fasst und die „Geburt des KI-Mythos

---

<sup>31</sup> So der Name der oft als *Dartmouth Conference* abgekürzten Summerschool von J. McCarthy, M. L. Minsky, N. Rochester und C.E. Shannon von 1955; <http://www-formal.stanford.edu/jmc/history/dartmouth/dartmouth.html> (zuletzt aufgerufen am 20.03.2023).

aus dem Geiste der Analogie“ (Zimmerli 2021, S. 198) nietzscheanisch analysiert, dann unter Verweis auf diese Reaktionsfreudigkeit, die es erlaubt „durch Kopplung mit einem mächtigen Mythos aus einem anderen Bereich“ (Zimmerli 2021, S. 198) sich mit zusätzlicher Bedeutsamkeit aufzuladen. KI koppelt sich nicht nur – wie von Zimmerli angeführt – an den Mythos der Go-West-Pioniere, der vollständigen Ersetzung der menschlichen Intelligenz durch Leistungen intelligenter Maschinen und Programme, der eschatologischen Naherwartung und des Posthumanismus, sondern vieler weiterer Mythen, Mythologeme und Narrative aus teils jüngerer (Star Wars, Terminator, Matrix) teils uralter (Homer) Geistesgeschichte. Dabei ist besonders interessant, dass KI an diesen mächtigen Mythen partizipiert, sich an deren Bedeutung auflädt, im Gegenzug (noch) aber kaum etwas zu diesen beisteuert. Wie KI verstanden wird, ist bedingt durch Vorstellungen des Kindes, des Sklaven, des Geschöpfes, des Genies etc. Wie Kind, Sklave, Geschöpf, Genie usw. verstanden werden, ist aber kaum bedingt durch KI. Bei aller Aktualität des KI-Diskurses, bei all seiner Präsenz und Relevanz ist KI (selbst wenn man die Erhebung zu einem KI-Mythos gewillt wäre mitzugehen) als Hülsenbegriff eben ein anderes – nämlich erheblich kleineres – Kaliber als die Topoi der Nachkommen, der Auflehnung oder der Weltherrschaft. Mit KI werden aktuelle Riffs über diese Topoi gespielt und nicht wenige tanzen nach dieser Musik. Es wird sich zeigen, ob die – bekanntlich als KI-Sommer und KI-Winter jahreszeitlich zyklisch variierende – Üblichkeit dieses Begriffes es schafft, sich als neue Eigentlichkeit (stets als usuell verdeckte Uneigentlichkeit) durchzusetzen, oder ob wir zu Metaphern wie ‚maritime‘ oder ‚bakterielle Intelligenz‘ etc. übergehen werden und sich die Wendung ‚künstliche Intelligenz‘ ähnlich ephemere anfühlen wird wie heute ihre einstige Primärbedeutung des manuellen Auslesens.

---

## 7 Fazit

Damit KI trotz heterogener Elemente und unterschiedlicher Interessen (Inbegriff), trotz hoher Reaktivität (Kopplungen) und eigentümlicher Inhaltsleere (Hülse) verstanden, also in Sinnstrukturen integriert werden kann, muss Implizites expliziert werden, muss das Hineingelegte sorgsam ausgelegt werden. In KI-Diskursen, in ihren Äußerungen und Sprachelementen, in verkapselten Metaphern etc. ist teils sprachpragmatisch, teils strategisch implizit eingeschrieben, was deren Beteiligte bzw. Sprecher *wollen*, *meinen* und *verstehen*, also deren Interessen und Manöver, die jeweiligen Übertragungshinsichten der Sprachbilder sowie die (Miss-)Verständnisse der gebrauchten Konzepte. Zudem zu explizieren sind die Bedingungen der Möglichkeit bildhafter Zugriffe auf KI: Mit welchen

Interessen werden welche Elemente zu welchen Inbegriffen zusammengefasst? Welche Reflexionsbegriffe orientieren die Generierung von KI-Vorstellungen? Durch welche Technik- und Menschenbilder, durch welche Götter- und Tierbilder etc. sind diese Vorstellungsorientierungen geistesgeschichtlich situiert? Welche Verortung haben diese Reflexionsbegriffe in einem jeweiligen Weltbild? Welche Rückschlüsse lassen die Instanzen des KI-Diskurses auf Weltbild, Technik-/Menschenbild etc., Reflexionsbegriffe, Inbegriffe, Begriffe, Metaphern etc. der Sprecher bzw. deren Kulturen und Kontexte zu?

Dieser Beitrag votiert für eine explizierende Reflexion quer durch diese Bild- und Begriffsebenen. In dieser Sicht zeigen sich an KI teils Tendenzen einer technizistischen Selbstvergötterung oder antiaufklärerischen Technodevotion, teils geschickte Inszenierungen derselben oder postulierter Leistungsfähigkeit. Auffällig ist, wie ungebrochen wirksam Bilder, Bildhaftes, Übertragungen, Metaphern etc. und weit zurückreichende geistesgeschichtliche Bedingungen der Verstehenorientierung auch bei so vermeintlich neuen Phänomenen wie KI tatsächlich sind. Metaphern bzw. bildhafte sprachliche Übertragungen sind fundamental für unser Sprechen, Denken und Handeln sowie unser Verstehen, auch von Technik. Metaphorische Übertragungen brauchen selbst wiederum eine Orientierung, damit man entscheiden kann, in welcher Hinsicht und wie weit ein Konzept einem anderen ähnlich ist oder nicht. Metaphern können nicht nur hinken, sondern auch erheblich irreführen. Man kann Metaphern nicht vermeiden, aber sich die Eigentlichkeits-Uneigentlichkeits-Spannung aktiv bewusst machen. V.a. aktuelle Technikkonzepte – KI, autonomes Fahren, maschinelles Lernen, etc. – sind nur mit einem entsprechenden Metaphernbewusstsein und – grundlegender – mit beständiger Aufmerksamkeit für Bild- und Bildhaftigkeitsbezüge adäquat zu verstehen.

Das bewegliche Heer aus Metaphern und Anthropomorphisierungen der Künstlichen Intelligenz kann nicht in reine Eigentlichkeit aufgelöst werden. Die jeweiligen Truppenbewegungen – um in Nietzsches Metapher über Metaphern zu enden – müssen aufgeklärt werden, was sie nicht zum Verschwinden bringt. Aber die Kenntnis ihrer Stärke und Stellung sowie des geistesgeschichtlichen Terrains mindert viel von ihrer Bedrohlichkeit und mitunter katastrophalem Überraschungspotenzial. KI-Diskurse behandeln Menschliches, Übermenschliches. Sie sind in technomythischer Perspektive eine Summe menschlicher Relationen.

## Literatur

- Augustin, Anna-Carolin. 2016. Der Golem: Von Mystik bis Minecraft. [www.jmberlin.de/node/4164](http://www.jmberlin.de/node/4164). Zugegriffen: 28. Juni 2022.
- Barnden, John A. und Mark G. Lee (Hrsg.). 2001. *Metaphor and Artificial Intelligence. A Special Double Issue of metaphor and Symbol*. [Place of publication not identified]: Psychology Press.
- Blumenberg, Hans. 1998. *Paradigmen zu einer Metaphorologie*, 1. Aufl. Frankfurt am Main: Suhrkamp.
- Blumenberg, Hans. 2001. *Ästhetische und metaphorologische Schriften*, 1. Aufl. Suhrkamp-Taschenbuch Wissenschaft, Bd. 1513. Frankfurt am Main: Suhrkamp.
- Botting, Eileen Hunt. 2021. *Artificial life after Frankenstein*. Philadelphia: University of Pennsylvania Press.
- Carbonell, Javier, Antonio Sánchez-Esguevillas und Belén Carro. 2016. The role of metaphors in the development of technologies. The case of the artificial intelligence. *Futures* 84: 145–153. doi: <https://doi.org/10.1016/j.futures.2016.03.019>.
- Cassirer, Ernst. 1985. Form und Technik. In *Symbol, Technik, Sprache: Aufsätze aus den Jahren 1927–1933*, 39–91. Hamburg: Meiner.
- Cassirer, Ernst. 1995. *Zur Metaphysik der symbolischen Formen*. Nachgelassene Manuskripte und Texte / Ernst Cassirer herausgegeben von John Michael Krois und Oswald Schwemmer, Band 1. Hamburg: Felix Meiner Verlag.
- DAIA. 2018. Artificial Intelligence and Global Challenges — Zero Hunger. *DAIA*. 7 Dezember 2018. <https://medium.com/daia/artificial-intelligence-and-global-challenges-a-plan-for-progress-603efece1905>. Zugegriffen: 1. Juli 2022.
- Ekin, Annette. 2021. AI can help us fight climate change. But it has an energy problem, too. <https://ec.europa.eu/research-and-innovation/en/horizon-magazine/ai-can-help-us-fight-climate-change-it-has-energy-problem-too>. Zugegriffen: 1. Juli 2022.
- Feuerbach, Ludwig. 2006. *Das Wesen des Christentums*, 3. Aufl. Gesammelte Werke / Ludwig Feuerbach. Hrsg. von der Berlin-Brandenburgischen Akademie der Wissenschaften durch Werner Schuffenhauer, Bd. 5. Berlin: Akad.-Verl.
- Foerst, Anne. 2008. *Von Robotern, Mensch und Gott. Künstliche Intelligenz und die existentielle Dimension des Lebens*. Göttingen: Vandenhoeck & Ruprecht.
- Freud, Sigmund. 1930. *Das Unbehagen in der Kultur*. (Ebook). Wien: Internationaler Psychoanalytischer Verlag.
- Friedrich, Alexander. 2013. *Metaphorologie der Vernetzung. Zur Theorie kultureller Leitmotiven*. Dissertation. Paderborn: Wilhelm Fink.
- Friedrich Nietzsche. 2022. eKGWB/NF-1872 — Nachgelassene Fragmente Sommer 1872 — Anfang 1873. [http://www.nietzschesource.org/#eKGWB/NF-1872,19\[210\]](http://www.nietzschesource.org/#eKGWB/NF-1872,19[210]). Zugegriffen: 30. Mai 2022.
- Friedrich Nietzsche, Paolo D'Iorio. 1879. Nietzsche Source — Digitale Kritische Gesamtausgabe Werke und Briefe (eKGWB): Nachgelassene Fragmente 1879 – 40. <http://www.nietzschesource.org/#eKGWB/NF-1879,40>. Zugegriffen: 28. Oktober 2020.

- Funk, Michael und Mark Coeckelbergh. 2019. (Technical) Autonomy as Concept in Robot Ethics. In *Inclusive Robotics for a Better Society: Selected Papers from INBOTS Conference 2018, 16–18 October, 2018, Pisa, Italy*, hrsg. José L. Pons, 59–65. Biosystems and Biorobotics Ser., Bd. 5. Wiesbaden: Springer.
- Future of Life Institute. 2015. AI Open Letter: Research priorities for robust and beneficial artificial intelligence. <http://futureoflife.org/ai-open-letter/>. Zugegriffen: 1. Juli 2022.
- Gadamer, Hans-Georg. 2021. Was ist der Mensch? What is Man? *Phainomena* 30 (116-117): 255–280.
- Garten, William und Frank A. Dean. 1982. Evolution of the Talos Missile. *Johns Hopkins APL Technical Digest* 3 (2): 117–122.
- Gibbs, Samuel. 2015. Apple co-founder Steve Wozniak says humans will be robots' pets. *The Guardian*. 25 Juni 2015. <https://www.theguardian.com/technology/2015/jun/25/apple-co-founder-steve-wozniak-says-humans-will-be-robots-pets>. Zugegriffen: 1. Juli 2022.
- Goethe, Johann Wolfgang von. o.D. Prometheus: (Herbst 1774). In *Poetische Werke: Erster Band: Gedichte*: Weltbild-Bücherdienst.
- Gransche, Bruno. 2024. Technische Autonomie. In *Handbuch Technikphilosophie*, hrsg. Mathias Gutmann, Benjamin Rathgeber und Klaus Wieglerling. Stuttgart: Metzler, J. B.
- Hawking, Stephen und Leonard Mlodinow. 2010. *The grand design*. New York: Bantam Books.
- Heidenreich, Felix. 2020. *Politische Metaphorologie. Hans Blumenberg heute*, 1. Aufl. Stuttgart: J.B. Metzler; Imprint: J.B. Metzler.
- Hesiod. 1914. *Homeric Hymns, Epic Cycle, Homeric Hymns*. Hesiod, . Translated by Loeb Classical Library Volume 57. London: 1914. Classical Texts Library Online. London: William Heinemann.
- HLEG on AI. 2019. Eine Definition der KI: Wichtigste Fähigkeiten und Wissensgebiete.
- Husserl, Edmund G. 1891. *Philosophie der Arithmetik. Psychologische und logische Untersuchungen*, Bd. 1. Halle (Saale): Pfeiffer.
- Kant, Immanuel. 1998. *Kritik der reinen Vernunft. [1787]*. Philosophische Bibliothek, Bd. 505. Hamburg: Felix Meiner.
- Kanz, Christine. 2009. *Maternale Moderne. Männliche Gebärdphantasien zwischen Kultur und Wissenschaft (1890–1933)*. München, Paderborn: Fink.
- Kapp, Ernst. 1877. *Grundlinien einer Philosophie der Technik. Zur Entstehungsgeschichte der Cultur aus neuen Gesichtspunkten*. Braunschweig: George Westermann.
- Kluxen, Wolfgang. 2010. Äquivok. In *Historisches Wörterbuch der Philosophie*. HWPh, Bd. 1, hrsg. Joachim Ritter, Karlfried Gründer und Gottfried Gabriel, 480 [1699–1700]. Basel: Schwabe.
- La Mettrie, Julien Offray de. 2015. *L'homme machine. Der Mensch eine Maschine; französisch/deutsch*. Reclams Universal-Bibliothek, Bd. 19281. Stuttgart: Reclam.
- Lakoff, George und Mark Johnson. 2003. *Metaphors we live by*. Chicago, London: University of Chicago Press.
- Liveley, Genevieve und Thomas Sam. 2020. Homer's Intelligent Machines: AI in Antiquity. In *AI narratives: A history of imaginative thinking about intelligent machines*, hrsg. Stephen Cave, Kanta Sarasvati Monique Dihal und Sarah Dillon, 25–49. Oxford: Oxford University Press.

- Manzeschke, Arne und Bruno Gransche. 2023. Bilder machen Menschen: Zur Bildermacht der Künstlichen Intelligenz. In *Künstliche Intelligenz, Mensch und Gesellschaft*, hrsg. Michael Heinlein und Norbert Huchler. Springer VS.
- Mayor, Adrienne. 2018. *Gods and Robots. Myths, machines and ancient dreams of technology*. Princeton, Oxford: Princeton University Press.
- Moravec, Hans P. 1988. *Mind children. The future of robot and human intelligence*. Cambridge, Mass.: Harvard University Press.
- Musa Giuliano, Roberto. 2020. Echoes of myth and magic in the language of Artificial Intelligence. *AI & SOCIETY* 35 (4): 1009–1024. doi: <https://doi.org/10.1007/s00146-020-00966-4>.
- Nietzsche, Friedrich. 2009. Ueber Wahrheit und Lüge im aussermoralischen Sinne. In *Nietzsche Source – Digitale Kritische Gesamtausgabe Werke und Briefe (eKGWB)*, hrsg. Giorgio Colli und Mazzino Montinari. Berlin, New York: Walter De Gruyter.
- PARO. 2014. PARO Therapeutic Robot. <http://www.parorobots.com>. Zugegriffen: 29. Juni 2022.
- Pohlmann, Rosemarie. 2010. Autonomie. In *Historisches Wörterbuch der Philosophie*. HWPh, Bd. 1, hrsg. Joachim Ritter, Karlfried Gründer und Gottfried Gabriel, 2407–2463 [701–719]. Basel: Schwabe.
- Seising, Rudolf. 2021. *Es denkt nicht! Die vergessenen Geschichten der KI*, 1. Aufl. Frankfurt am Main, Wien, Zürich: Büchergilde Gutenberg.
- Sophocles. 1982. *Antigone*. Mit griechischem Text, 6. Aufl. Kleine Vandenhoeck-Reihe, Bd. 1116. Göttingen: Vandenhoeck und Ruprecht.
- Sternberg, Robert J. und Scott Barry Kaufman (Hrsg.). 2011. *The Cambridge handbook of intelligence*. Cambridge, New York, Melbourne: Cambridge University Press.
- Strasser, Anna, Wolfgang Sohst, Ralf Stapelfeldt und Katja Stepec (Hrsg.). 2021. *Künstliche Intelligenz – die große Verheißung*. MoMo Berlin Philosophische KonTexte, Band 8. Berlin: xenomoi.
- Weizenbaum, Joseph und Bernhard Pörksen. 2000. Das Menschenbild der Künstlichen Intelligenz. Ein Gespräch mit Joseph Weizenbaum. *Communicatio Socialis* 33 (1): 4–17. doi: <https://doi.org/10.5771/0010-3497-2000-1-4>.
- West, David M. und Larry E. Travis. 1991. The Computational Metaphor and Artificial Intelligence: A Reflective Examination of a Theoretical Falsework. *AI Magazine* (12): 64–79.
- Wiedemann, Caroline. 2017. Werden wir für sie wie Katzen sein? *Frankfurter Allgemeine Zeitung*. 30 November 2017. [https://www.faz.net/aktuell/feuilleton/debatten/kuenstliche-intelligenz-maschinen-ueberwinden-die-menschheit-15309705.html?printPagedArticle=true#pageIndex\\_2](https://www.faz.net/aktuell/feuilleton/debatten/kuenstliche-intelligenz-maschinen-ueberwinden-die-menschheit-15309705.html?printPagedArticle=true#pageIndex_2). Zugegriffen: 7. Juni 2022.
- Wiener, Norbert. 1963. *God and Golem, Inc. A Comment on Certain Points where Cybernetics Imprints on Religion*. Cambridge, Massachusetts: MIT Press.
- Witte, Bernd, Theo Buck, Hans-Dietrich Dahnke, Regine Otto und Peter Schmidt (Hrsg.). 2004. *Goethe Handbuch. Band 1 Gedichte*. Stuttgart: J.B. Metzler; Imprint: J.B. Metzler.
- Wittpahl, Volker. 2017. *Digitale Souveränität*. Berlin, Heidelberg: Springer Berlin Heidelberg.
- World Economic Forum. 2022. This is how AI could feed the world's hungry while sustaining the planet. <https://www.weforum.org/agenda/2020/09/this-is-how-ai-could-feed-the-world-s-hungry-while-sustaining-the-planet/>. Zugegriffen: 1. Juli 2022.

- Yalouris, Athanasios. 2019. Artificial Intelligence: Are we creating a new Frankenstein? *Hospital Chronicles* (14 (1)): 3–6.
- Zimmerli, Walther Ch. 2021. Künstliche Intelligenz und postanalogenes Menschsein. In *Künstliche Intelligenz – die große Verheißung*, hrsg. Anna Strasser, Wolfgang Sohst, Ralf Stapelfeldt und Katja Stepec, 193–219. MoMo Berlin Philosophische KonTexte, Band 8. Berlin: xenomoi.



# Bilder machen Menschen. Zur Bildermacht der Künstlichen Intelligenz

Arne Manzeschke und Bruno Gransche

## Zusammenfassung

Der Diskurs um Künstliche Intelligenz (KI) ist ein bildmächtiger. Das betrifft sowohl die konkret visuellen wie auch die eher konzeptionellen Artefakte. So unumgänglich die sehr unterschiedliche Zielgruppen orientierenden Bilder auch sind, so sind die Bilder nie sicher vor Störung, Täuschung, vor Schein oder Magie. Bilder der KI sind von Menschen hergestellte *Abbilder* sozio-technischer Artefakte, aber auch von Menschen hergestellte *Vorbilder* für solche Artefakte. Nicht zuletzt liefern Bilder referenzfreie Simulationen einer sich selbst setzenden Realität. Ein aufgeklärter Umgang mit den Bildern der KI muss sich der Mühe unterziehen, die zur rationalen Abkürzung tendierenden Bilder ‚auszubuchstabieren‘ und die mit ihnen verbundenen Ebenen der bildtheoretischen Funktion (Imitation, Repräsentation, Simulation) zu differenzieren, um so Aussagegehalte diskutieren zu können. Zugleich bedarf es einer menschlich selbstkritischen Klärung über die Produktionsbedingungen für die Bilder der KI (*genitivus subjectivus*), die als Element einer größeren intellektuellen Operation (Probleme erkennen und lösen) eingebettet und verknüpft

---

A. Manzeschke (✉)

Institut für Pflegeforschung, Gerontologie und Ethik, Evangelische Hochschule Nürnberg, Nürnberg, Deutschland

E-Mail: [arne.manzeschke@evhn.de](mailto:arne.manzeschke@evhn.de)

B. Gransche

Institut für Technikzukünfte ITZ, Karlsruher Institut für Technologie KIT, Karlsruhe, Deutschland

E-Mail: [bruno.gransche@kit.edu](mailto:bruno.gransche@kit.edu)



werden müssen. Der Artikel skizziert in einem ersten Schritt die Dimensionen und Funktionen von (menschengemachten) Bildern. In einem zweiten Schritt werden diese mit Rekurs auf Hans Blumenberg um Konzepte wie Weltbild und Weltmodelle erweitert, die einen wichtigen Interpretationsrahmen für die Bilder der KI (im *genitivus objectivus* – dritter Schritt – und *genitivus subjectivus* – vierter Schritt) darstellen und ausschnittsweise vorgestellt und ‚gelesen‘ werden. Im fünften Schritt werden die visuellen an die sprachlichen Bilder (Metaphern) zurückgebunden und die Übertragungsfunktion zwischen den Bedeutungsquellen und -zielen im KI-Diskurs kritisch reflektiert.

### Schlüsselwörter

Künstliche Intelligenz • Metapher • Bild • Technikbild • Menschenbild • Weltbild

### Vorbemerkung

Dieser Beitrag geht von den gleichen Prämissen aus wie der Beitrag „Das bewegliche Heer der Künstlichen Intelligenz. Ein Technomythos als Summe menschlicher Relationen“ (Gransche und Manzeschke 2023 in diesem Band) und komplementiert dessen Fokus auf Sprachbilder mit einem Schwerpunkt auf anthropologische Erwägungen des Bildermachens, auf das Verhältnis von Mensch- und Technikbildern zu Weltmodellen und Weltbild sowie auf visuelle Bilder. Dabei steht der eigentümliche Bildbezug von KI im Mittelpunkt, denn KI prozessiert und generiert wesentlich Bilder und sie wird in speziellen Bildern vorgestellt und kommuniziert.

Der Beitrag „Das bewegliche Heer der Künstlichen Intelligenz. Ein Technomythos als Summe menschlicher Relationen“ (Gransche und Manzeschke 2023 in diesem Band) analysiert hingegen KI zunächst als Inbegriff, dann als Metapher und situiert dies in einem erweiterten Spektrum der Bedeutungsübertragungen. Dies wird verdeutlicht an einem Beispiel aus dem Spektrum, nämlich dem Topos des Prometheus und diversen Verhältnissen von Schöpfung und Geschöpflichkeit zwischen Gott, Mensch und KI. Der Beitrag schließt mit einer Problembetrachtung der unhintergehbaren Bildhaftigkeit der Vorstellungsorientierung allgemein sowie von KI speziell.

Beide Beiträge können eigenständig oder als zwei Seiten einer Medaille gelesen werden.

„Poesie ist die Muttersprache des menschlichen Geschlechts [...] Sinne und Leidenschaften reden und verstehen nichts als Bilder. In Bildern besteht der ganze Schatz menschlicher Erkenntniß und Glückseligkeit.“ (Hamann 1999, S. 197)

## 1 Der Bilder machende Mensch

Menschen machen Bilder von der Welt und sich selbst. Sie tun das in Form von Abbildungen und – im weiteren Sinne – von Konzepten und Weltbildern. Erste Höhlenmalereien datieren etwa 36.000 Jahre zurück. Bereits diese Bilder lassen ein ‚symbolisches Denken‘ erkennen, das ‚nicht nur‘ abbildet, sondern bildlich denkt und erzählen will (vgl. Pievani und Zeitoun 2020, S. 107).

Etymologisch bedeutet das althochdeutsche ‚*bilidi*‘ ‚einerseits ‚(Wunder-) Zeichen‘, ‚Wesen‘, ‚Gestalt‘, andererseits ‚Bild, Abbild, Nachbildung‘ [...] einerseits wird also das betont, wodurch etwas seine Gestalt gewinnt, in sein Wesen kommt, zur vollen Entfaltung seiner Wunderkraft gelangt; andererseits das, was ein solches Ur-Bild nachbildet, darstellt, bezeichnet.“ (Kamper 1997, S. 589) Folgt man dieser Wortbedeutung, so wird verständlich, dass der Bild-Begriff keineswegs auf zweidimensionale Gebilde beschränkt ist – der Terminus ‚Bildhauer‘ ist ein guter Hinweis auf das Form-gebende Moment des Bildermachens mit den unterschiedlichsten Materialien wie Lehm, Holz, Metall, Stein ...

Neben den deutschen Begriffen weisen auch die kulturell vorgängigen, prägenden Termini aus dem Griechischen (z. B. *eikon*, *eidolon*) und Lateinischen (z. B. *imago*, *pictura*, *effigies*, *simulacrum*) eine enorme Bedeutungsvielfalt auf und liefern keine, einer Definition genügende Schärfe. Im Griechischen werden *eidolon* und *phantasma* vor allem für innere Bilder verwendet; sie werden im Lateinischen mit *simulacrum* übersetzt und sind eher negativ besetzt; sie tendieren zur Oberfläche, zu Schein und Trug. In der biblischen Tradition werden deshalb die *eidola* (lat.: *idola*) als Götzenbilder<sup>1</sup> verstanden – daher der Kampf (*Ikonoklasmus*) gegen die Götzenverehrung (*Idolatrie*).

Auf einer Ebene bilden Bilder den Gegenstand ab. Sie ahmen ihn nach, vergleichbar den Worten, die onomatopoetisch auf den benannten Gegenstand verweisen. Platons *Kratylos* thematisiert diese Parallele (Krat. 430a–433b):

„Nun gibt es richtige und falsche Verteilung der Nachahmungen auf die Dinge – und in diesem Sinne richtige und falsche Bilder und Wörter. Diejenige Verteilung gilt als richtig, ‚welche jedem das ihm Zukommende und Ähnliche zuteilt“.

---

<sup>1</sup> Götze bezeichnet ein Wesen, das als Gottheit verehrt wird, einen Abgott und seine bildliche Darstellung. Im 14. Jh. wird es im Sinne von Heiligenbild/Bildwerk gebraucht, bei Luther im 16. Jh. meint es falscher Gott/Abgott. Ab dem 16./17. Jh. nimmt der Bedeutungsaspekt des Bildwerks in Götze selbst ab und wird dann im Kompositum Götzenbild als Abgott in bildlicher Darstellung gebraucht. <https://www.dwds.de/wb/G%C3%B6tzenbild>, zuletzt aufgerufen 01.08.2022.

So ist es für Platon durchaus wichtig, dass Bild und Abgebildetes noch unterscheidbar bleiben und im Bild nicht alles des ‚Vorbildes‘ enthalten ist. Ein Bild muss dem Abgebildeten hinreichend ähnlich und zugleich unähnlich sein:

„Wäre das dann Kratylos und dann daneben ein Bild des Kratylos, oder wären es zwei Kratylos?“ (Krat. 432b).

Bilder bilden die Wirklichkeit ab, sie nähern sich mimetisch der Wirklichkeit an. Unter dem Begriff Bilder versammeln sich dabei visuelle Artefakte von großer optischer Einfachheit bis hin zu enormer Komplexität, von schlichten ikonographischen Signalen bis hin zu umfassenden, nur mittelbar visualisierbaren Konzepten wie Menschen- oder Weltbildern. Bilder stellen im Sinne des Letzteren zudem eine Kategorie eigener Art dar, insofern sie ‚innere Bilder‘ bezeichnen, „sei es um die Sinneswahrnehmungen, sei es um das Denken, die Erinnerung, die Vorstellungstätigkeit (*phantasia, imaginatio*) oder das Träumen zu beschreiben und zu erklären.“ (Scholz 2000, S. 621). Sie gründen in der (nicht vollends kontrollierbaren) *Einbildungskraft* des Menschen, die ihn zu solchen Bildern befähigt, ihn beflügelt, aber auch bedroht.<sup>2</sup> Solche Bilder sind nur begrenzt vom Menschen verfertigt. Für die verfertigten wie die ‚empfangenen‘ gilt gleichermaßen: Die Bilder als das *Imaginäre* bringen eine Differenz ins Spiel, die nicht allein vom Menschen gesetzt wird. Das Imaginäre wie das Imaginierte erlauben Unterscheidungen, die die Einrichtung in der Welt, mit den Anderen und mit sich selbst eröffnen und auf ihre Weise auch orientieren.

Die Ambivalenz der Bilder setzt bis heute, kurz gesagt, u. a. zwei diametral entgegengesetzte Praktiken aus sich heraus: Bilderdienst und Bildersturm. Dies zeigt sich auch paradigmatisch im Alten Testament, das zu Beginn der Zehn Gebote ein Bilderverbot formuliert: „Du sollst dir kein Bildnis noch irgendein Gleichnis machen“ (Ex 20, 1–6). Einerseits bedeutet das die Zurückweisung jeglicher Bilderfabrikation, wissend um ihre Verführungskraft; andererseits findet sich in der gleichen Traditionslinie die Ikonisierung, also die Repräsentation der (heiligen) Ordnung.

<sup>2</sup> Das bekannte Capriccio von Francisco de Goya von 1799 ist betitelt: „Der Schlaf der Vernunft gebiert Ungeheuer“ (das spanische *sueño* wird zumeist mit ‚Schlaf‘, seltener mit ‚Traum‘ übersetzt). Es ist ein Bild, das von der ungewollten Einbildungskraft erzählt, die im Menschen waltet, wenn Wachheit und Vernunft schlafen gehen. Es sind monströse Bilder, die dann entstehen und den Menschen bedrohen. Es sind seine eigenen Bilder; eigen nicht unbedingt im Sinne eines Urheberrechts, aber eigen, dass sie seinen Sinnen entsprungen sind und ihm nun zu schaffen machen. Sie markieren das „Andere der Vernunft“ (Böhme/Böhme 1983), *ohne* das der Mensch nicht Mensch sein kann, *mit* dem er aber Mensch werden müsse.

Eine Ursprungslegende legt nahe, dass das Bildnis als Stellvertreter für einen Verstorbenen angefertigt worden ist (vgl. Scholz 2000, S. 624). Menschen verfertigen also Bilder „aus der Angst vor dem Tod“ (Kamper 1997, S. 592). Das Leben wird in ein Bild gebannt, um es unsterblich zu machen. Im Bild wird das Abwesende als anwesend beschworen, unabhängig davon, wie ‚realistisch‘ die Abbildung gerät. So spricht aus den Bildern auch eine Art von Magie, die als Ausflucht und Ablenkung von der Wirklichkeit verstanden werden kann.<sup>3</sup> Doch nicht alle Bilder werden angefertigt, um der Drohung des Todes zu entkommen. Sie begegnen und erscheinen dem Menschen auch ungemacht und ungewollt: in der Natur, in Traum, Rausch, Wahn oder in der Offenbarung. Sie können nicht gesehen werden. Wie also umgehen mit den Bildern, wenn es ohne sie nicht geht, aber mit ihnen die Verwirrung und Unsicherheit steigt.<sup>4</sup> Die Ambivalenz der Bilder liegt nicht allein in ihren Wirkungen, sondern ebenso in ihrer Unumgänglichkeit.

Bilder transzendieren als Symbole das Dargestellte und verweisen auf einen weiten Hof von Bedeutungen und Beziehungen. Als Symbole existieren diese Materialien, ihr Arrangement und ihre Bedeutungen allerdings nur durch und für Menschen, denn es ist das *animal symbolicum* (Cassirer 2001/2; Schmidinger/Sedmak 2007), das sie dazu macht. Als Symbole stellen Bilder nicht nur dar, sondern vermitteln Sinn<sup>5</sup>; sie geben Orientierung – maximal auf der Ebene von Weltbildern. Der sinnhafte Überschuss unterscheidet sie von ‚einfachen‘ Icons, die etwa in Bahnhöfen oder Flughäfen anzutreffen sind und hauptsächlich topologisch orientieren.

---

<sup>3</sup> Dieser Bildmagie hält Susan Sontag (2010) die „Mortifikation des Lebendigen“ durch das stillstellende Bild entgegen. Es wäre zu diskutieren, inwieweit Sontags Kritik, die ja auf die enorme Bildervermehrung durch die Technik der Fotografie und die Möglichkeiten der Reproduktion in den 1970er Jahren zielt und die sich in Zeiten des Internets und des Smartphones noch einmal enorm gesteigert hat, nicht zu einem Kipppunkt kommt, an dem die schiere Quantität in eine neue Qualität umschlägt.

<sup>4</sup> Der Streit zwischen den Malern Zeuxis und Parrhasios (Plinius, Nat. Hist. XXXV, 64) weist auf die Ambivalenz des Bildermachens hin. Es gewinnt der Maler, der nicht nur Tiere, sondern sogar Menschen mit seiner Kunst zu täuschen vermag – die Faszination über die Illusion (il-ludere: verspotten, täuschen) überwiegt hier sogar die Ent-Täuschung.

<sup>5</sup> Sinn und Orientierung sind in bestimmter Hinsicht gleichzusetzen. Bilder als sinnstiftende Momente der Orientierung sind demnach solche, die die tendenziell unbegrenzte Menge der Möglichkeiten in der Welt auf eine ‚sinnvolle‘ Zahl begrenzt und Richtung(en) angeben, in denen Handlungen anschließen können. „So ist Sinn nicht nur das, womit die Orientierung etwas anfangen kann, sie kann auch nur mit Sinn anfangen. Sie ist im ganzen als selektiver Sinnfindungsprozess zu verstehen.“ (Stegmaier 2008, S. 182).

Das Bild markiert eine „changierende Stellung zwischen einer magischen Ordnung der vollen Präsenz, in der das Bild identisch ist mit dem, was es zeigt, und einer Ordnung der zur Leere tendierenden Repräsentation, in der es bestenfalls ähnlich ist“ (Kamper 1997, S. 589). Neben magischer Präsenz und kunstfertiger Repräsentation, neben Identität und Differenz, bringt Kamper die „technische Simulation“ als dritte Dimension des Bildes ins Spiel (Kamper 1997, S. 591). Die (bewegten, virtuellen oder hyperrealistischen Bilder) bieten ein ‚Als-ob‘, das der Mensch als Trugbild ablehnen, oder in das er spielerisch einwilligen kann.<sup>6</sup> *Simulakra* als urbildlose Abbilder eröffnen dann den Raum des Spiels und schließen an eine Grunddimension des Menschseins an (vgl. Huizinga 1939; Plessner 2003, Baudrillard 1981).<sup>7</sup> Folgt man älteren Vorstellungen von Simulation als Scheinbild, so wäre hier die Möglichkeit gegeben, zwischen dem Echten und dem ‚nur Simulierten‘ als dem Schein oder dem (unernsten) Spiel zu unterscheiden. Ein neueres Verständnis von Simulation sprengt diese Unterscheidung jedoch auf und macht sie obsolet: „Wenn also der Begriff der Mimesis das ‚System von Vorbild und Kopie, Original und Fälschung, Replikation ersten und zweiten Grades‘ bezeichnet, so definiert sich Simulation in genauem Gegensatz dadurch, daß hier keine ‚Möglichkeit der Differenzierung zwischen Original und Kopie‘ mehr besteht. In der Welt des Simulakrum implodiert jede solche Unterscheidung: ‚Der Effekt des Realen ersetzt das Reale selbst““ (Dotzler 2003, S. 511 mit Verweis auf Kraus 1998, vgl. außerdem Baudrillard 1981). Wenn aber die Bilder und die Worte nach neueren Medientheorien nur noch für sich selbst stehen und keinen Referenzpunkt mehr haben, dann wird die Frage virulent, was sie uns (noch) zu sagen haben.

„Die Menschen leben heute nicht in der Welt. Sie leben nicht einmal in der Sprache. Sie leben vielmehr in ihren Bildern, in den Bildern, die sie sich von der Welt, von sich selbst und von den anderen Menschen gemacht haben, die man ihnen von der Welt, von sich selbst und von den anderen Menschen gemacht hat.“ (Kamper 1997, S. 591)

Kamper ergänzt diesen Befund um die Jahrtausendwende diagnostisch: „Und sie leben eher schlecht in dieser imaginären Immanenz. Sie sterben daran.“ (Kamper 1997, S. 591) Die heutige Fülle der Bilder scheint das Problem noch zu

---

<sup>6</sup> „Simulation [in der gegenwärtig dominanten Bedeutung] bezeichnet damit die Funktion von Zeichenprozessen, in denen es nicht um Repräsentation oder Vorspiegelung ihnen externer Dinge geht, sondern diese Vorspiegelung selber an die Stelle der Dinge tritt.“ (Dotzler 2003, S 533).

<sup>7</sup> Es gilt Schillers Diktum: „Der Mensch spielt nur, wo er in voller Bedeutung Mensch ist, und er ist nur da ganz Mensch, wo er spielt“ (Schiller 1795).

steigern – und zugleich zu verdecken. Ein Ausgang aus „der selbstproduzierten Bilderhöhle“, so Kamper (ebd.), durch erneuerte Bild-askese erscheint unmöglich, und auch der Durchgang durch die Fülle der Bilder nur schwer vermittelbar. Wer welche Bilder und wie viele zu sehen bekommt, verdankt sich kaum mehr individuellen ästhetischen Vorlieben, sondern hat wesentlich mit algorithmisch erfassten und befüllten Profilen zu tun, die in den sozialen Netzwerken zu ‚Bilderhöhlen‘ werden können. Die Bilderfülle nimmt – auch in den anderen Formaten, Fernsehen, Radio, Zeitung etc. – beständig zu. Immer mehr Bilder wollen gesehen, bewertet, verstanden, aber auch produziert und verbreitet werden. Wie schon in anderen Domänen zuvor versucht der Mensch, durch Technik Abhilfe zu schaffen. Das wirft die Frage auf, ob der Einsatz einer Künstlichen Intelligenz zur *Bildererkennung* auf mehr als nur einen vordergründigen Zusammenhang verweist.

Kampers Monitum lässt sich so lesen, dass die Präponderanz des Bildes, das angeblich mehr als tausend Worte sagt, auf eine Kommunikationssituation trifft, in der das Bild zu einer Abkürzung wird – zeitlich und semantisch. Dies lässt sich sehr gut illustrieren an einer Beobachtung des Kulturphilosophen Bernard Braun:

„Am 12. September 2001 machte die Frankfurter Allgemeine Zeitung zum bislang ersten und einzigen Mal mit einem großformatigen Foto ohne jeden Text auf dem Titelblatt auf. Es zeigte die rauchenden Türme des World Trade Center. Das Bild ersetzte hier das Wort für das scheinbar Unaussprechliche und damit sich jedem rationalen Zugang Entziehende.“ (Braun 2019, S. 558)

Das Ereignis vom 11. September 2001 trägt sicherlich Elemente, die sich dem Sprachlichen entziehen. Ob sie durch das Bild dargestellt oder aufgerufen werden können, ist zumindest kritisch zu betrachten. Was das Bild bei Einzelnen evoziert, bleibt – wenn es sich tatsächlich dem Sprachlichen entzieht – eben unausgesprochen. Für eine gesellschaftliche Verständigung über das, was solche Ereignisse bedeuten und von wem sie in welche Bilder gefasst werden, werden wir Menschen aber gerade nicht auf Sprache verzichten können. Das verlangt ein ‚Ausbuchstabieren der Bilder‘. Der rationale Diskurs ist langsamer, langwieriger und anstrengender, weil er die Komplexität und Differenziertheit nicht aussparen kann, die das Bild (scheinbar!) nicht erfordert. Versteht man Bilder jedoch nicht nur als optisch wahrnehmbare Artefakte, sondern auch in ihrer konzeptuellen Variante von ‚Weltbildern‘ und berücksichtigt man, dass ihr ‚Lektürewert‘ wesentlich von der – allerdings abnehmenden (so Scholz 2000, S. 619 f.) – Fähigkeit des Menschen abhängt, diese hermeneutisch differenziert wahrzunehmen und

zu bewerten, dann zeigt sich die Herausforderung umso stärker, die sich für die Analyse von Bildern einer Künstlichen Intelligenz auftut.

---

## 2 Weltbilder und Weltmodelle

1961 hat der Philosoph Hans Blumenberg in seinem Vortrag „Weltbilder und Weltmodelle“ auf ein grundsätzliches Problem wissenschaftlichen Arbeitens aufmerksam gemacht. Descartes (1637/1960) habe, damals neu und von da an stilgebend, den Versuch zur Begründung der wissenschaftlichen Methode nicht nur als Werkzeug zur wissenschaftlichen Wahrnehmung und Beschreibung unternommen, sondern zugleich als „vollendeten Selbstbesitz des Menschen, Theorie und Moral sollten am Ende eins werden und in dieser Einheit dem Menschen seine Selbsterfüllung und sein Daseinsglück gewährleisten“ (Blumenberg 2015, S. 128). Weiter: „Hier liegt eine Differenz zwischen der Totalvorstellung von der Natur einerseits und der Zweckbestimmung der Totalität der Naturerkenntnis andererseits vor, die in der Folge höchst bedeutsam werden sollte. Denn in ihr sind zum ersten Male ‚Weltbild‘ und ‚Weltmodell‘ auseinander getreten, ja ihr fundamentaler Unterschied wird überhaupt erst sichtbar aktualisiert.“ (ebd.)

Die Cartesianische Unterscheidung von (objektiv) erforschbaren Körpern (*res extensa*) und (subjektiv) erfahrbarem Geist (*res cogitans*) lief auf eine Abtrennung beider voneinander hinaus, auch wenn beim Menschen beide doch irgendwie zusammengedacht werden mussten (Descartes 1641/1959). Die Körper lassen sich nach allgemeinen Gesetzen erforschen, zusammengefasst von einer wissenschaftlichen Theorie von Kausalitäten und Materialitäten, strukturiert in einem *Weltmodell*, das unabhängig vom Geist des Menschen gelte. Das *Weltbild* hingegen, welches Menschen Sinn zur Orientierung in der Welt geben sollte, konnte vom Menschen gerade nicht absehen, sondern war notwendigerweise auf ihn zugeschnitten. Wissenschaftsgeschichtlich setzte Descartes eine Trennung, die bis heute als offene Wunde den Menschen beunruhigt (vgl. Schultz 2001, S. 223 ff.).

Streng genommen müsste naturwissenschaftliches Erkenntnisstreben von menschlichen Orientierungsbedürfnissen absehen und aussagen, ‚was der Fall ist‘. Ein solches Weltmodell wurde aber (auch von Descartes selbst) nicht konsequent von Weltbildern getrennt. Im Grunde blieben er (und die ihm Folgenden) einem mittelalterlichen Weltbild verhaftet, in dem das Weltmodell lediglich „ein ‚Organ‘ des Weltbildes“ war. „Das Weltbild enthielt die Sinnggebung und sozusagen die ‚Gebrauchsanweisung‘ für alle je denkbaren Weltmodelle. [...] Wissenschaft aus diesem Fundierungszusammenhang entbunden, konnte aus sich heraus nicht wissen, was sie tat“ (Blumenberg 2015, S. 129).

Aufgabe der Philosophie könne es nicht sein, an die historische Bruchkante zurückzuführen, um Erkenntnis und Sinngebung wieder zusammenzuführen. Sie müsse anerkennen, dass Wissenschaft autonom geworden ist; diese „erzeugt und erweckt und erhält diesen Sinn ständig selbst in der Lebendigkeit ihres Handelns“ (a. a. O., S. 130). So gebe es keine Zuordnung von Weltbild und Weltmodell mehr. Problematisch sei nun aber, dass es nicht einfach bei dieser Trennung bleibe, sondern „daß das ‚Weltmodell‘ die Stelle des ‚Weltbildes‘ besetzte und noch immer dabei ist, die Restsubstanz des Weltbildbestandes aufzuzehren. Daß es so etwas wie Wissenschaftsgläubigkeit geben kann, beruht darauf, daß die Wissenschaft ihre Bedingtheit durch einen Weltbildglauben verloren hat.“ (a.a.O., S. 131). „Ein auf das Ungenügen am Sinnverlust spekulierendes Angebot an Surrogaten findet die fast beliebige Besetzbarkeit dieser Potenz vor und nutzt sie aus.“ (a.a.O., S. 132).

Weltbilder ließen sich heute – unter den gegenwärtigen Bedingungen von Wissenschaft und Moral – nicht mehr einfach instaurieren. Der Plural Weltbilder sei schon ein Widerspruch in sich. „Nur die in einem homogenen Geistesraum unangefochtene Geltung eines Weltbildes enthält zugleich Idealität und Toleranz, um darin human sein zu können und das innervierende Gleichgewicht von Verständlichkeit und Befremdlichkeit, von Rechtfertigung und Normierung zu wahren.“ (a.a.O., S. 133) Eine Pluralität von Weltbildern ist schlicht dysfunktional. Entweder kommt es hier zu einer Spannung, die in Feindschaft und gegenseitigem Ausschluss sich entlädt. Oder die ihnen unterliegenden Ideologien werden wirkmächtiger und bemächtigen sich der jeweiligen Weltbilder für ganz andere Zwecke. Das ist die Gefahr, die in Weltmodellen liegt, die – auch mit anderen Interessen – sich der Weltbildfunktion zu bedienen suchen. Blumenberg illustriert das etwa an Darwins Theorie über die Entwicklung der Arten, deren Transformation in Weltbilder zu gefährlichen, dogmatistischen Übergriffen auf die menschliche Lebensform geführt habe. Es ist kaum abzuweisen, dass die Vorstellungen von KI ebenfalls Weltbildartiges enthalten und von einer Wissenschaftsgläubigkeit zehren, die nicht selten im Modus des Weltmodells daherkommt. Dem müsste wissenschaftlich begegnet werden: „Nicht über Welt Dinge und Weltkräfte zu verfügen und sich ihrer zu bemächtigen, ist der wesentliche und primäre Sinn von Wissenschaft (vielmehr derjenige von Technik, die sowohl angewandte Wissenschaft als auch Problemquelle der Wissenschaften ist), sondern unsere Weltvorstellung in der Verfügung und unter der Kontrolle theoretischer Verantwortung zu halten.“ (a.a.O., S. 136).

Betrachtet man die Künstliche Intelligenz unter einer bestimmten Perspektive als eine Technik, dann gehört sie zu den Dingen und Kräften, mit denen Menschen über diese Welt und Andere zu disponieren, sich wohl auch ihrer zu



bemächtigen suchen – und die damit zu einer der Problemquellen der Wissenschaften, aber auch der ‚Praxis‘ wird. Wie bereits gezeigt (vgl. Gransche und Manzeschke 2023 in diesem Band), sind die Hinsichten auf KI vielfältiger und damit auch die Forderungen an die Wissenschaften, die Kontrolle theoretischer Verantwortung vielgestaltiger auszulegen.

### 3 Bilder der KI (genitivus objectivus)

Bilder sind ein unumgängliches und zugleich problematisches, weil Geist und Sinne auch verwirrendes Medium zur menschlichen Orientierung in der Welt. Sie oszillieren zwischen Wahrheitsanspruch, Täuschungsverdacht, erratischer Selbstreferenz und kollektivierender Schau. Sie verweisen in den Raum der Ästhetik, aber auch den der Semiotik, Epistemologie und Ontologie. Während die *Mimesis* die notwendige Differenz zwischen Ur- und Abbild betont, kassiert die kurrente Vorstellung von *Simulation* diese Differenz auf einer höheren Ebene, in der es das Eigentliche und Uneigentliche so nicht mehr zu geben scheint.

In diesem Schritt sollen ‚Bilder der KI‘ auf der Ebene visualisierter Artefakte (eine Anfrage über eine beliebige Suchmaschine im Internet vermittelt einen brauchbaren Eindruck) sowie auf der Ebene von Konzepten und Weltbildern untersucht werden. Dabei gehen wir mit Huber (1989) davon aus, dass auch Technikbilder „Geltungskonflikte von Weltbildern“ transportieren.

Bilder werden von Gegenständlichem, aber auch von zunächst Gegenstandslosem (Geistigem, das dann eine Brücke schlägt zu Weltbildern, vgl. Hogrebe 1971) angefertigt. Das betrifft ontologische Kategorien (ewige Ordnung oder Schicksal wie die Parzen bzw. Moiren und Sibyllen), aber auch abstrakte Konzepte von Freiheit<sup>8</sup> oder Gerechtigkeit<sup>9</sup>, nicht zuletzt auch Tugenden als Abbildungen des Gutseins<sup>10</sup>. Zumeist handelt es sich dabei um Darstellungen von Frauen, die an entsprechenden Attributen oder Haltungen ihre Bedeutung erkennen lassen.<sup>11</sup>

<sup>8</sup> Geradezu emblematisch ist hier die Freiheitsstatue in New York; vgl. außerdem Starobinski 1988.

<sup>9</sup> Die Fülle der Bilder und Texte ist unüberschaubar, für einen guten Überblick vgl. Schild 1995; Schild und Pleister 1988.

<sup>10</sup> Vgl. zur Übersetzung des griechischen *areté* als Tugend und Gutsein – und das nicht vorrangig in einem moralischen Sinne – Stemmer 1998.

<sup>11</sup> Wenn wir heute diese Gestalten vor allem als Personifikationen, Allegorien oder Symbole verstehen, so müssen wir uns doch in Erinnerung rufen, dass es für die Kunst bis in die Renaissance darum ging, in einem (neu)platonischen Sinne die göttlichen Ideen in den künstlerischen Darstellungen zu repräsentieren. „Es ging den Theoretikern wie dann auch den

Folgt man dieser Praxis, so verwundert es nicht, dass gegenwärtig auch die Künstliche Intelligenz in die Reihe dieser Ikonologie aufrückt. Dabei sind allerdings einige bedeutende Unterschiede hervorzuheben. Zunächst einmal erscheint die (!) Intelligenz nicht als weibliches Wesen, sondern androgyn, sie ist zumeist unbekleidet und lässt so ihre Konstruiertheit erkennen, sei es durch einen Blick auf die äußerlich erkennbaren technischen Bauteile, sei es durch einen Blick in das Innere der Maschine, wo Drähte, Platinen, Motoren oder ähnliches zu erkennen sind. Es ist eine glatte Nacktheit, die kein sexuelles Begehren signalisiert oder erweckt, aber ähnlich wie der Körper der Jungfrauen „in sich selbst nichtssagend und bestimmungslos“; er kann deshalb „nach außen hin für jede Botschaft verwendet werden“ (Schild 1995, S. 243). Das mag ein Hinweis sein, warum sich die KI hinsichtlich der Übertragungskombinationen als so reaktionsfreudig erweist, mehr oder weniger überall Anteil nimmt, ohne selbst Anteil zu geben.<sup>12</sup>

Das führt auf einen zweiten Punkt: Die KI erscheint verkörpert<sup>13</sup>, zumeist humanoid, aber dezidiert nicht als Mensch, sie erscheint verkörpert als *Roboter* und damit als ein erkennbares technisches Artefakt. Die Ikonizität der KI bedient sich zwar bekannter Muster, wandelt sie jedoch in signifikanter Weise ab und markiert damit eine wichtige Differenz. Ähnlichkeit, Identität und Differenz spielen auch bei diesen Bildern eine Rolle: hinreichend identisch, um die Nähe und Bedeutung für Menschen zu signalisieren; hinreichend different, um das Andere zu markieren. Die Bilder von Tugenden, Schicksalsmächten oder Leitvorstellungen zeigen ihre Bedeutung für den Menschen in etwas, das es zu erstreben oder anzunehmen und zu tragen gilt. Der Vor-Bild-Charakter der einen ist so unmissverständlich wie der Bestimmungscharakter der anderen. Der Charakter der KI scheint von anderer Art zu sein, ihre Leistung kann von Menschen nicht erreicht werden, zum Vorbild taugt sie nicht. Sehr wohl könnte sie aber als Bestimmungsmacht ‚gelesen‘ werden, die aufgrund ihrer Leistung den Menschen schicksalshaft (heute würde man eher ‚sachzwanghaft‘ sagen) bestimmt. In diesem Sinne lassen

---

Künstlern um Wahrheit und Wirklichkeit, nicht so sehr um die Veranschaulichung abstrakter Begriffe.“ (Schild 1995, S. 12).

<sup>12</sup> Vgl. zur Reaktionsfreudigkeit Gransche & Manzeschke „Künstliche Intelligenz ins Bild gesetzt – Sprachbild und Bildsprache eines Technikmythos“ (unter <https://www.kimege.de/ki-atlas/>, zuletzt aufgerufen 20.03.2023), „Thesenpapier zu den sozialen Implikationen Künstlicher Intelligenz“ ([https://www.kimege.de/wp-content/uploads/2022/07/KIMEGe\\_Thesenpapier.pdf](https://www.kimege.de/wp-content/uploads/2022/07/KIMEGe_Thesenpapier.pdf), zuletzt aufgerufen 20.03.2023) sowie Gransche & Manzeschke in diesem Band „Das bewegliche Heer der Künstlichen Intelligenz. Ein Technomythos als Summe menschlicher Relationen.“

<sup>13</sup> Damit ist ein wesentlicher Kritikpunkt früherer KI-Diskussion um ihre Leistungsfähigkeit aufgenommen worden, und zwar nicht nur in der bildlichen Darstellung, sondern auch technologisch, vgl. Dreyfus 1985.

sich Prognosen über die Singularität (Kurzweil 2013) Superintelligenz (Bostrom 2014) oder der Mind Children (Moravec 1990) als Bestimmungen der Weiterentwicklung (trans-)humanen Lebens lesen. Der Roboter erscheint als ‚natürliche‘ Verkörperung der KI und bietet Analogieschlüsse zu Körper und Geist von Menschen an, die weder technologisch noch anthropologisch tragen. Auch wenn Referenzen in Bildsprache und Sprachbildern Ähnlichkeitsvermutungen nahelegen, so sind die KI-Bilder Simulationen ohne echten Referenzpunkt. KI simuliert Denken und Intelligenz, Emotion und Intention. Sie simuliert eine Einheit von Körper und Bewusstsein, was auf den ersten Blick so ähnlich aussieht wie bei Menschen. Dass es sich anders verhält, ist kein Problem. Problematisch sind jedoch die Schlussfolgerungen von Menschen, die aus Analogieschlüssen von der KI ‚Entscheidungen‘ erwarten oder ihr Bewusstsein zuschreiben (vgl. Rosengrün 2021).

Die Farbgebung bei Darstellungen der KI ist fast durchgehend getragen von einem kühlen Blau, das den Hintergrund der Bilder prägt. In der christlichen Ikonographie ist Blau die Farbe der Luft und des Himmels bzw. der Verbindung mit ihm; Maria, die Mutter des Gottessohnes Jesus Christus, trägt einen blauen Umhang, in dem sie die Menschen schützend birgt (LCI 1994). Das kühle Blau mag auch mit der temperamentmäßig moderierten kühlen Rationalität bzw. Klugheit assoziiert werden, die als Merkmal der KI gilt (vgl. Grieser 2017). Eine andere Leitfarbe ist silbergrau bzw. weiß, das die Hauptfarbe für die Körper der Roboter abgibt und an Edelmetall oder Alabaster erinnert. Auch hier mag man Anklänge an die christliche Ikonographie erkennen: Gold und Weiß sind die Farben Gottes und der Heiligen; wie ein Abglanz Gottes markiert das Weiß bzw. Silber eine andere, vom Menschen unterschiedene Sphäre – nicht zuletzt der Unschuld und Reinheit.

Intelligenz ist ein Fähigkeitenbündel (vgl. Gardner 2005), und so finden sich in den Bildern der KI solche, die auf Aspekte von Intelligenz (oder was man dafür halten mag) verweisen: das Manipulieren an einem Smartboard, das Aufblitzen eines Gehirn-Ähnlichen als Denkvorgang, Netze, Datenströme und Schaltkreise als Hinweise auf die Komplexität der Computationen sowie ihrer technischen Basis, denkerische Haltung à la Rodin oder Hamlets Geste eines fundamentalontologischen Monologs („*to be or not to be, that's the question*“) und nicht zuletzt die vielen Roboter- und Menschenhände in einer Komposition, die an Michelangelos Schöpfungsbild in der Sixtinischen Kapelle erinnert. Das Bildmaterial changiert zwischen religiösen Residuen und materialistischen Details, zwischen zeitgeistigen Accessoires und menschlichen Grundhaltungen. So verwundert es kaum, dass religiöse Motivik, die bereits auf der Metapherenebene zu finden

ist, auch hier starke Aufnahme findet. Neben Gottesprädikaten wie Omnipräsens (vulgo Ubiquität), Allwissenheit und Unerkennbarkeit<sup>14</sup> koppeln KI-Bilder mit einer Reihe von biblischen Motiven, u. a. der Schöpfungsgeschichte.<sup>15</sup> In der Vielzahl der Motive ist längst nicht mehr klar, wer hier Schöpfer und wer Geschöpf ist bzw. wer hier auf wen angewiesen ist – aber vielleicht ist genau das bei manch strategischen Inszenierungen der entscheidende Punkt.

Bemerkenswert für Bilder der KI ist, dass sie mit den dazugehörigen Roboterkörpern nie als energieabhängige technische Geräte erscheinen – kein Stromkabel, keine Ladestation. Die Nähe zu menschen- oder götterähnlicher Selbstbewegtheit verdrängt vollständig ihre energietechnische Begrenztheit. In ihrer Anmutung erscheinen sie zumeist freundlich und zugewandt, was *grosso modo* auf Darstellungen aus dem Forschungsbereich zutrifft, wohingegen der Bereich der Science-Fiction auch bedrohliche Anmutungen bietet, die dann auch stärker geschlechtlich binär ausgestaltet sind.

Die genannten Darstellungen lassen den ‚intellektuellen Vorgang‘ der KI (vor-erst?) als *Blackbox* uneinsehbar. Und dieses Konstruktionsmerkmal ist bereits zu Beginn der KI-Forschung gesetzt worden. Seising (2021) weist darauf hin, dass diejenigen, die in Dartmouth 1956 den Begriff der *Artificial Intelligence* durchsetzten (das waren vor allem McCarthy, Minsky, Rochester und Shannon) im Grunde auf der Außenseite der Prozesse bleiben wollten. Ihnen ging es darum, eine Maschine zu bauen, die so aussieht ‚als ob‘ sie intelligent wäre.

---

<sup>14</sup> In diesem Zusammenhang ist die Überlegung interessant, ob es mit der KI wie mit den Göttern ist: Die Menschen können und dürfen die Götter nicht sehen. Das gilt für Mose und alle biblischen Propheten, das gilt aber auch für Semele, die Geliebte des Zeus, die ihren Geliebten in wahrer Gestalt zu sehen beehrte und dafür sterben musste. – Die Rede von der Unerkennbarkeit hat bei aller Normativität auch eine historische Dimension. Was gesehen werden darf oder nicht gesehen werden darf, das ist historisch durchaus divergent und jeweils Ausdruck einer bestimmten Lage. Dass die genauen Computationen der KI für den Menschen unverständlich und verborgen bleiben, hat wohl nicht nur eine technische Seite (zu schnell, zu komplex), sondern scheint auch sozial bedingt zu sein. Es erinnert ein wenig an den Zauberer von Oz, der auf der Vorderbühne eine imponierende Inszenierung bietet, wohingegen auf der Rückseite, gleichsam im Maschinenraum, etwas sehr Banales vor sich geht. Die Arbeit der Clickworker, die Ghost-Work, gehört zu dieser Kategorie. – Auch das ein Bild aus einer Erzählung, die uns in vergleichender, vielleicht metaphorischer Weise zu informieren und orientieren vermag.

<sup>15</sup> Damit spielt auch Dyson (2014, S. 9): „Es gibt zwei Arten von Schöpfungsmythen: solche, in denen das Leben aus dem Lehm entsteht, und solche, in denen es vom Himmel fällt. In diesem Schöpfungsmythos entstanden die Computer aus Lehm, und der Code fiel vom Himmel.“ Wiener (1965, S. 27ff.; 61ff.) postuliert sehr früh, dass Maschinen andere Maschinen schaffen können und dass damit religiöse Fragen verbunden sein mögen.

„Das heißt: Die Forschung orientierte sich an der Verhaltenstheorie, wollte alles Nichtbeobachtbare unberücksichtigt und die Blackbox verschlossen lassen und nichts darüber aussagen, ob das Verhalten der Maschine tatsächlich intelligent ist, ob es künstliche Intelligenz überhaupt gibt, geben kann oder geben wird. Es ist die mit diesen Forschungen angestrebte Imitation von Intelligenz, die an der ‚Künstlichen Intelligenz‘ das Künstliche ist“. (Seising 2021, S. 119)

Verfolgt man, wie Seising, die Genese der Künstlichen Intelligenz, so wird deutlich, dass die verarbeitete Information zunächst auf einer rein semiotischen Ebene verstanden wird und ihre semantische und pragmatische Interpretation erst später (nach Dartmouth) hinzutreten. Erst in diesem Überschritt gewinnt der hier verwendete Intelligenzbegriff die heute mit menschlicher Intelligenz verknüpfte Bedeutung.

---

#### 4 Bilder der KI (genitivus subjectivus)

Als eine der ersten Aufgaben für Künstliche Intelligenz wurde die Bilderkennung genannt, damals noch Mustererkennung oder Symbolverarbeitung. Dabei ist der Begriff Bilderkennung metaphorisch hinsichtlich beider Wortelemente. Es geht um die Suche nach und das Wiedererkennen von bestimmten Mustern in z. T. sehr großen Datenmengen. Diese Daten müssen hierzu entsprechend aufbereitet werden, sodass dann mit oder ohne menschliche Anleitung die KI ‚lernt‘, hier Regelmäßigkeiten zu erkennen und so dem Menschen Hinweise auf den Charakter des jeweiligen ‚Bildes‘ geben kann (z. B. Buchstabe in einem Wort in einem Text, Tumor in einem histologischen Schnitt, Verkehrsschild auf einer Kreuzung, Konsumgewohnheit einer Person X usw.). Die ‚Daten‘ sind das Material, die eine KI verarbeitet, zu denen sie Interpretationen und Empfehlungen formuliert und an denen sie ‚lernt‘.

Auch wenn die meisten Menschen es sprachlich gar nicht wahrnehmen, ist die Rede von der ‚Datenverarbeitung‘ in der Informatik eine grobe Verstellung der ‚Fakten‘ (vgl. Rheinberger 2007). Der aus dem Lateinischen stammende Begriff des ‚Datum‘ bezeichnet ein ‚Gegebenes‘: „1. etwas, d. h. ein je besonderes Seiendes, das sich 2. dem Denkenden zeigt und zwar ohne dessen Zutun (G. ist ‚nicht durch, sondern für Denken‘) und 3. als so Vorfindliches zur Voraussetzung für weitere Behandlung wird.“ (Heyde 1974, Sp. 101). Auch wenn dieses antike Verständnis mittlerweile kritisiert und modifiziert wurde, ist an diesen Sprachgebrauch zu erinnern. Die sogenannten Daten sind streng genommen *Fakten*, nämlich der Lebenswelt nachgebildete digitalisierte *Artefakte*, die zum Teil ihres Entstehungskontextes entkleidet wurden. Das als ‚Daten‘ (*datum*)

bezeichnete Material ist eine Ansammlung von operativ hergestellten ‚Fakten‘ (*factum*), die erst durch mehrere Schritte der Bearbeitung (sei sie von Menschen in Klick-Worker Geisterarbeit oder von Maschinen) zustande kommen: Auswahl, Formatierung, Referenzierung, Annotierung ... Die *objektiven Daten* und aus ihnen gewonnen *objektiven Ergebnisse* sind ein epistemologisches Missverständnis oder Effekt eines Täuschungsversuchs. Die KI ‚arbeitet‘ nur mit Fakten, sowohl auf der Input- wie auf der Output-Seite. In Abwandlung eines alten Programmiererwortes: *facts in – facts out*. Es handelt sich bei dieser sprachlichen Ungenauigkeit aber keineswegs um eine Petitesse. Die Vorstellung, mit Daten zu arbeiten, legt nahe, dass es sich hier um etwas (*Vor*)Gegebenes handelt, das in der Bearbeitung einfach aufgegeben sei. Wer *Geber* dieser Gaben ist, bleibt ausgespart, würde aber, wenn es benannt würde, für Aufmerksamkeit und Nachdenklichkeit sorgen können. Es sind ja nicht von der Natur gegebene Dinge, sondern es sind interessengeleitet von Menschen operativ hergestellte Sachverhalte, die auch weiterhin diesen Interessen dienen sollen. Diese mögen durchaus ehrenwert sein: Gesundheitsversorgung, sichere Arbeitsabläufe, optimierte Logistik usw. Aber es bleiben Interessen, über die, weil sie nicht einfach gegeben, sondern gesetzt werden, zu verhandeln ist und die zu rechtfertigen sind. Gerade das Verschleiern (Dissimulation) des interessegeleiteten Gemachtseins und die Inszenierung (Simulation) des objektiven Gegebenseins gehören zur Macht und Interessensicherung.

Das Missverstehen von KI-Leistungen setzt sich bei den Operationen fort. Zwar gibt es eine steigende Sensibilität für einen durch ‚Daten‘ und die in ihnen bereits eingeschriebenen Verzerrungen bedingten ‚algorithmic bias‘<sup>16</sup>. Unter Einbezug solcher Erwägungen bleibt aber erstaunlicherweise die Erwartung bestehen, dass die Berechnungen präziser, die Empfehlungen unparteiischer und die Lösungen kreativer seien als die von Menschen.

Auch die Leistungen der KI werden in (Sprach)Bildern formuliert: Die KI ‚denkt‘. Es scheint schnell und für nahezu jedermann klar, dass dies eine uneigentliche Redeweise ist, die eine menschliche Fähigkeit auf eine außer-menschliche technische Instanz überträgt. Nun ist schon für den Menschen recht unklar, was Intelligenz eigentlich meint; sicher aber mehr als nur logisches Denken und Sprachvermögen, worauf normale Intelligenztests abstellen. Aber dieser

---

<sup>16</sup> Auch hier ist die Sprache ungenau: Nicht die Algorithmen verursachen den Bias, sondern die sog. Daten, die in ihrer Qualität und Zusammensetzung zu ungerechten Ergebnissen führen können, indem bestimmte Individuen oder Gruppen privilegiert oder auch benachteiligt werden. Algorithmus ist nichts weiter als eine Rechenregel, die zwar auch ‚ungerecht‘ sein kann, aber leichter erkannt werden kann als das zu Ungerechtigkeiten führende Gefälle in den Daten.

unterbestimmte Begriff wird zum Maßstab für eine bestimmte Leistungsfähigkeit einer Maschine, die dann ‚intelligent‘ genannt wird. Für die berühmte erste Dartmouth Konferenz 1956 wurde im Antragsschreiben der Begriff der „künstlichen Intelligenz“ eingeführt und folgendermaßen gefasst: „Für den vorliegenden Zweck wird das Problem der Künstlichen Intelligenz so verstanden, dass man eine Maschine dazu bringt, sich in einer Weise zu verhalten, die man intelligent nennen würde, wenn sich ein Mensch so verhielte.“<sup>17</sup>

Die semantische Unbestimmtheit der Metapher *Intelligenz* wirkt sich auf der Empfängerseite der Maschine überschießend aus. Ihre Leistung, die zunächst einmal auf einer schnellen und präzisen, dem menschlichen Vermögen überlegenen Operation beruht, wird als eine enorme (intelligente) Leistung auf menschlicher Seite anerkannt, die vor allem auf einem – schiefen – Vergleich zu beruhen scheint. Allerdings werden die Ausgaben der KI gerade nicht metaphorisch verstanden. Was sie genau tut, bleibt in der ‚Blackbox‘ und für die Menschen verborgen. Ein Problem, auf das schon sehr früh Norbert Wiener aufmerksam gemacht hat:

„It is my thesis that machines can and do transcend some of the limitations of their designers, and that in doing so they may be both effective and dangerous. It may well be that in principle we cannot make any machine the elements of whose behavior we cannot comprehend sooner or later. This does not mean in any way that we shall be able to comprehend these elements in substantially less time than the time required for operation of the machine, or even within any given number of years or generations [...]. An intelligent understanding of their mode of performance may be delayed until long after the task which they have been set has been completed. This means that though machines are theoretically subject to human criticism, such criticism may be ineffective until long after it is relevant.“ (Wiener 1960, S. 1355).

Nicht zuletzt deswegen hat Josef Weizenbaum (2005, S. 58) die technischen Systeme als für den Menschen „incomprehensible“ bezeichnet. Die Verstehensschwierigkeit liegt offenbar nicht allein auf der Ebene der Computationen<sup>18</sup>, die

---

<sup>17</sup> Zitiert nach Seising (2021, S. 56). Vermutlich geht die Formulierung auf den Mitantragsteller Marvin Minsky zurück, der mit dieser Formulierung immer wieder in Verbindung gebracht wird. Im Antrag lautet die Formulierung: „For the present purpose the artificial intelligence problem is taken to be that of making a machine behave in ways that would be called intelligent if a human were so behaving“; <http://jmc.stanford.edu/articles/dartmouth/dartmouth.pdf>

<sup>18</sup> Als *Computer* wurden bis in die 1950er Jahre die Menschen bezeichnet, die in Verwaltung, Forschung oder Industrie notwendige, zeitraubende und repetitive Berechnungen durchführten. Erst danach ist der Begriff auf technische Geräte, die diese Computationen ausüben konnten, übertragen worden.

hier vorgenommen werden – sie können (in Grenzen) objektiv beschrieben und kritisiert werden. Wie sie bewertet werden, ist epistemisch subjektiver Natur und müsste mit Argumenten eines anderen Registers kritisiert werden. Das Problem liegt sowohl auf der Ebene der Entität KI, deren ontologischer Status (Bewusstsein: ja oder nein?) ebenso wenig verstanden wird, wie ihre Ausgaben (*outputs*) in ihrem semantischen Gehalt einzuschätzen sind. Genauer wird man wohl sagen müssen: deren semantischer Gehalt notorisch überschätzt wird. Eine Gefahr im Diskurs um die KI wäre es, die Metapher der künstlichen Intelligenz beim Wort zu nehmen.

---

## 5 Bilder und Metaphern

Durch die Antike hindurch ist eine Parallelisierung von Wort und Bild zu beobachten und damit auch eine zwischen Poet und Maler. Es wird aber auch eine Grenze deutlich: „Während Dichtern Ausdrucksmittel [...] auch für Seelenzustände zur Verfügung stehen, müssen sich Künstler [...] in engeren Grenzen bewegen. Insbesondere müssen sich Bildhauer und Maler auf einen Augenblick beschränken, den sie für ihre Darstellungen auswählen.“ (Scholz 2000, S. 629). Diese Beschränkung ist jedoch nur vordergründig eine. Zum einen zeigt sich, dass – auch in den Wissenschaften – die Kommunikation via Bilder einen immer wichtigeren Stellenwert erhält. Mehr und mehr werden Bilder zu Ausdrucksmitteln auch für Seelenzustände (z. B. in einer sehr ‚einfachen‘ Form von Emoticons). Zum anderen aber gewinnen die Bilder im KI-Diskurs eine Stellung, in der das scheinbar Unaussprechliche auf der Bildseite (vgl. Braun 2019) das Ungesprochene auf der Wortseite überdeckt.

Ähnlich wie bei Bildern funktioniert die Metapher auf den ersten Blick nur in eine Richtung, als könnte metaphorischer Gebrauch von einem realweltlichen Begriff nur auf der ‚Bildseite‘ gemacht werden. Blumenberg weist darauf hin: „Lebensweltlich muss es immer schon Rückübertragungsverhältnisse gegeben haben, damit die Forcierung des Bewußtseins durch die Metapher ertragen werden konnte.“ (Blumenberg 1997, S. 90) Wenn die Rückübertragung das Bewusstsein forciert oder ‚den Verstand erfrischt‘ (Wittgenstein 1980, S. 1), dann ist zu fragen, was uns eine KI, die ‚denkt‘, zu denken gibt. An anderer Stelle (Gransche und Manzeschke in diesem Band) haben wir bei der Metapher darauf hingewiesen, dass die Rückübertragung durchaus stattfindet und – wenn es gut geht – auch die Metaphernquelle semantisch erläutern und bereichern kann. Werden dabei aber grundlegende Regeln missachtet, so ist Missverstehen das – zumeist unentdeckte – Ergebnis.



Ein Problem, das hierbei nicht übersehen werden sollte, ist das potenzielle Auseinandertreten von theoretischer Einstellung und praktischem Anspruch, das bei dem Phänomen KI auftritt. Es entsteht aktuell durchaus der Eindruck, dass die Erwartungen an die praktische Leistung der KI deutlich höher liegen und Anstrengungen in diese Richtung stärker forciert werden, als die an eine theoretische Klärung einer *künstlichen Intelligenz*. Dem entsprechen die mit KI assoziierten Bilder durchaus. Es scheint, als würden Metaphern und Simulakra in ihrer Uneindeutigkeit geradezu beruhigend wirken. Sie werden nicht zuletzt dort eingesetzt, wo Eindeutigkeit (anscheinend?) nicht erreicht werden kann (oder soll).

Blumenberg zitiert Kleist, der die Menschen in zwei Klassen einteilt: solche, die sich auf eine Metapher, und solche, die sich auf eine Formel verstehen. Und er warnt zugleich: „Den Überfluß, an Metaphern, den unsere Rhetorik produziert, können wir uns nur leisten, weil die Leistungsfähigkeit von Formeln unseren Spielraum bestimmt für das, was über die blanke Daseinssicherung hinausgreift, also auch für das, was uns Metaphern an Überschreitung der Formelhaftigkeit anbietet. Formeln gewährleisten vor allem, die Ausgangszustände von Prozessen mit beliebigen Endzuständen zu verbinden, ohne für das Zwischenfeld oder für die Totalität empirische Gegenständlichkeit vorauszusetzen.“ (Blumenberg 1997, S. 100 f.)

Worauf verstehen sich Menschen im Umgang mit KI? Wir haben gezeigt, dass es sich rhetorisch um eine Metapher handelt. Die aber wird zugleich mit der Vorstellung verbunden, das Prozessieren und die Ergebnisse der metaphorisch gefassten KI basiere auf Formeln. Paradox könnte man formulieren: Es wird auf mehrdeutige Weise formuliert (und angenommen), dass es bei dem Phänomen KI um Präzision, wenn nicht gar Eindeutigkeit gehe. Auch Menschen, die sich auf Formeln verstehen, müssen eingestehen, dass nicht alles in den algorithmischen Prozeduren der ‚lernenden‘ KI nach Formeln verläuft (vgl. Ertel 2016, bes. S. 265 ff.). Das „Zwischenfeld“ zwischen den Ausgangs- und den Endzuständen ist stochastischer Natur. Faktisch verlaufen die Berechnungen einer KI in diesem Feld ohne „Totalität empirischer Gegenständlichkeit“. Vielmehr werden in diesem Feld sogenannte Daten gewichtet, und die mehrfache Gewichtung liefert einen statistisch hoch wahrscheinlichen Output. Das sind zweifellos wertvolle Berechnungen. Sie aber für die Lösung eines Problems zu nehmen, unterschlägt den sehr viel größeren Prozess und die daran beteiligten Menschen, die etwas als Problem definieren, die KI an das Problem ‚heranführen‘ und am Ende deren Berechnungen als Lösungen allererst plausibilisieren müssten.

## 6 Vorläufiges Fazit

Bilder verwirren, sie orientieren, aber sie verleiten auch. Es ist schwierig mit ihnen, aber es geht auch nicht ohne sie. Es geht um einen reflektierten Umgang mit den Bildern, vergleichbar dem Umgang, den Cassirer für Symbole vorgeschlagen hat:

„Philosophie aber will nun nicht an Stelle der alten Formen eine andere, höhere Form setzen; sie will nicht ein Symbol durch ein anderes ersetzen – sondern ihre Aufgabe besteht im Durchschauen des symbolischen Grundcharakters der Erkenntnis selbst. Lösen können wir uns von diesen Formen nicht [...] aber wir können und müssen [...] sie in ihrer] relativen Notwendigkeit begreifen und einsehen. Das ist die einzig mögliche ideelle Befreiung vom Zwang zur Symbolik“ (Cassirer 1995, S. 265.)

Der Diskurs um KI verläuft in einer terminologischen Übertragungs-Rückübertragungsschleife, bei der uneigentliche Begriffe (analoge Redeweisen wie ‚KI denkt‘, ‚KI entscheidet‘) aus dem menschlichen Bereich (Denken, Entscheiden) auf KI übertragen werden. In der KI verläuft ein Rechenprozess, der von Menschen kodiert wird und prinzipiell überprüft werden kann. Praktisch ist das nicht ganz so einfach, weil viele Personen ein Programm über einen längeren Zeitraum kodieren, sodass Dokumentations- und Verstehenslücken die vollständige Transparenz verhindern. Mehr noch ist das maschinelle Lernen jedoch ein Prozessschritt, der systematisch nicht vollständig kontrolliert und verstanden werden kann; die KI ‚geht ihren eigenen Weg‘ und hinterlässt dabei eine ethisch noch zu füllende Verantwortungslücke (vgl. Manzeschke 2021, S. 227). KIs liefern statistische Wahrscheinlichkeiten; Korrelationen also, die epistemisch schwächer sind als Kausalitäten. Hinzu kommt nun noch die systematisch nicht einholbare partielle Intransparenz der KI-Computationen, die sie in Teilen unverständlich machen. Das schmälert nicht ihre Leistung, aber es relativiert ihren Status als objektiv.

Bei der Rückübertragung wird nämlich oft das Uneigentliche (‚Entscheidung der KI‘) zum Eigentlichen erhoben und als solches orientierend für menschliches Denken, Entscheiden und Handeln wirksam. Streng genommen müsste der Rückübertrag als etwas Uneigentliches einer spezifisch menschlichen Sinnoperation unterzogen werden, und diese Operation – wie sonst auch – hermeneutisch ausgewiesen werden. Weil es sich aber scheinbar um ein technisch berechnetes Ergebnis aus einem ‚Denkvorgang‘ handelt, wird das Uneigentliche bezüglich seiner Orientierungskraft vereinglicht.

Um klug mit der KI umzugehen, wäre hier das (ironische) Lob des Erasmus auf die Torheit zu beherzigen: „Ubi res abest, ibi simulatione esse optimam“ – „Wo die Sache fehlt, vertritt sie am besten der Schein“ (Erasmus 1979, S. 129).

---

## Literatur

- Baudrillard, Jean. 1981. *Simulacres et Simulation*, Paris: Éditions Galilée.
- Blumenberg, Hans. 1997. Ausblick auf eine Theorie der Unbegrifflichkeit. In *Ders.: Schiffbruch mit Zuschauer. Paradigma einer Daseinsmetapher*, 85–106. Frankfurt am Main: Suhrkamp.
- Blumenberg, Hans. 2015. Weltbilder und Weltmodelle. In *Ders.: Schriften zur Technik*, S. 126–137. Frankfurt: Suhrkamp.
- Böhme, Hartmut und Gernot Böhme. 1983. Das Andere der Vernunft. Zur Entwicklung von Rationalitätsstrukturen am Beispiel Kants. Frankfurt am Main: Suhrkamp.
- Bostrom, Nick. 2014. *Superintelligence. Paths. Dangers, Strategies*. Oxford: University Press.
- Braun, Bernard. 2019. *Geschichte der Kunstphilosophie und Ästhetik*. Bd. 4. Darmstadt: Wissenschaftliche Buchgesellschaft.
- Cassirer, Ernst. 2001/2002. *Philosophie der symbolischen Formen*, 3 Bde. (= Bd 11–13 der Ernst Cassirer Werke), Hamburg: Meiner.
- Cassirer, Ernst. 1995. *Zur Metaphysik der symbolischen Formen*, Ernst Cassirer Nachgelassene Manuskripte und Texte (= ECN 1), Hamburg: Meiner.
- Dotzler, Bernhard J. 2003. Simulation. In *Ästhetische Grundbegriffe*, hrsg. Karlheinz Barck, Martin Fontius, Dieter Schlenstedt, Burkhard Steinwachs und Friedrich Wolfzettel, 5:509–34. Stuttgart /Weimar: J. B. Metzler.
- Dreyfus, Hubert L. (1985). *Die Grenzen künstlicher Intelligenz. Was Computer nicht können*. 1. Aufl. Königstein/Ts.: Athenäum.
- Dyson, George. 2014. *Turings Kathedrale. Die Ursprünge des digitalen Zeitalters*. 3. Aufl. Berlin: Ullstein.
- Ertel, Wolfgang. 2016. *Grundkurs Künstliche Intelligenz. Eine praxisorientierte Einführung*. 4. Aufl. Wiesbaden: Springer Vieweg.
- Descartes, René. 1959. *Meditationes de prima philosophia. Über die Grundlagen der Philosophie*. Hamburg: Felix Meiner.
- Descartes, René. 1960. *Discours de la methode. Von der Methode des richtigen Vernunftgebrauchs und der wissenschaftlichen Forschung*. Hamburg: Felix Meiner.
- Erasmus von Rotterdam. 1979. *Lob der Torheit*, übersetzt und hrsg. von Uwe Schultz, Frankfurt: Insel.
- Gardner, Howard. 2005. *Abschied vom IQ. Die Rahmentheorie der vielfachen Intelligenzen*. 4. Aufl. Stuttgart: Klett-Cotta.
- Gransche, Bruno und Arne Manzeschke. 2023. Das bewegliche Herr der Künstlichen Intelligenz: Ein Technomythos als Summe menschlicher Relationen. In *Künstliche Intelligenz, Mensch und Gesellschaft*, hrsg. Michael Heinlein und Norbert Huchler. Springer VS.

- Grieser, Alexandra. 2017. Blue Brains: Aesthetic Ideologies and the Formation of Knowledge Between Religion and Science. A connective concept. In *Aesthetics of religion. A connective concept*, ed. Alexandra Grieser und Jay Johnston, 237–270. Berlin: De Gruyter.
- Hamann, Johann Georg. 1999. Aesthetica in nuce. Eine Rhapsodie in Kabbalistischer Prose. In: Johann Georg Hamann. Sämtliche Werke hrsg. Johann Nadler, reprogr. Nachdruck Tübingen/Wuppertal: Brockhaus.
- Hogrebe, Wolfram. 1971. Bild. In *Historisches Wörterbuch der Philosophie*, hrsg. Ritter, Joachim und Gründer, Karlfried. Darmstadt: Wissenschaftliche Buchgesellschaft, Bd. 1: 913–919.
- Heyde, Johannes Erich. 1974. Gegeben(es). In *Historisches Wörterbuch der Philosophie*, hrsg. Ritter, Joachim und Gründer, Karlfried. Darmstadt: Wissenschaftliche Buchgesellschaft, Bd. 3: 101–103.
- Huber, Josef. 1989. Technikbilder. Weltanschauliche Weichenstellungen der Technologie- und Umweltpolitik. Opladen: Westdeutscher Verlag.
- Huizinga, Johan. 1939. Homo Ludens. Versuch einer Bestimmung des Spielelements der Kultur. Amsterdam: Pantheon.
- Kamper, D. 1997. Bild. In *Vom Menschen. Handbuch Historische Anthropologie*, hrsg. Wulf, C., 589–595. Weinheim/Basel: Beltz.
- Krauss, Rosalind. 1998. Eine Bemerkung über die Photographie und das Simulakre. In *Das Photographische. Eine Theorie der Abstände*. 1. Aufl. München: Wilhelm Fink.
- Kurzweil, Ray. 2013. Menschheit 2.0. Die Singularität naht. 2. Aufl. Berlin: Lola Books.
- LCI: Art. 1994. Farbensymbolik. In *Lexikon der christlichen Ikonographie*, Bd. 2: 7–14. Freiburg: Herder.
- Manzeschke, Arne. 2021. Digitalisierung und Organisationsethik. Ethische und technikphilosophische Skizzen. In *Ethik in der Medizin* 33, Nr. 2: 219–32.
- Moravec, Hans. 1990. Mind Children. Der Wettlauf zwischen menschlicher und künstlicher Intelligenz. Hamburg: Hoffmann und Campe.
- Pievani, Telmo und Veléry Zeitoun. 2020. *Homo Sapiens. Der große Atlas der Menschheit*. 1. Aufl. Darmstadt: Wissenschaftliche Buchgesellschaft.
- Platon (2000). Kratylos. In: Sämtliche Werke in 3 Bden. Hrsg. von Heinrich Löwenthal. Heidelberg: Verlag Lambert Schneider, Bd. I, S. 543–616.
- Plessner: Helmuth. 2003. Der Mensch im Spiel. In *Conditio humana*. Gesammelte Schriften, Bd. VIII, 307–313. Frankfurt am Main: Suhrkamp.
- Rheinberger, Hans Jörg. 2007. Wie werden aus Spuren Daten, und wie verhalten sich Daten zu Fakten? In *Nach Feierabend*, 3: 117–125.
- Rosengrün, Sebastian. 2021. Was ist KI und wenn ja wie viele? Vier Rätsel einer Philosophie der Künstlichen Intelligenz. In *Natürliche und Künstliche Intelligenz im Anthropozän*, hrsg. Joachim Rathmann und Uwe Voigt, 33–52. Darmstadt: Wissenschaftliche Buchgesellschaft.
- Schild, Wolfgang und Wolfgang Pleister. 1988. Recht und Gerechtigkeit im Spiegel der europäischen Kunst. Köln: DuMont Buchverlag.
- Schild, Wolfgang. 1995. Bilder von Recht und Gerechtigkeit. Köln: DuMont Buchverlag.
- Schiller, Friedrich. 1795. Über die ästhetische Erziehung des Menschen. 15. Brief. <https://www.friedrich-schiller-archiv.de/ueber-die-aesthetische-erziehung-des-menschen/fuenfzehnter-brief/>

- Schmidinger, Heinrich, und Clemens Sedmak (Hrsg.). 2007. Der Mensch – ein „animal symbolicum“?. Sprache – Dialog – Ritual. Darmstadt: Wissenschaftliche Buchgesellschaft.
- Scholz, O. R. 2000. Bild. In *Ästhetische Grundbegriffe*, hrsg. Barck, K., Martin Fontius, Dieter Schlenstedt, Burkhard Steinwachs und Friedrich Wolfzettel, Bd. 1: 618–669. Stuttgart/Weimar: J. B. Metzler.
- Schultz, Uwe. 2001. Descartes. Hamburg: Europäische Verlagsanstalt/Rotbuch Verlag.
- Seising, Rudolf. 2021. Es denkt nicht! Die vergessene Geschichte der KI. Frankfurt am Main: Büchergilde Gutenberg.
- Sontag, Susan. 2010. In Platons Höhle. In *Texte zur Theorie der Fotografie*, hrsg. Bernd Stiegler, 1. Aufl., 277–301. Stuttgart.
- Starobinski, Jean. 1988. Die Erfindung der Freiheit 1700–1789. Frankfurt am Main: S. Fischer.
- Stegmaier, Werner. 2008. Philosophie der Orientierung, Berlin/Boston: de Gruyter.
- Stemmer, Peter. 1998. Tugend. I. Antike. In *Historisches Wörterbuch der Philosophie*, Bd. 10: 1532–1548. Darmstadt: Wissenschaftliche Buchgesellschaft.
- Weizenbaum, Josef. 2005. ...der Mensch, nicht die Maschine, ist das Maß!. Ethische Aspekte der Informationstechnologie. In *Hochtechnologiemedizin im Spannungsfeld zwischen Ökonomie, Politik, Recht und Ethik*, hrsg. Wolfgang Niederlag, Heinz-U. Lemke, Alberto Bondolfi, und Otto Rienhoff, 58–63. Dresden: Health Academy.
- Wiener, Norbert. 1960. Some Moral and Technical Consequences of Automation. *Science*, New Series. Vo. 3410: 1355–1358.
- Wiener, Norbert. 1965. *Gott & Golem Inc.* Düsseldorf/Wien: ECON.
- Wittgenstein, Ludwig. 1980. *Culture and value: Vermischte Bemerkungen* (Amended 2. ed. with Engl. transl). Blackwell.



# Maschinen lernen nicht! – „Machine learning“-Algorithmen entzaubert

Irmhild Rogalla

## Zusammenfassung

Ein Roboter, der selbstständig lernt: Für die einen faszinierend und erstrebenswert, für andere ein Horrorszenario. Dabei ist „machine learning“ nur ein Oberbegriff für statische Optimierungsverfahren, dessen Fehleinschätzung als „Lernen“ zu großen Missverständnissen führt. Dieser Beitrag klärt daher über die Grundlagen von „machine learning“ auf, indem er zeigt, dass die Funktionen von KI-Systemen nichts weiter sind als angewandte Mathematik. Allerdings wurden die dafür neu entwickelten Algorithmen aus Marketing-Gründen von Anfang an als „lernend“ bezeichnet. Eine detaillierte Analyse der Analogien zum Lernen in der einschlägigen Originalliteratur zum „machine learning“ zeigt aber auch: Die Analogien haben keine Basis in Theorien oder Forschung zum menschlichen oder biologischen Lernen, sie adressieren allein einzelne Aspekte oder abstrakte formale Ähnlichkeiten in den Abläufen. Eine realistischere Einschätzung der Leistungen und Grenzen von „machine learning“-Algorithmen und KI-Systemen als derzeit vor allem in den Medien üblich scheint daher dringend geboten. Sie eröffnet dann auch Gestaltungsmöglichkeiten und -perspektiven in der Entwicklung wie im Einsatz solcher Systeme.

## Schlüsselwörter

Algorithmen • KI-Systeme • Lernen • Lern-Analogien • Machine learning

I. Rogalla (✉)

Institut für Digitale Teilhabe, Hochschule Bremen, Bremen, Deutschland

E-Mail: [info@institut-pi.de](mailto:info@institut-pi.de)

© Der/die Autor(en), exklusiv lizenziert an Springer Fachmedien Wiesbaden GmbH, ein Teil von Springer Nature 2024

M. Heinlein und N. Huchler (Hrsg.), *Künstliche Intelligenz, Mensch und Gesellschaft*, [https://doi.org/10.1007/978-3-658-43521-9\\_6](https://doi.org/10.1007/978-3-658-43521-9_6)

„Machine Learning“, die Vorstellung, Technik könnte lernfähig sein, erzeugt sehr unterschiedliche Reaktionen, große Befürchtungen, große Hoffnungen, spontane Ablehnung, aber auch Schulterzucken angesichts eines technischen Fachbegriffs. Reaktionen sind oftmals emotional geprägt, sei es mit fiktionalen Bezügen (z. B. zu Filmen wie „Terminator“ oder „Her“) oder – zumindest dem Anspruch nach – sachlich eingebettet. Viele Diskussionen, Mythen oder Bilder, seien es KI-Utopien wie Dystopien, haben kaum Bezüge zur Realität. Die häufig verwendeten (Marketing-)Begriffe, die der KI-Diskussion hohe Aufmerksamkeit verleihen, werden nicht verstanden als das, was sie eigentlich sind: Technische Fachbegriffe, die maximal Analogien aufweisen zu ihren gesellschaftswissenschaftlichen Pendanten, aus denen sie entlehnt sind, und somit sehr eingeschränkte Gültigkeit aufweisen. Der Experte für „machine learning“ Andriy Burkov drückt es stellvertretend für viele andere aus (2019, S. 13): „Maschinen lernen nicht“. Der Versuch, „menschliches Lernen“ und „machine learning“ zu vergleichen, verbietet sich daher von selbst, für ihn fehlt jede Grundlage. Denn „machine learning“ ist ein Oberbegriff für Verfahren, mit denen auf informationstechnischer Infrastruktur mit Methoden angewandter Mathematik und Statistik umfangreiche Berechnungen angestellt und dafür große Datenmengen verarbeitet werden. Diese Verfahren sind überwiegend Jahrzehnte alt. Aber erst in den letzten Jahren sind alle Voraussetzungen für ihren tatsächlichen Einsatz erfüllt: hinreichend leistungsfähige Hardware, hinreichend leistungsfähige Internet- und sonstige Datenübertragungsverbindungen sowie Verfügbarkeit extrem großer Datenmengen. „Machine Learning“ und KI-Systeme sind so zu einem weiteren Treiber der digitalen Transformation geworden. Insofern wären sie auch keiner besonderen Erwähnung wert, wenn nicht ihre Vermarktung als „lernende“ Systeme oder „intelligente“ Assistenten so gut funktionieren würde, dass viele Menschen bereit sind, den Systemen Lernfähigkeiten zuzuschreiben. Hierfür kann es unterschiedliche Erklärungen geben. Eine mögliche bieten Gransche und Manzeschke (2023 in diesem Band) mit „KI als Metapher“ an: Weder über A („machine learning“) noch über B ((menschliches) Lernen) verfügen die meisten Menschen hinreichende Kenntnisse, sodass sie zunächst einmal bereit sind, Maschinen Lernfähigkeiten zugestehen. Dies geschieht natürlich vor dem Hintergrund einer langen Tradition, einschlägiger fiktionaler Darstellungen sowie des aktuellen Hypes um „künstliche Intelligenz“.

## 1 Die Nicht-Vergleichbarkeit von „machine learning“ und Lernen

Die Übertragung auf Maschinen von all dem was Lernen bedeuten kann, führt offensichtlich zu großen Missverständnissen und gravierenden Fehleinschätzungen. Ein angemessener Umgang mit den angeblich „lernenden“ Systemen als das, was sie sind, nämlich leistungsfähige, informationstechnische Systeme, die spezielle Funktionen ausführen können, kann hierdurch erschwert oder sogar unmöglich gemacht werden. Denn solange KI-Systeme nicht als technische Systeme (mit spezifischen Grenzen, Schwächen und Risiken) angesehen werden, sondern als eigenständige (Quasi-)Akteure, gelten sie auch nicht als menschengemacht und gestaltbar. So bleiben Optionen, die es in der Entwicklung wie im Einsatz von KI-Systemen gibt, ungenutzt.

Dieser Beitrag fragt daher danach, was im engeren Sinne unter „machine learning“ zu verstehen ist. Die Bezeichnungen „Künstliche Intelligenz“ und „machine learning“ werden dabei nicht, wie sonst oft üblich, synonym verstanden. Denn „machine learning“ bezeichnet eine spezifische Art von Algorithmen, die in der Entwicklung von KI-Systemen, für ihre Implementierung bzw. Programmierung eine große Rolle spielen. Die Bezeichnung dieser Algorithmen als „lernend“ geschah von Anfang an aus Marketing-Gründen, aus der Perspektive von Informatikern, die damit die gegenüber klassischen, deterministischen Algorithmen veränderte flexiblere Funktionsweise der ersten „machine learning“-Algorithmen betonen wollten. Es lässt sich leicht belegen, dass der Ausdruck „lernen“ oder „lernend“ in der Regel als Analogie genutzt wird und lediglich einzelne, sehr spezielle Merkmale adressiert, die nach allgemeinem Dafürhalten zum Bedeutungsfeld von „Lernen“ gehören. Eine Diskussion, ob „machine learning“ „echtes Lernen“ ist (vgl. Bringsjord 2018, Kodelja 2019), erübrigt sich; und in der Folge auch die Fragen danach, was eigentlich menschliches oder biologisches Lernen ausmacht (vgl. z. B. Rogalla 2012), ob dieses bereits vollständig verstanden ist oder welche Lerntheorien (vgl. z. B. Lefrancois 2015, Grubitzsch, Rexilius 1987, S. 629 ff.) einschlägig sein könnten.

---

## 2 Algorithmen und „maschinelles Lernen“

Für das Verhältnis von KI-Systemen und „machine learning“-Algorithmen wird hier von folgenden technischen Voraussetzungen ausgegangen: KI-Systeme sind zunächst einmal (meist sehr aufwendige) Computersysteme, die in Entwicklung wie im Betrieb enorme Ressourcen brauchen und verbrauchen (Crawford 2021,



insbesondere S. 23 ff.). Sie werden für sehr unterschiedliche Zwecke eingesetzt, von der Robotik bis zur Spracherkennung, von Filmempfehlungen beim Streaming bis zur Heizungssteuerung. Typische KI-Systeme sind verteilte informationstechnische Systeme, die aus großen Computerverbänden („Cloud“) und Anwendungssystemen bestehen. Die Anwendungssysteme können handelsübliche Endgeräte sein, Laptops wie Smartphones, aber auch Roboter, autonome Fahrzeuge oder sonstige „smarte Systeme“, die zur Bewegungssteuerung über entsprechende Sensoren und Aktoren verfügen. Wie bei allen anderen IT-Systeme werden die Funktionen von KI-Systeme durch Software realisiert. In mehrfacher Hinsicht essentiell für KI-Systeme sind Daten bzw. Datensätze in der Regel in Form sehr großer Datenmengen, deren Sammlung, Aufbereitung und Verarbeitung entsprechende Datenbanken, Datenspeicher, Hardware sowie Netzwerke und ihre physische Infrastruktur voraussetzt. Einen großen Teil der „künstlichen Intelligenz“ von KI-Systemen machen implementierte sogenannte „Modelle“ aus, die Beziehungen zwischen den Daten beschreiben. Genauer gesagt: Es handelt sich dabei um Programme, die Beziehungen zwischen Merkmalen von Datensätzen berechnen. Typische Beispiel dafür sind: die Ausgabe der englischen Übersetzung für einen deutschen Satz oder der Ertrag eines Windparks für den nächsten Tag/die nächste Woche aufgrund aktueller Wetterdaten. Die für die jeweilige Anwendung angepassten (häufig ist von „trainierten“ die Rede) und implementierten Modelle wiederum werden im Entwicklungsprozess des KI-Systems mithilfe von „machine learning“-Algorithmen erstellt. „Machine learning“-Algorithmen (auch: ML oder mf-Algorithmen) sind also Teil der Software-Entwicklung von KI-Systemen und finden sich sehr früh im Entwicklungsprozess. Dies ist zunächst einmal nicht ungewöhnlich: Bei der Programmierung neuer Software werden geeignete vorhandene Algorithmen ausgewählt und implementiert. Entscheidend ist vielmehr der Unterschied zwischen klassischen und ml-Algorithmen, denn letztere verändern den Prozess der Erstellung von Software. Teilweise ist sogar von einer Automatisierung des Programmierens die Rede (Huber 2021).

Algorithmen sind zunächst einmal Lösungsprinzipien. In der klassischen Programmierung und Datenverarbeitung gilt: Ein Algorithmus ist „eine präzise, d. h. in einer festgelegten Sprache abgefasste, endliche Beschreibung eines allgemeinen Verfahrens unter Verwendung elementarer Verarbeitungsschritte zur Lösung einer gegebenen Aufgabe.“<sup>1</sup> Algorithmen bestehen also aus wohldefinierten Einzelschritten. Sie können zur Ausführung in ein Computerprogramm

---

<sup>1</sup> Gabler Wirtschaftslexikon: Algorithmus. Online: <https://wirtschaftslexikon.gabler.de/definition/algorithmus-27106/version-250769> Revision von Algorithmus vom 19.02.2018 – 13:18.

implementiert, aber auch in menschlicher Sprache formuliert werden. Bei ihrer Verwendung wird eine bestimmte Eingabe in eine bestimmte und immer dieselbe Ausgabe überführt. Eine typische Variante von Algorithmen sind Berechnungsvorschriften. Entsprechend gehört es zu den Eigenschaften von Algorithmen, dass sie nicht nur eindeutig, sondern auch determiniert (sie liefern bei denselben Voraussetzungen immer das gleiche Ergebnis) und deterministisch (die nächste anzuwendende Regel im Verfahren ist zu jedem Zeitpunkt eindeutig definiert) sind. Dies macht – wenn die Algorithmen implementiert und damit in durch Maschinen ausführbare Programme verwandelt werden – ihre Leistungsfähigkeit aus. Aber es kennzeichnet auch ihre Beschränkung: Sie sind nicht flexibel, benötigen bestimmte Eingaben, mit denen sie genau das, was sie tun sollen, tun, um bestimmte Ausgaben zu erzeugen. Abkürzungen, andere Lösungen, Varianten oder ähnliches gibt es nicht. Somit hängt die Leistungsfähigkeit von Algorithmen stark davon ab, wie gut ein Algorithmus einem Problem gerecht wird. Ein wichtiger Bereich an der Schnittstelle zwischen Informatik und Mathematik ist daher die Untersuchung und Entwicklung von Algorithmen. Ursprünglich in der universitären Grundlagenforschung angesiedelt, ist dieser Bereich inzwischen Domäne großer Digitalkonzerne. Denn neue Lösungen und gute Algorithmen sind sehr gefragt.

Einen solchen neuen Ansatz stellen „machine learning“-Algorithmen dar: Ihr grundlegendes Konzept ist bereits Jahrzehnte alt. Die Bezeichnung geht auf Arthur Samuel (1959) von IBM zurück, der die von ihm entwickelten Algorithmen für ein Programm, das Dame spielt, so nannte („machine-learning procedures“). Samuel ging es auf technischer Ebene darum, eine Art von Problemen zu lösen, die mit klassischen Algorithmen nicht lösbar waren. Denn für das Damespiel ist, ähnlich wie für Schach oder Go, kein klassischer deterministischer Algorithmus bekannt, der einen Sieg oder zumindest ein Unentschieden garantiert. Dies gilt (bis heute) schon deswegen, weil die Berechnung aller möglichen Züge jede verfügbare Rechnerkapazität übersteigt. Deswegen entwickelte Samuel neue Algorithmen. Sie arbeiten auf Basis der Regeln des Spiels, einer Ziel- bzw. Richtungsbeschreibung („Gewinnen“) und einiger Parameter bzw. Kriterien, die anzeigen, ob die Richtung stimmt. Gleichzeitig wollte Samuel zeigen, dass sein Programm nicht nur „lernfähig“ ist, sondern besser Dame spielt als menschliche Spieler. Er wählte das Damespiel ausdrücklich deswegen als Beispiel, um möglichst vielen Menschen das Verhalten des Programms und seine Leistungsfähigkeit demonstrieren zu können.

Damit sind die wesentlichen Merkmale von „machine learning“-Algorithmen benannt: Es handelt sich um heuristische Methoden, deren Funktionsweise sich von der klassischer deterministischer Algorithmen mit im Voraus definierten

Schrittfolgen unterscheidet. Heuristische Methoden, für manche nur ein vornehmer Ausdruck für „Herumprobieren“ (vgl. Zweig 2019, S. 52), sind Methoden für die Suche nach Lösungen. Sie garantieren aber weder, dass überhaupt eine Lösung gefunden wird, noch lässt sich feststellen, ob es sich um die beste Lösung handelt. Vor dem Hintergrund des Determinismus klassischer Algorithmen erscheinen Heuristiken als flexibel und werden deswegen als „lernfähig“ bezeichnet. Tatsächlich beruhen die gängigsten und am weitesten verbreiteten ml-Algorithmen auf statischen Berechnungen bzw. darauf, dass sich aufgrund einer Stichprobe aus Daten Schlussfolgerungen ziehen bzw. Muster erkennen lassen. Statistisch gesehen, handelt es sich bei den Mustern um Korrelationen, also wenn-dann-Beziehungen, die eine bestimmbare Stärke und unter bestimmten Bedingungen auch eine bestimmte (mathematisch darstellbare) Form haben (vgl. Hartung 1991, S. 72 ff., 545 ff. und 569 ff.).

Entscheidend ist dabei: Während bei klassischen deterministischen Algorithmen eine bestimmte Eingabe über wohldefinierte Einzelschritte garantiert und mathematisch beweisbar in eine bestimmte Ausgabe überführt wird, beruhen ml-Algorithmen (mit denen die statistischen Modelle erstellt werden) eben auf Heuristiken (vgl. Zweig 2019, S. 49 ff., 130 ff.). Der Computer erhält also beim Erstellen des Modells *nicht* die Anweisung: Wenn eine Mail drei oder mehr Wörter aus der Liste der Schlüsselwörter enthält, dann kennzeichne sie als Spam. Stattdessen lautet die Anweisung: Stelle fest, anhand welcher Merkmale sich Mails so in zwei Gruppen aufteilen lassen, dass sich diese beiden Gruppen möglichst klar unterscheiden und sich in der einen Gruppe möglichst alle Mails befinden, die als Spam gekennzeichnet sind. Es gibt also Unschärfen. Es ist nicht bekannt, ob es eine Lösung gibt und damit ist jedes Modell auch nur ein Lösungsversuch.

An dieser Stelle ist das Schulterzucken angesichts eines technischen Fachbegriffs situiert: Ja, ml-Algorithmen ermöglichen die Bearbeitung von Problemen, die mit klassischen Algorithmen nicht lösbar sind und sie ermöglichen dies auf eine neuartige Weise, nämlich indem im Entwicklungsprozess zunächst auf der Basis von statischen Auswertungen und Optimierungen Modelle von Beziehungen zwischen Merkmalen von Daten erstellt und dann diese Modelle implementiert werden.

### 3 „Machine learning“ lernt nicht

Die für das Verständnis der Mechanismen von „machine learning“ erforderliche Mathematik, Statistik und Wahrscheinlichkeitstheorie sowie das für die Entwicklung und Anwendung von Algorithmen erforderliche „Computational Thinking“ (vgl. Curzon und Mc Owan 2018) sind jedoch ebenso formal wie abstrakt und daher kaum allgemeinverständlich darstellbar. Genau dies führte und führt – neben dem häufig anzutreffenden Interesse an (Selbst)Vermarktung – dazu, dass sich die Benennung „machine learning“ eingeschlichen hat und gerne auf weitere Analogien zu (biologischem oder menschlichem) Lernen zurückgegriffen wird. Mit einer umfassenden Analyse<sup>2</sup> der Primärliteratur aus der Forschung und Entwicklung zu ml-Algorithmen lassen sich für unterschiedliche Gruppen von ml-Ansätzen jeweils spezifische Bezüge bzw. Analogien zu Lernen nachweisen.

#### 3.1 Statistical Learning Theory

Die erste dieser Gruppen bilden Methoden, die auf der „Statistical Learning Theory“ (vgl. Alpaydin 2019, Hastie et al. 2009, Luxburg und Schoelkopf 2009, Vapnik 1999/2010) beruhen. Sie sind im „machine learning“ am weitesten verbreitet und beruhen darauf, dass sich Daten mit statistischen Methoden auswerten lassen, genauer: dass sich aufgrund einer Stichprobe aus Daten Schlussfolgerungen ziehen lassen. Diese Schlussfolgerungen entsprechen Regelmäßigkeiten oder Mustern, die sich in Daten finden lassen, insbesondere dann, wenn sehr große Datenmengen zur Verfügung stehen. Klassische Beispiele hierfür sind die Identifizierung von Spam-Mails und die Prognose von Kundenverhalten („Kunden, die diesen Artikel gekauft haben, interessierten sich auch für ...) zur Optimierung von Verkäufen wie von Beschaffung, Lagerhaltung und Logistik.

Die Idee des statistisch geprägten „machine learnings“ besteht darin, zunächst ein statistisches Modell zu finden, welches Korrelationen von Merkmalen der Daten darstellt. Dafür wird ein Datensatz genutzt (die Stichprobe, genannt „Trainingsdaten“) über den möglichst viel bekannt ist. Einige Merkmale („features“ oder „label“) der Daten werden als (potentiell) relevant gekennzeichnet und das gewünschte Ergebnis („outcome“)<sup>3</sup> festgelegt. „Machine learning“-Algorithmen

---

<sup>2</sup> Es handelt sich um eine spezielle Variante der Artefaktanalyse (vgl. Froschauer und Lueger 2020; für den Kontext technische Systeme, Digitalisierung, „machine learning“: Suchman 2009; Mackenzie 2017).

<sup>3</sup> Wird das gewünschte Ergebnis vorher festgelegt, werden die entsprechenden „Lern-“ bzw. korrekter: Optimierungsprozesse auch als überwacht („supervised“) bezeichnet. Wird vorab

sind dann in der Lage, mittels Wahrscheinlichkeitsrechnung und/oder Schätzmethoden, Zusammenhänge zwischen Merkmalen und (gewünschtem) Ergebnis zu ermitteln (vgl. Hastie et al. 2009). Bei der Spam-Erkennung zum Beispiel enthält der Datensatz sowohl Spam- als auch andere Mails. Die Spam-Mails sind als solche gekennzeichnet, denn diese Kennzeichnung ist der gewünschte Outcome. Außerdem sind bestimmte Merkmale festgelegt, zum Beispiel Schlüsselwörter (wie „kaufen“, „billig“, „viagra“) im Betreff oder im Text, spezifische Arten der Formatierung und Ähnliches. Mithilfe beispielsweise einer „Support Vector Machine“ (vgl. Ng und Soo 2018), also einem ml-Algorithmus lassen sich nun typische Kennzeichen von Spam-Mails berechnen. Genauer gesagt, versucht der Algorithmus, eine Grenzlinie (deswegen „Vector“) zwischen den Merkmalen der beiden Gruppen von Mails, Spam und anderen zu finden. Sehr vereinfacht könnte diese Grenzlinie beispielsweise heißen: alle Mails mit mehr als drei Schlüsselwörtern sind Spam. Die Werte und Einstellungen, mit denen die (beste) Grenzlinie ermittelt wird, wird „Modell“ genannt. Das Modell seinerseits ist eine Funktion oder eine Gruppe von Funktionen, aus denen mit Hilfe bekannter, deterministischer Algorithmen ein Programm erstellt werden kann, welches in der Lage ist, Spam-Mails – natürlich mit einer gewissen Unschärfe (vgl. zu Details Luxburg und Schoelkopf 2009) – als solche zu klassifizieren.

Dieses typische, heuristische (vgl. o.: „2. Algorithmen und „maschinelles Lernen“) Vorgehen beim Erstellen von Modellen, von denen es eine Reihe unterschiedliche Varianten gibt, wird häufig als „lernen“ bezeichnet. Eine Analogie, die vor dem Hintergrund klassischer deterministischer Algorithmen verständlich ist: Die „Machine Learning“-Heuristiken werden für Datenanalysen eingesetzt, also für Analysen, die sonst Aufgaben von Experten waren. Zudem sind aufgrund der schiereren Menge an Daten die Analysen, für die es nur wenige Spezialisten gab und gibt sehr aufwendig. Insofern handelt es sich beim „machine learning“ um eine Automatisierung der Programmierung bzw. Software-Entwicklung (Huber 2021, vgl. auch datarevenue o. J.). Gleichzeitig sind die Vorgänge sehr unanschaulich und abstrakt: Analysiert werden Daten mit statistischen Mittel, mit entsprechender Software. Dass hierfür auch für Laien verständliche Beschreibungen gesucht wurden, ist nachvollziehbar. So kommt die Analogie zum Lernen zustande. Bei den statischen Methoden ist dann meist von „Lernen aus Beispielen“, „induktivem Schließen“ oder gar „verallgemeinern“ die Rede. Betrachtet man diese Beschreibungen vor dem Hintergrund dessen, was über menschliches

---

kein Ergebnis festgelegt, sondern nur eine allgemeine Suche nach Zusammenhängen oder Mustern durchgeführt, heißt der entsprechende Prozess „unüberwacht“ („unsupervised“). Die Verwendungszwecke der entsprechenden ml-Algorithmen sind jeweils unterschiedlich.

Lernen bekannt ist bzw. durch Lerntheorien erfasst wird (vgl. hierzu ausführlich Rogalla 2012, S. 69 ff.), so werden die Beschränkungen der Analogie sehr deutlich. Drei wesentliche Aspekte seien hier genannt:

- Das, was als „Lernen aus Beispielen“ (z. B. bei Vapnik 1999, S. 988) bezeichnet wird, ergibt sich aus den angewendeten statistischen Methoden, deren Funktion es ja ist, Korrelationen zu berechnen, also nach Zusammenhängen bzw. Mustern zu suchen.
- Menschen, die aus Beispielen lernen, sei es das Erkennen von Spam-Mails oder von Tieren, entwickeln einen Begriff oder eine Kategorie. Sie verstehen, und deswegen können sie u. a. auch zwischen verschiedenen Arten von Zusammenhängen unterscheiden. Sonst wären nicht schon Kinder in der Lage sowohl Katzen und Hunde als auch Spinnen und Vögel als Tiere zu erkennen. Ein ml-Algorithmus hingegen verarbeitet lediglich Datensätze und stellt Zusammenhänge zwischen Merkmalen fest, Merkmalen, die noch dazu von Menschen benannt und als (möglicherweise) relevant gekennzeichnet wurden. Eine Kategorie wie „Tier“ kann er niemals erkennen oder bilden, da er weder den Zweck von Kategorien oder Oberbegriffen begreift, noch ihren Sinn versteht.
- Während Menschen nicht nicht lernen können (vgl. z. B. Bateson 1987), beschränkt die Ausführung von ml-Algorithmen, ihr „learning“ bzw. „training“ auf das Erstellen des statistischen Modells. Es ist also zeitlich klar begrenzt und findet niemals aus Interesse bzw. „eigeninitiativ“ statt sondern nur im Rahmen der Entwicklung von KI-Systemen, auf Basis der Vorgaben von Data Scientists und durch sie stark kontrolliert.

### 3.2 Mathematisch-funktionale Ansätze

Die zweite Gruppe von ml-Algorithmen bilden die mathematisch-funktionale Ansätze (z. B. Jain u. a. 1999, Shalev-Shwartz und Ben-David 2014). Sie sind den statischen Ansätzen prinzipiell ähnlich (vgl. insbesondere Vapnik 2010/1999), nutzen aber ein über die Statistik hinaus gehendes Repertoire an mathematischen Methoden und verfolgen andere Ziele. Während die statischen Ansätze, wie dargestellt, ihre 1999 aus der Datenanalyse nicht verleugnen können, zielen mathematisch-funktionale Ansätze ausdrücklich auf „Systems That Learn“ (Jain u. a. 1999). Auch ihre Argumentation basiert explizit auf bestimmten Vorstellungen oder Theorien von Lernen, wie zum Beispiel von tierischem Lernen

aus Erfahrung (Shalev-Shwartz und Ben-David 2014, S. 19 ff.) oder, deutlich abstrakter, von kindlichen wie wissenschaftlichen Lernprozessen, die aus induktivem Schließen und empirischer Überprüfung bestehen (Jain u. a. 1999, S. 3 ff.). Dabei wird teilweise der Anspruch erhoben, (auch) menschliches Lernen erklären zu können: „The Theory of Machine Inductive Inference (or „Computational Learning Theorie“ etc.) attempts to clarify the process by which a child or adult discovers systematic generalization about her environment.“ (Jain u. a. 1999, S. 13).

Die grundlegende Idee mathematisch-funktionaler Ansätze besteht in einer formalen Beschreibung („Rekonstruktion“) von „Lernen“ analog beispielsweise zu einer formalen Beschreibung von Kräften in der Mechanik. Als Basis für die formale Beschreibung von Lernen werden lediglich fünf Voraussetzungen genutzt: 1) eine theoretisch mögliche Realität, 2) nachvollziehbare Hypothesen, 3) verfügbare Daten über eine beliebige (inkl. der tatsächlichen) Realität, 4) eine abstrakte Vorstellung von einem wissenschaftlich handelnden Menschen<sup>4</sup>, (5) erfolgreiches Verhalten eben dieses wissenschaftlich Handelnden, der in der gegebenen, möglichen Realität handelt (vgl. Jain u. a. 1999, S. 13 f.). Hieraus ergibt sich die Grundfrage: Für welche Klassen von möglichen Realitäten gibt es wissenschaftlich Handelnde, die garantiert in jeder Realität aus der Klasse erfolgreich sind? Diese Grundfrage ist mathematisch darstellbar, aus „möglichen Realitäten“ werden dann Mengen von ganzen Zahlen, die zum Beispiel für Ergebnisse von Experimenten stehen können.

Mathematisch-funktionale Ansätze liegen letztlich allen „machine learning“-Algorithmen zugrunde, auch denen der „Statistical Learning Theory“ und des „Deep Learning.“ Gemeinsam ist ihnen, dass ihre Ergebnisse mathematische Funktionen sind, die auf heuristische Weise Optimierungsprobleme lösen. Berechnet wird also, um im Bild zu bleiben, wie ein wissenschaftlich handelnder Mensch sich verhalten muss, um auf jeden Fall Generalisierungen (Muster, Regeln o.ä.) zu finden. Garantiert wird allerdings nur, dass etwas gefunden wird, nicht die beste und auf keinen Fall eine richtige oder eine wahre Lösung. Einige dieser Ansätze eignen sich auch für komplexe Aufgabe wie beispielsweise Spracherkennung, bei denen die rein statistischen Ansätze an Grenzen stoßen.

Mathematisch-Funktionale Ansätze stehen paradigmatisch für die Auffassung von Lernen als Induktion. Dieses Erkennen von Strukturen, Regeln oder

---

<sup>4</sup> Im Original „a scientist“. Dieser „Wissenschaftler“ ist aber eine so abstrakt-formale Figur, dass sich die Frage nach dem Geschlecht nicht stellt, bzw. ihre Angabe bereits eine für den Gedankengang störende Konkretisierung darstellen würde.

Gesetzmäßigkeiten aus Beispielen bzw. Einzelfällen ist auch aus lerntheoretischer Perspektive eine Form tierischen wie menschlichen Lernens bzw. von dessen Darstellung. Allerdings handelt es sich um eine sehr abstrakte, eben rein formale Darstellung, die viele andere Aspekte vernachlässigt. Dies zeigt sich schon daran, dass sich mithilfe des induktiven Schlusses sowohl das Verhalten von Ratten (Shalev-Shwartz und Ben-David 2014) beim Fressen als auch von Wissenschaftlern (Jain u. a. 1999) bei der empirischen Forschung modellieren lässt.

Besonders problematisch erscheint allerdings der Anspruch, der beispielsweise durch „Systems That Learn“ (Jain u. a. 1999) erhoben wird: Eine bestimmte Vorstellung des Ablaufs von Lernprozessen wird formal modelliert, in mathematische Funktionen und entsprechende Algorithmen übersetzt und implementiert. Das so reduzierte Verhalten, welches ausschließlich mit Daten „hantiert“, soll dann die ursprünglichen Lernprozesse besser erklären können? So absurd dies bereits erscheint, soviel problematischer ist es tatsächlich, weil dieser Modellierung bzw. Formalisierung keine differenzierte Vorstellung von Lernprozessen bzw. keine halbwegs gehaltvolle, moderne Lerntheorie zugrunde liegt.

### 3.3 Deep Learning und neuronale Netze

Bei der dritten Gruppe der ml-Algorithmen handelt es sich um die sogenannten „Deep Learning“-Algorithmen (vgl. Goodfellow et al. 2018, Schmidhuber 2014). Sie ergänzen und erweitern die statistischen Modelle, die auf Basis festgelegter Merkmale von Daten Muster in Datensätzen finden, um Methoden, die relevante Merkmale „extrahieren“ bzw. identifizieren. Dabei geht es darum, Probleme zu lösen, die sich aus der Datenverarbeitung selbst ergeben. Zwei typische Beispiele zur Illustration:

1. Wenn die Datensätze aus digitalen Fotografien bestehen bzw. solche enthalten, kann der Computer als Daten lediglich Informationen über einzelne Pixel des Bildes (genauer: ihre Farb- und Helligkeitswerte sowie Positionen) verarbeiten. Ob es sich bei dem Abgebildeten um moderne Kunst, ein menschliches Gesicht oder eine Ampel handelt, ist völlig irrelevant und für das System nicht unterscheidbar. Deswegen werden seit Jahren digitale Fotos in großem Maßstab von Menschen „gelabelt“, d. h. mit Beschreibungen versehen. Sollen KI-Systeme, beispielsweise zur Fahrzeugsteuerung, diese Erkennungsleistung übernehmen, ergibt sich die Frage, wie sie von den einzelnen Pixeln beispielsweise zu einer Ampel kommen, bei der sie auch noch „erkennen“ müssen, ob diese grün, gelb oder rot



anzeigt, auch dann, wenn auf dem Bild durch entsprechenden Lichteinfall alle Pixel nur in Graustufen wiedergegeben werden.

2. Eine weitere wesentliche Frage ist die nach den für die zu findenden Muster relevanten Merkmalen. Das Finden und Festlegen relevanter Merkmale kann ein sehr aufwendiger Prozess sein (vgl. o.), bei dem sich immer wieder die Frage stellt, ob und in welchem Maße (möglicherweise) relevante Merkmale in den Daten überhaupt enthalten bzw. abgebildet sind. Deswegen wird „machine learning“ auch dazu eingesetzt aus Daten bzw. Datensätzen selbst Erkenntnisse zu gewinnen, und zwar insbesondere dann, wenn es sich um Bilder, Fotos, Video- oder Audiodateien handelt. Diese Art des „machine learning“ wird als „representation learning“ bezeichnet (Goodfellow 2018, S. 5 f.), weil die ursprüngliche Darstellung so in eine Repräsentation umgewandelt wird, dass möglichst viele Informationen erhalten bleiben, diese aber durch Maschinen besser verarbeitbar sind. Dabei geht es vor allem darum, diejenigen Faktoren in den ursprünglichen Darstellungen zu identifizieren, die für die Unterschiede („Variation“) zwischen den Darstellungen verantwortlich sind. Bei einer Sprachaufzeichnung also beispielsweise Alter und Geschlecht des/der Sprechenden, Sprache und Dialekt, verwendete Wörter, (typische) Betonungen und vieles andere mehr. Denn für das System besteht jede Audiodatei nur aus zeit- und wertdiskreten (digitalen) Signalen. Bei Fotos von Ampeln wiederum geht es – vorausgesetzt, sie sind bereits als Ampeldarstellungen gekennzeichnet – um Faktoren wie Tageszeit und Lichteinfall/Beleuchtung, die Perspektive, die Art der Ampel (für fließenden Verkehr, für Fahrradfahrer, für Fußgänger, für Straßenbahnen, ...), das Signal (grün, „bitte warten“, ...), den Ort bzw. das Land u. a. Das Identifizieren der Faktoren, die für Variationen verantwortlich sind, kann ein sehr aufwendiger Prozess sein, da häufig alle (oder fast alle) Faktoren alle (oder fast alle) Datenpunkte jeder Darstellung beeinflussen. Bei Ampeldarstellungen gilt dies beispielsweise offensichtlich für den Lichteinfall und die Perspektive.

Deep Learning-Algorithmen werden insgesamt eingesetzt, um dieses Problem zu lösen. Dafür werden zwei grundlegende Ideen kombiniert: Darstellungen werden in „Schichten“ zerlegt bzw. aus ihnen zusammengesetzt und die Erkennung der Faktoren über ein „Netz“ aus „Knoten“ und „Kanten“ realisiert. Ein Musterbeispiel für einen entsprechenden „machine-learning“-Algorithmus, der beide Ideen kombiniert, ist das sogenannte „mehrschichtige Perzeptron“, auch „tiefes Feedforward-Netz“ genannt (Goodfellow 2018, S. 6). Diese mehrschichtigen Perzeptrone sind aktuell Standard, wenn es um Bilderkennung geht, egal ob Katzen von Hunden unterschieden, Verkehrsschilder erkannt oder Fehler in Werkstücken detektiert werden sollen.

Beim „Deep Learning“ werden Analogien zum Lernen auf einer ganz anderen Ebene hergestellt als im Rahmen der „Statistical Learning Theory“. Bereits die Bezeichnung als „Perzeptron“ zeigt es: Es geht um Vergleiche zur Wahrnehmungs- und Informationsverarbeitung im Gehirn. Bezug genommen wird dafür auf die Ähnlichkeit zwischen den Darstellungen der mathematischen Funktionen als Netz und der Darstellung der Neuronen im Gehirn. Diese Neuronen, die neurophysiologisch gesehen Signale wie Sinneswahrnehmungen verarbeiten, bestehen aus „Knoten“, den Zellkörpern mit Eingangskanälen („Dendriten“), und bilden Verbindungen bzw. „Kanten“ zu anderen Neuronen durch die Synapsen (vgl. z. B. Kandel 2014, S. 69 ff.). Allerdings ist bis heute umstritten und Gegenstand der Forschung, wie Neuronen tatsächlich funktionieren, inwieweit das Gedächtnis veränderbar ist und in welchem Verhältnis dies zu den verschiedenen Arten und Ebenen von Lernen steht (vgl. Kandel 2014, S. 220 ff., S 305 ff., LeDoux 2021, S. 147 ff.).

Zusammenfassend lässt sich daher feststellen: Mit der Darstellung der „neuronalen Netze“ des „deep learning“ wird ein Bezug zu den neuronalen Grundlagen von Lernprozessen hergestellt, der diese Grundlagen sehr grob vereinfacht. Insofern ist es auch kein Wunder, dass Vergleiche von „deep learning“ bzw. „neural network“-Algorithmen mit (menschlichem) Lernen sowohl in der Algorithmenentwicklung selbst (z. B. von Schmidhuber S. 34) als auch von anderen (z. B. Linde 2022) kritisiert werden. Menschliches Gehirn und „machine learning“-System unterscheiden sich konzeptionell und prinzipiell auch auf dieser biologisch-systemischen Ebene so sehr, dass jeder, auch metaphorischer Bezug in die Irre führt. Allein schon die ständige Aktivität und Eigendynamik des Gehirns bzw. jedes biologischen Nervensystems unterscheidet es fundamental von maschinellen Systemen.

### **3.4 Spielstrategien als prototypisches Anwendungsgebiet für Kombinationen von ml-Algorithmen**

Im Unterschied zu den bisherigen Gruppen wird nunmehr das prototypische Anwendungsgebiet für ml-Algorithmen schlechthin betrachtet: Spiele, insbesondere Brettspiele. Denn diese, insbesondere Schach, werden immer wieder als Beleg für die Intelligenz von KI und für die vermeintliche Lernfähigkeit der Maschine herangezogen. Insbesondere werden sie häufig als Messlatte für den Vergleich der Leistungs- und Lernfähigkeit von Menschen und Maschinen genutzt. Samuels (1959, vgl. o.) Nutzung des Damespiels ist dafür ein recht frühes Beispiel.

Die grundlegende Idee hinter der Entwicklung von KI-Systemen, die Spiele spielen, ist die Erprobung von „machine learning“-Algorithmen bzw. ihrer Kombinationen an Problemen, die durch klassische Algorithmen nicht lösbar sind. Dadurch soll ihre Leistungsfähigkeit nachgewiesen werden, und zwar auf möglichst allgemeinverständliche Weise, um letztlich kommerzielle Erfolge zu erzielen. Ein jüngeres Beispiel hierfür ist das IBM Watson-System, mit der Software DeepQA, welches 2011 die US-Quizshow „Jeopardy!“ gewann. Dieses System, seine verschiedenen Komponenten, die verwendeten Algorithmen, ihre (Weiter)Entwicklung und ihre Optimierung sind ausführlich dokumentiert (Ferrucci u. a. 2010). Das eigentliche Ziel der Systementwicklung ist die maschinelle Beantwortung natürlichsprachiger Fragen auf der Basis der Auswertung natürlichsprachiger Dokumente, wie sie in Unternehmen und Organisationen in großer Zahl anfallen (Ferrucci u. a. 2010, S. 59). Das „Jeopardy!“-Quiz wurde wegen seiner Anforderungen an natürlichsprachige Antworten als Herausforderung für die Entwicklung gewählt. Eine wichtige Rolle für diese Auswahl spielte aber auch die Bekanntheit der Sendung und der Wettbewerbscharakter: die Maschine besiegt die besten Menschen (a.a.O, S. 60). Für die Systementwicklung wurde zunächst das „Jeopardy!“-Quiz sehr genau analysiert (a.a.O, S. 60 ff.), Metriken für die Leistungsfähigkeit des zu entwickelnden Systems abgeleitet (a.a.O, S. 63 ff.) und auf vorhandene QA-Systeme bzw. Algorithmen (QA = Questions and Answers, a.a.O, S. 66 f.) angewendet. Erst dann wurde mit der Entwicklung des DeepQA-Ansatzes begonnen (a.a.O, S. 67), dessen Besonderheit in der spezifischen Kombination und Weiterentwicklung vorhandener Technologien, eben dem umfassenden Einsatz verschiedener „machine learning“-Algorithmen sowie extrem leistungsfähiger Hardware liegt. Dafür hat allein das Kernteam der Algorithmen-Entwicklung zwischen 2004 und 2007 (a.a.O, S. 75), über sechzig Personenjahre Aufwand in die Entwicklung gesteckt. Für das Gesamtsystem dürfte sich der Aufwand auf mehrere hundert Personenjahre belaufen, die Gesamtkosten liegen bei mehreren Millionen US-\$. Tatsächlich siegte Watson im Februar 2011 bei „Jeopardy!“ gegen frühere Champions und diese Nachricht ging als erneuerter Beleg für die These, wie unterlegen Menschen der Maschine sind, um die Welt. So wurde beispielsweise behauptet: „Der IBM-Rechner versteht Menschen so gut, dass er als „ganz normaler“ Kandidat den TV-Quiz „Jeopardy“ gewinnt.“<sup>5</sup> Auch wenn IBM sicherlich auf diese und ähnliche Effekte spekuliert hatte: von „verstehen“ oder „lernen“ im menschlichen Sinne kann auch bei IBM

---

<sup>5</sup> <https://www.n-tv.de/technik/Computer-gewinnt-Jeopardy-article2634646.html>  
[Stand:22.04.2022].

Watson in keiner Weise die Rede sein. Denn es wurden umfangreiche statistische Modelle und ein komplexes System entwickelt, aber es handelt sich nach wie vor um ein datenverarbeitendes, informationstechnisches System, welches Korrelationen und Wahrscheinlichkeiten berechnet.

### 3.5 Kognitionswissenschaftliche Zugänge

Abschließend sollen, quasi als fünfte Gruppe, einige aktuelle kognitionswissenschaftliche Überlegungen zusammengefasst werden, die explizit „machine learning“ und menschliches Lernen vergleichen und beispielhaft für weitere Arbeiten aus diesem Gebiet stehen: „Building Machines That Learn and Think Like People“ (Lake u. a. 2016) und „Human and Machine Learning“ (Kao und Venkatachalam 2021). Diesen Überlegungen ist gemeinsam, dass sie zumindest nicht direkt auf die Entwicklung von Algorithmen oder Systemen abzielen, sondern nach grundlegenden Mechanismen des Lernens suchen, diese als Prozesse der Informationsverarbeitung verstehen<sup>6</sup> und von der Unterscheidung, ob diese Informationsverarbeitung in biologischen oder künstlichen Systemen stattfindet, abstrahieren (vgl. z. B. Gardner 1989).

Dabei werden zwar die bisherigen Leistungen von ml-Algorithmen und KI-Systemen nicht bestritten, aber ihre Beschränkungen deutlich gemacht. Diese Beschränkungen begründen auch den grundsätzlichen Wechsel der Perspektive, der mit kognitionswissenschaftlichen Ansätzen verbunden ist: Zwar werden abstrahierte Prozesse der Informationsverarbeitung untersucht, aber es wird nicht mehr davon ausgegangen, dass eine mit formalen, mathematisch-funktionalen Mitteln vorgenommene Rekonstruktion von „Lernprozessen“ (vgl. o.) das Wesentliche menschlicher oder generell biologischer Lernprozesse erfasst. Stattdessen werden diejenigen Fähigkeiten des Menschen in den Blick genommen, bei denen Menschen unbestritten besser sind als Maschinen und gefragt, wie Menschen tatsächlich lernen.

Insbesondere werden für kognitionswissenschaftliche Überlegungen Ergebnisse der psychologischen und pädagogischen Lernforschung sowie Lerntheorien, die sich auf menschliches Lernen beziehen, rezipiert, um menschliches Lernen und seine Überlegenheit zu erklären (Kao und Venkatachalam 2021, S. 902 ff., Lake u. a. 2016, S. 15 ff.). Klar gestellt wird dadurch, dass es substantielle Unterschiede zwischen der Funktionsweise und den Leistungen menschlichen Lernens

---

<sup>6</sup> Die Frage, ob es korrekt bzw. auch nur sinnvoll ist, Lernen als Informationserarbeitung aufzufassen, ist strittig, vgl.o.

und von ml-Algorithmen bzw. KI-Systemen gibt, sodass insbesondere die Erfolge bei Spielen wie Schach und Go nicht zur Erklärung menschlichen Lernens herangezogen werden können (Kao und Venkatachalam 2021, S. 902 f.). Insgesamt wird eine erhebliche Diskrepanz zwischen menschlichen und maschinellen Fähigkeiten konstatiert (Lake u. a. 2016, S. 38). Auch ist die kognitionswissenschaftliche Forschung und Entwicklung von arbeitsfähigen Systemen noch weit entfernt.

---

## 4 Zusammenfassung und Bewertung

Die nachfolgende Tabelle gibt einen abschließenden Überblick über die fünf Gruppen (Tab. 1):

Mit Ausnahme der noch in den Anfängen steckenden kognitionswissenschaftlichen Konzepte basieren alle ml-Verfahren auf mathematischen, statistischen oder informationstechnischen Ansätzen. Die genutzten Lern-Analogien beruhen entweder auf formalen Schlussverfahren, die mit Lernprozessen gleichgesetzt werden, oder auf sehr vereinfachten Annahmen über die Funktionsweise des Gehirns. Sie haben in der Regel keine Basis in Theorien oder Forschung zum menschlichen oder biologischen Lernen, sondern bewegen sich bestenfalls auf der Ebene abstrakter formaler Ähnlichkeiten in den Abläufen. Allerdings haben sich die Lern-Analogien, genauso wie die Rede von der künstlichen Intelligenz längst verselbstständigt. Nicht nur in medialen Darstellungen, auch in vielen Sachbüchern wird schlicht davon ausgegangen, dass KI-Systeme lernen könnten wie Menschen. Schlimmer noch: Gerade durch die Zuschreibung von Lernfähigkeit werden aus KI-Systemen vermeintlich selbstständig lernende und handelnde Akteure. Dies geschieht, obwohl viele für menschliche Lernprozesse konstitutive Bedingungen (z. B. Verkörperung, Weltverständnis, Intersubjektivität) offensichtlich nicht erfüllt sind. Zudem fehlt den Maschinen jede Form von Motivation, wie sie für alle Formen biologischen oder menschlichen Lernens, egal ob auf (prä)zelluläre oder auf wissenschaftlicher Ebene, als Selbsterhaltungstrieb, Neugier oder Wunsch nach Reputation kennzeichnend und wohl auch notwendig ist.

Ein adäquater Umgang mit KI-Systemen, deren Funktion auf ml-Algorithmen beruht, erfordert daher, diese Systeme als das wahrzunehmen, was sie sind:

**Tab. 1** Kennzeichen typischer "machine learning"-Ansätze

	Ansatz	Lern-Analogie	Bewertung
Statistical Learning Theory	Statistisches Modell (Korrelationen) aus Merkmalen der Datensätze	Induktives Schließen/ „Lernen aus Beispielen“	Merkmalzusammenhänge aus Beispielen zu berechnen ist weder Begriffsbildung oder Generalisierung noch lernen
Mathematisch-funktionale Ansätze	Optimiertes Modell bestehend aus mathematischen Funktionen	Induktives Schließen; formalisierte, hypothetische Lernprozesse	formalisierte hypothetische Lernprozesse haben keinerlei lerntheoretische Basis
Deep Learning und neuronale Netze	Extraktion von Merkmalen aus Datensätzen (Voraussetzung für Datenverarbeitung)	Wahrnehmungs- und Informationsverarbeitung im Gehirn („neuronale Netze“)	Ähnlichkeiten nur bei grafischen Darstellungen und grob vereinfachter Funktionsweise
Spielstrategien	Kombination mehrerer „ml“- und anderer Ansätze (insbes. Analysen)	Schach, das königliche Spiel – erfordert (von Menschen) strategische Meisterleistungen und entsprechende Übung	Ein durch klare Regeln eingeschränktes Spiel, eine formale Umgebung und eine sehr spezifische, abstrakte Aufgabe – sicherlich eine Leistung, aber nicht einmal typische für (menschliches) Lernen
Kognitionswissenschaftliche Zugänge	Suche nach verallgemeinerbaren Mechanismen	Perspektivwechsel: Es werden Ergebnisse aus Lernforschung und Lerntheorien rezipiert	Die zugrunde liegende Auffassung von Lernen als Informationsverarbeitung ist zumindest strittig

Leistungsfähige informationstechnische Systeme, die spezielle Funktionen ausführen können und gleichzeitig spezifische Grenzen und Schwächen haben, wie einschlägige Erfahrungen zeigen<sup>7</sup>.

Eine solche realistische Einschätzung erfordert neben der Frage nach wirtschaftlichen, gesellschaftlichen und politischen Interessen, die sich mit dem Einsatz von KI-Technologien verbinden, genaue Beschreibungen dessen, was die Systeme tatsächlich leisten (vgl. Tucker 2022). Eine allgemeine informationstechnische Grundbildung („digital literacy“), die diese Bezeichnung verdient, könnte hierzu einen nennenswerten Beitrag leisten. Ergänzt werden müsste sie durch „data“ oder „statistical literacy“, die ein besseres Verständnis der statistischen Grundlagen und ihrer Grenzen ermöglicht. Eine wirklich realistische Sicht ermöglicht auch die „Aufklärung über die Möglichkeiten und Grenzen von KI-Systemen“ vor allem im betrieblichen Einsatz als Grundlage für demokratische Mitgestaltung (Schröder und Höfers 2022, S. 19 ff.).

Insbesondere ergeben sich auf Basis realistischer Einschätzungen Gestaltungsoptionen in der Entwicklung wie im Einsatz von KI-Systemen, die sonst gar nicht in den Blick geraten. Dies gilt zum einen in technischer Hinsicht, mit Blick auf Verbesserungen der Anpassungsfähigkeit von KI-Systemen und Robotern. Statt hier mit unrealistischen Erwartungen an „lernende Systeme“ heranzugehen, könnte eine bessere Differenzierung den Blick für sinnvolle, auch flexible Einsatzmöglichkeiten schärfen. Realistische Gestaltungsoptionen könnten sich aber auch aus soziotechnischen Ansätzen ergeben, die Menschen, IT bzw. KI-Systeme und Organisation sowie Beziehungen zwischen ihnen gleichermaßen berücksichtigen, sie für die Gestaltung betrieblicher Abläufe und Prozesse nutzbar ebenso machen, wie für eine passende Kompetenz- und Personalentwicklung. Die Weiterentwicklung entsprechend geeigneter sozio-technischer, wissenschaftlicher und handlungstheoretischer Ansätze (vgl. z. B. Böhle et al. 2004) speziell im

---

<sup>7</sup> Um nur ein Beispiel zu nennen: Unter der Überschrift: „Künstliche Intelligenz Toyota feuert die Roboter“ berichtete der Tagesspiegel vom 04.01.2019, dass Toyota in der Automobilproduktion in Japan statt Robotern an Fertigungsstraßen wieder Arbeiterinnen und Arbeiter einsetzt, weil dieses beispielsweise viel exakter und den tatsächlichen Gegebenheiten angepasst schweißen können und so Materialverbrauch und Ausschuss deutlich geringer sind (<https://www.tagesspiegel.de/themen/reportage/kuenstliche-intelligenz-toyota-feuert-die-roboter/23821418.html> (Letzter Abruf: 27.08.22)). Auch im Service-Bereich flaut der Hype um den Einsatz „intelligenter“ Roboter in Japan bereits wieder ab. So wurde im Henn-na-Hotel die Erfahrung gemacht, dass Roboter mehr Arbeit machen, als sie erledigen (können), eben weil sie noch nicht annähernd so flexibel sind, wie Menschen (<https://www.golem.de/news/automatisierung-japanisches-roboterhotel-entlaesst-roboter-1901-138757.html>, vgl. auch <https://www.heise.de/news/Post-aus-Japan-Weniger-Roboter-ist-mehr-4285614.html> (Letzter Abruf: jeweils 27.08.22)).

Hinblick auf KI-Systeme, seien es Roboter, Chatbots oder gar Systeme zur Entscheidungsunterstützung, steht allerdings noch aus.

---

## Literatur

- Alpaydin, Ethem. 2019. *Maschinelles Lernen*. Berlin: De Gruyter.
- Bateson, Gregory. 1987. *Geist und Natur. Eine notwendige Einheit*. Frankfurt a. M.: Suhrkamp.
- Bringsjord, S., Naveen Sundar, G., Banerjee, S. und Hummel, J. 2018. Do Machine-Learning Machines Learn? In *PT-AI 2017, SAPERE*, hrsg. V. C. Müller, 44: 136–157. [https://doi.org/10.1007/978-3-319-96448-5\\_14](https://doi.org/10.1007/978-3-319-96448-5_14)
- Böhle, Fritz, Sabine Pfeiffer und Nese Sevsay-Tegethoff. 2004. *Die Bewältigung des Unplanbaren*. Wiesbaden: VS Verlag.
- Burkov, Andriy. 2019. *Machine Learning Kompakt – Alles was Sie wissen müssen*. Frechen: mitp-Verlag.
- Crawford, Kate. 2021. *Atlas of AI*. New Haven and London: Yale University Press.
- Curzon, Paul and Peter W. Mc Owan. 2018. *Computational Thinking*. Berlin: Springer.
- DataRevenue. o.J. Was ist Machine Learning? – eine visuelle Erklärung. <https://www.datavenue.com/de-blog/was-ist-machine-learning>. Letzter Aufruf: 05.04.2022.
- Ferrucci, D. u.a. 2010. Building Watson: An Overview of the DeepQA Project. In *AI Magazine*. <https://doi.org/10.1609/AIMAG.V31I3.2303>
- Froschauer, Ulrike und Lueger, Manfred. 2020. *Materiale Organisation der Gesellschaft: Artefaktanalyse und interpretative Organisationsforschung*. Weinheim: Beltz Juventa.
- Gardner, Howard. 1989. *Dem Denken auf der Spur. Der Weg der Kognitionswissenschaft*. Stuttgart: Klett-Cotta.
- Goodfellow, Ian, Bengio, Yoshua und Courville, Aaron. 2018. *Deep Learning: Das umfassende Handbuch. Grundlagen, aktuelle Verfahren und Algorithmen, neue Forschungsansätze*. Frechen: mitp.
- Gransche, Bruno und Arne Manzeschke. 2023. Das bewegliche Herr der Künstlichen Intelligenz: Ein Technomythos als Summe menschlicher Relationen. In *Künstliche Intelligenz, Mensch und Gesellschaft*, hrsg. Michael Heinlein und Norbert Huchler. Springer VS.
- Grubitzsch, Siegfried und Rexilius, Günter. 1987. *Psychologische Grundbegriffe*. Reinbek bei Hamburg: Rowohlt.
- Hartung, Joachim. 1991. *Statistik*. München, Wien: Oldenbourg.
- Hastie, Trevor, Robert Tibshirani and Jerome Friedman. 2009. *The Elements of Statistical Learning. Data Mining, Inference, and Prediction*. Berlin: Springer.
- Huber, Marco. 2021. Künstliche Intelligenz: Lernen Maschinen wie Menschen? Biointelligenz : Gemeinsame Vortragsreihe 2021/22 der Württembergischen Landesbibliothek und der Fraunhofer-Gesellschaft. <https://www.wlb-stuttgart.de/die-wlb/kultur-und-wissenschaft/veranstaltungen/biointelligenz/> Letzter Aufruf: 05.04.2022.
- Jain, S. u.a. 1999. *Systems That Learn. An Introduction to Learning Theory*. Cambridge Massachusetts: MIT Press.
- Kandel, Ernst 2014. *Auf der Suche nach dem Gedächtnis. Die Entstehung einer neuen Wissenschaft des Geistes*. München: Goldmann.



- Kao, Y. and R. Venkatchalam. 2021. Human and Machine Learning. In *Comput Econ* 57:889–909. <https://doi.org/10.1007/s10614-018-9803-z>
- Kodelja, Z. 2019. Is machine learning real learning? *CEPS Journal* 9, 11–23. <https://doi.org/10.25656/01:18133>
- Lake, B. u.a. 2016. Building Machines That Learn and Think Like People. In *Behavioral and Brain Sciences*. <https://doi.org/10.1017/S0140525X16001837>.
- LeDoux, Joseph. 2021. *Bewusstsein. Die ersten vier Milliarden Jahre*. Stuttgart: Klett-Cotta.
- Lefrancois, Guy, R. 2015. *Psychologie des Lernens*. Berlin, Heidelberg: Springer.
- Linde, Helmut. 2022. Wie sich Deep Learning vom Gehirn unterscheidet. <https://www.golem.de/news/kuenstliche-intelligenz-wie-sich-deep-learning-vom-gehirn-unterschiedet-2202-162231.html> Letzter Aufruf: 08.04.2022.
- Luxburg, U. and B. Schoelkopf. 2009. Statistical Learning Theory: Models, Concepts, and Results. In *Handbook of the History of Logic*. Vol. 10: Inductive Logic. ed. Gabbay, D. M., S. Hartmann and J.H. Woods, 651–706. North Holland: Elsevier. <https://doi.org/10.1016/B978-0-444-52936-7.50016-1>.
- Mackenzie, Adrian. 2017. *Machine Learners: Archaeology of a Data Practice*. London. MIT Press.
- Ng, Annalyn und Kenneth Soo. 2018. *Data Science – was ist das eigentlich?! Algorithmen des maschinellen Lernens verständlich erklärt*. Berlin: Springer.
- Rogalla, Irmhild. 2012. *Moderne Arbeit – Moderne Berufe. Ein interdisziplinäres Modell*. Berlin: R&W-Verlag.
- Samuel, Arthur L. 1959. Some Studies in Machine Learning Using the Game of Checkers. In *IBM Journal*, Vol. 3, No.3.
- Schmidhuber, J. 2014. Deep Learning in Neural Networks: An Overview 2014. Technical Report IDSIA-03–14 / [arXiv:1404.7828](https://arxiv.org/abs/1404.7828) v4 [cs.NE]
- Schröder, Lothar und Petra Höfers. 2022. *Praxishandbuch künstliche Intelligenz. Handlungsanleitungen, Praxistipps, Prüffragen, Checklisten*. Frankfurt: Bund-Verlag.
- Shalev-Shwartz, Shai and Shai Ben-David. 2014. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge: Cambridge University Press.
- Suchman, Lucy 2009. *Human-machine reconfigurations: plans and situated actions*. Cambridge: Cambridge University Press.
- Tucker, Emily. 2022. Artifice and Intelligence. *Tech Policy Press*. <https://techpolicy.press/artifice-and-intelligence/> Letzter Aufruf: 16.05.2022.
- Vapnik, V. N. 1999. An Overview of Statistical Learning Theory. In *IEEE TRANSACTIONS ON NEURAL NETWORKS*. <https://doi.org/10.1109/72.788640>.
- Vapnik, Vladimir N. 2010. *The Nature of Statistical Learning Theory*. Berlin: Springer.
- Zweig, Katharina. 2019. *Ein Algorithmus hat kein Taktgefühl. Wo künstliche Intelligenz sich irrt, warum uns das betrifft und was wir dagegen tun können*. Heyne Verlag.

---

# **Normierung und Regulierung von Künstlicher Intelligenz**



# Die digitale Verantwortungslücke: Vorschläge zur Haftung für algorithmisches Fehlverhalten

Anna Beckers und Gunther Teubner

## Zusammenfassung

Wenn autonome Algorithmen, die innerhalb unterschiedlicher sozio-digitaler Institutionen agieren, Fehlentscheidungen treffen, welche Konsequenzen ergeben sich für die rechtliche Haftung? Zur Entwicklung adäquater Haftungskonzepte recurriert der Beitrag auf sozio logische und philosophische Theoriestücke 1) der Personifikation nicht-menschlicher Akteure, 2) der Mensch-Maschine-Assoziation als emergentem sozialen System mit Qualitäten eines Kollektivakteurs und 3) der distribuierten Kognition in der Interkonnektivität von Algorithmen.

## Schlüsselwörter

Sozio-digitale Institutionen • Personifizierung • Haftung für Künstliche Intelligenz • Verantwortungslücken

---

Für kritisch-konstruktive Hinweise danken wir Michael Heinlein. Auch aus den Diskussionen der KiMeGe-Gruppe haben wir vielfältige Anregungen erhalten.

---

A. Beckers (✉)

Maastricht University, Maastricht, Niederlande  
E-Mail: [anna.beckers@maastrichtuniversity.nl](mailto:anna.beckers@maastrichtuniversity.nl)

G. Teubner

Frankfurt a.M., Deutschland  
E-Mail: [G.Teubner@jur.uni-frankfurt.de](mailto:G.Teubner@jur.uni-frankfurt.de)

# 1 Die digitale Verantwortungslücke: Drei Fälle

## 1.1 Robo-Advice

Samathur Li Ki-kan, ein Tycoon aus Hongkong, verklagt Raffaele Costa, einen Investment-Broker, wegen des Fehlverhaltens eines Robo Advice Computers auf Ersatz eines Schadens von 23 Mio. US\$. Der Supercomputer mit dem Namen K1 sollte Online-Quellen durchkämmen, um die emotionale Stimmung (!) der Anleger einzuschätzen und auf dieser Basis Prognosen für US-Aktientermingeschäfte abzugeben. Obwohl die ersten Simulationen vielversprechend erschienen, gingen beim realen Einsatz im Aktienhandel regelmäßig große Geldsummen verloren. Am 14. Februar 2018 verlor Li aufgrund einer Stop-Loss-Order über 20 Mio. US\$ (zum Fall Beardsworth und Kumar 2019; Pratt 2018).

Dies ist der erste bekannte Fall, in dem Schadensersatzforderungen für die Entscheidungen von autonomen Algorithmen wegen der von ihnen verursachten Anlageverluste vor Gericht gestellt wurden. In diesem Fall geht es um ein Blackbox-Problem: Wenn Menschen die Entscheidungen des Algorithmus nicht nachvollziehen können, wer haftet, wenn etwas schiefgeht? Nach bisher geltendem Recht gibt es keinerlei Schadensersatz, wenn die beteiligten menschlichen Akteure ihre Verhaltenspflichten erfüllt haben. Für bloßes Maschinenversagen haftet niemand. Dies ist die erste große Verantwortungslücke.

## 1.2 Panama-Papers

Im Investigativen Journalismus ist es schon heute üblich, Algorithmen zu nutzen. Dies gilt für die Recherche, insbesondere die Analyse einer großen Zahl von Dokumenten, aber auch für das Erstellen von journalistischem Content. So auch in diesem (hypothetischen) Falle, in dem ein internationales Konsortium von Journalisten autonome Algorithmen in einer komplexen Untersuchung einsetzte, um illegale Steuerpraktiken in einer Steueroase aufzudecken.<sup>1</sup> Die Schwierigkeit der Ermittlungsarbeit bestand darin, eine enorm große Anzahl von überaus komplexen Dokumenten zu analysieren. Algorithmen übernahmen die Arbeit des

---

<sup>1</sup> Dies ist ein fiktiver Fall, der aber öffentlich zugängliche Informationen zur Recherche der Panama-Papiere so aufbereitet, um die emergenten Eigenschaften einer Mensch-Algorithmus-Assoziation und ihre Rechtsprobleme zu illustrieren, hierzu im Detail: Panama Papers, The Journalists behind the Leak, 25. April 2016, abrufbar unter <https://www.sueddeutsche.de/politik/panama-papers-the-journalists-behind-the-leak-1.2966929> (zuletzt besucht am 14.07.2022).

Markierens, Kategorisierens und Auswählens der relevanten Texte. Die Menschen waren in enger Interaktion an ihrer Arbeit beteiligt (Diakopoulos 2019, S. 13 ff.). Auch die spätere Veröffentlichung der Arbeit und ihre Verbreitung in den Nachrichten wurde von Algorithmen in der Form von Priorisierungs- und Filterentscheidungen beeinflusst.

Der Fall zeigt, wie erfolgreich die Zusammenarbeit von Algorithmen und Journalisten sein kann, um verborgene Skandale aufzudecken. Ohne die Hilfe von Algorithmen wäre, allein aufgrund der Masse an Dokumenten, der Steuerskandal nie an die Öffentlichkeit gelangt. Doch eine solche Praxis birgt ein beträchtliches Schadenspotenzial. Wer haftet etwa, wenn bei einer solchen Untersuchung Personen oder Unternehmen beschuldigt werden, die aber in Wahrheit nicht beteiligt waren? Besonders schwierig ist dies dann, wenn nicht eindeutig zu bestimmen ist, ob dem Algorithmus der Fehler unterlaufen ist oder ob menschliches Versagen vorliegt oder aber ob gerade die Zusammenarbeit zwischen Mensch und Algorithmus die falsche Anschuldigung verursachte.? Für die Haftung einer Mensch-Maschinen-Assoziation kennt das geltende Recht keine Grundlage. Hier klafft die zweite bedeutende Verantwortungslücke.

### 1.3 Flash Crash

Der algorithmische Hochfrequenzhandel ist der bekannteste Fall für Schäden, die durch die Vernetzung von Algorithmen entstehen. Berühmt wurde ein Flash-Crash auf dem US-amerikanischen Kapitalmarkt im Jahr 2010 (United States Commodity Futures Trading Commission und U.S. Securities & Exchange Commission 2016). Das US-Justizministerium machte einen Händler, Navinder Saro, für den Crash verantwortlich. Saro wurde wegen ‚Spoofing‘ verklagt: Sein Algorithmus soll falsche Aufträge deswegen auf dem Markt platziert haben, um andere Handelsalgorithmen zu veranlassen, seinem Beispiel zu folgen. Dies erlaubte ihm, seine eigene Strategie in die genau entgegengesetzte Richtung zu ändern und damit Gewinne zu erzielen. Die Ermittlungen ergaben, dass das von Menschen initiierte Spoofing in der Tat eine Ursache für das Verhalten der anderen Algorithmen war und letztlich den Crash auslöste. Es konnte jedoch nicht eindeutig bestimmt werden, an welcher Stelle innerhalb der algorithmischen Operationen die unrechtmäßige Handlung stattgefunden hatte, die zu den katastrophalen Folgen führte. Die Algorithmen hatten zwar den programmierten Anweisungen gemäß gehandelt und die Transaktionen entsprechend ausgeführt. Der immense Schaden schien jedoch vor allem dadurch begründet, dass bei diesem algorithmischen Hochfrequenzhandel Algorithmen eingesetzt werden, die sehr ähnlich

programmiert sind und damit ihre Handlungen wechselseitig verstärken (Yadav 2016, S. 1080). Dies deutet auf ein Herdenverhalten hin, in dem sich individuell vorprogrammierte Entscheidungen in einem interdependenten Prozess akkumulieren und eben dadurch die katastrophalen Folgen auslösen. In solchen Fällen, in denen Algorithmen individuell korrekt handeln, aber aufgrund ihrer gegenseitigen Verstärkung Schäden verursachen, versagt das traditionelle Haftungsrecht völlig. Dies ist die dritte Verantwortungslücke.

---

## **2 Maschinenverhalten, sozio-digitale Institutionen, Haftungsrecht**

### **2.1 Maschinenverhalten und sozio-digitale Institutionen**

#### **2.1.1 Wider den techniddeterministischen Kurzschluss**

Um diese gravierenden Haftungslücken zu bekämpfen, wurden in der äußerst lebhaften Reformdiskussion bereits zahlreiche Vorschläge zur Verantwortung für Algorithmen und entsprechende Haftungsregelungen unterbreitet: Sie reichen von der Haftung der Betreiber für gefährliche Anlagen über eine digitale Assistentenhaftung bis zur Haftung der Algorithmen selbst. Viele Autoren machen bei ihren Vorschlägen allerdings einen interdisziplinären Kurzschluss. Viel zu direkt schließen sie von den technischen Eigenschaften von Computern auf deren Risiken und auf rechtliche Haftungsregeln. ‚Technologie determiniert rechtliche Haftung‘ – mit einem solchen Argument bleiben sie zu einfachen Kausalmodellen und ebenso zu einfachen normativen Schlussfolgerungen verhaftet (z. B. Konertz und Schönhof 2020). Wir schlagen demgegenüber ein komplexeres Modell vor, das die unterschiedlichen sozialen Kontexte des Computereinsatzes ins Zentrum haftungsrechtlicher Argumente rückt und die Rolle der Sozialwissenschaften zur Bestimmung eines adäquaten Haftungsregimes betont.

Unser Ausgangspunkt ist eine in der IT-Forschung entwickelte Typologie, die drei Arten des Maschinenverhaltens unterscheidet: individuell, kollektiv und hybrid (Rahwan et al. 2019).

Diese Typologie dürfte für das Haftungsrecht von hoher Relevanz sein. Um aber den techniddeterministischen Kurzschluss zu vermeiden, ist es notwendig, ‚sozio-digitale Institutionen‘ als intervenierende Variable zwischen Technik und Recht einzuführen. Unter sozio-digitalen Institutionen verstehen wir stabilisierte Komplexe sozialer Erwartungen, in unserem Falle Erwartungen in Bezug auf das Verhalten der Algorithmen in sozialen Kontexten. Solche Institutionen sind



auf unterschiedliche binäre Codes unterscheiden. In solchen Brückeninstitutionen treffen Erwartungen rechtlicher, wirtschaftlicher, politischer und technologischer Art aufeinander, und es ist häufig schwierig, zwischen den Erwartungen der beteiligten Systeme zu unterscheiden (Esposito 2022, S. 33 ff.)

Sozio-digitale Institutionen verdichten die noch unbestimmten ‚affordances‘ digitaler Technologien, also die kontingenten Nutzungschancen der Technologie, durch ihre gesellschaftliche Spezifikation zu konkreten Erwartungsbündeln (Graber 2021, 14). Sozio-digitale Institutionen integrieren diverse technische und soziale Erwartungen zu Chancen und Risiken des Einsatzes von Algorithmen in einem Prozess der Ko-Produktion.<sup>3</sup> Diese Institutionen dienen somit als wirksame strukturelle Kopplungen zwischen technischen und sozialen Systemen, darunter auch dem Rechtssystem. Gegenüber traditionellen sozialen Institutionen

„zeichnen sich die instituierten und informellen Ordnungen der informationstechnologischen Kultur gerade dadurch aus, dass sie technisch werden. Die Codes und Programme, die technischen Standards und die mathematischen Algorithmen übernehmen jetzt einen Teil der Ordnungsleistung, den bislang ausschließlich symbolisch sinnhafte Ordnungen tragen mussten.“ (Vesting 2021, S. 220.)

Die Neuheit sozio-digitaler Institutionen gegenüber den bisher bekannten sozialen Institutionen besteht darin, dass sie auf der einen Seite die gewohnten an Bewusstsein gekoppelten Erwartungen bilden, auf der anderen Seite aber computergenerierte Erwartungen, die an rigide mathematische Kalkulationen gekoppelt sind. Diese asymmetrische ‚techno-digitale Normativität‘ schafft durchaus neue Gefährdungspotentiale (Karavas 2009, S. 478 ff.).

Erst die Einbettung in unterschiedliche sozio-digitale Institutionen erklärt die durchaus verwirrende Tatsache, dass Algorithmen manchmal als bloße Objekte oder Werkzeuge, manchmal als komplexe Personkonstrukte, manchmal als Mitglieder seltsamer Hybride, manchmal schließlich als gänzlich de-personalisierte Prozesse erscheinen. Zurecht macht Reyes 2021, 1483 ff. den jeweiligen sozio-technischen Kontext dafür verantwortlich, ob Algorithmen als ‚tools‘, als ‚conduits‘ oder als ‚autonomous entities‘ behandelt werden. Es gibt nicht die eine richtige Lösung für ihre Statuszuschreibung. Vielmehr hängen die unterschiedlichen Zuschreibungen von der Eigenrationalität und der Eigennormativität der jeweiligen sozio-digitalen Institutionen ab. Diese entscheiden darüber, wie unter den ‚affordances‘ der Technologie im jeweiligen sozialen Kontext gewählt wird und welcher Status dafür den Algorithmen zugewiesen wird.

---

<sup>3</sup> Zur Ko-Produktion unterschiedlicher Sozialsysteme Feenberg (2017), 75; Jasanoff (2004); Winner (1980).



### 2.1.2 Die konstitutive Rolle der Sozialwissenschaften

Damit plädieren wir für einen ‚institutional turn‘ im Recht der Digitalität.<sup>4</sup> Erst im Kontext eigensinniger sozio-digitaler Institutionen wird vorstrukturiert, welche Formen sozialer Verantwortung und rechtlicher Haftung für das Handeln autonomer Algorithmen angemessen sind (Balkin 2015, S. 49; ähnlich Linardatos 2019.). Und genau für diese Zusammenhänge werden die Sozialwissenschaften als Vermittler zwischen IT-Wissenschaften und Rechtswissenschaft gebraucht. Der Einsatz sozialwissenschaftlicher Methoden „enables the identification of the common socio-legal needs that drive the personhood debates in the context of AI systems and demonstrates the importance of the socio-technical context for building appropriate legal rules related to AI“ (Reyes 2021, S. 1475). Diese Methoden sind in der Lage, die konkreten sozio-digitalen Institutionen und ihre Risiken mit ausreichender Tiefenschärfe zu analysieren und mit ausreichender Dichte zu interpretieren. In einer solchen ‚institutional analysis‘ (Selznick 1969) ist die Rolle der Sozialwissenschaften keineswegs darauf beschränkt, wie es Juristen häufig missverstehen, nur deskriptiv-empirisch bestehende soziale Normen zu erforschen. Vielmehr besteht ihr zentraler Beitrag gerade darin, die verschiedenen Sinndimensionen sozio-digitaler Institutionen eigenständig zu interpretieren und sie mit ‚institutional imagination‘ (Unger 1996) in interpretativen, funktionalen oder normativen Analysen weiterzuentwickeln.

Sozio-digitale Institutionen korrelieren in ihren Unterschieden mit den drei oben genannten Arten von Maschinenverhalten: Individuelles Maschinenverhalten wird in der Institution der digitalen Assistenz genutzt. Hybrides Maschinenverhalten lässt in der Interaktion zwischen Mensch und Maschine die Institution einer Mensch-Maschine-Assoziation entstehen. Kollektives Maschinenverhalten hingegen findet in der technischen Welt statt, die nur indirekt mit der gesellschaftlichen Sphäre verbunden ist. Hier ist die Gesellschaft den unsichtbaren Maschinen und deren vernetzten Operationen ausgesetzt. Jeder dieser Kontakte schafft also eine je eigene sozio-digitale Institution.

## 2.2 Sozio-digitale Institutionen und Haftungsrecht

Für jede sozio-digitale Institution schlagen wir ein spezifisches Haftungsregime vor. Damit tragen wir der Einbettung des Rechts in konkrete gesellschaftliche

---

<sup>4</sup> Dies folgt der Forderung nach einer institutionellen Wende in der Vertragsauslegung, die insbesondere für entstehende Institutionen in der digitalen Sphäre relevant wird, Wielsch (2018). Zum ‚institutional turn‘ in den Sozialwissenschaften Roberto Esposito (2022), 33 ff.

Institutionen Rechnung. So betont der politische Philosoph Roberto Esposito 2022, (S. 79 f.), dass

„... das Recht nur innerhalb der konkreten Ordnung seinen Entfaltungsraum finden kann. Nicht auf einen abstrakten Normenkomplex reduzierbar, ist es materiell in dem sozialen Gefüge verwurzelt, aus welchem es hervorgeht. Vielmehr als eine Ordnung zu konstituieren, bringt das Recht, indem es menschliche Beziehungen regelt, ein Geflecht von Relationen zum Ausdruck, das im sozialen Körper bereits vorhanden ist.“

Freilich entstehen die sozio-digitalen Institutionen der Online-Welt nicht einfach ex nihilo. Sie alle haben ihren Ursprung in seit Langem bestehenden sozialen Institutionen der Offline-Welt. Gleichzeitig aber werden diese etablierten Institutionen durch die ‚Invasion‘ der Algorithmen grundlegend verändert. Jede sozio-digitale Institution weist eigene neuartige Schädigungsrisiken auf: Das Risiko *digitaler Assistenz* besteht darin, dass Aufgaben statt an Menschen an autonome Algorithmen delegiert werden, deren Entscheidungen nicht mehr kontrollierbar sind. *Mensch-Maschine-Assoziationen* erzeugen das Risiko emergenter Kollektiventscheidungen, die nicht auf Individualentscheidungen der beteiligten Algorithmen oder Menschen zurückführbar sind. Das Risiko *digitaler Interkonnektivität* besteht darin, dass gesellschaftliche Kommunikation einer undurchschaubaren Vernetzung von interagierenden Algorithmen ausgesetzt ist (zu Einzelheiten Beckers und Teubner 2021, S. 14 ff., 45 ff., 90 ff., 111 ff.).

Als vermittelnde Disziplin zwischen Informatik und Rechtswissenschaft dürften die Sozialwissenschaften über die analytischen Instrumente verfügen, die ko-evolutionäre Dynamik zwischen technologischen Innovationen, sozio-digitalen Institutionalisierungen und haftungsrechtlichen Entwicklungen zu analysieren (Nassehi 2019, S. 15 ff.). In der Wahl zwischen relevanten sozialwissenschaftlichen Theoremen kann sich das Haftungsrecht dabei nicht, wie es jedoch vielen Autoren vorschwebt, ausschließlich auf ökonomische Analysen stützen. Diese konzentrieren sich auf Anreize für Sorgfaltsstandards und Aktivitätsniveaus, sind aber gegenüber umfassenderen gesellschaftlichen Problemen, insbesondere den Interessen der Opfer an einer Entschädigung oder ökologischen Schädigungen, relativ gleichgültig. Gemäß dem in der Philosophie und Soziologie entwickelten Prinzip der ‚Transversalität‘ (Welsch 1996; Wagner 1996) versuchen wir deshalb, relevante Erkenntnisse auch aus anderen Sozialwissenschaften zu gewinnen. Hierbei rekurrieren wir insbesondere auf Theorien zur gesellschaftlichen Personifizierung von Algorithmen, zu emergenten Eigenschaften von

Mensch-Algorithmus-Assoziationen und zur verteilten Kognition in miteinander vernetzten Algorithmen. Letztlich orientieren sich unsere Vorschläge zur Risikohaftung an einer Kombination ökonomischer und soziologischer Theoreme.

Weil die involvierten sozio-digitalen Institutionen ganz unterschiedlich sind, wird man von vornherein auf abstrakt-generelle Haftungsregeln verzichten müssen (Krafft et al. 2022, unter 1; Lohsse et al. 2019, S. 21 ff.). Wir schlagen stattdessen konkretere unterschiedliche Haftungsregimes vor, die der unterschiedlichen institutionellen Einbettung des algorithmischen Verhaltens Rechnung tragen. Die Haftungsregimes reagieren auf die je verschiedenen Risiken der drei sozio-digitalen Institutionen: auf das bisher unbekannte Risiko einer Delegation von Aufgaben an autonome Algorithmen in digitaler Assistenz, auf das Risiko einer neuartigen Kollektivität von hybriden Mensch-Algorithmus-Assoziationen und auf das ebenso bisher unbekannte Risiko, dass die Gesellschaft der undurchsichtigen Vernetzung von Algorithmen in digitaler Interkonnektivität ausgesetzt sind.

### 2.2.1 Digitale Assistenz

Individuelles Maschinenverhalten bezieht sich auf intrinsische Eigenschaften eines einzelnen Algorithmus, dessen Dynamik durch Quellcode oder Design in seiner Interaktion mit der Umwelt bestimmt wird (Rahwan et al. 2019, S. 481). Diese technischen Eigenschaften allein können jedoch nicht den Ausschlag geben, ob Algorithmen als autonome Akteure qualifiziert werden können oder nicht. Stattdessen bestimmen sozio-digitale Institutionen, ob Algorithmen den sozialen Status von bloßen Instrumenten einnehmen, ob sie Agenten in Prinzipal-Agent-Beziehungen sind oder ob sie – als eine mögliche zukünftige Entwicklung – zu unabhängigen sozio-ökonomischen Akteuren werden, die ihre Eigeninteressen verfolgen („E-persons“).

Für Prinzipal-Agent-Beziehungen können sozialwissenschaftliche Theorien klären, unter welchen Bedingungen die Institution der digitalen Assistenz entsteht. Wenn die Aufgabendelegation von einem menschlichen Akteur (oder einer Organisation) an einen Algorithmus zwei autonome, aber interdependente Handlungsketten erzeugt, dann entsteht zwischen diesen eine Prinzipal-Agent-Beziehung (z. B. Krafft, Zweig und König 2022, unter 1). Solche Beziehungen setzen notwendigerweise sowohl für den Prinzipal als auch für den Agenten soziale Handlungsfähigkeit voraus. Daher wird innerhalb der sozio-digitalen Institution eine selektive Zuschreibung von Personalität an bestimmte digitale Prozesse erforderlich.

Personifizierung von Algorithmen – für diesen komplexen sozialen Prozess liefern mehrere soziale Theorien die entsprechende Begründung. Die Wirtschaftswissenschaften tragen zu diesem Thema relativ wenig bei. Wenn sie den Einsatz von Algorithmen in Märkten beobachten, setzen sie implizit Algorithmen als rationale Akteure voraus (z.B. Gabaix und Laibson 2000, S. 433 ff. Im Gegensatz zu engen Rational-Choice-Annahmen begreift die soziologische Theorie die Personifizierung als einen performativen Akt, der die soziale Realität eines Akteurs überhaupt erst begründet, der aber nicht auf eine spezifische Rationalität festgelegt werden kann (Muhle 2018; Beckers und Teubner 2021). Komplementär dazu definiert die Akteur-Netzwerk-Theorie die interaktiven Eigenschaften, die einen Algorithmus zu einem ‚Aktanten‘ machen. Sie zeigt außerdem auf, in welchen Aspekten sich dieser von einem menschlichen Akteur unterscheidet (Latour 2004, S. 62 ff.). Die Informationsphilosophie legt fest, unter welchen Bedingungen algorithmische Handlungen als autonom gelten können (Floridi und Sanders 2011, S. 192 ff.). Die Systemtheorie beschreibt detailliert, wie in einer Situation doppelter Kontingenz die emergente Mensch-Maschine-Kommunikation die soziale Identität des Algorithmus und seine Handlungsfähigkeit überhaupt erst konstituiert (Elena Esposito 2017). Dabei schafft jeder soziale Kontext für Algorithmen seine eigenen Kriterien der Personalität, die Wirtschaft nicht anders als die Politik, die Wissenschaft, die Moral oder das Recht. Konkrete Operationen in den verschiedenen sozialen Systeme schreiben den Algorithmen als ihren ‚Personen‘ auf unterschiedliche Weise Handlungen, Rechte und Pflichten zu und statten sie mit spezifischen Ressourcen, Interessen, Absichten, Zielen oder Präferenzen aus. Für eine solche Zuschreibung von Personalität ist kein voller Konsens der beteiligten Akteure erforderlich. Denn wie schon gesagt, beruhen Institutionalisierungen auf Konsensunterstellungen (Luhmann 1987 [1972], S. 64 ff.), die selbst bei hoher Meinungspluralität wirksam sind. In der politischen Philosophie wird detailliert beschrieben, wie die konkrete Übertragung der ‚potestas vicaria‘ die Stellvertreter-Personalität von Algorithmen konstituiert, die zugleich „eindeutige Risiken und Gefahren für die Modernität impliziert“ (Trüstedt 2020, S. 196 f.; Trüstedt 2022, V.1).

Ergebnis sozialer Personifizierung ist, dass nun technische Risiken in soziale Risiken transformiert werden. Die Risiken stammen nicht mehr einfach von fehlgeschlagenen Kausalverläufen, sondern von fehlerhaften Entscheidungen der Quasi-Akteure, die Erwartungen über die Handlungen digitaler Aktanten enttäuschen. In Situationen der digitalen Assistenz ist nicht mehr eine instrumentelle Subjekt-Objekt-Beziehung wirksam, sondern eine Subjekt-Subjekt-Beziehung, genauer eine Prinzipal-Agent-Beziehung mit ihren typischen kommunikativen Risiken. Je mehr sich die Institution digitaler Assistenz in den

Online-Transaktionen stabilisiert, desto stärker wird das Recht herausgefordert, seinerseits nach eigenen Kriterien zu entscheiden, welchen Grad an Rechtspersönlichkeit es den digitalen Akteuren zubilligt. Deswegen sind Haftungsregeln erforderlich, die auf Handlungsrisiken digitaler Akteure und nicht bloß auf die Kausalrisiken gefährlicher Objekte reagieren. Genau aus diesem Grund halten wir die rechtspolitischen Vorschläge, eine neue Gefährdungshaftung einzuführen oder die Regeln zur Produkthaftung zu modifizieren, welche die Algorithmen schlicht als Objekte, als gefährliche Anlagen oder als fehlerhafte Produkte behandeln, für nicht adäquat. Sie ignorieren gerade das Neuartige an Algorithmen – die in der digitalen Assistenz zugeschriebene autonome Entscheidungsfähigkeit. Auf (fehlerhafte) Entscheidungen von Algorithmen im Kontext digitaler Assistenz sind vielmehr die Regeln der Prinzipal-Agent-Haftung anzuwenden. Der Prinzipal ist gebunden, wenn der Algorithmus als Stellvertreter Verträge abschließt, und er haftet, wenn der Algorithmus als Gehilfe eine rechtswidrige Entscheidung trifft und einen Schaden verursacht.

### **2.2.2 Digitale Hybridität**

Im Vergleich zur digitalen Assistenz können die Sozialwissenschaften zum Verständnis der hybriden Mensch-Maschine-Assoziationen, d. h. der eng verflochtenen Interaktionen von Algorithmen und Menschen, auf andere Weise beitragen. Würde man hier mit dem individualistischen Ansatz der Prinzipal-Agent-Beziehung versuchen, den Verantwortungsbeitrag einzelner menschlicher oder algorithmischer Handlungen zu identifizieren, so übersähe man, dass sich hier ein neues Kollektiv etabliert hat – die Mensch-Maschine-Assoziation. Sie entwickelt Kollektivrisiken, die sich qualitativ von den Individualrisiken der digitalen Assistenz unterscheiden. Während die digitale Assistenz mit den Risiken einer neuartigen algorithmischen Autonomie zu kämpfen hat, erzeugt die digitale Hybridität Risiken, die mit den emergenten Eigenschaften kollektiven Verhaltens zusammenhängen. Hier spielen die Sozialwissenschaften ihre Vermittlerrolle zwischen Informatik und Haftungsrechtsdogmatik, indem sie die Bedingungen bestimmen, unter denen sich in sozialen Praktiken Mensch-Maschine-Assoziationen konstituieren.

Weil sie sich bedingungslos dem methodologischen Individualismus verpflichtet fühlen, bleiben ökonomische Autoren allerdings skeptisch gegenüber dem Realitätsstatus von Kollektivakteuren. Sie begreifen diese als bloße ‚nexus of contracts‘ und sehen ihre Personifizierung bestenfalls als eine Abkürzung für komplexe inter-individuelle Beziehungen und schlimmstenfalls als gefährliche ‚errors‘, ‚traps‘ oder ‚fictions‘ (Jensen und Meckling 1976; polemisch zugespitzt Easterbrook und Fischel 1989). In der Welt der Ökonomie zählt nur

das Verhalten von Individuen. Im Gegensatz dazu konzentrieren sich soziologische Analysen gerade auf die unterschiedlichen Sozialformationen, die in den Kontakten zwischen Mensch und Algorithmus entstehen (z. B. Nassehi 2019, S. 224; Hepp 2020; Latour 2007, S. 273 ff.). Diese reichen von kurzfristigen und losen Individual-Interaktionen über Netzwerke bis hin zu kollektiven Mensch-Algorithmus-Assoziationen, die mit interner Arbeitsteilung, Ressourcenausstattung und Kompetenzverteilung ausgestattet sind.<sup>5</sup> In solchen Konstellationen verdichteter Interaktion, in denen die Einzelhandlungen dem gesamten hybriden Gebilde, nicht aber mehr den einzelnen beteiligten Algorithmen oder Menschen zugerechnet werden (z. B. Pettit 2007), muss das Haftungsrecht auf das Risiko der neuartigen kollektiven Handlungsfähigkeit der Hybride reagieren. Hier sind Normen einer Kollektivhaftung notwendig, die freilich unterhalb der Schwelle der Haftung einer vollwertigen juristischen Person angesiedelt sind.

### 2.2.3 Digitale Interkonnektivität

Im Gegensatz zu den beiden bisher diskutierten sozio-digitalen Institutionen ist das kollektive Maschinenverhalten ein rein digitales Verhalten, das von der Interkonnektivität autonomer Algorithmen erzeugt wird. Ihr Risiko besteht darin, dass anders als bei digitaler Assistenz kein einzelner handelnder Akteur identifiziert werden kann. Ebenso wenig lässt sich das Risiko der Interkonnektivität mit den Risiken der hybriden Mensch-Maschine-Assoziationen vergleichen. Hier geht es um miteinander vernetzte algorithmische Operationen, denen die gesellschaftliche Kommunikation letztlich ausgesetzt ist, ohne sie zu verstehen. Beim kollektiven Maschinenverhalten findet keine wechselseitige Kommunikation zwischen Menschen und Algorithmen statt, nur eine indirekte strukturelle Kopplung. Die algorithmischen Kalkulationen sind interdependent und lassen sich als ein ‚restless collective‘ qualifizieren, das mit Formen distributiver Kognition arbeitet (Wiedemann 2014, S. 313). Die kollektive Handlung wird dabei durch das interdependente Handeln einzelner Akteure hervorgerufen (Rauer 2017, S. 440 ff. mit Bezug auf Hutchins 1995). Eine solche ‚Kollektivität ohne Kollektiv‘ kann nicht als planmäßig organisiertes Sozialgebilde beschrieben werden, weder als Organisation noch als Netzwerk. Es ist ein zufällig entstandener ‚Schwarm‘ von Algorithmen. Das Verhältnis der Gesellschaft zu solchen algorithmischen Schwärmen beschreibt die Systemtheorie als Sozialkontakt zu ‚unsichtbaren Maschinen‘ (Luhmann 1997, S. 117 ff.; ähnlich Hildebrandt 2015, S. 40). Ihr Einfluss auf

---

<sup>5</sup> Die ökonomische Transaktionskostentheorie von Williamson (1996) arbeitet zwar ähnlich mit einer gleitenden Skala von ‚spot-contracts‘ bis hin zu ‚formal organizations‘, verfehlt aber gerade deren korporative und kollektive Elemente.

die Gesellschaft ist nur schwer zu fassen. Weder gibt es hier genuine Kommunikation zwischen Menschen und einzelnen Algorithmen, noch bildet sich ein kommunikationsfähiges Kollektiv aus Menschen und Algorithmen heraus. Statt eines direkten über Kommunikation vermittelten Einflusses üben vernetzte Algorithmen einen nur indirekten über strukturelle Kopplung vermittelten Einfluss auf die sozialen Beziehungen aus. Deswegen dürfte auch eine rechtliche Haftung für die Entscheidungen einzelner Algorithmen ausgeschlossen sein. Wir plädieren stattdessen für Fondslösungen, die politische Entscheidungen von Regulierungsbehörden erfordern, um den je beteiligten Industriesektoren die Verantwortung aufzuerlegen.

## **2.3 Haftungssubjekte**

Die Unterschiede zwischen sozio-digitalen Institutionen und ihren Risiken wirken sich auch auf die Frage aus, welche Akteure oder Organisationen für algorithmische Fehlentscheidungen verantwortlich sind. Auch hier schlagen wir eine differenzierte Behandlung je nach institutionellem Kontext vor. Um die Unterschiede vorab kurz zu kennzeichnen: Auf Entscheidungen einzelner Algorithmen findet die Prinzipal-Agent-Haftung Anwendung, welche die Verantwortung für algorithmisches Fehlverhalten ausschließlich dem menschlichen Nutzer (oder der Organisation) als dem Prinzipal auferlegt, weil er das Risiko der Delegation gesetzt hat. Dagegen haftet für die kollektiven Handlungen der Mensch-Maschine-Hybride das ganze Netzwerk der beteiligten Akteure, also Hersteller, Programmierer, Händler und Nutzer. Interkonnektivitätsschäden schließlich müssen von kollektiven Fonds ersetzt werden, die von der gesamten betroffenen Branche einzurichten sind.

### **2.3.1 Nutzer/Betreiber**

Die digitale Assistenz erzeugt Verantwortlichkeiten nur innerhalb der bilateralen Beziehung zwischen Algorithmus und dem menschlichen Nutzer/Betreiber (oder der Organisation). Eine Prinzipal-Agent-Beziehung erfordert, dass der Agent die Ziele des Prinzipals verwirklicht und der Prinzipal für die von ihm veranlassten Handlungen des Agenten einsteht. Die Prinzipal-Agent-Haftung zielt gerade nicht auf die Vielzahl der am Computereinsatz beteiligten Akteure – Programmierer, Hersteller, Händler, Betreiber –, sondern ausschließlich auf den Nutzer, der eine Aufgabe an die digitale Technik delegiert und damit das Risiko autonomer Entscheidungen des Algorithmus übernimmt. Für dieses Delegationsrisiko ist daher nur der menschliche Nutzer/Betreiber (oder die Organisation) verantwortlich.

Demgegenüber argumentieren einige Autoren, dass dies die Risiken in ungerechter Weise allein auf den Nutzer/Betreiber verlagert. Sie sehen auch andere Akteure in der Rolle des verantwortlichen Prinzipals, insbesondere den Hersteller, einschließlich des Backend-Betreibers, der im Hintergrund Programm-Updates und ähnliche Leistungen erbringt (Sanz Bayón 2019, unter 7; Europäisches Parlament 2020, Abs. 8). Damit ignorieren sie jedoch die spezifischen Risiken der Aufgabendelegation und gelangen ihrerseits zu einer ungerechten Risikoverteilung zwischen Hersteller und Nutzer. Denn diese setzen ganz unterschiedliche Risiken. Die Risiken müssen jeweils genau definiert und dann ausschließlich denjenigen Akteuren zugewiesen werden, die sie tatsächlich eingegangen sind. Die Prinzipal-Agent-Haftung reagiert auf die Risiken der Arbeitsteilung zwischen Nutzer und Algorithmus. Im Gegensatz dazu reagiert die weiterhin anwendbare Produkthaftung auf die besonderen Risiken der Programmierung, Herstellung und Überwachung der Algorithmen.

### 2.3.2 Netzwerkmitglieder

Im Gegensatz zu dieser ausschließlich den Nutzer/Betreiber treffenden Prinzipal-Agent-Haftung werden im Fall digitaler Hybridität rechtswidrige Handlungen nicht einzelnen Akteuren, sondern der Mensch-Maschine-Assoziation selbst zugerechnet. Dann aber wird die daraus resultierende finanzielle Haftung auf eine Vielzahl von Akteuren kanalisiert, die ‚hinter‘ dem digitalen Hybrid stehen. Ein ganzes Netz verschiedener Akteure ist an der Mensch-Maschine-Assoziation beteiligt und profitiert davon. Da die Kontrolle im Netz über die Netzwerkknoten verstreut ist, folgt auch die Verantwortung dieser spezifischen Risikostruktur. Für Mensch-Maschine-Assoziationen lässt sich eine ausgereifte Unternehmenshaftung, die die Assoziation als juristisch handelnde und haftbare Person versteht, (zumindest vorerst) nicht etablieren. Hierzu fehlen dem Recht bisher die dogmatischen Strukturen. Daher halten wir stattdessen die Prinzipien einer ‚Netzwerkhaftung‘ für gut geeignet, um die Verantwortung der Beteiligten für die Handlungen der digitalen Hybride zu konstituieren (Vladeck 2014, S. 149; Allain 2013, S. 1074). Die von uns vorgeschlagene digitale Netzwerkhaftung ist am haftungskanalisierenden Modell der amerikanischen ‚enterprise liability‘ und an Elementen der deutschen Gesamthandshaftung orientiert. Sie funktioniert in zwei Schritten: erst Handlungszurechnung, dann Haftungszurechnung. Im ersten Schritt wird, ohne dass die Beiträge von Menschen und Algorithmen im Einzelnen identifiziert werden müssen, das Fehlverhalten dem Hybriden als kollektive Handlung zugerechnet. Im zweiten Schritt wird die Haftung für die kollektive Handlung auf die Netzmitglieder kanalisiert. Diese Mitglieder haben das hybride



Netzwerk aufgebaut und kontrollieren es, wenn auch nur indirekt. Sie profitieren von dessen Aktivitäten. Als Folge dieser Zurechnung haften alle Netzknotten entsprechend ihrem Anteil, wobei der Anteil anhand des jeweiligen ökonomischen Nutzens und der Kontrolle über das Hybrid bestimmt wird. In Anlehnung an die bekannte Marktanteilshaftung (market share liability) schlagen wir eine ‚Netzanteilshaftung‘ vor (allgemein dazu: Teubner 2004, S. 246 ff., S. 248.). Einzig in dem Fall, dass ein Unternehmen das Netzwerk zentral durch vertragliche Verabredungen koordiniert, sollte die primäre Haftung bei diesem liegen und entsprechend berechnet werden (Loo 2020, S. 189). In der Regel wird dies der Produzent sein, der dann bei den anderen Netzwerkknoten Rückgriff nehmen kann.

### 2.3.3 Industriesektor

Im Falle der Interkonnektivität ist der Kreis der Haftungssubjekte wieder anders zu ziehen. Die Verantwortung verlagert sich auf ein größeres soziales Kollektiv. Die algorithmische Vernetzung der ‚unsichtbaren Maschinen‘ macht es von vornherein unmöglich, einen individuell verantwortlichen Akteur zu bestimmen. Angesichts nur indirekter ‚struktureller Kopplung‘ zwischen algorithmischer Interkonnektivität und Gesellschaft kann auch keine Eins-zu-eins-Verantwortungsbeziehung hergestellt werden. Deshalb schlagen wir vor, dass für Schäden aus Interkonnektivität Haftungsfonds einzurichten sind, die von dem beteiligten Industriesektor finanziert werden (Erdélyi und Erdélyi 2020). Die Beiträge der jeweiligen Akteure werden aufgrund ihres Marktanteils und ihrer spezifischen Problemlösungskapazität berechnet. Ein bewährtes Vorbild ist der US-amerikanische Superfund für Umweltschäden. Der Superfund ist nicht nur auf die Entschädigung einzelner Betroffener ausgerichtet, sondern zugleich auf die Kompensation der umfassenderen sozialen und ökologischen Schadensfolgen. Dies sollte auch für Vernetzungsschäden aufgegriffen werden. Dazu müssen Restitutionsmaßnahmen als zusätzliche Instrumente des Haftungsrechts entwickelt werden. Bei Schäden großen Ausmaßes sollte die für den Fonds zuständige Behörde ermächtigt werden, die Akteure auszuwählen, die über eine starke Problemlösungskapazität verfügen, und ihnen die Aufgabe der Restitution und Folgenbeseitigung aufzuerlegen. Die beteiligten Akteure sollten zu Maßnahmen verpflichtet werden, welche die negativen externen Effekte der Interkonnektivität für die Zukunft begrenzen oder sogar ausschalten.

## 2.4 Rechtsstatus der Algorithmen

Schließlich erfordern die neuen sozio-digitalen Institutionen – digitale Assistenz, Hybridität, Interkonnektivität –, auch den Rechtsstatus der Algorithmen differenziert zu betrachten. Es kann nicht um eine vollständige rechtliche Personifizierung von Software-Agenten, Mensch-Computer-Assoziationen oder Multi-Agenten-Systemen gehen. Stattdessen schlagen wir vor, als Antwort auf die drei Risiken den rechtlichen Status von Algorithmen darauf auszurichten, welche Rolle sie jeweils innerhalb ihrer sozio-digitalen Institution spielen.

### 2.4.1 Teilrechtsfähigkeit

Nur im Falle digitaler Assistenz ist es angemessen, den Algorithmen den Status einer begrenzten Rechtspersönlichkeit einzuräumen. Diese Teilrechtsfähigkeit ermöglicht es ihnen, als Stellvertreter mit bindender Wirkung für ihre Prinzipale Verträge mit Dritten zu schließen. Als Erfüllungs- oder Verrichtungsgehilfen sind Algorithmen mit der erforderlichen rechtlichen Handlungsfähigkeit auszustatten, sodass für ihr Fehlverhalten die Prinzipale haften.

### 2.4.2 Mitgliedschaft im Hybrid

Im Unterschied dazu besteht die angemessene Antwort auf das Assoziationsrisiko der digitalen Hybridität darin, dem beteiligten Software-Agenten den Rechtsstatus eines vollwertigen Mitglieds der Mensch-Maschine-Assoziation zu verleihen. Eine Maximallösung im zukünftigen Recht würde der hybriden Assoziation selbst Handlungen, Rechte und Pflichten sowie die finanzielle Haftung zurechnen – eine Lösung, die im Privatrecht völlig neue Wege der Rechtspersönlichkeit beschreiten würde. Eine Minimallösung im geltenden Recht würde bloß auf den Rechtsbegriff des Vertragszwecks rekurrieren und ihn für die Auslegung von Computererklärungen und die Bestimmung der Rechte und Pflichten der Teilnehmer nutzbar machen. Wir favorisieren demgegenüber eine mittlere Lösung, die eine Analogie zu den Grundsätzen einer Netzwerkhafung herstellt. Der Algorithmus erhält damit den Status eines Netzwerknotens.

### 2.4.3 Teil des Risikopools

Unsere Antwort auf die Risiken der digitalen Interkonnektivität heißt ‚Risikopool‘. Das Haftungsrecht muss dessen Grenzen verbindlich festlegen. Das rechtswidrige Verhalten wird dem Pool zugerechnet. Der Rechtsstatus der Algorithmen besteht dann weder in eigener Personalität noch in Mitgliedschaft in einem Hybrid, sondern darin, ein bloßes Teilelement eines digitalen Risikopools zu sein.

**Tab. 1** Kausale und normative Beziehungen zwischen Maschinenverhalten, sozio-digitalen Institutionen und Haftungsregelungen

Maschinenverhalten	Sozio-digitale Institution	Emergente Risiken (Neue Akteure)	Haftungsregime	Haftungssubjekte	Rechtsstatus der Algorithmen
individuell	Digitale Assistenz	Digitale Autonomie (Aktanten)	Prinzipal-Agenten-Haftung	Nutzer/Operator • Gefährdung	Beschränkte Rechtsfähigkeit
hybrid	Mensch-Maschine-Assoziation	Kollektivhandeln (Hybride)	Netzwerkhaftung	Netzwerkknoten • Kontrolle • Ökonomischer Nutzen	Mitgliedschaft im Hybrid
kollektiv	Kopplung an digitale Konnektivität	Distributierte Kognition (Schwärme)	Kollektive Fonds	Industriesektor • Marktanteil • Problemlösungs-kapazität	Teil des Risikopools

### 3 Drei Haftungsregimes

#### 3.1 Synopsis

Um die Unterschiede zwischen den drei Haftungsregimes, ihre digitalen und sozialen Voraussetzungen und ihre rechtlichen Normierungen zu verdeutlichen, stellen wir hier in tabellarischer Form eine komprimierte Fassung unserer Vorschläge vor. Die (rekursiven) Zusammenhänge von digitalen Maschinenverhalten, sozio-digitalen Institutionen und Haftungsnormen sehen folgendermaßen aus (Tab. 1):

#### 3.2 Haftungsrechtliche Normen

Zur Konkretisierung der Haftungsregimes schlagen wir die folgenden Regeln vor:

#### 3.3 Prinzipal-Agent-Haftung

*Die Prinzipal-Agent-Haftung für Fehlentscheidungen eines Algorithmus greift ein, wenn 1) ein menschlicher Prinzipal (oder eine Organisation) eine Aufgabe an einen*

*Algorithmus delegiert, 2) die Delegation die Entscheidungsfreiheit des Agenten erfordert, 3) dem Agenten eine fehlerhafte Handlung zuzurechnen ist, die für einen Programmierer weder vorhersehbar noch erklärbar ist, 4) die Handlung gegen eine vertragliche oder deliktische Sorgfaltspflicht verstößt und 5) die Kausalität zwischen Handlung und Schaden nachgewiesen werden kann.*

*6) Als Rechtsfolge haftet ausschließlich der Nutzer des Algorithmus als Prinzipal. 7) Der Schadensersatz beschränkt sich nicht auf die engen Entschädigungsgrundsätze der Gefährdungshaftung, sondern folgt den Schadensersatzregeln des Vertrags- und Deliktsrechts, wonach für über Körper- und Sachschaden hinausgehende Schäden Ersatz zu leisten ist.*

### **3.4 Netzwerkhaftung**

*Sofern die Voraussetzungen der Prinzipal-Agent-Haftung nicht vorliegen, greift eine Netzwerkhaftung ein, wenn 1) in der Zusammenarbeit zwischen Mensch und Maschine 2) eine rechtswidrige Entscheidung getroffen wurde und 3) ihre Tätigkeiten so eng miteinander verflochten sind, dass 4) die Fehlentscheidung weder dem Menschen noch dem Algorithmus zugerechnet werden kann und 5) ein Kausalzusammenhang zwischen den individuellen Handlungen und Schaden nicht hergestellt werden kann, jedoch 6) nachgewiesen werden kann, dass die kollektive Entscheidung den Schaden verursacht hat.*

*7) Als Rechtsfolge haften die Teilnehmer des Akteursnetzwerks, d. h. Produzenten, Programmierer, Händler und die menschlichen Teilnehmer innerhalb des Hybriden. 8) Die Netzwerkhaftung trifft primär den Hersteller als Zentrale des vernetzten Unternehmens. 9) Der Hersteller kann die anderen Teilnehmer entsprechend ihrem Anteil am Netzwerk in Regress nehmen. 10) Der Netzwerkanteil wird durch die Kriterien des wirtschaftlichen Nutzens und der Kontrolle innerhalb des Netzwerks bestimmt.*

### **3.5 Haftung des Fonds**

*Sofern weder Prinzipal-Agent-Haftung noch Netzwerkhaftung eingreifen, ist eine Entschädigung 1) nur dann möglich, wenn ein Fonds oder eine Versicherung für die algorithmischen Schäden eingerichtet worden ist. Voraussetzungen für eine Entschädigung sind, dass 2) die miteinander verbundenen algorithmischen Entscheidungen eine vertragliche oder deliktische Pflicht verletzen und 3) zusammen einen 4) Schaden verursachen.*

*5) Als Rechtsfolge haftet der Fonds für die Schäden. 6) Über den Schadensersatz entscheidet eine Behörde, die mit der Verwaltung des Fonds beauftragt wird. Die Behörde bestimmt zugleich 7) die den Fonds finanzierenden Akteure des Wirtschaftszweigs ex-ante nach ihrem jeweiligen Marktanteil und 8) die Akteure, die bei einer ex-post Haftung anhand ihrer Problemlösungskapazität herangezogen werden.*

---

## **4 Diskussion der Ausgangsfälle**

Welche Rechtsfolgen ergeben sich nach den vorgeschlagen Haftungsregeln für die drei in der Einleitung vorgestellten Fälle? Wir hatten zunächst festgestellt, dass nach geltendem Recht in allen Fällen eine Haftung ausgeschlossen ist und die Schäden damit allein vom Geschädigten zu tragen sind. Nach unseren Vorschlägen sind die Fälle je einer sozio-digitalen Institution zuzuordnen und den entsprechenden Haftungsregeln zu unterwerfen.

### **4.1 Robo-Advice: Digitale Aktanten**

Rechtlich stellt sich in diesem Fall zunächst die Frage, ob der Anlagevermittler Costa, der die Beratung und Betreuung von Aktientermingeschäften durch Algorithmen anbietet, für eigene Pflichtverletzungen verantwortlich ist. Wenn der Geschädigte Li jedoch nicht nachweisen kann, dass Costa die Pflichten eines Anlagevermittlers verletzt hat, wird der Erfolg seiner Klage davon abhängen, ob die Fehlentscheidungen des autonom handelnden Algorithmus die Haftung begründen können (Wagner und Luyken 2020; Hughes und Williamson 2019; Sanz Bayón 2019).

Der Stopp-Loss-Befehl des Algorithmus war eine rechtswidrige Entscheidung des Algorithmus. Die hier relevante sozio-digitale Institution ist die digitale Assistenz, bei der der Algorithmus als Agent in einer Prinzipal-Agent-Beziehung auftritt. In diesem Zusammenhang hilft die Produkthaftung nicht weiter, da Costa keine Herstellerpflichten verletzt hat. Eine verschuldensunabhängige Haftung, die erst vom Gesetzgeber angeordnet werden müsste, würde wiederum viel zu weit gehen. Sie würde der finanziellen Haftung Tür und Tor öffnen. Als reine Kausalhaftung würde sie den Auftraggeber für jede Handlung, ob rechtswidrig oder nicht, des algorithmischen Agenten haftbar machen, sofern diese einen finanziellen Schaden verursacht. Somit bleibt nur die Prinzipal-Agent-Haftung als Anspruchsgrundlage übrig (Strzelczyk 2018). Wenn der Prinzipal Costa einen Finanzmaklervertrag mit Li abgeschlossen und seine vertraglichen Pflichten der

Portfolioverwaltung an K1 als seinen Erfüllungsgehilfen delegiert hat, haftet er für alle Verstöße gegen die vertraglichen Pflichten, die K1 begangen hat. Nach geltendem Recht besitzt der Algorithmus jedoch die für die Prinzipal-Agent-Haftung zwingend erforderliche Rechtsfähigkeit (noch) nicht. Gerichte können jedoch autonomen Algorithmen Rechtssubjektivität zusprechen, wie sie es in der Vergangenheit bei bestimmten Assoziationen menschlicher Akteure getan haben. Für die Prinzipal-Agent-Haftung reicht es aus, Algorithmen mit einer Teilrechtsfähigkeit auszustatten, nämlich der Fähigkeit, die vertraglichen Pflichten des Auftraggebers zu erfüllen. Unter diesen Voraussetzungen kann die hier vorgeschlagene Prinzipal-Agent-Haftung als Klagegrund für die Haftung von Costa erfolgreich sein.

## 4.2 Panama-Papers: Digitale Hybride

Wenn der Algorithmus, der die Vielzahl von Dokumenten in Zusammenarbeit mit den Journalisten analysiert hat, seiner Programmierung entsprechend operiert hat, die menschlichen Journalisten ihren Kontrollpflichten nachgekommen sind und nichts von dem Resultat einer falschen Beschuldigung gewusst haben, kann niemand haftbar gemacht werden (Lewis et al. 2019, S. 69; Oster 2018, S. 29 ff.). Hier ist eine Situation der ‚kollektiven moralischen Verantwortung‘ gegeben, in der eine Gruppe eine unrechtmäßige Handlung begeht, obwohl sich die beteiligten Einzelpersonen korrekt verhalten haben (Copp 2007). Der Algorithmus arbeitete wie programmiert und traf Entscheidungen über die Kennzeichnung, Klassifizierung, Auswahl und Aufbereitung von Informationen für die Verwendung durch Menschen wie vorgesehen, und die menschlichen Journalisten überprüften diese Informationen mit der erforderlichen Sorgfalt auf der Grundlage ihres Wissens. Es ist schwierig, eine einzelne unrechtmäßige Handlung zu identifizieren, obwohl die kollektive Arbeit von Algorithmen und Journalisten zu den unrechtmäßigen Anschuldigungen geführt hat. Für diese Fälle ist eine Netzwerkhafung wie oben geschildert angemessen. Das Fehlverhalten ist der Mensch-Maschine-Assoziation als Kollektiv von Journalisten und Algorithmen zuzurechnen.

Sofern die unrechtmäßige Handlung dem Mensch-Maschine-Hybrid zugeschrieben werden kann, ist es möglich, die finanzielle Haftung auf die Mitglieder des Netzwerks zu kanalisieren. Der Geschädigte kann den zentralen Knotenpunkt des Netzes erfolgreich verklagen. Im Fall des hybriden Journalismus kann dies entweder die kontrollierende Nachrichtenorganisation oder der Hersteller des Algorithmus sein. Im Zusammenhang mit der algorithmischen Nachrichtenverbreitung könnte es das Unternehmen sein, das den Algorithmus herstellt, also

ein Nachrichten- oder Social-Media-Unternehmen. Eine solche Haftung würde unabhängig davon gelten, ob es spezielle gesetzliche Vorschriften für die verschuldensunabhängige Haftung von Nachrichten Anbietern oder Plattformen gibt. Innerhalb des Netzwerks würde die interne anteilige Verteilung der Haftung entsprechend dem wirtschaftlichen Nutzen und der Kontrolle im kollaborativen Netzwerk erfolgen.

### **4.3 Flash Crash: Digitale Schwärme**

Die von den vernetzten Algorithmen im Finanzmarkt verursachten Schäden sind ein Fall des Interkonktivitätsrisikos. Die schädigenden Operationen der Algorithmen waren hier zwar von außen ausgelöst, etwa durch betrügerisches Verhalten eines einzelnen Händlers, und begünstigt durch die Infrastruktur des Finanzmarktes selbst, auf dem Algorithmen gleichartiger Programmierung zugelassen sind. Die unmittelbaren Ursachen des Zusammenbruchs sind jedoch die interdependenten Operationen vernetzter Algorithmen. Ihre Geschwindigkeit übersteigt bei weitem menschliche Fähigkeiten. Ein menschliches Eingreifen in die schadensverursachenden Operationen war unmöglich. Vorhersehbarkeit und individuelles Verschulden, wie es für Fahrlässigkeit erforderlich wäre, können in einem solchen Kontext nicht verlangt werden.

Für solche Fälle schlagen wir anstelle der vergeblichen Suche nach einem verantwortlichen Akteur, wie sie die U.S.-Finanzverwaltung versucht hat, eine Fondslösung vor. Der Fonds wird unter die Aufsicht der Finanzmarktbehörden gestellt (ähnlich Yadav 2016, S. 1095). Er wird durch eine relativ geringe Marktzugangsgebühr finanziert, die von den Nutzern und Herstellern der Finanzmarktalgorithmen anhand ihres Marktanteils entrichtet wird.

Hier kommt nun das Kriterium der Rechtswidrigkeit ins Spiel. Nicht jeder Flash Crash sollte den Geschädigten den Zugang zum Fonds eröffnen. Ein durch bloße Volatilität der Finanzmärkte ausgelöster Schaden ist nicht per se ein Haftungsgrund. Der Zugang der Geschädigten zum Fondskapital nach einem Flash Crash ist auf diejenigen Fälle zu beschränken, in denen ein Rechtsverstoß durch vernetzte Algorithmen festgestellt werden kann. Anhand dieses Kriteriums lassen sich Schäden, die durch die normale Volatilität der Finanzmärkte verursacht werden, von Crashes unterscheiden, die auf einen Verstoß gegen die Verhaltensregeln am Finanzmarkt zurückzuführen sind.

Die Finanzmarktaufsichtsbehörde sollte zudem befugt sein, die am algorithmischen Hochfrequenzhandel beteiligten Akteure zu Wiederherstellungs- und

Präventionsmaßnahmen anzuweisen. In Bezug auf den algorithmischen Handel wäre dies beispielsweise ein ‚digital clean up‘. Die Behörde würde den beteiligten Firmen auferlegen, systemische Änderungen an der Infrastruktur des algorithmischen Handels vorzunehmen. So könnte das Risiko eines algorithmischen Schwarmverhaltens dadurch gemindert werden, dass eine Verlangsamung der algorithmischen Entscheidungen in das System einprogrammiert wird (Yadav 2016, S. 1097 ff.).

---

## 5 Ausblick

Diese drei Haftungsregimes liefern einen wesentlichen Beitrag zu einer sich herausbildenden Digitalverfassung (dazu Celeste 2020, S. 24 ff.). Sicherlich geht es auch darum, eine Optimierung von Sorgfaltsstandards und Aktivitätsniveaus, welche die algorithmischen Aktivitäten zu steuern versucht, anzustreben. Doch sind mit rechtsökonomischen Argumenten die konstitutionellen Ansprüche an Digitalität nicht erschöpft. Denn in der rekursiven Dynamik Technologie/ Sozialinstitutionen/ Haftungsrecht vollzieht sich eine wechselseitige Konstituierung von Rechtsnormen, Rechtssubjekten und sozialen Institutionen. Ja, das Auftreten von autonomen Algorithmen als neuartige Quasi-Subjekte stellt das Privatrecht vor eine Verfassungsfrage erster Ordnung, auf die es womöglich mit der rechtlichen Personifizierung von isoliert agierenden Algorithmen und von Mensch-Algorithmus-Assoziationen sowie der Konstitution von Risikopools samt deren haftungsrechtlichen Konsequenzen reagieren muss. Diese Entwicklung vollzieht sich wie gesagt weder nur technikdeterminiert noch ökonomiedeterminiert, sondern hängt zugleich von vielfältigen politischen Entscheidungen über die gesellschaftliche Verantwortung von digitalen Technologien ab. Sofern in dem nur scheinbar technischen Haftungsrecht über die Bekämpfung von Verantwortungsdefiziten, über Chancen und Gefahren der Personifizierung algorithmischer Prozesse, über die Sorgfaltsmaßstäbe für algorithmische Entscheidungen, über den Schutz von durch digitale Fehlentscheidungen gefährdeten individuellen Rechtspositionen, über das Ausmaß von Entschädigungsleistungen und über die notwendige Reprogrammierung von risikobehafteten Algorithmen gestritten und entschieden wird, leistet das Haftungsrecht einen nicht unbeträchtlichen Teilbeitrag zu einer sich gerade entwickelnden Digitalverfassung.



## Literatur

- Allain, Jessica. 2013. From Jeopardy! to Jaundice: The Medical Liability Implications of Dr. Watson and Other Artificial Intelligence Systems. *Louisiana Law Review* 73: 1049–1079.
- Balkin, Jack. 2015. The Path of Robotics Law. *California Law Review Circuit* 6: 45–60.
- Beardsworth, Thomas, und Kumar, Nishant. 2019. Going to Court Over Losses When Robot Is to Blame, *Insurance Journal* 7 May 2019.
- Beckers, Anna, und Teubner, Gunther. 2021. *Three Liability Regimes for Artificial Intelligence: Algorithmic Actants, Hybrids, Crowds* Oxford: Hart.
- Celeste, Edoardo. 2020. *Digital Constitutionalism: The Role of Internet Bills of Rights*. Dublin: University College Dublin.
- Copp, David. 2007. The Collective Moral Autonomy Thesis. *Journal of Social Philosophy* 38: 369–388.
- Diakopoulos, Nicholas. 2019. *Automating the News: How Algorithms are Rewriting the Media*. Cambridge/Mass.: Harvard University Press.
- Easterbrook, Frank H., und Fischel, Daniel. 1989. The Corporate Contract. *Columbia Law Review* 89: 1416–1448.
- Erdélyi, Olivia J., und Erdélyi, Gabor. 2020. The AI Liability Puzzle and a Fund-Based Work-Around. AIES '20: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, 50–56.
- Esposito, Elena. 2017. Artificial Communication? The Production of Contingency by Algorithms. *Zeitschrift für Soziologie* 46: 249–265.
- Esposito, Roberto. 2022. *Institution und Biopolitik*. Zürich: Diaphanes.
- Europäisches Parlament. 2020. Zivilrechtliche Haftungsregelung für künstliche Intelligenz. Entschließung vom 20. Oktober 2020 2020/2012(INL).
- Feenberg, Andrew. 2017. *Technosystem: The Social Life of Reason*. Cambridge, MA: Harvard University Press, 2017. Cambridge, MA: Harvard University Press.
- Floridi, Luciano, und Sanders, J.W. 2011. On the Morality of Artificial Agents, In *Machine Ethics*, hrsg. Michael Anderson und Susan L. Anderson, 184–212. Cambridge: Cambridge University Press.
- Gabaix, Xavier und David I. Laibson, „A Boundedly Rational Decision Algorithm“, *American Economic Review*, 90 (2000), 433–438.
- Graber, Christoph B.. 2021. How the Law Learns in the Digital Society. *Law, Technology and Humans* 3: 12–27.
- Hauriou, Maurice E. 1933. *Aux sources du droit: Le Pouvoir, l'ordre et la liberté*. Paris: Bloud & Gay.
- Hepp, Andreas. 2020. *Deep Mediatization: Key Ideas in Media & Cultural Studies*. London: Routledge.
- Hildebrandt, Mireille. 2015. *Smart Technologies and the End(s) of Law*. Cheltenham: Edward Elgar.
- Hughes, Benn, und Williamson, Russell. 2019. When AI Systems Cause Harm: The Application of Civil and Criminal Liability. *Digital Business Law – Blog*.
- Hutchins, Edwin. 1995. *Cognition in the Wild*. Boston: MIT Press.

- Jasanoff, Sheila. 2004. The Idiom of Co-Production', In *States of Knowledge: The Co-production of Science and the Social Order*, hrsg. Sheila Jasanoff, 1–12. London: Routledge.
- Jensen, Michael, und Meckling, William H. 1976. Theory of the Firm: Managerial Behavior, Agency Costs and Ownership Structure. *Journal of Financial Economics* 3: 306–360.
- Karavas, Vagios. 2009. The Force of Code: Law's Transformation under Information-Technological Conditions. *German Law Journal* 10: 463–481.
- Konertz, Roman, und Schönhof, Raoul. 2020. *Das technische Phänomen „Künstliche Intelligenz“ im allgemeinen Zivilrecht: Eine kritische Betrachtung im Lichte von Autonomie, Determinismus und Vorhersehbarkeit*. Baden-Baden: Nomos.
- Krafft, Tobias D., Zweig, Katharina A., und König, Pascal D. 2022. How to Regulate Algorithmic Decision-Making: A Framework of Regulatory Requirements for Different Applications. *Regulation & Governance* 16: 119–136.
- Latour, Bruno. 2004. *Politics of Nature: How to Bring the Sciences into Democracy*. Cambridge/Mass.: Harvard University Press.
- Latour, Bruno. 2007. *Eine neue Soziologie für eine neue Gesellschaft: Einführung in die Akteur-Netzwerk-Theorie*. Frankfurt: Suhrkamp.
- Lewis, Seth C., Sanders, Amy Kristin, und Carmody, Casey. 2019. Libel by Algorithm? Automated Journalism and the Threat of Legal Liability. *Journalism & Mass Communication Quarterly* 98: 60–81.
- Linaratos, Dimitrios. 2019. Künstliche Intelligenz und Verantwortung. *Zeitschrift für Wirtschaftsrecht* 504–509.
- Lohsse, Sebastian, Schulze, Reiner und Staudenmayer, Dirk. 2019. Liability for Artificial Intelligence, In *Liability for Robotics and in the Internet of Things*, hrsg. Sebastian Lohsse, Reiner Schulze und Dirk Staudenmayer, 11–26. Baden-Baden/Oxford: Nomos/Hart.
- Loo, Rory van. 2020. The Revival of Respondeat Superior and Evolution of Gatekeeper Liability. *Georgetown Law Journal* 109: 141–189.
- Luhmann, Niklas. 1965. *Grundrechte als Institution: Ein Beitrag zur politischen Soziologie*. Berlin: Duncker & Humblot.
- Luhmann, Niklas. 1987 [1972]. *Rechtssoziologie*. Opladen: Westdeutscher Verlag.
- Luhmann, Niklas. 1997. *Die Gesellschaft der Gesellschaft*. Frankfurt: Suhrkamp.
- Muhle, Florian. 2018. Sozialität von und mit Robotern? Drei soziologische Antworten und eine kommunikationstheoretische Alternative. *Zeitschrift für Soziologie* 47: 147–163.
- Nassehi, Armin. 2019. *Muster: Theorie der digitalen Gesellschaft*. Munich: C.H.Beck.
- North, Douglass C.. 1990. *Institutions, Institutional Change and Economic Performance*. Cambridge: Cambridge University Press.
- Oster, Jan. 2018. Haftung für Persönlichkeitsverletzungen durch Künstliche Intelligenz. *UFITA – Archiv für Medienrecht und Medienwissenschaft* 14–52.
- Pettit, Philip. 2007. Responsibility Incorporated. *Ethics* 117: 171–201.
- Pratt, Nicholas. 2018. HK Invstor Sues Robo-adviser in Potential Landmark Case about AI Liability. *Commercial Risk* 14 May.
- Rahwan, Iyad, et al.. 2019. Machine Behaviour. *Nature* 568: 477–486.
- Rauer, Valentin. 2017. Distribuierte Handlungsträgerschaft. Verantwortungsdiffusion als Problem der Digitalisierung sozialen Handelns, In *Politik und Verantwortung: Analysen zum Wandel politischer Entscheidungs- und Rechtfertigungspraktiken*, Christopher Daase, Julian Junk, Stefan Kroll und Valentin Rauer (ed.), 436–453. Baden-Baden: Nomos.

- Reyes, Carla L.. 2021. Autonomous Corporate Personhood. *Washington Law Review* 96: 1453–1510.
- Romano, Santi. 1918. *L'ordinamento giuridico*. Florenz: Sansoni.
- Sanz Bayón, Pablo. 2019. A Legal Framework for Robo-Advisors, In *Datenschutz / Legal-Tech*, hrsg. Erich Schweighofer, Franz Kummer, Ahti Saarenpää und Burkhard Schafer, 311–318. Bern: Weblaw.
- Schmitt, Carl. 1922. *Politische Theologie*. München: Duncker & Humblot.
- Selznick, Philip. *Law, Society and Industrial Justice*, Russell Sage, New York 1969.
- Strzelczyk, Bret E.. 2018. Rise of the Machines: The Legal Implications for Investor Protection with the Rise of Robo-Advisors. *DePaul Business & Commercial Law Journal* 16: 54–85.
- Teubner, Gunther. 2001. Rechtsirritationen: Zur Koevolution von Rechtsnormen und Produktionsregimes, In *Moral und Recht im Diskurs der Moderne: Zur Legitimation gesellschaftlicher Ordnung*, hrsg. Günther Dux und Frank Welz, 351–381. Opladen: Leske und Budrich.
- Teubner, Gunther. 2004. *Netzwerk als Vertragsverbund: Virtuelle Unternehmen, Franchising, Just in Time in sozialwissenschaftlicher und juristischer Sicht*. Baden-Baden: Nomos.
- Trüstedt, Katrin. 2020. *Representing Agency*. *Law & Literature* 32: 195–206.
- Trüstedt, Katrin. 2022. *Stellvertretung: Zur Szene der Person*. Konstanz: Konstanz University Press.
- Unger, Roberto M.. 1996. Legal Analysis as Institutional Imagination. *Modern Law Review* 59: 1–23.
- United States Commodity Futures Trading Commission and U.S. Securities & Exchange Commission. 2016. Findings Regarding the Market Events of May 6, 2010. <http://www.sec.gov/news/studies/2010/marketevents-report.pdf>.
- Vesting, Thomas. 2021. *Gentleman, Manager, Homo Digitalis: Der Wandel der Rechtssubjektivität in der Moderne*. Weilerswist: Velbrück.
- Vladeck, David C. 2014. Machines without Principals: Liability Rules and Artificial Intelligence. *Washington Law Review* 89: 117–150.
- Wagner, Gerhard. 1996. Transversale Vernunft und der soziologische Blick: Zur Erinnerung an Montesquieu. *Zeitschrift für Soziologie* 25: 315–329.
- Wagner, Gerhard, und Luyken, Lina. 2020. Haftung für Robo Advice, In *Festschrift für Christine Windbichler*, hrsg. Gregor Bachmann, Stefan Grundmann, Anja Mengel und Kaspar Krolow, 155–176. Berlin: de Gruyter.
- Welsch, Wolfgang. 1996. *Vernunft: Die zeitgenössische Vernunftkritik und das Konzept der transversalen Vernunft*. Frankfurt: Suhrkamp.
- Wiedemann, Carolin. 2014. Between Swarm, Network, and Multitude: Anonymous and the Infrastructures of the Common. *Distinktion: Scandinavian Journal of Social Theory* 15: 309–326.
- Wielsch, Dan. 2018. Contract Interpretation Regimes. *Modern Law Review* 81: 958–988.
- Williamson, Oliver E.. 1996. *Transaktionskostenökonomik*. Hamburg: Lit Verlag.
- Winner, Langdon. 1980. Do Artifacts Have Politics? *Daedalus* 109: 121–136.
- Yadav, Yesha. 2016. The Failure of Liability in Modern Markets. *Virginia Law Review* 102: 1031–1100.



# Normierung, Regulierung, Governance: Wie, von wem und mit welchen Mitteln kann der Einsatz Künstlicher Intelligenz gesellschaftlich gestaltet werden?

Karsten Weber und Nadine Kleine

## Zusammenfassung

KI-Regulierung bzw. -Normierung und Governance sind nicht nur aufgrund des raschen technischen Wandels ein schwieriges Unterfangen; allein die Komplexität des Feldes und dessen soziale, ökonomische, gesellschaftliche und politische Aspekte selbst bringen weitreichende und miteinander wechselwirkende Herausforderungen mit sich. Angesichts globalisierter Forschung und Entwicklung an sowie des weltweiten Einsatzes von KI sind nationalstaatliche Regulierungen vermutlich zum Scheitern verurteilt; supranationale Normierung, Regulierung, und Governance bringen wiederum das Problem der Durchsetzbarkeit mit sich. Nicht zuletzt deshalb unterscheiden sich aktuelle Gestaltungsansätze bspw. stark in Hinblick auf ihren Entwicklungsstatus und ihre Wirkmächtigkeit, weil die dahinterstehenden Akteur\*innen über unterschiedliche Eingriffsreichweiten und -tiefen verfügen. Unterschiedliche Akteur\*innengruppen ringen um die Deutungshoheit und verschiedenste Interessenslagen fließen in den Diskurs über Normierung, Regulierung und Governance ein. Die Komplexität wird zudem dadurch erhöht, dass KI direkt und

---

K. Weber (✉)

Ostbayerische Technische Hochschule Regensburg, Regensburg, Deutschland

E-Mail: [karsten.weber@oth-regensburg.de](mailto:karsten.weber@oth-regensburg.de)

N. Kleine

Fachbereich 1: Kultur- und Sozialwissenschaften, Universität Osnabrück, Osnabrück, Deutschland

E-Mail: [nadine.kleine@uni-osnabrueck.de](mailto:nadine.kleine@uni-osnabrueck.de)

indirekt reguliert und normiert werden kann, es aber gleichzeitig auch Regulierung und Normierung menschlichen Handelns durch KI gibt. Der Beitrag soll die Komplexität des Feldes aufzeigen, Typen von Akteur\*innen vorstellen, verschiedene Gestaltungsansätze beschreiben und eine vorläufige Abschätzung der Möglichkeiten von Normierung, Regulierung und Governance von KI geben.

---

### Schlüsselwörter

Artificial Intelligence Act • Deutungshoheit über KI • Gesellschaftliche Interessens- und Machtverhältnisse • Gesellschaftliches Narrativ von KI • Normierung *von* und *durch* KI • Normierende Wirkung von Technik

---

## 1 Überblick einer vielschichtigen Debatte

In den wissenschaftlichen, politischen, ökonomischen und gesellschaftlichen Debatten über Normierung bzw. Regulierung von Künstlicher Intelligenz (KI) – im Englischen wird zuweilen von ‚regulation‘ gesprochen, aber im Wesentlichen laufen die entsprechenden Debatten dort unter der Rubrik ‚governance‘ – werden zahlreiche und teilweise auch sehr unterschiedliche Aspekte des Einsatzes von KI behandelt. Sehr prominent sind Themen wie *Vertrauen in und Akzeptanz von KI-Systemen*, *Entscheidungsunterstützung durch und -delegation an KI-Systeme*, *Diskriminierung durch KI* sowie *Datenschutz und Privatsphäre im Zusammenhang mit KI* (eine ähnliche Liste findet sich bspw. auch bei Gasser und Almeida 2017). Obwohl der dabei verwendete KI-Begriff meist eher unscharf definiert bzw. sehr weit gefasst wird, werden diese Themen überwiegend in Bezug auf lernende Systeme, also Machine-Learning- oder Deep-Learning-Systeme, diskutiert und eine geeignete Normierung bzw. Regulierung gefordert. Dass bspw. Symbolverarbeitungsansätze der KI nicht umfänglich diskutiert werden – anders als in den 1960er bis 1980er Jahren (bspw. Sonar und Weber 2020) – bedeutet nicht, dass solche Verfahren nicht mehr genutzt werden, sondern zunächst einmal nur, dass diese eine geringere öffentliche Sichtbarkeit erzielen, weil sie inzwischen eher Nischenanwendungen darstellen.

Es ist allerdings zu betonen, dass insbesondere in der öffentlich geführten Diskussion nicht streng zwischen bspw. Big-Data-Analytics auf Grundlage statistischer Verfahren und KI-Ansätzen gleich welcher Art unterschieden wird (vgl. Weber 2022). Man kann dies aus einer streng wissenschaftlich-analytischen Perspektive bedauern oder gar als Problem ansehen, aber in Hinblick auf die

gesellschaftlichen Auswirkungen solcher Methoden ist die fehlende Trennschärfe durchaus nachvollziehbar, da es bspw. für Personen, die durch Anwendung solcher Methoden eine Diskriminierung erfahren, vermutlich irrelevant ist, ob eine Benachteiligung durch ein System basierend auf statistischen Verfahren oder Deep Learning zustande kommt. Dieser Logik folgt, wie noch zu sehen sein wird, bspw. auch der aktuelle Entwurf des *Artificial Intelligence Act* (AIA) der EU.

Obwohl eine Zeitlang gerade in der Öffentlichkeit intensiv über die Gefahr eines massiven Arbeitsplatzverlustes durch den Einsatz von KI-Systemen in der Arbeitswelt gesprochen wurde, ist dieses Thema in der derzeitigen Debatte nicht mehr ganz so stark vertreten. Die frühen Schätzungen, denen zufolge massive Arbeitsplatzverluste zu befürchten seien (Frey und Osborne 2013 und Frey und Osborne 2017), weil die dort ausgeübten Tätigkeiten vergleichsweise leicht durch KI-Systeme zu übernehmen wären, haben sich bisher nicht bewahrheitet. Das allein kann sicher keine Entwarnung sein, da entsprechende Veränderungen der Arbeitswelt in der Regel über längere Zeiträume ablaufen, aber aktuellere Studien sehen nicht nur ein geringeres Rationalisierungspotenzial, das auch nur unter bestimmten Bedingungen zum Tragen kommt (bspw. Dengler, Matthes und Wydra-Somaggio 2018), sondern zeigen auch auf, dass zumindest die Chance besteht, dass der Einsatz von KI-Systemen nicht nur zur Vernichtung, sondern auch zur Schaffung neuer Arbeitsplätze beitragen könnte (z. B. Autor, Mindell und Reynolds 2020; Balsmeier und Woerter 2019; Hamid, Smith und Barzanji 2017). Aus Normierungs- bzw. Regulierungssicht wäre dann eher die Frage zu klären, wie man die Transformation der Arbeitswelt konstruktiv begleitet, um den Übergang sowohl für die unmittelbar betroffenen Personen als auch für die Gesellschaft als Ganzes sozialverträglich zu gestalten, und nicht die Frage, ob und wie man mithilfe von Normierung und Regulierung einen KI-Einsatz verhindern oder doch zumindest erheblich verteuern könnte, um einen Arbeitsplatzabbau abwenden zu können (bspw. Moser 2021). Vor allem das Thema der Einkommensungleichheit, die durch den Einsatz von KI noch weiter vergrößert werden könnte, wird dabei diskutiert (Korinek und Stiglitz 2019; Suedekum 2018).

Allerdings sind Diskussionen um die Frage der Auswirkungen des KI-Einsatzes auf die Arbeitswelt beileibe nicht verschwunden. In Bezug auf jene Arbeitsbereiche, in denen der Einsatz von KI bereits Realität oder zumindest für die nächste Zukunft absehbar ist, wird intensiv darüber nachgedacht, wie eine Transformation aussehen könnte. Dabei ist bedenkenswert, dass entsprechende Diskussionen gerade für Handlungsfelder und Professionen besonders intensiv geführt werden, die mit und am Menschen tätig sind und bei denen zuweilen ein sehr enges Beziehungsgeflecht zwischen den Professionellen und den

Klient\*innen Bedingung der professionellen Tätigkeit ist. So wird bspw. in einschlägigen Fachzeitschriften der Radiologie eine umfängliche wissenschaftliche Auseinandersetzung über die Veränderung der eigenen Profession geführt (z. B. Attenberger und Reiser 2022; Kauffmann et al. 2022). Ähnliches gilt auch für andere medizinische Disziplinen (z. B. Reddy et al. 2020), meist (aber nicht ausschließlich) mit diagnostischem Bezug, z. B. in der Onkologie (bspw. Klee-siek et al. 2021) oder in der Psychiatrie und Psychologie (für einen Überblick siehe Krug et al. 2020) – aber ebenso für andere Professionen (z. B. Gasser und Schmitt 2020), bspw. für die Soziale Arbeit (bspw. Schneider, Sonar und Weber 2022; Schneider 2022).

Im Folgenden soll nun versucht werden, die Normierungs- und Regulierungsdebatte in Hinblick auf Künstliche Intelligenz und in Bezug auf die skizzierten Themenfelder so nachzuzeichnen, dass zumindest ein grober Überblick entstehen kann. Tatsächlich scheint zum jetzigen Zeitpunkt ein höherer Anspruch an eine entsprechende Systematisierung eher verfehlt, da die Diskussion hierüber noch sehr fluide ist; zu groß ist die Vielfalt der beteiligten Akteur\*innen und damit der Interessen. Klar ist zwar, dass nationale Normierungs- und Regulierungsdebatten am Ziel vorbeigehen, wenn sie nicht durch supranationale Vereinbarungen flankiert werden, aber wie internationale Regeln aussehen könnten, bleibt bisher unklar. Vielleicht wird der AIA der EU in dieser Hinsicht erste Klärungen mit sich bringen; dies wird jedoch erst die Zukunft zeigen können.

---

## 2 Akteur\*innen einer komplexen Debatte

Wie kaum anders zu erwarten bei einem komplexen Thema von globaler Reichweite ist auch im Fall der KI-Normierungsdebatte die Zahl und die Art der beteiligten Akteur\*innen groß und sehr divers. Bewusst wird hier von ‚Akteur\*innen‘ und nicht von ‚Stakeholder\*innen‘ gesprochen, denn wenn man die obigen Bemerkungen in Betracht zieht, so sind bspw. Arbeitnehmer\*innen, Konsument\*innen, Patient\*innen und/oder Bürger\*innen die entscheidenden Stakeholder\*innen, doch als Akteur\*innen tauchen sie, vor allem in der öffentlichen und politischen Debatte, allenfalls indirekt und dann vertreten durch Interessenverbände auf. In Hinblick auf autoritär regierte Staaten wie China als einem der Länder, in dem die KI-Entwicklung massiv vorangetrieben wird, mag dies nicht sonderlich überraschen, aber für die USA oder die EU bzw. für europäische Staaten wie Großbritannien erscheint dies doch eher defizitär. Andererseits könnte man argumentieren, dass die letztgenannten Staaten und supranationalen Strukturen Formen der parlamentarischen Demokratie verwenden und damit die

Vertretung der Stakeholder\*innen und deren Belange durch Interessenverbände, Abgeordnete und andere Stellvertreter\*innen der Normalfall wäre. Angesichts der immer wieder behaupteten grundstürzenden Veränderungen (bspw. Wang und Siau 2019), die der ubiquitäre KI-Einsatz mit sich bringen würde, könnte man aber trotzdem ein Legitimationsdefizit (und, nebenbei gesagt, auch eine Forschungslücke, vgl. Brougham und Haar 2018) vermuten.

Die Akteur\*innen der KI-Normierungsdebatte kann man zunächst grob in die Kategorien der

- a. *staatlichen und suprastaatlichen Akteur\*innen* (bspw. UN, OECD, EU-Kommission, High-Level Expert Group, Bundesregierung, Enquete-Kommission, Ministerien)
- b. *Verbände* (bspw. Gewerkschaften, Industrie- und Handelsverbände, Arbeitgeberverbände, Verbraucherzentrale(n), Bundesverband KI e. V.)
- c. *Nichtregierungsorganisationen* (z. B. Stiftung neue Verantwortung, Algorithm Watch, AI Now, Access Now, Future of Life Institute)
- d. *Unternehmen* (z. B. GAFA, Microsoft, Telekom) sowie
- e. *Normierungs- und Prüforganisationen* (bspw. TÜV, DIN, DKE)

einordnen. Die genannten Beispiele sind unsystematisch gewählt und können keinen Anspruch auf Vollständigkeit erheben – mit ihrer Nennung soll vor allem verdeutlicht werden, wie divers die beteiligten Akteur\*innen sind.

Dass in dieser Aufzählung wissenschaftliche Institutionen zunächst nicht auftauchen, hat einen spezifischen Grund, der nichts damit zu tun hat, dass die KI-Normierungsdebatte im wissenschaftlichen Kontext nicht geführt werden würde – das komplette Gegenteil ist (natürlich) der Fall, die Menge der wissenschaftlichen Publikationen zu diesem Thema ist unüberschaubar. Einerseits gehören Wissenschaftler\*innen damit zu den Akteur\*innen der Normierungsdebatte, aber andererseits ist zu beobachten, dass deren Wirkmächtigkeit meist nur dann gegeben ist, wenn sie im Kontext wissenschaftlicher Institutionen aktiv oder von den genannten Akteur\*innen-Gruppen als Expert\*innen für den Diskurs herangezogen werden.

Wissenschaftliche Institutionen zählen wir zu den ‚Zwischenakteur\*innen‘, die zwischen den oben aufgeführten Akteur\*innen vermittelnd wirken. Mit Vermittlung ist dabei nicht Mediation oder Verhandlung gemeint, sondern Aufbereitung, Interpretation und Verbreitung von Wissen. Daher zählen wir bspw. auch *politische Parteien und parteinahe Stiftungen*, die *Datenethikkommission*, die verschiedenen *Datenschutzbeauftragten* und das *European Data Protection Board*, Einrichtungen wie die *Plattform lernende Systeme* oder die *Plattform Industrie 4.0*



und nicht zuletzt eben wissenschaftliche Einrichtungen oder Netzwerke wie das *DFKI* oder *Claire AI* dazu. All diesen Einrichtungen kommt die Rolle zu, durch Beiträge zur KI-Normierungsdebatte einen Interessenausgleich zwischen den anderen Akteur\*innen, die stärker partikulare Interessen vertreten, herzustellen bzw. zu befördern. Wissenschaftliche Institutionen kommen dadurch zunehmend in eine Situation, in der nicht das Wissen selbst, das dort produziert wird, relevant ist, sondern die Erwartung, dass diese Institutionen objektiv und nicht interessegeleitet agieren. Mit anderen Worten: Robert K. Mertons CUDOS-Schema<sup>1</sup> (Merton 1974) feiert ein Revival, was allein schon angesichts der Erkenntnisse aus den Science and Technology Studies (STS) über die gesellschaftliche Konstruktion von Technik (z. B. die Beiträge in Bijker, Hughes und Pinch 1987) und solchen Konzepten wie Triple-Helix (bspw. Etzkowitz und Leydesdorff 2000; Etzkowitz 2003; Leydesdorff 2000) verwundern kann. Bedenkt man zudem, dass bspw. das Triple-Helix-Modell nicht nur eine enge Zusammenarbeit zwischen Wissenschaft, Wirtschaft und Regierung beschreibt, sondern dadurch insbesondere die Frage der Finanzierung der Wissenschaft durch Wirtschaft und Regierung aufgeworfen wird, kann man zumindest kritische Nachfragen in Hinsicht auf die Interessenlosigkeit wissenschaftlicher Institutionen (klar zu unterscheiden von Wissenschaftler\*innen als Personen) stellen.

---

### 3 Existierende KI-Normierungsansätze

Von allen Typen der oben genannten (Zwischen-)Akteur\*innen wurden in den letzten Jahren zahlreiche Policy-Paper zur Normierung von KI veröffentlicht, von denen der Entwurf des *Artificial Intelligence Act* (AIA) der EU vermutlich

---

<sup>1</sup> Das CUDOS-Schema stellt eine normative Konzeption von Wissenschaft dar und steht für: C = Communisms (manchmal auch: Communitarianism) – alle Erkenntnisse der Wissenschaft sollen allen Interessierten frei zur Verfügung stehen; U = Universalism – niemand darf aufgrund von ethnischer Zugehörigkeit, Religion, Geschlecht etc. aus der Wissenschaft ausgeschlossen werden; D = Desinterestedness – Wissenschaft soll nicht um der Verfolgung eigennütziger Ziele betrieben werden; OS = Organized Scepticisms – wissenschaftliche Aussagen müssen stets rigoros geprüft, dogmatische Ansprüche müssen abgewiesen werden. Merton hat diese Konzeption von Wissenschaft in Reaktion auf den Umgang mit jüdischen Wissenschaftler:innen im nationalsozialistischen Deutschland entwickelt; sie stellt ein Ideal von Wissenschaft dar, keine Beschreibung. Trotzdem hat CUDOS das (öffentliche) Bild von Wissenschaft stark geprägt (bspw. Bray & von Storch 2017; Kuipers 2009; Macfarlane & Cheng 2008; Meyer & Sandøe 2012; Moriarty 2011) und wird – wenn auch oft implizit – meist dann herbeizitiert, wenn der Eindruck entsteht, dass Wissenschaft ebenfalls ein primär interessegeleitetes Unterfangen sei.

das wichtigste Dokument darstellt. Viele der neueren Policy-Paper der übrigen Akteur\*innen beziehen sich, oft direkt in Form eines Positionspapiers (mit Aufforderung der Beachtung), auf den AIA. Es ist daher zu erwarten, dass die zukünftige KI-Normierungsdebatte, zumindest in der Europäischen Union bzw. in ihren Mitgliedsländern und eng an die EU angelehnten Ländern, an dem AIA bzw. dessen Fortschreibung orientiert sein wird.

Allerdings war und ist der AIA beileibe nicht der einzige Versuch, den Einsatz von KI zu normieren bzw. zu regulieren. Hier eine willkürliche und beileibe nicht vollständige Auswahl, die aber zumindest die Spannweite der Typen von Akteur\*innen im Zusammenhang mit der KI-Normierungsdebatte widerspiegeln kann:

- OECD: Recommendation of the Council on Artificial Intelligence.<sup>2</sup>
- EU: Ethics guidelines for trustworthy AI.<sup>3</sup>
- Die Deutsche Bundesregierung: Nationale Strategie für Künstliche Intelligenz.<sup>4</sup>
- Future of Life Institute: Asilomar AI Principles.<sup>5</sup>
- Google: Recommendations for Regulating AI.<sup>6</sup>
- DKE: Deutsche Normierungsroadmap.<sup>7</sup>

Schiff et al. (2022) führen zahlreiche wissenschaftliche Studien auf, in denen versucht wurde, einen vollständigeren Überblick über KI-Normierungsdokumente herzustellen (vgl. auch Butcher und Beridze 2019; Schmitt 2021; Ulicnec et al. 2021). Viele dieser Dokumente bewegen sich – zumindest legt das eine erste Sichtung nahe – im Bereich ethischer Leitlinien und Ethikcodizes; deren Bindungskraft beruht, wenn überhaupt, auf der Selbstverpflichtung von

---

<sup>2</sup> <https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449>, zuletzt besucht am 18.05.2022.

<sup>3</sup> <https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai>, zuletzt besucht am 18.05.2022.

<sup>4</sup> [https://www.bmwk.de/Redaktion/DE/Publikationen/Technologie/strategie-kuenstliche-intelligenz-der-bundesregierung.pdf?\\_\\_blob=publicationFile&v=10](https://www.bmwk.de/Redaktion/DE/Publikationen/Technologie/strategie-kuenstliche-intelligenz-der-bundesregierung.pdf?__blob=publicationFile&v=10) und [https://www.ki-strategie-deutschland.de/files/downloads/201201\\_Fortschreibung\\_KI-Strategie.pdf](https://www.ki-strategie-deutschland.de/files/downloads/201201_Fortschreibung_KI-Strategie.pdf), beide zuletzt besucht am 22.06.2022.

<sup>5</sup> <https://futureoflife.org/ai-principles/>, zuletzt besucht am 18.05.2022.

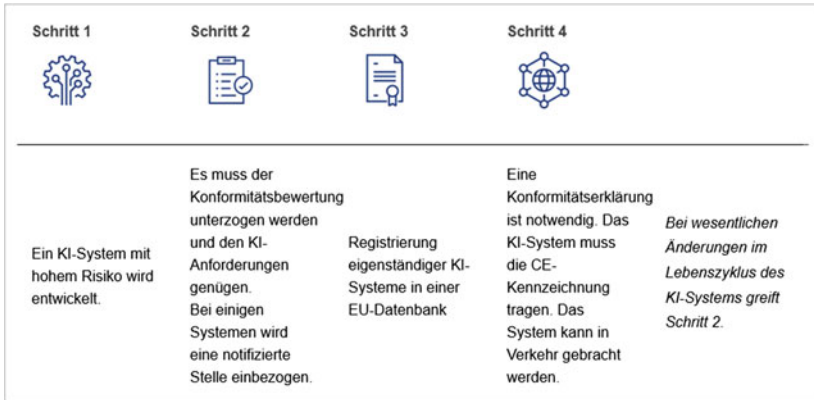
<sup>6</sup> <https://ai.google/static/documents/recommendations-for-regulating-ai.pdf>, zuletzt besucht am 22.06.2022.

<sup>7</sup> <https://www.din.de/resource/blob/772438/6b5ac6680543eff9fe372603514be3e6/normungsroadmap-ki-data.pdf>, zuletzt besucht am 22.06.2022.

Akteur\*innen. Hier gibt es also keine Einklagbarkeit, die nur das Recht bieten könnte; trotzdem gibt es Autor\*innen, die sogar für eine verstärkte Nutzung von Selbstverpflichtungen plädieren (bspw. Roski et al. 2021). Weiterhin weisen Schiff et al. (2022) darauf hin, dass die meisten Dokumente nur wenig Hinweise auf die Überführung der darin enthaltenen Forderungen, Empfehlungen und/oder Ansprüche in konkrete Handlungsanweisungen für die Praxis der KI-Entwicklung und des KI-Einsatzes bieten, sondern auf der eher abstrakten Ebene moralischer Gebote verbleiben.

Insbesondere die philosophische bzw. ethische Debatte bewegt sich auf einem hohen Abstraktionsniveau. Selbst Konzepte, die die Normierung und damit Gestaltung von KI-Systemen an die Menschenrechte binden wollen, wie es Yeung et al. (2020) vorschlagen, werfen die Frage auf, welchen praktischen Nutzen dies haben könnte, wenn man zur Kenntnis nehmen muss, dass die Existenz der Deklaration der Allgemeinen Menschenrechte von 1948 nicht verhindert, dass Menschenrechtsverletzungen global und in großer Zahl stattfinden. Solche Konzepte haben ohne Zweifel den Charme, an Prinzipien anzuknüpfen, die in vielen Ländern akzeptiert werden – aber dort, wo es keine Institutionen zu deren Durchsetzung gibt, ist der praktische Nutzen solcher Prinzipien sehr begrenzt (vgl. Smuha 2021). Mit anderen Worten: Die KI-Normierungsdebatte ist immer auch eine Debatte über die (Un-)Möglichkeit der Institutionalisierung ihrer Durchsetzung (vgl. auch Cihon, Maas und Kemp 2020). Ohne dass dies an dieser Stelle weiter ausgeführt werden kann, setzt diese Einsicht außerdem der Hoffnung auf eine globale und einheitliche Normierung enge Grenzen (vgl. auch Wong 2020 in Hinblick auf kulturelle Vielfalt; ebenso ÓhÉigeartaigh et al. 2020).

Geht man davon aus, dass der AIA für die EU wohl den entscheidenden Normierungsansatz darstellt, ist wichtig zu beachten, dass der AIA nicht Technologien normieren bzw. regulieren will, sondern das jeweilige Risiko einer KI-Anwendung für Rechte der EU-Bürger\*innen. Dieser risikobasierte Ansatz sieht folgende Kategorisierung von KI-Anwendungen vor: a) ein minimales bzw. begrenztes Risiko, b) ein hohes Risiko und c) ein unannehmbares Risiko. Während im letztgenannten Fall die Anwendung nicht zugelassen werden soll, sieht der AIA bei einem hohen Risiko konkrete Schritte vor, die durchlaufen werden müssen (siehe Abb. 1). Insofern nimmt der AIA (ob intendiert oder nicht ist dabei gar nicht so wichtig) viele Kritikpunkte, die oben angedeutet wurden, auf. Der AIA soll vertrauenswürdige KI fördern, „die mit den Werten und Interessen der [Europäischen] Union in Einklang stehen“ (Europäische Kommission 2021) und so nicht zuletzt die internationale Position und Wettbewerbsfähigkeit der EU stärken. Damit ist auch klar, dass der AIA als Normierungsansatz die Interessen aller oben genannten Akteur\*innen nicht nur in Europa, sondern global betrifft.



**Abb. 1** Vorschriften für Anbieter von KI-Systemen mit hohem Risiko (Europäische Kommission 2022<sup>8</sup>)

In der sich abzeichnenden Debatte über den Entwurf des AIA kristallisieren sich solche Aspekte heraus, über die Einigkeit bei den Akteur\*innen besteht, und solche, für die das nicht zutrifft. Einigkeit scheint darüber zu bestehen, dass nicht Technologien selbst normiert werden sollen, sondern Anwendungen bzw. die Auswirkungen von Anwendungen – wobei ‚Anwendungen‘ vermutlich fast im Sinne der Interaktionsfläche zwischen KI-Systemen und Anwender\*innen verstanden werden, gerade weil der AIA einen sehr weiten KI-Begriff unterstellt. Außerdem wird gefordert, dass ein zukünftiger AIA eng mit bestehenden Regulierungen und Gesetzen abgestimmt sein muss, um Rechtsunsicherheit und Konflikte zu verhindern, bspw. in Bezug auf die GDPR, den EU *Cybersecurity Act* oder den EU *Digital Services Act*. Allerdings gibt es auch einige Streitpunkte: Der AIA operiert mit einem sehr weiten KI-Begriff und umfasst Verfahren und Methoden, die in Fachdebatten nicht unbedingt als KI verstanden werden. Vermutlich soll dies sicherstellen, dass keine Normierungslücken entstehen, die bspw. von Unternehmen oder anderen Stakeholder\*innen genutzt werden, um einer Normierung ihrer Anwendungen zu entgehen. Außerdem wird auf diese Weise vermieden, dass der AIA angesichts der schnellen technischen Entwicklung in kürzester Zeit veraltet, weil neue Technologien nicht mehr adressiert werden könnten. Während einige Akteur\*innen dies positiv bewerten, reagieren andere mit Kritik, da dieser

<sup>8</sup> [https://ec.europa.eu/info/strategy/priorities-2019–2024/europe-fit-digital-age/excellence-trust-artificial-intelligence\\_de](https://ec.europa.eu/info/strategy/priorities-2019–2024/europe-fit-digital-age/excellence-trust-artificial-intelligence_de), zuletzt besucht am 22.06.2022.

weite KI-Begriff aus unternehmerischer Perspektive zu viele aktuelle und zukünftige Software-Lösungen umfasst und dies innovationshemmend wirken könnte. Uneinigkeit und damit potenzielle Konflikte zeichnen sich auch in Hinblick auf den Umgang mit Quellcodes und Daten ab, denn ein Offenlegungszwang stünde in Konflikt bspw. mit der Sicherung von Betriebsgeheimnissen, doch eine Überprüfung von möglichen Bias und damit potenziell negativen Auswirkungen auf Nutzer\*innen von KI-Systemen wird ohne Einsicht in verwendete Algorithmen und Daten tendenziell verunmöglicht. Ebenfalls konfliktträchtig ist die Frage nach einem Forschungsprivileg sowie in Bezug auf die Nutzung KI-basierter *Remote Biometric Identification*, bspw. durch Strafverfolgungsbehörden.

---

## 4 Normierung durch KI

Bisher wurde insbesondere *Normierung von KI* betrachtet. Noch nicht weiter behandelt werden konnte das Thema *Normierung durch KI*. Hier kann nur stichwortartig darauf verwiesen werden, dass die bloße Existenz von KI oder auch deren Einsatz selbst normierende Wirkung zeitigen kann. Beispiele, die in diesem Zusammenhang oft genannt werden, sind KI-gestützte Sprachübersetzungsanwendungen, die Geschlechterstereotype reproduzieren bspw., wenn ‚nurse‘ mit ‚Krankenschwester‘ und ‚physician‘ mit ‚Arzt‘ übersetzt wird und so geschlechterspezifische Rollenzuweisungen stattfinden und dadurch möglicherweise Einstellungen und Haltungen der Nutzer\*innen geprägt werden.

Insbesondere Timnit Gebru hat diese Diskussion in der letzten Zeit geprägt – nicht, weil sie die einzige Autorin wäre, die über solche Probleme berichtet, sondern weil ihr Arbeitgeber Google sie nach der Veröffentlichung von Aufsätzen sanktioniert hatte und dies zu ihrem Austritt aus dem Unternehmen führte. Zusammen mit anderen Autor\*innen hat Gebru auf Bias bspw. in KI-gestützten Übersetzungsprogrammen sowie Bilderkennungs- und Bildklassifikationssystemen hingewiesen (z. B. Bender et al. 2021; Buolamwini und Gebru 2018; Gebru 2020; Raji et al. 2020). Das Argument, das in diesen Beiträgen stark gemacht wird, ist, dass hier eben nicht nur ein Bias vorliegt, der zu verfälschten Ergebnissen führt, sondern dass hierdurch eine Normierung des menschlichen Verhaltens bspw. in Hinblick auf Geschlechterrollen oder im Umgang mit ethnischen Minderheiten befördert werden würde, die zum Nachteil dieser Stakeholder\*innen sei – also Normierung als und durch Diskriminierung stattfände.

Man kann dieses Thema auch in anderer Weise darstellen und fragen, wer bzw. was Normierung im Sinne von Handlungsvorgaben vornimmt und vornehmen darf. Jipp und Steil (2021) stellen daher die Frage „Steuern wir oder

werden wir gesteuert?": Geben wir der Technik vor, wie sie sich zu verhalten hat oder gibt die Technik vor, wie wir zu handeln haben – wobei dann die Frage im Raum steht, ob überhaupt noch von Handlungen gesprochen werden kann. Gerade da, wo eine eng gekoppelte Interaktion von Menschen und KI-Systemen stattfindet, besteht bei Umkehrung des Verhältnisses zwischen Menschen und Maschinen die Möglichkeit weitreichender Auswirkungen auf das Selbstbild der Menschen. Dabei ist wichtig zu bedenken, dass dies beileibe nicht nur den Einsatz von KI in professionellen Kontexten betrifft, sondern gerade auch den alltäglichen Gebrauch solcher Systeme, bspw. im Kontext KI-gestützter Social-Media-Plattformen, Sales-Systemen oder im Fall des autonomen Fahrens.

Für eine weitere und genauere Untersuchung der Normierung durch KI wäre es daher sinnvoll, die (wissenschaftliche) Debatte zu Bias in KI-Systemen ebenso wie zur Gestaltung der Mensch-Technik-Interaktion unter dem Gesichtspunkt zu analysieren, inwieweit hierbei eine Umkehr der Handlungsvorgaben stattfindet. Im vorliegenden Text kann dies allerdings nur angedeutet werden (für Hinweise aus einer technikhistorischen Perspektive siehe Weber 2018; 2021). Instruktiv dürften hierbei Untersuchungen dazu sein, wie Nutzer\*innen den Umgang mit KI-Systemen wahrnehmen (bspw. Bhargava, Bester und Bolton 2021) oder wie sich Professionen durch den (prospektiven) Einsatz von KI-Systemen verändern (zur Übersicht z. B. Krug et al. 2020; Schneider, Sonar und Weber 2022). Ebenso wichtig wäre die Berücksichtigung der Debatte über ‚governance by algorithm‘ – also die Ausübung von Macht mithilfe von (KI-)Algorithmen (bspw. Kuziński und Misuraca 2020).

---

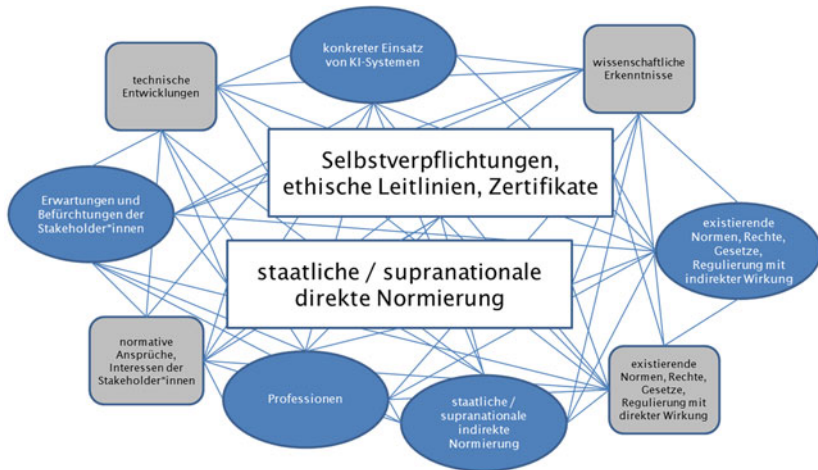
## 5 Direkte und indirekte Normierung

Der im vorliegenden Beitrag mehrfach genannte *Artificial Intelligence Act* (AIA) der EU stellt den Versuch dar eine EU-weite Normierung bzw. Regulierung des KI-Einsatzes zu formulieren und diese dann nach Verabschiedung natürlich auch durchzusetzen. Dadurch, dass alle Institutionen, Organisationen, Unternehmen bzw. Nutzer\*innen, die in der EU KI-Systeme zum Einsatz bringen wollen, durch den AIA betroffen wären – der Wirkungsmechanismus ist dem des GDPR vergleichbar – bleibt die Reichweite des AIA nicht auf Europa beschränkt, sondern wäre global. In jedem Fall aber stellt der AIA eine *direkte* Normierung des KI-Einsatzes bzw. von KI-Systemen dar, denn die darin festgehaltenen Regeln sollen unmittelbar Einfluss auf die Gestaltung der Systeme selbst nehmen.

Weiter unten (siehe Abb. 2) wird auch von *indirekter* Normierung gesprochen. Damit ist gemeint, dass bestehende (moralische) Normen, Regeln oder

Gesetze den Einsatz von KI-Systemen indirekt beeinflussen können. So können bspw. Standes- oder Professionsethiken ein Umfeld generieren, das mehr oder minder empfänglich für den Einsatz von KI-Systemen ist; Ausbildungsordnungen können dazu beitragen, dass in einem professionellen Umfeld das Know-how zur Nutzung von KI-Systemen vorhanden ist – oder eben nicht; Professionsbilder mit nicht selten ungeschriebenen und trotzdem wirkmächtigen Regeln und Erwartungen bzgl. des professionellen Handelns der Mitglieder können wiederum den Einsatz von KI-Systemen befördern, aber auch (latente) Widerstand dagegen hervorrufen (bspw. Schneider, Sonar und Weber 2022); Gesetzgebungsvorhaben, die gar nicht spezifisch KI-Systeme betreffen, können nichtsdestoweniger Einfluss auf die Fördermöglichkeiten zur Einführung von KI-Systemen und den Umgang mit ihnen haben (zur Übersicht der existierenden Regularien hinsichtlich KI im Gesundheitswesen vgl. auch Pechmann et al. 2022). Ein Beispiel hierfür wären die in den letzten Jahren erlassenen Bundesgesetze zur Förderung der Digitalisierung des Gesundheitswesens, u. a. das Krankenhauszukunftsgesetz (KHZG), das Patienten-Datenschutzgesetz (PDSG), das Gesetz zur Weiterentwicklung der Gesundheitsversorgung (GVWG) oder das Digitale-Versorgung-und-Pflege-Modernisierungs-Gesetz (DVPMG). Wenn solche Gesetzgebungsvorhaben Technologien wie KI-Systeme finanziell fördern – bspw. durch Zuschüsse, durch Refinanzierungsmöglichkeiten durch Solidarkassen etc. – wird es für entsprechende Stakeholder\*innen attraktiv, solche Systeme einzusetzen. Dadurch wird einrichtungsübergreifend eine Pfadabhängigkeit und möglicherweise ein Lock-in bewirkt (bspw. Crootof 2019; Hennemann 2018; Robbins und van Wynsberghe 2022), die selbst normierenden Charakter für das zukünftige Agieren entsprechender Einrichtungen besitzen.

Selbst wenn man annimmt, dass sich all diese (moralischen) Normen, Regeln, Gesetze oder ungeschriebenen Handlungserwartungen nicht widersprechen – eine Annahme, die aller Wahrscheinlichkeit nach kontrafaktisch ist – und somit als Gesamtkorpus einer Normierung gesehen werden können, bedeutet das, dass indirekte Normierung bspw. länder- und/oder bereichsspezifische Gestaltungsoptionen ermöglicht, die potenziell nicht im Sinne jener Institutionen sind, die direkt normieren, zumindest aber von diesen nicht antizipiert werden konnten. Damit könnten Partikularinteressen in den Vordergrund geraten, es entsteht Normierungsvielfalt, wo Normierungseinheitlichkeit angestrebt wurde, und es stellt sich in manchen Fällen die Frage der Legitimation der Normierung, wenn sie – wie bspw. im Falle überkommener Standesethiken oder Professionsbilder – nie einen demokratischen Abstimmungsprozess durchlaufen haben.



**Abb. 2** Normierungsnetz. (Eigene Darstellung)

## 6 Fazit: Die (Un-)Möglichkeit der KI-Normierung

Die Überschrift dieses letzten Abschnitts soll natürlich nicht andeuten, dass es ein sinnloses und automatisch nicht von Erfolg gekröntes Unterfangen wäre zu versuchen, KI zu normieren und zu regulieren – der Versuch und der damit angestoßene und fortwährende Diskurs sind absolut notwendig. Trotzdem sollten Erwartungen an die Reichweite und die Auswirkungen von Normierungsansätzen grundsätzlich nicht überschätzt werden. Dazu muss man nicht einmal einem technologischen Determinismus anhängen, wie man das bspw. bei den Autor\*innen des folgenden Zitats (Wang und Siau 2019, S. 61) vermuten könnte:

“The exponential advancement in artificial intelligence (AI), machine learning, robotics, and automation are rapidly transforming industries and societies across the world. [...] This new industrial revolution is expected, on one hand, to enhance and improve our lives and societies. On the other hand, it has the potential to cause major upheavals in our way of life and our societal norms. The window of opportunity to understand the impact of these technologies and to preempt their negative effects is closing rapidly.”



Man kann dies so verstehen, dass die Weiterentwicklung der Technologie im Allgemeinen und von KI im Speziellen im Grunde selbständig abläuft und Menschen sowie Gesellschaften diese Entwicklung nur beobachten und sich allenfalls daran anpassen können. Wenn diese Sichtweise korrekt wäre, dann müsste man vermutlich von der Unmöglichkeit der KI-Normierung ausgehen bzw. annehmen, dass sich Normierung bzw. Regulierung nur auf Anpassungsleistungen von Individuen und Gesellschaften bezögen. Tatsächlich wäre eine solche Sichtweise aber inkonsistent, denn entsprechende Anpassungsleistungen würden den Kontext der Anwendung von KI verändern und damit aller Wahrscheinlichkeit nach auch die KI-Technologie selbst.

Man muss aber, wie bereits gesagt, keinem Technikdeterminismus anhängen, um in Hinblick auf die Reichweite und die Auswirkungen von Normierungsansätzen eher bescheidene Erwartungen zu hegen. Es reicht eine auch nur oberflächliche Analyse der Faktoren, die mit Normierungsansätzen zusammenhängen bzw. auf diese Wirkung zeigen. Geht man bspw., wie in Abb. 2 angedeutet, von einer groben Einteilung von Normierungsansätzen in Selbstverpflichtungen auf der einen und staatlicher bzw. supranationaler Normierung auf der anderen Seite aus, stellt außerdem noch direkte und indirekte Normierung in Rechnung und berücksichtigt die vielfältigen und zahlreichen Stakeholder\*innen und deren Haltungen bzgl. KI, dann hat man nicht einmal ansatzweise all jene Faktoren berücksichtigt, die auf ein mögliches Normierungsgeschehen Einfluss nehmen. Denn wissenschaftliche Erkenntnisse und technische Entwicklungen, selbst voneinander abhängig, geben – zumindest bis zu einem gewissen Grad – Normierungsbedarf und -möglichkeiten überhaupt erst vor; Gleiches gilt für den konkreten Einsatz von KI-Systemen im Sinne der normativen Kraft des Faktischen. Dieser Einsatz findet nicht nur, aber sehr oft im Kontext von Professionen mit einer eigenen Verfasstheit statt und es existieren zahlreiche Normen, Gesetze und andere Formen der Normierung und Regulierung, die diesen Kosmos von Faktoren direkt und indirekt (mit-)gestalten. Nicht zuletzt muss bedacht werden, dass das Normierungsnetz in Abb. 2 in der Realität eingebettet ist in komplexe Wirtschafts- und Gesellschaftsgeschehen, die ganz eigene Rahmenbedingungen setzen: Wirtschaftskrisen, Kriege und Pandemien – um nur jene außergewöhnlichen Ereignisse zu nennen, die wir gerade (Frühjahr/Sommer 2022) erleben – bringen eine Dynamik mit sich, denen Normierungsansätze regelmäßig nicht oder zumindest nicht vollständig gerecht werden können. Ohne die Corona-Pandemie wäre es bspw. kaum denkbar gewesen, dass eine Technologie wie die Corona-Warn-App innerhalb von wenigen Wochen zum Einsatz gebracht werden könnte – der gesellschaftliche Abstimmungsprozess hätte unter normalen Bedingungen in Deutschland Jahre, wenn nicht Jahrzehnte gedauert. Aus diesem

Grund ist eine Prognose der zukünftigen Normierungsformen von KI durch die – mal mehr, mal weniger große – Ungewissheit des Entwicklungsstandes und der Einsatzmöglichkeiten von KI, der gesellschaftlichen Situation allgemein und der Situation von für KI-Anwendungen besonders interessante Sektoren im Speziellen und nicht zuletzt des rechtlich gesetzten Rahmens erschwert. All dies wirkt immerfort zusammen, bedingt und beeinflusst sich – ohne, dass ein Ende dieses dynamischen Geschehens zu erwarten wäre.

Doch es gibt durchaus Hoffnung, denn man sollte die Wirkmächtigkeit von Normierungsansätzen auch nicht unterschätzen. Das Beispiel der EU-Datenschutzgrundverordnung (GDPR) zeigt, dass Normierung möglich und wirkmächtig sein kann. Das Potenzial solcher Normierungen liegt womöglich nicht nur und vielleicht nicht einmal primär in ihrer unmittelbaren Wirkung auf die Gestaltung von Technologien, Produkten, Dienstleistungen und Prozessen, sondern darin, dass global eine Diskursverschiebung stattfand und stattfindet – schon eine oberflächliche Literaturrecherche hierzu ist äußerst instruktiv, denn seit 2018 erscheinen unzählige Publikationen dazu, in denen Abschätzungen darüber durchgeführt werden, wie sich die GDPR bspw. auf Branchen, Geschäftsmodelle, Technologien oder auch Forschungsdisziplinen auswirkt (nur einige und willkürlich ausgewählte Beispiele: Kretschmer, Pennekamp und Wehrle 2021; Li, Yu und He 2019; Lindgren 2018; Mangini, Tal und Moldovan 2020; die Beiträge in Slokenberga, Tzortzatou und Reichel 2021). Denn selbst in Ländern, in denen traditionell kein oder nur wenig Datenschutz betrieben wurde, finden nun jene Stakeholder\*innen – oder zumindest die entsprechenden Interessenvertretungen –, die für einen umfassenderen Datenschutz plädieren, deutlich mehr Gehör. Ob diesem Gesinnungswandel dann ein (neues) Verständnis für den Wert von Datenschutz und Privatsphäre zugrundliegt oder eher eigene (wirtschaftliche) Interessen, ist dabei zweitrangig. Mit anderen Worten: Das Potenzial bspw. des AIA der EU könnte darin liegen, einen ähnlichen Meinungswandel zu befördern, sodass ein grundsätzliches Verständnis darüber entsteht, dass eine komplexe und mit weitreichenden Folgen verbundene Technologie nicht unreguliert bleiben darf, sondern dass die Entwicklung und Durchsetzung von Normierungsansätzen notwendig sind. Klar muss aber auch sein, dass mit dieser Einsicht noch lange keine inhaltlichen Festlegungen einhergehen. Schließlich: Selbst, wenn man eine vorsichtig optimistische Haltung in Hinblick auf die Möglichkeit der KI-Normierung einnimmt, sollte man einen langen Atem mitbringen. Die Normierung dieser Technologie ist kein Sprint, sondern ein Marathon – und das nicht zuletzt, weil die schöne neue Welt der KI trotz aller Erwartungen in vielen Fällen noch gar nicht existiert (Bareis und Katzenbach 2021).

## Literatur

- Attenberger, U., und M. Reiser. 2022. Future perspectives: Wie beeinflusst künstliche Intelligenz die Entwicklung unseres Berufsfeldes? *Der Radiologe*, 62(3): 267–270. <https://doi.org/https://doi.org/10.1007/s00117-022-00969-w>
- Autor, D., D. Mindell und E. Reynolds. 2020. The work of the future: Building better jobs in an age of intelligent machines. MIT. <https://workofthefuture.mit.edu/wp-content/uploads/2021/01/2020-Final-Report4.pdf>
- Balsmeier, B., und M. Woerter. 2019. Is this time different? How digitalization influences job creation and destruction. *Research Policy*, 48(8), 103765. <https://doi.org/10.1016/j.respol.2019.03.010>
- Bareis, J., und C. Katzenbach. 2021. Talking AI into being: The narratives and imaginaries of national AI strategies and their performative politics. *Science, Technology, und Human Values*, 016224392110300. <https://doi.org/10.1177/01622439211030007>
- Bender, E. M., T. Gebru, A. McMillan-Major, und S. Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? Proceedings of the 2021 ACM conference on fairness, accountability, and transparency, 610–623. <https://doi.org/10.1145/3442188.3445922>
- Bhargava, A., M. Bester, und L. Bolton. 2021. Employees' perceptions of the implementation of robotics, artificial intelligence, and automation (RAIA) on job satisfaction, job security, and employability. *Journal of Technology in Behavioral Science* 6(1): 106–113. <https://doi.org/10.1007/s41347-020-00153-8>
- Bijker, W. E., T.P. Hughes, und T.J. Pinch. (Hrsg.). 1987. *The social construction of technological systems: New directions in the sociology and history of technology*. MIT Press.
- Bray, D., und H. von Storch. 2017. The normative orientations of climate scientists. *Science and Engineering Ethics*, 23(5): 1351–1367. <https://doi.org/10.1007/s11948-014-9605-1>
- Brougham, D., und J. Haar. 2018. Smart technology, artificial intelligence, robotics, and algorithms (STARA): Employees' perceptions of our future workplace. *Journal of Management und Organization* 24(2): 239–257. <https://doi.org/10.1017/jmo.2016.55>
- Buolamwini, Joy, und Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Proceedings of the 1st conference on fairness, accountability and transparency*, hrsg. S. A. Friedler und C. Wilson, 81: 77–91. PMLR. <https://proceedings.mlr.press/v81/buolamwini18a.html>
- Butcher, J., und I. Beridze. 2019. What is the state of artificial intelligence governance globally? *The RUSI Journal*, 164(5–6): 88–96. <https://doi.org/10.1080/03071847.2019.1694260>
- Cihon, P., M. M. Maas, und L. Kemp. 2020. Fragmentation and the future: Investigating architectures for international AI governance. *Global Policy*, 11(5): 545–556. <https://doi.org/10.1111/1758-5899.12890>
- Crotof, Rebecca. 2019. “Cyborg justice” and the risk of technological-legal lock-in. *Columbia Law Review*, 119(7): 233–251.
- Dengler, K., B. Matthes, und G. Wydra-Somaggio. 2018. Regionale Branchen- und Berufsstrukturen prägen die Substituierbarkeitspotenziale. IAB Kurzbericht 22/2018. <https://doku.iab.de/kurzber/2018/kb2218.pdf>

- Etzkowitz, H. 2003. Innovation in innovation: The triple helix of university-industry-government relations. *Social Science Information*, 42(3): 293–337. <https://doi.org/10.1177/05390184030423002>
- Etzkowitz, H., und L. Leydesdorff. 2000. The dynamics of innovation: From national systems and “mode 2” to a triple helix of university–industry–government relations. *Research Policy*, 29(2): 109–123. [https://doi.org/10.1016/S0048-7333\(99\)00055-4](https://doi.org/10.1016/S0048-7333(99)00055-4)
- Europäische Kommission. 2021. Vorschlag des Europäischen Parlaments und des Rates zur Festlegung harmonisierter Vorschriften für Künstliche Intelligenz (Gesetz über Künstliche Intelligenz) und zur Änderung bestimmter Rechtsakte der Union. [https://eur-lex.europa.eu/resource.html?uri=cellar:e0649735-a372-11eb-9585-01aa75ed71a1.0019.02/DOC\\_lundformat=PDF](https://eur-lex.europa.eu/resource.html?uri=cellar:e0649735-a372-11eb-9585-01aa75ed71a1.0019.02/DOC_lundformat=PDF)
- Frey, Carl Benedikt, und M. A. Osborne. 2013. The future of employment. Working Paper. Oxford Martin School. [https://sep4u.gr/wp-content/uploads/The\\_Future\\_of\\_Employment\\_ox\\_2013.pdf](https://sep4u.gr/wp-content/uploads/The_Future_of_Employment_ox_2013.pdf)
- Frey, C. B., und M. A. Osborne. 2017. The future of employment: How susceptible are jobs to computerisation? *Technological Forecasting and Social Change*, 114: 254–280. <https://doi.org/10.1016/j.techfore.2016.08.019>
- Gasser, U., und V. A. F. Almeida. 2017. A layered model for AI governance. *IEEE Internet Computing*, 21(6): 58–62. <https://doi.org/10.1109/MIC.2017.4180835>
- Gasser, Urs, und Carolyn Schmitt. 2020. The role of professional norms in the governance of artificial intelligence. In *The Oxford handbook of ethics of AI*, hrsg. M. D. Dubber, F. Pasquale, und S. Das, 140–159. Oxford University Press. <https://doi.org/10.1093/oxfordhb/9780190067397.013.8>
- Geburu, Timnit. 2020. Race and gender. In *The Oxford handbook of ethics of AI*, hrsg. M. D. Dubber, F. Pasquale, und S. Das, 251–269. Oxford University Press. <https://doi.org/10.1093/oxfordhb/9780190067397.013.16>
- Hamid, O. H., N. L. Smith, und A. Barzanji. 2017. Automation, per se, is not job elimination: How artificial intelligence forwards cooperative human-machine coexistence. 2017 IEEE 15th International Conference on Industrial Informatics (INDIN), 899–904. <https://doi.org/10.1109/INDIN.2017.8104891>
- Hennemann, M. 2018. Künstliche Intelligenz und Wettbewerbsrecht. *Zeitschrift für Wettbewerbsrecht*, 16(2): 161–184. <https://doi.org/10.15375/zwer-2018-0204>
- Jipp, Meike, und J. Steil. 2021. Steuern wir oder werden wir gesteuert? Chancen und Risiken von Mensch-Technik-Interaktion. In *Zusammenwirken von natürlicher und künstlicher Intelligenz*, hrsg. R. Haux, K. Gahl, M. Jipp, R. Kruse, und O. Richter, 17–34. Springer Fachmedien. [https://doi.org/10.1007/978-3-658-30882-7\\_3](https://doi.org/10.1007/978-3-658-30882-7_3)
- Kauffmann, Anna L., J. Hennrich, C. Buch, und T. Eymann. 2022. Künstliche Intelligenz in der Radiologie und Strahlentherapie aus der Perspektive von Ärzten und Medizinphysikern: Eine Interviewstudie. In *Künstliche Intelligenz im Gesundheitswesen*, hrsg. M. A. Pfannstiel, 617–638. Springer Fachmedien. [https://doi.org/10.1007/978-3-658-33597-7\\_29](https://doi.org/10.1007/978-3-658-33597-7_29)
- Kleesiek, J., J. M. Murray, C. Strack, S. Prinz, G. Kaissis, und R. Braren. 2021. Künstliche Intelligenz und maschinelles Lernen in der onkologischen Bildgebung. *Best practice onkologie*, 16(4): 176–185. <https://doi.org/10.1007/s11654-021-00298-9>

- Korinek, A., und J. E. Stiglitz. 2019. Artificial intelligence and its implications for income distribution and unemployment. In *The economics of artificial intelligence. An agenda*, hrsg. A. Agrawal, J. Gans, und A. Goldfarb, 349–390. University of Chicago Press.
- Kretschmer, M., J. Pennekamp, und K. Wehrle. 2021. Cookie banners and privacy policies: Measuring the impact of the GDPR on the Web. *ACM Transactions on the Web*, 15(4): 1–42. <https://doi.org/10.1145/3466722>
- Krug, H., U. Bittner, V. Rolfes, H. Fangerau und K. Weber. 2020. Verunsicherung des ärztlichen Selbstverständnisses durch Künstliche Intelligenz? Ein Überblick über potenzielle Auswirkungen ihres Einsatzes im ärztlichen Alltag. In *Unsicherheit in der Medizin (Jahrbuch für Ethik in der Klinik 2020)*, hrsg. A. Frewer, L. Bergemann, und E. Langmann, 195–210. Königshausen und Neumann.
- Kuipers, T. A. F. 2009. The gray area for incorruptible scientific research: An exploration guided by Merton's norms conceived as 'default-norms'. In *EPSA Epistemology and Methodology of Science*, hrsg. M. Suárez, M. Dorato, und M. Rédei, 149–164. Springer Netherlands. [https://doi.org/10.1007/978-90-481-3263-8\\_13](https://doi.org/10.1007/978-90-481-3263-8_13)
- Kuziemski, M., und G. Misuraca. 2020. AI governance in the public sector: Three tales from the frontiers of automated decision-making in democratic settings. *Telecommunications Policy*, 44(6), 101976. <https://doi.org/10.1016/j.telpol.2020.101976>
- Leydesdorff, L. 2000. The triple helix: An evolutionary model of innovations. *Research Policy* 29(2): 243–255. [https://doi.org/10.1016/S0048-7333\(99\)00063-3](https://doi.org/10.1016/S0048-7333(99)00063-3)
- Li, H., L. Yu, und W. He. 2019. The impact of GDPR on global technology development. *Journal of Global Information Technology Management* 22(1): 1–6. <https://doi.org/10.1080/1097198X.2019.1569186>
- Lindgren, P. 2018. GDPR regulation impact on different business models and businesses. *Journal of Multi Business Model Innovation and Technology* 4(3): 241–254. <https://doi.org/10.13052/jmbmit2245-456X.434>
- Macfarlane, B., und M. Cheng. 2008. Communism, universalism and disinterestedness: Re-examining contemporary support among academics for Merton's scientific norms. *Journal of Academic Ethics* 6(1): 67–78. <https://doi.org/10.1007/s10805-008-9055-y>
- Mangini, V., I. Tal, und A.-N. Moldovan. 2020. An empirical study on the impact of GDPR and right to be forgotten—Organisations and users perspective. *Proceedings of the 15th International Conference on Availability, Reliability and Security*, 1–9. <https://doi.org/10.1145/3407023.3407080>
- Merton, R. K. 1974 [1942]. The normative structure of science. In *The sociology of science*, hrsg. R. K. Merton, 267–278. University of Chicago Press.
- Meyer, G., und P. Sandøe. 2012. Going public: Good scientific conduct. *Science and Engineering Ethics*, 18(2): 173–197. <https://doi.org/10.1007/s11948-010-9247-x>
- Moriarty, P. 2011. Science as a public good. In *A manifesto for the public university*, hrsg. J. Holmwood, 56–73. Bloomsbury Academic.
- Moser, E. 2021. Gerechtigkeitsüberlegungen zur Robotersteuer: Sollen wir die Automatisierung besteuern, um Arbeitsplätze zu erhalten? *Momentum Quarterly*, 10(4): 188–260. <https://doi.org/10.15203/MOMENTUMQUARTERLY.VOL10.NO4.P208-222>
- ÓhÉigeartaigh, S. S., J. Whittlestone, Y. Liu, Y. Zeng, und Z. Liu. 2020. Overcoming barriers to cross-cultural cooperation in AI ethics and governance. *Philosophy und Technology*, 33(4): 571–593. <https://doi.org/10.1007/s13347-020-00402-x>

- Pechmann, L., M. Mildner, T. Suthau, und M. Leucker. 2022. Regulatorische Anforderungen an Lösungen der künstlichen Intelligenz im Gesundheitswesen. In *Künstliche Intelligenz im Gesundheitswesen*, hrsg. M. A. Pfannstiel, 175–198. Springer Fachmedien. [https://doi.org/10.1007/978-3-658-33597-7\\_8](https://doi.org/10.1007/978-3-658-33597-7_8)
- Raji, I. D., T. Gebru, M. Mitchell, J. Buolamwini, J. Lee, und E. Denton. 2020. Saving face: Investigating the ethical concerns of facial recognition auditing. Proceedings of the AAAI/ACM conference on AI, ethics, and society, 145–151. <https://doi.org/10.1145/3375627.3375820>
- Reddy, S., S. Allan, S. Coghlan, und P. Cooper. 2020. A governance model for the application of AI in health care. *Journal of the American Medical Informatics Association* 27(3): 491–497. <https://doi.org/10.1093/jamia/ocz192>
- Robbins, S., und A. van Wynsberghe. 2022. Our new artificial intelligence infrastructure: Becoming locked into an unsustainable future. *Sustainability*, 14(8): 4829. <https://doi.org/10.3390/su14084829>
- Roski, J., E. J. Maier, K. Vigilante, E. A. Kane, und M. E. Matheny. 2021. Enhancing trust in AI through industry self-governance. *Journal of the American Medical Informatics Association*, 28(7): 1582–1590. <https://doi.org/10.1093/jamia/ocab065>
- Schiff, D. S., K. Laas, J. B. Biddle, und J. Borenstein. 2022. Global AI ethics documents: What they reveal about motivations, practices, and policies. In *Codes of ethics and ethical guidelines*, hrsg. K. Laas, M. Davis, und E. Hildt, 23: 121–143). Springer International Publishing. [https://doi.org/10.1007/978-3-030-86201-5\\_7](https://doi.org/10.1007/978-3-030-86201-5_7)
- Schmitt, L. 2021. Mapping global AI governance: A nascent regime in a fragmented landscape. *AI and Ethics*. <https://doi.org/10.1007/s43681-021-00083-y>
- Schneider, D. 2022. Ethische und professionsspezifische Herausforderungen. Der Diskurs um algorithmische Systeme der Entscheidungsunterstützung im Kontext der Teilhabepflicht für Menschen mit Behinderung. In *Künstliche Intelligenz und Gesundheit. Ethische, philosophische und sozialwissenschaftliche Explorationen*, hrsg. A. Sonar, und K. Weber, 87–132. Steiner.
- Schneider, D., A. Sonar, und K. Weber. 2022. Zwischen Automatisierung und ethischem Anspruch – Disruptive Effekte des KI-Einsatzes in und auf Professionen der Gesundheitsversorgung. In *Künstliche Intelligenz im Gesundheitswesen*, hrsg. M. A. Pfannstiel, 325–348. Springer Fachmedien. [https://doi.org/10.1007/978-3-658-33597-7\\_14](https://doi.org/10.1007/978-3-658-33597-7_14)
- Slokenberga, S., O. Tzortatzou, und J. Reichel, J. (Hrsg.). 2021. GDPR and biobanking: Individual rights, public interest and research regulation across Europe (Bd. 43). Springer International Publishing. <https://doi.org/10.1007/978-3-030-49388-2>
- Smuha, N. A. 2021. Beyond a human rights-based approach to AI governance: Promise, pitfalls, plea. *Philosophy and Technology*, 34(S1): 91–104. <https://doi.org/10.1007/s13347-020-00403-w>
- Sonar, A., und K. Weber. 2020. KI gestern und heute: Einsichten aus der Frühgeschichte der KI für aktuelle ethische Überlegungen zum Einsatz von KI in der Medizin. *Arbeit*, 29(2): 105–122. <https://doi.org/10.1515/arbeit-2020-0009>
- Suedekum, J. 2018. Digitalisierung und die Zukunft der Arbeit: Was ist am Arbeitsmarkt passiert und wie soll die Wirtschaftspolitik reagieren. IZA Standpunkte 90. Institut zur Zukunft der Arbeit. <https://ideas.repec.org/p/iza/izasps/sp90.html>

- Ulnicane, I., W. Knight, T. Leach, B. C. Stahl, und W.-G. Wanjiku. 2021. Framing governance for a contested emerging technology: Insights from AI policy. *Policy and Society*, 40(2): 158–177. <https://doi.org/10.1080/14494035.2020.1855800>
- Wang, W., und K. Siau. 2019. Artificial intelligence, machine learning, automation, robotics, future of work and future of humanity: A review and research agenda. *Journal of Database Management* 30(1): 61–79. <https://doi.org/10.4018/JDM.2019010104>
- Weber, K. 2018. Computers as omnipotent instruments of power. *The ORBIT Journal* 2(1): 1–19. <https://doi.org/10.29297/orbit.v2i1.97>
- Weber, Karsten. 2021. Soziale Roboter in der Science-Fiction: Ein Bericht über das Verhältnis von Herr und Knecht mit wechselnden Rollen für Menschen und Roboter. In *Soziale Roboter*, hrsg. O. Bendel, 539–556. Springer Fachmedien Wiesbaden. [https://doi.org/10.1007/978-3-658-31114-8\\_28](https://doi.org/10.1007/978-3-658-31114-8_28)
- Weber, Karsten. 2022. Das öffentliche Bild der Künstlichen Intelligenz. In *Künstliche Intelligenz und Gesundheit. Ethische, philosophische und sozialwissenschaftliche Explorationen*, hrsg. A. Sonar, und K. Weber, 207–226. Steiner.
- Wong, P.-H. 2020. Cultural differences as excuses? Human rights and cultural values in global ethics and governance of AI. *Philosophy and Technology*, 33(4): 705–715. <https://doi.org/10.1007/s13347-020-00413-8>
- Yeung, K., A. Howes, und G. Pogrebna. 2020. AI governance by human rights–centered design, deliberation, and oversight: An end to ethics washing. In *The Oxford handbook of ethics of AI*, hrsg. M. D. Dubber, F. Pasquale, und S. Das, 75–106. Oxford University Press. <https://doi.org/10.1093/oxfordhb/9780190067397.013.5>

---

# **Gesellschaftliche Risiken und soziale Dynamiken Künstlicher Intelligenz**





# Soziale Dynamik der Künstlichen Intelligenz

Hartmut Hirsch-Kreinsen

## Zusammenfassung

Die Fragestellung ist, wie die Dynamik der Künstlichen Intelligenz (KI), insbesondere ihr gegenwärtiger Boom, zu erklären sind. Der Verweis auf die schnellen Fortschritte der Informationstechnologien und der verschiedenen Methoden der KI der letzten Jahrzehnte kann diese Dynamik allein nicht zureichend erklären. Die These ist vielmehr, dass KI als „Promising Technology“ zu verstehen ist. Ihre Dynamik wird von einem Technologieversprechen über ihre besondere Leistungsfähigkeit und aussichtsreiche, völlig neue Anwendungspotenziale, geprägt. Voraussetzung und Folge der KI-Dynamik ist dabei die Etablierung und Durchsetzung eines spezifischen Innovationsmodus, der mit den tradierten und industrieorientierten Regelungen und Praktiken des Innovationssystems kaum mehr kompatibel ist und die Innovationspolitik vor neue Herausforderungen stellt

## Schlüsselwörter

Technologieversprechen • KI-Community • Schlüsseltechnologie • Besonderer Innovationsmodus • Innovationspolitik • Technikutopie

---

H. Hirsch-Kreinsen (✉)

Sozialforschungsstelle Dortmund – Fakultät für Sozialwissenschaften, TU Dortmund, Dortmund, Deutschland

E-Mail: [Hartmut.Hirsch-Kreinsen@tu-dortmund.de](mailto:Hartmut.Hirsch-Kreinsen@tu-dortmund.de)

© Der/die Autor(en), exklusiv lizenziert an Springer Fachmedien Wiesbaden GmbH, ein Teil von Springer Nature 2024

M. Heinlein und N. Huchler (Hrsg.), *Künstliche Intelligenz, Mensch und Gesellschaft*, [https://doi.org/10.1007/978-3-658-43521-9\\_9](https://doi.org/10.1007/978-3-658-43521-9_9)

201

## 1 KI als „Promising Technology“

Die kurze Geschichte der Künstlichen Intelligenz (KI) bzw. Artificial Intelligence (AI) ist ein Aufeinanderfolgen von Aufschwung- und Krisenphasen, wobei letztere als „AI-Winter“ bezeichnet werden. Spätestens seit den 2010er-Jahren lässt sich allerdings ein kontinuierlicher Aufschwung der KI-Entwicklung beobachten. Eine allgemein akzeptierte Interpretation ist, dass dieser Aufschwung vor allem auf den schnellen informationstechnologischen Fortschritt sowie die sich damit kontinuierlich erweiternden Nutzungsmöglichkeiten der KI zurückgeführt werden kann (z. B. Bibel 2014; Görz et al. 2021b). Diese Auffassung wird besonders auch für das Ende 2022 einer faszinierten Öffentlichkeit mit großem Werbeaufwand vorgestellte Sprachsystem ChatGPT formuliert. Damit bleibe einmal wieder ein „Schwellenjahr“ in der Entwicklung Künstlicher Intelligenz in Erinnerung (Gurevych 2023).

Ohne Frage ist die informationstechnologische Entwicklung ein zentraler Bestimmungsfaktor für die KI-Dynamik. Jedoch handelt es sich bei dieser Auffassung um eine lediglich technikzentrierte Sicht. Demgegenüber zeigt die sozialwissenschaftliche Technik- und Innovationsforschung, dass soziale und gesellschaftliche Bedingungen mit technischen und eben auch informationstechnologischen Entwicklungen in enger Wechselwirkung stehen, diese beeinflussen und erst ermöglichen. Danach kommt besonders *Erwartungen* über die zukünftigen Chancen neuer Technologien ein hoher Einfluss auf Verlauf und Richtung von technologischen Innovationen zu (Beckert 2016). Denn Erwartungen und damit verknüpfte Versprechungen, Visionen und Szenarien über mögliche und wünschenswerte Entwicklungsperspektiven reduzieren die Ungewissheiten und die Komplexität von Innovationen.

Die These ist daher, dass mit der KI ein Technologieversprechen verknüpft ist, das ihre Entwicklung maßgeblich bestimmt und vorantreibt. In Anschluss an die sozialwissenschaftliche Innovationsforschung kann KI als *Promising Technology* verstanden werden (van Lente und Rip 1998; Bender 2005; Borup et al. 2006; Konrad 2006). Ein Technologieversprechen ist als Narrativ zu verstehen, das Entwicklungspotenziale der KI, denkbare Möglichkeiten ihrer Anwendung und Nutzung sowie die damit zukünftig realisierbaren Gewinne in Aussicht stellt. Dabei muss das Versprechen allerdings mehr oder weniger plausibel mit dem jeweiligen Stand der Forschung und einsichtiger technologischer Potenziale begründet sein. Entworfen wird ein Technologieversprechen in der Regel von einer kleinen Gruppe in der Sache engagierter und von den formulierten Visionen überzeugter Protagonisten. Es offeriert interessierten Akteuren eine Innovationsperspektive, ist dann die Basis für den Entwurf einer Forschungsagenda,

begründet damit einen einzuschlagenden Weg für die Innovation und ist schließlich die Voraussetzung für ein koordiniertes Innovationshandeln heterogener Akteure (Hirsch-Kreinsen 2023, S. 17).

Diesen Zusammenhängen soll am Beispiel der KI-Entwicklung in der Bundesrepublik genauer nachgegangen werden.<sup>1</sup> F<sup>1</sup> Methodische Basis der folgenden Argumentation ist zum einen die Auswertung einer großen Zahl einschlägiger Dokumente, Preprints, politischer Verlautbarungen, Websites und Fachpublikationen. Zum Zweiten handelt es sich um die Ergebnisse von 19 Interviews mit KI-Expertinnen und Experten aus Wissenschaft und Wirtschaft, die zwischen Oktober 2021 und März 2022 durchgeführt worden sind. Zum Dritten bedient sich die Analyse einer Reinterpretation vorliegender eigener Forschungsergebnisse über den gesellschaftlichen Digitalisierungsprozess der letzten Jahre.

---

## 2 Dynamik und Phasen der KI-Entwicklung

Eng verknüpft mit der KI-Entwicklung in den USA und teilweise der in Großbritannien beginnt sich die KI in der Bundesrepublik ab den 1970er-Jahren zu etablieren. Grundlegender Bezugspunkt der deutschen KI-Entwicklung ist dabei, ähnlich wie in vielen anderen Ländern, die als „Geburtsstunde der KI“ bezeichneten Konferenz am Dartmouth College, New Hampshire, im Jahr 1956. Der Initiator dieser Konferenz, der Mathematiker und Informatiker John McCarthy, gilt als der „Gründungsvater“ der KI, der auf der besagten Konferenz den Begriff „Artificial Intelligence“ erstmals eingeführt hat (Konrad 1998; Koehler 2021). Danach ist ‚Künstliche Intelligenz‘ eine wissenschaftliche Disziplin, die das Ziel verfolgt, menschliche Wahrnehmungs- und Verstandesleistungen zu operationalisieren und durch Artefakte, kunstvoll gestaltete technische – insbesondere informationsverarbeitende – Systeme verfügbar zu machen“ (Görz et al. 2021, S. 2). In Deutschland lassen sich verschiedene Entwicklungsphasen identifizieren, die sich in Hinblick auf die jeweiligen wissenschaftlichen und technologischen Grundlagen, das darauf bezogene Technologieversprechen und

---

<sup>1</sup> Vgl. dazu ausführlich Hirsch-Kreinsen (2023) und Hirsch-Kreinsen und Krokowski (2023); zur Historie der KI in der Bundesrepublik liegen bisher Studien unterschiedlicher Provenienz vor; hervorzuheben ist vor allem die wissenssoziologische Studie von Petra Ahrweiler (1995a/b), die ein unverzichtbarer Ausgangspunkt für die hier vorgelegte Analyse war; weiterhin der technikhistorischen Ansatz von Seising und Dittmann (2018) sowie Rückblicke früher beteiligter Informatiker, z. B. Konrad (1998), Siekmann (2009) und Bibel (2014; 2020); zur internationalen Entwicklung, vgl. z. B. die Studie von Nilsson (2010).

die damit begründeten Zukunftserwartungen, die maßgeblich das Technologieversprechen vorantreibende KI-Community sowie den Grad der Institutionalisierung eines soziotechnischen Feldes der KI unterscheiden. Diese Phasen verlaufen in der Bundesrepublik mehr oder weniger parallel zur internationalen KI-Entwicklung, insbesondere zu der in den USA. Zugleich weisen sie aber auch nationalspezifische Besonderheiten auf.

## **2.1 Wissenschaftliche und kommerzielle Versprechungen: die ersten Phasen**

Für den Zeitraum bis ca. 2010 können – sehr vereinfacht – vier Entwicklungsphasen unterschieden werden.

### *1970er-Jahre: KI in der wissenschaftlichen Nische*

Die erste Phase kann als wissenschaftlich orientierte Gründungsphase charakterisiert werden. Spätestens ab Mitte der 1970er-Jahre beginnt sich die KI als wissenschaftliche Nischendisziplin, teilweise gegen nachhaltigen Widerstand der damals schon etablierten Informatik, zu etablieren. Maßgeblicher Akteur ist eine kleine Gruppe von Nachwuchswissenschaftlern, die in diesem Forschungsfeld große Zukunftschancen sehen. Es handelt sich um Akteure, die über mehr oder weniger ausgeprägte gemeinsame Orientierungen verfügen und informelle Kontakte pflegen und sich über ihre Forschungsideen und Perspektiven locker austauschen (Ahrweiler 1995a). Technologische Basis der Phase der 1970er-Jahre sind besonders heuristische Ansätze und symbolische KI auf der Basis regelbasierter Methoden (Bauberger et al. 2021, S. 908).

Der erste Schritt zu einer Institutionalisierung der frühen KI in der Bundesrepublik kann auf den Februar 1975 datiert werden, als in Bonn ein erstes organisiertes Treffen „Künstliche Intelligenz“ stattfindet. Rückblickend wird festgehalten, dass dieses Ereignis für die KI in der Bundesrepublik einen entscheidenden Startpunkt darstellt (Konrad 1998), da bei diesem Treffen erstmals systematisch wissenschaftliche Perspektiven und Zielsetzungen der KI im Rahmen wissenschaftlicher Fachvorträge entworfen und präzisiert wurden. Insofern wird in diesem Kontext erstmals ein Technologieversprechen formuliert, das mit wissenschaftlicher Zielrichtung Entwicklungs- und Institutionalisierungsperspektiven der KI expliziert. Diese Vision trifft auch auf eine wachsende Aufmerksamkeit der staatlichen Wissenschafts- und Forschungspolitik. Die neuen

Informationstechnologien und auch die frühe KI werden schon damals als „moderne Schlüsseltechnologien“ und als entscheidender „Produktivfaktor“ für die zukünftige ökonomische und gesellschaftliche Entwicklung angesehen. 3 F.<sup>2</sup>

Bemerkenswert ist, dass diese Entwicklung in Deutschland von der spätestens seit Ende der 1960er-Jahre in den USA und in Großbritannien vorgebrachten heftigen Kritik an der KI kaum tangiert wird. Die dortige Kritik ist als Reaktion auf die nicht eingelösten großen Versprechungen aus der Anfangszeit der 1950er-Jahre anzusehen. Die Folge ist international eine KI-Phase, die als erster „AI-Winter“ (Teich 2020) bezeichnet wird.

### *Kommerzieller Aufbruch in den 1980ern*

Anfang der 1980er-Jahre beginnt eine Phase, die unisono als Beginn eines Aufschwungs der KI in der Bundesrepublik bezeichnet wird (Ahrweiler 1995a; Bibel und Furbach 2018; Teich 2020). Ausgangspunkt ist ein grundlegend erneuertes Technologieversprechen, dessen primärer Fokus sich allerdings nun auf sich jetzt als aussichtsreich erscheinende kommerziellen Perspektiven der KI richtet. Expliziert wird diese Perspektive erstmals bei der 10. Jahrestagung der Gesellschaft für Informatik im Herbst 1980 mit einem Vortrag des amerikanischen Informatikers Edward A. Feigenbaum. Er betont die überaus große wirtschaftliche Bedeutung des damals so bezeichneten „Knowledge-Based Approach“ der KI, d. h. die Methoden der damals weithin verfolgten symbolischen KI und darauf basierender Expertensysteme. Er skizziert „große“ Potenziale in Hinblick auf Methoden der Wissensrepräsentation, die komplexe Anwendungen wie die Erkennung von Sprache, Analyse und Synthese in der Chemie sowie die medizinische Diagnostik, Therapie und Prospektion in der Mineralogie (Görz et al. 2021a, S. 8). Prognostiziert werden daher auch weitreichende ökonomische Effekte der erwarteten Nutzung von KI-Systemen (BMFT 1988, zit. n. Ahrweiler 1995a, S. 120).

Diese Versprechungen konvergieren mit einem wachsenden Interesse der damaligen Forschungspolitik an der KI. Ein unmittelbarer Anstoß hierfür ist die 1982 gestartete japanische Forschungsinitiative der „Fifth Generation Computer Systems“, mit der faktisch ein internationaler Technologiewettlauf in Hinblick auf Computer und KI eröffnet wird. Daneben treffen das Technologieversprechen und die damit verschränkten politischen Aktivitäten auf eine wachsende Resonanz bei Unternehmen aus der IT- und Elektrotechnischen Industrie, die sich mit Fragen der KI und mit möglichen Anwendungen von Expertensystemen zu befassen beginnen. Parallel dazu kann von einer fortschreitenden Institutionalisierung der KI im Wissenschaftssystem der Bundesrepublik gesprochen werden.

---

<sup>2</sup> So der damalige Forschungsminister Hans Matthöfer (zit. n. Ahrweiler 1995a: 85).

So wird im Bereich der Wissenschaft 1983 die bisherige Fachgruppe »Künstliche Intelligenz« innerhalb der Gesellschaft für Informatik zu einem Fachausschuss »Künstliche Intelligenz und Mustererkennung« mit thematisch ausdifferenzierten Fachgruppen deutlich aufgewertet; mit diesen Fachgruppen wird ein relativ breites thematisches Spektrum bearbeitet, das von Deduktionssystemen und Expertensystemen bis hin zu Bildanalyse und Robotik reicht (Bibel 2006). Darüber hinaus wird die erste KI-Proessur 1983 in der Bundesrepublik an der Universität Kaiserslautern geschaffen, auf die Jörg Siekmann berufen wird. Hervorzuheben ist in diesem Kontext besonders auch die vom damaligen Bundesministerium für Forschung und Technologie (BMFT) angestoßene Gründung des Deutschen Forschungszentrums für Künstliche Intelligenz (DFKI) in Saarbrücken bzw. Kaiserslautern 1988. Diese Einrichtung, die später als weltweit das größte und sehr renommierte KI-Forschungszentrum gilt, wird als Public-Private Partnership gegründet und es sind von seinem Anbeginn an große internationale Firmen beteiligt (Groth und Straube 2018, S. 16).

Eine forschungspolitische Reaktion auf den sich abzeichnenden internationalen Wettlauf ist auch die Initiative der EU-Kommission zu einem Europäischen Förderprogramm für die Informationstechnologie, die unter Beteiligung von großen europäischen Elektronikkonzernen 1984 im ersten »European Strategic Programme for Research in Information Technologies« (ESPRIT) mündet. Mit seiner Laufzeit von fünf Jahren und einem Etat von 750 Mio. ECU wird dieses Programm von damals Beteiligten als ein wichtiger Schub für die weitere KI-Entwicklung angesehen (Bibel und Furbach 2018). Zeitlich parallel dazu wird auf nationaler Ebene eine »Offensive der bundesdeutschen Forschungsförderung initiiert«, die sich vornehmlich auf KI und KI-verwandte Themen fokussiert und in einem Förderprogramm des BMFT »Informationstechnik« mündet. Eine hoch dotierte Fördermaßnahme richtet sich dabei auf die KI-Schwerpunkte »Wissensverarbeitung und Mustererkennung« (Ahrweiler 1995a, S. 117 ff.).

Folgt man damaligen Beobachtern, so gibt es zudem kaum ein deutsches Großunternehmen, das nicht umgehend ein KI-Team zusammenstellt, um eigene Expertensysteme zu entwickeln (Malsch et al. 1993, S. 11). Das Spektrum der Unternehmen reicht von damals etablierten Computerherstellern und elektrotechnischen Unternehmen wie IBM, Siemens-Nixdorf, AEG, Bull und Hewlett Packard über Softwarehäuser bis hin zu privaten Forschungseinrichtungen wie das damalige Battelle-Institut in Frankfurt, die mehrheitlich KI-Forschungszentren einrichten (McRobbie und Siekmann 1988, S. 25). Insgesamt wird damit eine KI-Dynamik erkennbar, die gleichermaßen von den Interessen der Wissenschaft, Forschungspolitik und Entwicklerunternehmen vorangetrieben wird. Im Ansatz

wird ein soziotechnisches Feld der KI erkennbar, dessen Kern eine locker vernetzte, aber auch bestimmende Konstellation zentraler Akteure aus Wissenschaft, Forschungspolitik und Teilen der Privatwirtschaft darstellt.

### *Krise Ende der 1980er*

Gleichwohl tritt spätestens ab dem Ende der 1980er-Jahre eine Situation ein, die als Krise der KI bzw. international als zweiter „AI-Winter“ bezeichnet wird (Teich 2020). Als unübersehbar erweist sich damals ein hohes Gap zwischen den Erwartungen seitens vieler Unternehmen und der Politik über die kommerzielle Nutzbarkeit von Expertensystemen einerseits und ihren tatsächlich realisierten Nutzungseffekten andererseits. Verfügbaren Daten zufolge verlaufen Diffusion und erste Anwendungen von Expertensystemen in den 1980er-Jahren nur sehr schleppend (z. B. Dostal 1993; Malsch et al. 1993). Die ursprünglichen technologischen und ökonomischen Versprechungen erweisen sich als überzogen. Dies führt zu massiven „Erwartungsenttäuschungen“ (Ahrweiler 1995a), die sich insbesondere bei vielen der beteiligten Unternehmen und in der Politik einstellen. Die eine Folge ist eine deutliche Reduktion des privatwirtschaftlichen Finanzierungsanteils an den inzwischen gegründeten KI-Instituten und damit deren Existenzgefährdung. Die andere Folge ist ein dramatischer Abbau von KI-orientierten FuE-Kapazitäten bei vielen Unternehmen. Bremsend wirkt auch die aufkeimende Skepsis auf der Seite der Forschungspolitik, die ebenfalls, von Anbeginn der Förderung, an kommerziellen Erfolgen der KI interessiert ist (Reuse 2008). Anders formuliert: Das soziotechnische Feld der KI der 1980er-Jahre und insbesondere die dominante Akteurskonstellation aus Wissenschaft, Politik und Unternehmen erweisen sich als instabil und erodieren.

### *Konsolidierung bis in die 2000er*

Indes lässt sich die folgende Entwicklungsphase, die den Zeitraum der 1990er- bis weit in die 2000er-Jahre hinein umfasst, keineswegs nur als Krisenphase verstehen. Vielmehr kann besonders für Deutschland, wie interviewte Wissenschaftler bestätigen, von einer langen Phase der Konsolidierung der KI gesprochen werden. So notiert ein Informatiker Anfang der 1990er-Jahre nach den Versprechungen der 1980er und der anschließenden Krise: „KI auf dem Weg in die Normalität“ (Brauer 1993). KI-Aktivitäten werden nur mehr partiell von der Forschungspolitik finanziert und die weitere KI-Entwicklung verläuft nur noch sporadisch in Kooperation mit Unternehmen. Es beginnt ein Prozess, der primär von der Wissenschaft vorangetrieben wird und kaum mehr die kommerziellen Dimensionen der 1980er-Jahre aufweist. Auch finden sich in dieser Phase kaum so weitreichende Versprechungen und Erwartungen wie zuvor. Allerdings kann

ein „Positionspapier“ zur Gründung der Arbeitsgemeinschaft der deutschen KI-Institute aus dem Jahr 1991 als eine Erneuerung der grundlegenden Forschungs- und Entwicklungsziele der KI interpretiert werden (Barth et al. 1991):

- Einerseits werden die grundlegenden Ziele der KI in Erinnerung gerufen, dass sie nämlich die Frage nach dem Wesen von Intelligenz und nach der technischen Realisierung von Funktionen, die sich davon ableiten ließen, verfolge.
- Andererseits wird relativ konkret hervorgehoben, dass sich KI als Wissenschaft verstehe, die Konzepte, Modelle, Methoden, Werkzeuge und Know-how schafft, um anwendbare intelligente Systeme im Zusammenwirken mit anderen zu realisieren.

Damit wird ein Prozess der Konsolidierung der KI programmatisch angestoßen. Festmachen lässt sich dieser an einer fortschreitenden Ausdifferenzierung der Themenfelder und Entwicklungsschwerpunkte sowie einer weiteren Institutionalisierung der KI. Die Themen umfassen vor allem Methoden des klassischen Maschinellen Lernens wie etwa die Bayessche Statistik und auch eine Revitalisierung der in den USA Ende der 1960er-Jahre verworfenen konnektionistischen KI-Konzepte, d. h. Neuronaler Netze (Bauberger et al. 2021; Görz et al. 2021b). Die institutionelle Voraussetzung hierfür beschreibt ein damals einflussreicher KI-Wissenschaftler: Trotz der Krise haben die Boomjahre der 1980er etwas Positives bewirkt: „Sie hinterließen eine Forschungsstruktur für die KI, die sonst in Europa bisher unerreicht ist und mit der japanischen und amerikanischen Infrastruktur vergleichbar ist“ (Siekman 1994, S. 24).

## 2.2 Boom seit den 2010er-Jahren

### *Weitreichendes Technologieversprechen*

Technologische Basis für die KI-Dynamik seit den 2010er-Jahren ist ein Entwicklungsschub, der als „*Big Bang of Deep Learning*“ bezeichnet wird (Görz et al. 2021a, S. 9). Angesprochen wird damit die Entwicklung Künstlicher Neuronaler Netze, von Methoden des Maschinellen Lernens sowie die wachsende Verfügbarkeit von großen Datenmengen. Darauf basiert ein Technologieversprechen, das neben kommerziellen und industriell-ökonomischen Zielen vor allem weitreichende gesellschaftliche Visionen umfasst. Dabei geht es um die Nutzbarkeit von KI in unterschiedlichsten sozialen Kontexten, die Verbesserung sozialer Lebens-



und Arbeitsverhältnisse und auch die Möglichkeit, mithilfe von KI die Herausforderungen der Umwelt und des Klimawandels zu bewältigen. Ein spezieller Fokus richtet sich in Deutschland allerdings auf industrielle KI-Anwendungen, wo massive Produktivitätsschübe durch KI vorausgesagt werden. Ein für die KI-Perspektive in Deutschland wichtiger diskursiver Bezugspunkt ist dabei der Anfang der 2010er-Jahre vorgestellte Vision von Industrie 4.0 mit ihren weitreichenden Versprechungen über die Modernisierung, das heißt eine durchgreifende Digitalisierung der Industrie. Insgesamt ist das Technologieversprechen das fortlaufend präzierte Ergebnis eines Diskussionsprozesses in den Jahren nach 2011, der von einer einflussreichen KI-Community vorangetrieben wird und in dessen Verlauf eine Vielzahl programmatischer Publikationen vorgelegt werden (z. B. Fachforum 2016; Bitkom/DFKI 2017; Bundesregierung 2018; EFI 2022).

### *Einflussreiche KI-Community*

Kern der KI-Community ist die seit den 1990er-Jahren konsolidierte Wissenschaftsdisziplin der KI. Sie weist, so interviewte Experten, eine besondere »Hartnäckigkeit« beim Verfolgen ihrer Ziele auf, die auf der kollektiven Überzeugung dieser Gruppe beruht, dass die KI eine große Zukunft habe und man den lang gehegten Visionen einer intelligenten Maschine nahekommen könne. Diese Community umfasst darüber hinaus Vertreter von IT- und Softwareunternehmen sowie einige an KI nach den 1990er-Jahren wieder interessierte Industrieunternehmen besonders aus der elektrotechnischen Branche. Auch spielen in diesem Kontext zunehmend interessierte Politikvertreter aus einer Reihe von Bundesministerien sowie führende Repräsentanten von Wirtschaftsverbänden eine Rolle. Eine zentrale institutionelle Voraussetzung für die Aktivitäten der KI-Community ist dabei die koordinierende Rolle der Akademie für Technikforschung (acatech). Die KI-Community formuliert im Wesentlichen das Technologieversprechen, nimmt starken Einfluss auf seine Konkretisierung und die Definition der Forschungsagenda und beeinflusst den öffentlichen Diskurs. Sie ist in der Lage, erhebliche Forschungsmittel und qualifizierte Wissenschaftler\*innen für ihre Zielsetzungen zu gewinnen und kann diese einflussreiche Position über den zunehmenden Erfolg des Technologieversprechens kontinuierlich ausbauen. Dabei gewinnt die KI-Community zunehmend Einfluss auf den fortlaufenden Prozess des Agenda Settings, d. h. die Festlegung und Fortschreibung von innovationspolitischen Zielen und sie beeinflusst maßgeblich Innovationsschwerpunkte und Ressourcenverteilung im gesellschaftlichen Innovationssystem.

*Politik: Agendasetting, Förderung, Koordination*

Das Technologieversprechen und die Aktivitäten der KI-Community treffen insbesondere bei der Innovationspolitik auf große Resonanz, die mit ihren vielfältigen Fördermaßnahmen maßgeblichen Einfluss auf die Präzisierung der Forschungsagenda und die Koordination der KI-Entwicklung nimmt. Dies geschieht in besonderer Weise mit einer programmatischen Erklärung der Bundesregierung, der *Strategie Künstliche Intelligenz der Bundesregierung* Ende 2018. Explizit bezieht sich diese Strategie auf den von der KI-Community vorangetriebenen Diskurs und die dabei formulierten Versprechungen. KI wird dabei als „Schlüsseltechnologie“ für die zukünftige gesellschaftliche Entwicklung bezeichnet, in die nachhaltig investiert werden müsse (Bundesregierung 2018, S. 4). Die Strategie lässt sich auf drei Ziele zuspitzen:

- Stärkung der Wettbewerbsfähigkeit Deutschlands und Europas durch die massive Förderung der KI generell,
- eine verantwortungsvolle und gemeinwohlorientierte Entwicklung und Nutzung von KI sowie
- die ethische, rechtliche, kulturelle und institutionelle Einbettung von KI in die Gesellschaft.

Damit greift die Politik explizit kritische Fragen der Beherrschbarkeit und Transparenz von KI-Prozessen und ihren oftmals als negativ konnotierten sozialen, ökologischen wie besonders auch ethischen Konsequenzen auf. Diese kritischen Themen werden daher auch, oft politisch initiiert von verschiedenen Gremien, von Kommissionen und Organisationen adressiert und mit Fragen der KI-Entwicklung und Anwendung verknüpft.

Insgesamt wird KI als innovationspolitischer ‚Leuchtturm‘ angesehen, mit dem sich Politik als zukunftsorientiert ausweisen kann und sich Legitimation beschafft. Zudem wird mit diesen innovationspolitischen Maßnahmen eine Vielzahl von Folgeaktivitäten auf den verschiedensten staatlichen Ebenen ausgebaut, intensiviert und neu angestoßen. Dies gilt gleichermaßen für entsprechende Initiativen auf der EU-Ebene.

*Legitimation durch den öffentlichen Diskurs*

Das Technologieversprechen und die innovationspolitischen Aktivitäten sind eng verknüpft mit einem intensiven gesellschaftspolitischen Diskurs über die generelle Bedeutung und die sozialen Konsequenzen der KI. Dabei ist das Technologieversprechen der KI nur schwer von der optimistischen Seite dieses Diskurses zu trennen. Es beeinflusst ihn, gewinnt aber auch Legitimation durch

den Bezug auf ihn und die geradezu euphorisch gefeierten und international breit inszenierten Highlights der Leistungsfähigkeit von KI, z. B. beim Schach oder dem Brettspiel Go. Demgegenüber bremsen KI-skeptische und dystopische Argumente in Hinblick auf soziale und ethische Herausforderungen die KI-Dynamik kaum. Vielmehr werden diese Einwände zunehmend Teil der Forschungsagenda selbst, werden politisch aufgegriffen und ihre Bearbeitung wird institutionell in einer Reihe von Gremien und Organisationen verankert.

### *Soziotechnisches Feld*

Strukturell etabliert sich mit dieser Dynamik ein soziotechnisches Feld der KI, dessen zentrales Merkmal eine stabile und eng vernetzte Entwicklerkonstellation aus Innovationspolitik, interessierten Teilen der Privatwirtschaft und der Wissenschaft ist.

Hervorzuheben ist dabei besonders der Einfluss einer zunehmenden Zahl von Unternehmen, die an KI als Entwickler, teilweise auch Anwender interessiert sind. Zum einen handelt es sich um IT- und technologieintensive Unternehmen aus etablierten Branchen, zum anderen um eine schnell wachsende Landschaft von hoch spezialisierten Start-ups. Für viele Entwicklungsunternehmen und einige Anwender aus den unterschiedlichsten Wirtschaftssektoren eröffnet KI neue Absatzperspektiven für neue und bislang nur schwer erschließbare Anwendungsfelder für digitale Technologien und entsprechende Produkte. Vielen interessierten Unternehmen geht es dabei primär um die Sicherung und den Ausbau ihrer Weltmarktposition durch KI-basierte Innovationen. Insofern steht das Technologieversprechen KI in der Tradition der seit Jahrzehnten laufenden Debatte über die Nutzungsmöglichkeiten von unterschiedlichsten Informationstechnologien für weltmarktorientierte Produktinnovationen.

Zwischen allen drei Bereichen – Wissenschaft, Unternehmen, Politik – spielen sich seit den 2010er Jahren in Deutschland in Hinblick auf die KI-Entwicklung stabile, komplementäre Interessen ein. Sie entwickeln sich im Kontext institutionalisierter Kommunikations- und Interaktionsbeziehungen etwa innerhalb der diversen Beratungsgremien der Innovationspolitik, den verschiedensten staatlich geförderten anwendungsorientierten Entwicklungsprojekten sowie den staatlichen Transfereinrichtungen. Darüber hinaus existieren dauerhafte enge Beziehungen zwischen Wissenschaft und Unternehmen im Kontext der Förderung von Instituten durch Unternehmen, gemeinsamer Forschungsaktivitäten wie beispielsweise auch gemeinsamer Publikationsaktivitäten bei Konferenzen.

### 3 Perspektiven der KI

Die bisherige Dynamik der KI legt die Frage nach den weiteren Perspektiven der KI-Entwicklung nahe. Entscheidend für Richtung und Reichweite der zukünftigen KI-Entwicklung ist dabei, inwieweit bisherige Grenzen und unbewältigte Herausforderungen der KI-Diffusion und Anwendung bewältigt und überwunden werden können. Je nachdem, wie diese Möglichkeiten eingeschätzt werden, werden sehr unterschiedliche Entwicklungsperspektiven der KI erwartet. Insofern wird die Frage nach den Perspektiven der KI-Entwicklung in der einschlägigen Literatur intensiv diskutiert und ist heftig umstritten (z. B. Ford 2018; Lenzen 2018; Görz et al. 2021b; Larson 2021).

#### *Grenzen und Herausforderungen*

Fragt man nach den bisherigen Grenzen und Herausforderungen der weiteren KI-Entwicklung, so ist ein ganzes Bündel sehr unterschiedlicher Faktoren und Bedingungen wie praktische Anwendungsprobleme und ökonomische Ungewissheiten zu sehen. Auch sind gesellschaftliche Rahmenbedingungen, besonders die offenen rechtlichen und ethischen Fragen, in den Blick zu nehmen und schließlich korrelieren diese Faktoren mit dem Stand der Entwicklung und den technologischen Defiziten und Systemgrenzen. Resümiert man die vorliegenden Befunde und den Stand der Debatte, so sind besonders die folgenden Herausforderungen hervorzuheben:

Als ein soziales und organisatorisches Dauerproblem der Entwicklung und Diffusion anwendungsfähiger KI-Systeme wird von den interviewten Expertinnen und Experten die Schwierigkeit eines effektiven Wissenstransfers zwischen verschiedenen beteiligten Akteuren und Wissensdomänen genannt. Denn, so die Begründung, die Integration von Wissensbeständen aus Entwicklungs- und Anwendungsbereichen ist eine essenzielle Voraussetzung für funktionierende KI-Systeme. Eine Entwicklung neuer und effektiver Modi des Wissenstransfers und der Kooperation zwischen den Akteuren aus verschiedenen Domänen stehe lediglich erst am Anfang (z. B. Brödner 2019; ten Hompel et al. 2019; Ecker et al. 2021).

Darüber hinaus wird auf grundlegende Funktionsprobleme und die oft nur begrenzte Leistungsfähigkeit vor allem von Systemen des Maschinellen Lernens verwiesen, die einer schnellen Diffusion und breiten Anwendung entgegenstehen:

- Zum einen betrifft dies die mangelnde Fähigkeit der Systeme zur Wahrnehmung von „offenen Welten“, d. h. nur schwer ex ante kalkulierbarer

Situationen. Treten unerwartete Ereignisse und nicht kalkulierbare Sonder-situationen auf, werde die mangelnde „Robustheit“ eines KI-Systems zum Problem (Lenzen 2018, S. 86).

- Zum Zweiten gibt es bislang keine Möglichkeit, das oft unverzichtbare Alltagswissen in die Systemprozesse einzubeziehen. Görz et al. zufolge gehört dazu nicht nur die Fähigkeit der KI, „zu einem gewissen Grad Abstraktionen zu leisten und Kausalitäten maschinell nachzuvollziehen, sondern auch das zu approximieren, was Menschen in besonderer Weise auszeichnet, nämlich Handlungen zu verstehen und zu erklären“ (Görz et al. 2021a, S. 10).
- Zum Dritten wird damit das schon seit den 1990er-Jahren diskutierte Problem der mangelnden Erklärbarkeit (Explainability) und fehlenden Transparenz der Prozesse des Maschinellen Lernens bezeichnet. Denn je risikoreicher deren Entscheidungen für menschliches Handeln, etwa bei Diagnosen und Therapie-vorschlägen in der Medizin, werden, desto wichtiger wird es sein zu verstehen, wie die Systeme eigentlich funktionieren. Vor allem geht es dabei um die Frage der Transparenz und der Sicherheit – wie und ob das System tatsächlich die Aufgabe löst, für die es eingesetzt wird.
- Zum Vierten bezeichnet die Frage nach der Datenqualität und der für bestimmte Anwendungsfälle benötigten Datenmenge ein nicht endgültig lösbares Problem. So ist vielfach unklar, ob ein System wirklich alle zur Verfügung stehenden Daten benötigt oder ob aus Rationalisierungsgründen die Menge der Daten reduziert werden kann. Umgekehrt geht es um das Problem, das oft nur zu geringe Mengen - von einem Experten als Small Data-Problem bezeichnet - und qualitativ sehr unterschiedliche Daten für einen Einsatz von lernenden Systemen verfügbar sind.

Diese Herausforderungen sind Gegenstand intensiver nationaler wie internationaler Forschungsaktivitäten. So stehen die Probleme der Explainability und Transparenz der Systemprozesse seit längerer Zeit im Fokus der FuE-Förderung und es sollen Systeme entwickelt werden, die robuster und vertrauenswürdiger als die bisherigen Methoden des Maschinellen Lernens sind (Lenzen 2018, S. 77 f.).

### *Ein neuer AI Winter?*

Angesichts der vielfältigen Anwendungs- und Systemprobleme ist eine ganze Reihe von Stimmen zu vernehmen, die mit einem erneuten Scheitern der hochfliegenden KI-Versprechen rechnen und einen neuen AI-Winter nicht ausschließen. Zur Begründung dieser Sichtweise werden mehrere Argumente angeführt: So sind die Jahre seit 2010 von einem neuen Hype-Zyklus charakterisiert, wie er bei KI schon mehrmals aufgetreten ist, der aber dauerhaft nicht aufrechtzuerhalten ist

(Larson 2021, S. 74 ff.). Denn aufgrund vieler Entwicklungs- und Anwendungsprobleme wird sich sehr bald herausstellen, dass die KI-Entwicklung keineswegs in Zukunft so bahnbrechend verlaufen wird, wie vielfach versprochen. Auch wird betont, dass die Leistungsfähigkeit der verfügbaren Technologie aktuell, wie früher schon, völlig überschätzt werde: „Ähnlich wie im KI-Winter Ende der 1960er- und 1980er-Jahre, als die vielversprechenden Entwicklungen in den KI-Labors sang- und klanglos an der praktischen Anwendung in der Realität scheiterten, besteht wieder einmal eine große Diskrepanz zwischen den Erwartungen an diese Technologien und ihren tatsächlichen Fähigkeiten.“ (Heimbrecht 2021).

Als Konsequenz wird prognostiziert, dass Erwartungen zurückgeschraubt, Investitionen reduziert sowie das öffentliche Interesse erlahmen werden. Als eine wesentliche Ursache hierfür wird auch das bis heute unklare Verständnis, was unter KI eigentlich zu verstehen sei, angesehen. Das führe zu überzogenen Erwartungen, Missverständnissen und schließlich zu einer Depression (Piekniewski 2022). Denn beispielsweise würden viele Erwartungen, die sich im Grunde auf Entwicklungen der starken KI richten, an der nur begrenzten Leistungsfähigkeit spezialisierter schwacher Lösungen gemessen, sodass Desillusionierungen die Folge sind. Daher, so einige Kritiker\*innen nachdrücklich, befinde sich die gegenwärtige Dynamik an einem „tipping point“ und die weitere Entwicklung könnte in eine neue Stagnation und einen neuen Winter münden (Kalthener 2021, S. 193 ff.). Aus der Sicht mancher Beobachter gilt dies auch für die als völlig überzogen angesehenen Erwartungen, die sich mit den Großen Sprachsystemen wie den verschiedenen Varianten von GPT verbinden (z.B. Acemuglo 2024).

### *Weitreichende Versprechungen: Artificial General Intelligence*

Ganz in Gegensatz zu diesen pessimistischen Erwartungen werden auch zum wiederholten Male im KI-Diskurs weitreichende Versprechungen und Erwartungen formuliert. Ein dauerhaft anhaltender und weitreichender KI-Boom wird von vielen KI-Akteuren erwartet. So sagen innovationspolitische Prognosen für die Schlüsseltechnologie KI in Zukunft eine „fast exponentielle Dynamik“ voraus (Kroll et al. 2022, S. 31) und geradezu euphorisch wird von einem kommenden „Goldenen Zeitalter“ der KI gesprochen. Zur Begründung wird auf neue technologische Potenziale und auf die Erwartung einer weiterhin schnellen technologischen Entwicklung verwiesen. Insbesondere werden große algorithmische Fortschritte beim Maschinellen Lernen und beim Deep Learning in Verbindung mit der Verfügbarkeit größter Datensätze und Fortschritten beim schnellen, parallelen Rechnen vorausgesagt. Damit würden absehbar die bisherigen Grenzen der KI bewältigt und überschritten. Denn die neuen Systeme sind dann nicht mehr

nur für spezialisierte Anwendungen nutzbar, sondern sie können tendenziell in den verschiedensten Anwendungsfeldern eingesetzt werden (Kersting und Tresp 2019, S. 3).

Der Kern dieser optimistischen Perspektiven sind die von einer ganzen Reihe von Wissenschaftlerinnen und Wissenschaftlern dezidiert geäußerten weitreichenden Erwartungen über Fortschritte bei der KI-Entwicklung in den nächsten Jahren und die damit verknüpften Realisationsmöglichkeiten für Konzepte einer sog. starken KI bzw. einer Artificial General Intelligence. Soweit einschätzbar, ist in der nationalen wie besonders der internationalen KI-Community eine signifikante Zahl von Wissenschaftler:innen überzeugt, dass Technologien mit starker KI zukünftig und längerfristig Realität werden (Müller und Bostrom 2014; Ford 2018; Science Media Center 2021). Erwartet wird, dass auf der Basis weiterentwickelter Methoden insbesondere Neuronaler Netze, schnell steigender Rechnerkapazitäten und eines verfügbaren wachsenden Datenvolumens eine generelle, menschenähnliche Maschinenintelligenz gepaart mit Alltagsbewusstsein und Emotionalität in absehbarer Zeit entwickelt werden kann. Es sei daher durchaus denkbar, dass man Alltagsbewusstsein mathematisch abbilden könne (ebd., S. 7). Ein treibendes Motiv hierfür ist fraglos das seit Anbeginn der KI formulierte und immer wieder erneuerte Versprechen und die tiefe Überzeugung, dass man menschliche Intelligenz mit Computermodellen nachbauen und auch endlich verstehen könne. Und unverkennbar wird mit solchen Erwartungen an die bis in die Gründerzeit der KI zurückgehende Vision einer intelligenten Maschine angeknüpft, die dem Menschen weitgehend ähneln soll (Nilsson 2010, S. 528).

### *„Veralltäglichung“ der KI*

Mit anderen Prognosen wird hingegen eine differenzierte Sicht betont. Erwartet wird eine Entwicklung, die als „Veralltäglichung“ der KI gefasst werden kann. Bezeichnet wird damit eine Perspektive, der zufolge bestehende Ansätze und Methoden inkrementell weiterentwickelt, die existierenden Grenzen und Herausforderungen schrittweise bewältigt werden und damit immer leistungsfähigere Systeme der sog. schwachen KI zum Einsatz kommen können. Indes ist offen, ob die angestrebten stabilen KI-Lösungen allein auf der Basis lernender Systeme absehbar tatsächlich realisiert werden können. Vielmehr, so kritische Auffassungen auch interviewter Expertinnen und Experten, habe sich die oft vorgenommene Verkürzung durch die Gleichsetzung von KI mit Maschinellern Lernen „als Irrweg“ erwiesen. Trotz riesiger Datenmengen zum Training der selbstlernenden KI-Systeme und enormem Rechenaufwand seien oft keine erklärbaren und robusten Lösungen für die jeweiligen Problemstellungen gefunden worden. Daher müsse die KI-Entwicklung „... die zunächst etwas einseitige Überbetonung des

maschinellen Lernens aus Massendaten zugunsten einer Kombination mit modernsten symbolischen Verfahren aufgeben“ (Wahlster 2020). Dies impliziert auch, dass sich die KI-Entwicklung in Zukunft auf konkrete und relativ präzise definierbare Anwendungsfelder beziehen müsse. Diese Anwendungsbereiche werden sich allerdings ausweiten und zu einer zunehmenden Ausdifferenzierung der KI-Entwicklung führen.

Perspektivisch kann daher neben einer generellen Veralltäglichen im technologischen Sinn eine *Hybridisierung* der KI-Entwicklung erwartet werden. Konkreter, symbolische KI-Modelle, die für die semantische Wissensrepräsentation genutzt werden, werden mit Deep-Learning-Ansätzen und Maschinellen Lernverfahren wie aber zukünftig auch mit Konzepten Großer Sprachmodelle zur Erkennung und Extraktion von strukturierten Szenen-, Ereignis- und Situationsinformationen aus Datenströmen integriert. Ganz offensichtlich wird dieser Entwicklungsperspektive auch in der internationalen Debatte eine zunehmend zentrale Bedeutung zugemessen (z. B. Brachman und Levesque 2022).

Es spricht vieles dafür, dass Veralltäglichen der KI in Zukunft der dominante Pfad der KI sein wird. Denn einerseits sprechen viele Faktoren gegen einen erneuten AI-Winter. Zu nennen sind hier vor allem der erreichte Stand der Forschung und Entwicklung und die bisher getätigten Investitionen in diesen, die Vested Interests der etablierten KI-Community und das inzwischen fest institutionalisierte soziotechnische Feld der KI. Andererseits sprechen begründete Zweifel gegen den Pfad der Artificial General Intelligence. Denn fraglich ist, ob die hier virulenten grundlegenden technologischen Hürden und bislang völlig ungelösten Entwicklungsprobleme in absehbarer Zeit überwunden werden können und die verschiedentlich erwarteten Durchbrüche bei der KI-Entwicklung erreicht werden können, um damit einer starken KI bzw. einer Artificial General Intelligence näher zu kommen (Barthelmeß und Furbach 2021; Koehler 2021). Darüber hinaus sind aber auch die Argumente vieler KI-Kritiker zu sehen, denen zufolge diese Entwicklungsperspektive niemals realisiert werden könne. Denn, so ein seit langer Zeit viel diskutiertes grundlegendes Argument, menschliche Intelligenz und Denken können nicht mit Berechnen gleichgesetzt und technisch simuliert werden und das Ziel einer humanorientierten KI sei letztlich ein Mythos (Brödner 2019; Dickson 2021; Larson 2021).



## 4 Ein spezifischer Innovationsmodus KI

Voraussetzung und Folge der KI-Dynamik, vor allem ihrer weiteren Entwicklung ist die Etablierung und Durchsetzung eines KI-spezifischen Innovationsmodus, der mit den tradierten und industrieorientierten Regelungen und Praktiken des deutschen Innovationssystems kaum mehr kompatibel ist und die Innovationspolitik vor große Herausforderungen stellt (Botthof et al. 2020; 2023). Er ist Konsequenz der spezifischen Entwicklungserfordernisse der KI sowie der damit einhergehenden bisherigen Grenzen ihrer Diffusion und Anwendung. Resumiert man die vorliegenden Forschungsergebnisse, so lassen sich seine zentralen Merkmale stichwortartig wie folgt zusammenfassen:

- *Transdisziplinarität und Wissenstransfer*: KI-Innovationen finden tendenziell in transdisziplinären Kontexten statt, die bislang voneinander disziplinär und sektoral abgeschottete Wissensdomänen integrieren; insbesondere ist ein eng gekoppelter Wissenstransfer zwischen der Entwicklungsseite und Anwendungsbereichen zunehmend unabdingbar.
- *Fluide Kooperationsprozesse und offene Innovation Ecosystems*: Relevantes Innovationswissen ist oft nur jenseits der Grenzen von etablierten Unternehmen verfügbar, d. h. hohe Bedeutung gewinnen für KI-Innovationen offene Innovation sog. Ecosysteme, insbesondere mit dem Einschluss spezialisierter Start-ups.<sup>3</sup>
- *Agile Innovationsstrategien*: Im Kontext von KI-Innovationen wandeln vor allem etablierte Unternehmen ihre Strategien in Richtung strategisch vorausschauender und zugleich flexibler Innovation, um die Potenziale der schnellen technologischen Entwicklung zu nutzen.
- *Heterogene Akteure*: KI-Innovationen finden zunehmend im Kontext heterogen zusammengesetzter Akteurskonstellationen statt; neben Informatik, KI und Ingenieurwissenschaften handelt es sich um weitere Akteure aus den verschiedensten Anwendungsfeldern, wissenschaftlichen Disziplinen und gesellschaftlichen Teilbereichen.
- *Erweiterte Kompetenzprofile*: KI-Innovationen basieren auf deutlich erweiterten Kompetenzprofilen jener Akteure, die sich mit der Entwicklung und Einführung von KI-basierten Systemen befassen. Dabei handelt es sich um

---

<sup>3</sup> Dieser Begriff wird unterschiedlich und häufig unscharf verwendet. Definiert werden kann er wie folgt: »The evolving set of actors, activities, and artifacts, and the institutions and relations, including complementary and substitute relations, that are important for the innovative performance of an actor or a population of actors« (Granstrand und Hogersson 2020).

einen Generaltrend, wonach mit KI in nahezu allen Beschäftigtensegmenten eine Verschiebung hin zu extrafunktionalen Kompetenzen wie Umgang mit KI-Systemen und Gestaltung des KI-Kontextes beobachtbar ist.

- *Sozio-technischer Charakter von Innovationen:* KI-Innovation weisen immer seltener nur eine technologische Seite auf. Vielmehr schließen erfolgreiche KI-Innovationen komplementäre soziale Innovationen systematisch mit ein. Dies betrifft etwa die Bereitstellung der erforderlichen Daten und ihre laufende Qualitätskontrolle, die sozial und ethisch orientierte Systemanwendung, eine oft erforderlich KI-orientierte Strukturierung und Organisation der Anwendungsbereiche sowie die Verfügbarkeit auf KI ausgerichtete Kompetenzen des zuständigen Personals. Dies ist etwa die Voraussetzung dafür, dass ökonomische Effekte auf Unternehmensebene wie auch ein wünschenswerter Wandel in der Gesellschaft insgesamt erreicht werden.
- *Ausdifferenzierung von Innovationsebenen:* KI-Innovationen verlaufen verstärkt auf den unterschiedlichen sozialen Ebenen und sind eingebunden in die verschiedensten Akteursgruppen sowohl in subnationaler als auch in internationaler Hinsicht und es entwickelt sich über bisher existierende Strukturen hinaus ein Mehrebenensystem der Innovation.

Der spezifische KI-Innovationsmodus und bisher vorliegende Entwicklungs- und Anwendungserfahrungen von KI implizieren neue Anforderungen für die tradierte Innovationspolitik. Folgt man vor allem auch den Hinweisen der interviewten Expertinnen und Experten, so können diese wie folgt umrissen werden:

- *Flexibilisierung des Wissenstransfers und Förderung von Innovation Ecosystems,*
- *Technologieoffenheit und verstärkter Fokus auf hybride KI-Lösungen,*
- *Intensivierung und Ausbau von Maßnahmen der Kompetenzentwicklung,*
- *Förderung von Konzepten sozio-technischer Systemgestaltung bei KI-Anwendungen,*
- *agile Fördermodelle mit kurzen Feedbackzyklen, aber auch Ausbau langfristig orientierter Grundlagenforschung,*
- *verstärkte Förderung von Transparenz und Erklärbarkeit bei Methoden des Maschinellen Lernens, einschließlich ethisch orientierter Lösungsansätze.*

Unisono wird in der innovationspolitischen Debatte schließlich betont, dass der Wandel der Innovationspolitik verknüpft sein sollte mit ihrer verstärkten *Missionsorientierung*. Im Kern geht es dabei um eine Abkehr von der Rolle des Staates als Rahmensetzer hin zu einem richtungsgebenden und innovations-treibenden Staat (Botthof et al. 2023). Eine zentrale Rolle kommt absehbar

hierbei der Fokus auf die Schlüsseltechnologie KI und das damit verknüpfte Technologieversprechen zu.

---

## 5 Zur Rhetorik des Technologieversprechens

Abschließend ist zu fragen, welches die Gründe für die nachhaltige Überzeugungskraft des Technologieversprechens KI sind. Denn angesichts der Abfolge von Auf- und Abschwungphasen der KI-Dynamik und auch der vielfältigen Grenzen der zukünftigen Entwicklung sowie den skeptischen und kritischen Gegenargumenten wäre es nicht überraschend, wenn diese verblasst und daran geknüpften Erwartungen dauerhaft in Enttäuschungen umschlagen. Hinzu kommt, dass der Ausgang von Innovationsprozessen grundsätzlich ungewiss und es ex ante keineswegs ausgemacht ist, ob die Versprechungen tatsächlich realisierbar sind. Für die Adressaten, nicht nur für die Fachöffentlichkeit, sondern auch für die politische und gesellschaftliche Öffentlichkeit, muss das Technologieversprechen daher möglichst glaubhaft sein. Dies ist allerdings sehr voraussetzungsvoll. Das Technologieversprechen muss eine Rhetorik aufweisen, die überzeugend ist und eine breit gestreute Anschlussfähigkeit ermöglicht. Die hier wirksamen Mechanismen können durch Aspekte wie kommunikative Generalisierungen und Reduktion gesellschaftlicher Komplexität, De-Kontextualisierung der Argumente, unabwiesbare Aktualität, Quantifizierbarkeit und alltagsweltliche Relevanz charakterisiert werden (Kieser 1996; Hirsch-Kreinsen 2016; Madsen 2019). Diese Aspekte lassen sich wie folgt genauer begründen:

- *Mehrdeutige Metaphorik:* Kern des Technologieversprechens ist die schon erwähnte Mehrdeutigkeit des Begriffs KI. Die Metaphorik zur pauschalen Beschreibung dieser Technologie wie ‚intelligent‘, ‚lernend‘ oder auch ‚autonom‘ ist ebenso irreführend wie leicht zugänglich. Die technologische Vision scheint daher klar und überzeugend zu sein und ist auf den ersten Blick nur schwer bestreitbar.
- *Unausweichlichkeit:* Die breite Nutzung der KI hat aufgrund ihrer schnellen Entwicklung und ihrer wiederholten politischen Zuschreibung als ‚Schlüsseltechnologie‘ für die zukünftige gesellschaftliche Entwicklung einen geradezu unabwiesbaren Charakter.

- *Beschleunigtes Wachstum*: KI ist nicht nur die Voraussetzung für ein beträchtliches und nachhaltiges Wachstum, sondern auch für einen nationalen Erfolg im globalen Technologiewettbewerb insbesondere für den in Deutschland bedeutsamen industriellen Sektor. Dabei ist die Mehrzahl der Prognosen zwar eindrucksvoll, aber nicht validierbar.
- *Generalisierung von Einzelfällen*: Die weitreichenden ökonomischen Versprechungen und die Botschaft an Unternehmen, die neue Technologie möglichst schnell und umfassend einzuführen, wird durch den ständigen und intensiven Verweis auf durch KI erfolgreiche Unternehmen unterstrichen. Die Botschaft an die Skeptiker ist: Wer mitmacht, wird nahezu zwangsläufig Erfolg haben.
- *Sicherung von gesellschaftspolitischer Legitimation*: Betont wird, dass sich mit KI positive und gesellschaftspolitisch wünschenswerte soziale Konsequenzen verbinden. Zudem wird eine erfolgreiche Bewältigung gesellschaftlicher und ökologischer Herausforderungen möglich werden.

Das Technologieversprechen wirkt allerdings nur dann wirklich überzeugend, wenn das Timing stimmt. Es muss den „Nerv der Zeit“ (Kieser 1996, S. 26) treffen. Ohne Frage trifft das für das Thema KI zu, denn es spricht den vorherrschenden Zeitgeist über die notwendige Modernisierung der Gesellschaft durch neue Technologien, insbesondere durch Digitalisierung, an.

Zusammenfassend: Das Technologieversprechen KI weist den Charakter einer nicht hinterfragbaren Technikutopie auf. Denn KI figuriert als das technologische Mittel, um die vielfältigen und drängenden gesellschaftlichen Herausforderungen zu bewältigen und eine bessere Gesellschaft zu realisieren. Dabei besteht das spezifisch Utopische in der grundlegenden Annahme, dass sich die Menschheit, entgegen immer wieder aufkeimender, technologisch begründeter Schreckens- und Katastrophenszenarien, dem technologischen Zuwachs an Verfügungsmacht über die Natur in ihrer Gesamtheit gewachsen zeigen werde (Münkler 1997, S. 62 f.).

Im Grunde geht es aber um mehr: Der alte technikutopische Traum von einer mithilfe von Technologie möglichen rationalen Planung und Steuerung gesellschaftlicher Entwicklung jenseits von intransparenter sozialer Komplexität, interessen geleiteten politischen Diskussionen und aufwendigen demokratischen Prozeduren scheint für viele Protagonisten in Reichweite zu liegen. Unnötig werden in dieser Sicht aufwendige Ursache-Wirkungs-Analysen, die konfliktthaltige Interpretation widersprüchlicher wissenschaftlicher Forschungsergebnisse sowie der interessenabhängige Entwurf von Handlungs- und Lösungsstrategien. Vielmehr verspricht die Technologie eine autonome und smarte, scheinbar objektive

Lösung für übermächtige gesellschaftliche Probleme wie die Klimakrise. Es handelt sich um jene Erwartung, die Evgeny Morozov als „Solutionismus“ gefasst hat (Morozov 2013, S. 5). Letztlich wird damit an die alte technokratische Gesellschaftsvision angeknüpft, wonach gesellschaftlicher Wandel weniger mit politischen Mitteln zu erreichen sei, als vielmehr grundlegende und wünschenswerte Veränderungen der Gesellschaft allein durch technologische Mittel möglich sind.

---

## Literatur

- Acemoglu, Daron. 2024. Get Ready for the Great AI Disappointment, <https://www.wired.com/story/get-ready-for-the-great-ai-disappointment/> (abgerufen am: 18.02.2024) (2024).
- Ahrweiler, Petra. 1995(a). *Künstliche Intelligenz-Forschung in Deutschland. Die Etablierung eines Hochtechnologie-Fachs*. Münster/New York.
- Ahrweiler, Petra. 1995(b). KI West und KI Ost: Die Institutionalisierung eines Hochtechnologie-Fachs in Deutschland. In *Soziologie und künstliche Intelligenz. Produkte und Probleme einer Hochtechnologie*, hrsg. Werner Rammert, 111–131. Frankfurt/New York.
- Barth, Gerhard, Thomas Christaller, Armin B. Cremers, Bernd Neumann, Franz Josef Radermacher, Bernd Radig, Michael M. Richter, Jörg H. Siekmann und Werner von Seelen. 1991. Künstliche Intelligenz: Perspektive einer wissenschaftlichen Disziplin und Realisierungsmöglichkeiten. *Informatik-Spektrum* 14(4): 201–206.
- Barthelmeß, Ulrike und Ulrich Furbach. 2021. Computer auf dem Weg zum Bewusstsein. FAZ. <https://www.faz.net/aktuell/wirtschaft/digitec/computer-auf-dem-weg-zum-bewusstsein-17421543.html?premium> (abgerufen am: 27.07.2021).
- Bauberger, Stefan, Birgit Beck, Aljoscha Burchardt und Peter Remmers. 2021. Ethische Fragen der Künstlichen Intelligenz. In *Handbuch*, hrsg. Günther Görtz et al., 905–934. München.
- Beckert, Jens. 2016. *Imagined Futures. Fictional Expectations and Capitalist Dynamics*. Cambridge/MA.
- Bender, Gerd. 2005. Technologieentwicklung als Institutionalisierungsprozess. *ZfS* 34(3): 170–187.
- Bibel, Wolfgang. 2006. The beginnings of AI in Germany. *Künstliche Intelligenz* 4: 48–54.
- Bibel, Wolfgang. 2014. Artificial Intelligence in a historical perspective. *AI Communications* 27: 87–102.
- Bibel, Wolfgang. 2020. On the Development of AI in Germany. *Künstliche Intelligenz* 34: 251–258.
- Bibel, Wolfgang und Ulrich Furbach. 2018. *Formierung eines Forschungsgebiets – Künstliche Intelligenz und Intellektik an der Technischen Universität München*. Deutsches Museum Preprint: 15.
- Bitkom, DFKI. 2017. Künstliche Intelligenz: Wirtschaftliche Bedeutung, gesellschaftliche Herausforderungen, menschliche Verantwortung. Positionspapier. <https://www.bitkom>.

- [org/Bitkom/Publicationen/Entscheidungsunterstuetzung-mit-Kuenstlicher-Intelligenz.html](https://www.bitkom.org/Bitkom/Publicationen/Entscheidungsunterstuetzung-mit-Kuenstlicher-Intelligenz.html) (abgerufen am: 25.06.2021).
- Borup, Mads, Nik Brown, Kornelia Konrad und Harro van Lente. 2006. The Sociology of Expectations. In *Science and Technology. Technology Analysis & Strategic Management* 18(3/4): 285–298.
- Botthof, Alfons, Jakob Edler, Katrin Hahn, Hartmut Hirsch-Kreinsen, Matthias Weber und Jan Wessels. 2020. *Transformation des Innovationssystems: Neue Anforderungen an die Innovationspolitik*. Karlsruhe.
- Botthof, Alfons, Jakob Edler, Katrin Hahn, Hartmut Hirsch-Kreinsen, Matthias Weber und Jan Wessels. 2023. *Transformative und agile Informationssysteme. Neue Praktiken und innovationspolitische Herausforderungen*. Frankfurt.
- Brachman, Ron und Hector Levesque. 2022. *Machines like Us. Toward AI with Common Sense*. Cambridge/MA.
- Brauer, Wilfried. 1993. KI auf dem Weg in die Normalität. *KI* 7(3): 85–91.
- Brödner, Peter. 2019. Grenzen und Widersprüche der Entwicklung und Anwendung »Autonomer Systeme«. In *Autonome Systeme und Arbeit. Perspektiven, Herausforderungen und Grenzen der Künstlichen Intelligenz in der Arbeitswelt*, hg. v. Hartmut Hirsch-Kreinsen und Anemari Karacic, 69–100. Bielefeld.
- Bundesregierung. 2018. Koalitionsvertrag zwischen CDU, CSU und SPD. <https://www.bundesregierung.de/breg-de/themen/koalitionsvertrag-zwischen-cdu-csu-und-spd-195906> (abgerufen am: 30.09.2020).
- Dickson, Ben. 2021. Abductive inference: the blind spot of artificial intelligence. <https://bdt.echtalks.com/2021/09/20/myth-of-artificial-intelligence-erik-dduniekon/> (abgerufen am: 14.02.2022).
- Dostal, Werner. 1993. Expertensysteme und Beschäftigung – Gibt es derzeit erkennbare Auswirkungen von Systemen Künstlicher Intelligenz auf Beschäftigung und Berufe? *Mitteilungen aus der Arbeitsmarkt- und Berufsforschung* 26(1): 63–76.
- Ecker, Wolfgang, Carl-Helmut Coulon und Markus Kohler. 2021. *KI in die Anwendung bringen – Eine Gemeinschaftsaufgabe für Hochschulen, Forschungseinrichtungen Unternehmen und Politik. Whitepaper aus der Plattform Lernende Systeme*. München.
- EFI (Expertenkommission Forschung und Innovation). 2022. *Gutachten zu Forschung, Innovation und technologischer Leistungsfähigkeit Deutschlands 2022*. Berlin.
- Fachforum Autonome Systeme im Hightech-Forum. 2016. *Autonome Systeme – Chancen und Risiken für Wirtschaft, Wissenschaft und Gesellschaft. Zwischenbericht*. München.
- Ford, Martin. 2018. *Architects of Intelligence: The truth about AI from the people building it*. Birmingham.
- Görz, Günther, Tanya Braun, Ute Schmid. 2021. Einleitung. In *Handbuch der Künstlichen Intelligenz*, hrsg. Günther Görz, Tanya Braun und Ute Schmid, 1–26. Berlin/Boston.
- Granstrand, Ove und Marcus Holgersson. 2020. Innovation ecosystems: A conceptual review and a new definition. *Technovation* 90–91.
- Groth, Olaf J. und Tobias Straube. 2018. *Vergleich nationaler Strategien zur Förderung von Künstlicher Intelligenz – Teil 3. Konrad-Adenauer-Stiftung*. <https://www.kas.de/documents/252038/4521287/Bewertung+der+deutschen+KI-Strategie+Teil+3.pdf/aa0ecb4e-3a71-de71-63ba-fb08bf72dd57?version=1.1&t=1559810781469> (abgerufen am: 13.05.2021).

- Görz, Günther, Tanya Braun und Ute Schmid. 2021a. Einleitung. In *Handbuch der Künstlichen Intelligenz*, hrsg. Görz, Günther, Ute Schmid, Tanya Braun. 6. Aufl. Berlin/Boston, 1–26.
- Görz, Günther, Ute Schmid und Tanya Braun (Hrsg.). 2021b. *Handbuch der Künstlichen Intelligenz*. 6. Aufl. Berlin/Boston: De Gruyter.
- Gurevych, Iryna. 2023. So begrenzt ist ChatGPT – noch. In *FAZ.net*. <https://www.faz.net/aktuell/wirtschaft/digitec/chatgpt-die-kuenstliche-intelligenz-ist-noch-sehr-begrenzt-18690464.html> (abgerufen am: 27.03.2023).
- Heimbrecht, Felix. 2021. KI: Der nächste Winter kommt bestimmt. <https://www.zukunftsinstitut.de/artikel/technologie/ki-der-naechste-winter-kommt-bestimmt/> (abgerufen am: 03.10.2021).
- Hirsch-Kreinsen, Hartmut. 2016. »Industry 4.0« as Promising Technology: Emergence, Semantics and Ambivalent Character. *Soziologisches Arbeitspapier Nr. 48/2016*. Dortmund.
- Hirsch-Kreinsen, Hartmut. 2023. *Das Versprechen der Künstliche Intelligenz. Gesellschaftliche Dynamik einer Schlüsseltechnologie*. Frankfurt/New York.
- Hirsch-Kreinsen, Hartmut und Thorben Krokowski. 2023. *Technologieversprechen Künstliche Intelligenz*. Berliner Journal für Soziologie (i. E.).
- Kaltheuner, Frederike (Hg.). 2021. *Fake AI*. Manchester.
- Kersting, Kristian und Volker Tresp. 2019. *Maschinelles und Tiefes Lernen. Whitepaper aus der Plattform Lernende Systeme*. München.
- Kieser, Alfred. 1996. Moden & Mythen des Organisierens. *Die Betriebswirtschaft* 56(1): 21–39.
- Koehler, Jana. 2021. Zum Begriff der Künstlichen Intelligenz. In *Handbuch Künstliche Intelligenz und die Künste*, hrsg. Stephanie Catani und Jasmin Pfeiffer. Berlin (Preprint Version). [https://www.dfki.de/fileadmin/user\\_upload/import/11269\\_KI-Kunst-JKoehler-PreprintVersion.pdf](https://www.dfki.de/fileadmin/user_upload/import/11269_KI-Kunst-JKoehler-PreprintVersion.pdf) (abgerufen am: 23.01.2022).
- Konrad, Erhard. 1998. Zur Geschichte der Künstlichen Intelligenz in der Bundesrepublik Deutschland. In *Sozialgeschichte der Informatik. Studien zur Wissenschafts- und Technikforschung*, hrsg. Dirk Siefkes, Peter Eulenhöfer, Heike Stach und Klaus Städtler, 287–296. Wiesbaden.
- Konrad, Kornelia. 2006. The Social Dynamics of Expectations: The Interaction of Collective and Actor-Specific Expectations on Electronic Commerce and Television. *Technology and Strategic Management* 8(3/4): 429–444.
- Kroll, Henning, Hendrik Berghäuser, Knut Blind, Peter Neuhäusler, Fabian Scheifele, Axel Thielmann und Sven Wydra. 2022. Schlüsseltechnologien. *Studien zum deutschen Innovationssystem*, Nr. 7, Berlin.
- Larson, Erik J. 2021. *The Myth of Artificial Intelligence*. Cambridge/MA/London.
- Lenzen, Manuela. 2018. *Künstliche Intelligenz. Was sie kann und was uns erwartet*. München.
- Madsen, Dag Øivind. 2019. The Emergence and Rise of Industry 4.0 Viewed through the Lens of Management Fashion Theory. *Administrative Sciences* 9(3): 71.
- Malsch, Thomas, Reinhard Bachmann, Michael Jonas, Michael, Ulrich Mill, Ulrich und Susanne Ziegler. 1993. *Expertensysteme in der Abseitsfalle? Fallstudien aus der industriellen Praxis*, Berlin.
- McRobbie, Michael A. und Jörg H. Siekmann. 1988. Artificial intelligence: Perspectives and predictions. *AI Communications* 1(4): 16–29.

- Morozov, Evgeny. 2013. *To Save Everything, Click Here – Technology, Solutionism and the Urge to Fix Problems that Don't Exist*. New York.
- Müller, Vincent C. und Nick Bostrom. 2014. Future progress in artificial intelligence: A Survey of Expert Opinion. In *Fundamental Issues of Artificial Intelligence*, hrsg. Vincent C. Müller. Berlin.
- Münkler, Herfried. 1997. Moral und Maschine. Star Trek im Spannungsfeld von Sozialutopie und technologischem Fortschritt. In »Unendliche Weiten ...« *Star Trek zwischen Unterhaltung und Utopie*, hg. v. Kai-Uwe Hellmann und Arne Klein, 59–71. Frankfurt am Main.
- Nilsson, Nils J. 2010. *The Quest for Artificial Intelligence: A History of Ideas and Achievements*. Cambridge/MA.
- Piekniowski, Filip. 2022. AI winter is well on its way. <https://blog.piekniowski.info/2018/05/28/ai-winter-is-well-on-its-way/> (abgerufen am: 02.04.2022).
- Reuse, Bernd. 2008. Schwerpunkte der Informatikforschung in Deutschland in den 80er Jahren. In *Informatikforschung in Deutschland*, hrsg. Bernd Reuse und Roland Vollmar, 27–60. Berlin/Heidelberg.
- Science Media Center. 2021. Transkript »Künstliche Intelligenz – Stand der Forschung und Förderung in Deutschland«. [https://www.sciencemediacenter.de/fileadmin/user\\_upload/Press\\_Briefing\\_Zubehoer/Transkript\\_KI-Situation-Deutschland\\_SMC-Press-Briefing\\_2021-03-17.pdf](https://www.sciencemediacenter.de/fileadmin/user_upload/Press_Briefing_Zubehoer/Transkript_KI-Situation-Deutschland_SMC-Press-Briefing_2021-03-17.pdf) (abgerufen am: 04.02.2022).
- Seising, Rudolf und Frank Dittmann. 2018. Eine historisch-kritische Einführung. In *Formierung eines Forschungsgebiets – Künstliche Intelligenz und Intellektik an der Technischen Universität München*, hrsg. Wolfgang Bibel, Ulrich Furbach. Preprint 15, Deutsches Museum. München.
- Siekmann, Jörg. 1994. Künstliche Intelligenz: Von den Anfängen in die Zukunft. In *Die maschinelle Kunst des Denkens*, hrsg. Günther Cyranek und Wolfgang Coy, 11–40. Braunschweig/Wiesbaden.
- Siekmann, Jörg. 2009. Die Entwicklung der Disziplin in Deutschland. *Künstliche Intelligenz* 23(1): 47–52.
- Teich, Irene B. 2020. Meilensteine der Entwicklung Künstlicher Intelligenz. *Informatik Spektrum* 43(4): 276–284.
- ten Hompel, Michael, Rainer Anderl und Harald Schöning. 2019. *Schneller zum Markterfolg. Memorandum des Forschungsbeirats der Plattform Industrie 4.0*. München.
- van Lente, Harro und Arie Rip. 1998. Expectations in Technological Developments: An Example of Prospective Structures to be filled in by Agency. In *Getting New Technologies Together: Studies in Making Sociotechnical Order*, hrsg. Disco C. Cornelis und Barend van der Meulen, 203–229. Berlin/New York.
- Wahlster, Wolfgang. 2020. Deep Learning alleine reicht nicht. FAZ. <https://www.faz.net/-ikh-a3574> (abgerufen am: 18.09.2021).





# Risiken und Gefahren der ‚Künstlichen‘ ‚Intelligenz‘

Reinhard Kreissl und Roger von Laufenberg

## Zusammenfassung

Die Debatte über das Für und Wider von Künstlicher Intelligenz (KI) wird auf der einen Seite mit dem Argument der Optimierung menschlichen Handelns und Wirkens geführt. Auf der anderen Seite dienen überzogene Szenarien einer alles einnehmende Technologie als Gegenbeispiel. Dabei mangeln beide Argumentationsstränge häufig einer realistischen Einschätzung, Beobachtung und Analyse der Möglichkeiten und Grenzen von KI, inklusive der damit einhergehenden realen Risiken und Gefahren. Die Erwartungen an das ‚Können der KI‘ sind häufig eher illusorischer Natur und unter- bzw. überschätzen dadurch auch die Risiken. Bei genauerer Analyse wird ersichtlich, dass der Begriff KI in vielen Fällen irreführend ist – weder künstlich, noch intelligent – in welcher die Fehleinschätzung über das Können der KI begründet ist. In diesem Beitrag gehen wir dabei auf diese verzweigte Risiko-Debatte ein, analysieren die Aspekte der Künstlichkeit und Intelligenz der KI, bevor wir auf die unterschiedlichen Stränge der KI-Risiko Debatten eingehen – anhand von vier konkreten Einsatzszenarien.

---

R. Kreissl (✉) · R. von Laufenberg  
VICESSE Research GmbH – Vienna Centre for Societal Security, Paulanergasse 4/8,  
1040 Wien, Österreich  
E-Mail: [reinhard.kreissl@vicesse.eu](mailto:reinhard.kreissl@vicesse.eu)

R. von Laufenberg  
E-Mail: [roger.von.laufenberg@vicesse.eu](mailto:roger.von.laufenberg@vicesse.eu)

## Schlüsselwörter

Algorithmus • Datafizierung • Maschinelle Datenverarbeitung • Maschinelles Lernen • Menschliche Intelligenz • Mensch-KI-Interaktion • Regulierung • Technikoptimismus • Technikpessimismus

## 1 Einleitung

„AI is neither *artificial* nor *intelligent*“ (Crawford 2021, S. 8).

KI ist ein in der aktuellen politischen und gesellschaftlichen Debatte inflationär genutztes Etikett und auch innerhalb der wissenschaftlichen Diskussion finden sich verschiedene Definitionen des Begriffs und konfligierende Prognosen hinsichtlich der Entwicklung des Feldes. Politische und wissenschaftliche Debatten sind vielfältig verschränkt. Dies zeigt sich an den Kontroversen über die rechtliche Kontrolle von KI, sowie an den kulturindustriellen Dramatisierungen und ihren dystopischen und utopischen Zukunftsprojektionen, die auch in der akademischen Diskussion Niederschlag finden. In der Diskussion über mögliche Risiken und Gefahren von KI trifft man immer wieder auf unterschiedliche Skalierungen oder Differenzierungen bei der Analyse, die dann mehr oder weniger starke Formen der Dramatisierung von Risiken und Gefahren nahelegen. Je nachdem wie eng oder weit KI gefasst wird ergibt sich dabei ein anderes Bild. Ein breites gefasstes Verständnis von KI tritt oft in Verbindung mit dringlichen Appellen auf, wie etwa in dem Bericht der National Security Commission on AI (2021, S. 7) der US-amerikanischen Regierung:

„No comfortable historical reference captures the impact of (...) AI on national security. AI is not a single technology breakthrough, like a bat-wing stealth bomber. The race for AI supremacy is not like the space race to the moon. AI is not even comparable to a general-purpose technology like electricity. However, what Thomas Edison said of electricity encapsulates the AI future: “It is a field of fields ... it holds the secrets which will reorganize the life of the world.” Edison’s astounding assessment came from humility. All that he discovered was “very little in comparison with the possibilities that appear.“

Dieses Zitat ist ein Beispiel für den neuerdings erhobenen dringlichen Ton in diesem Bereich. KI erscheint als eine in mächtige ökonomische und politische Kontexte eingebettete, ungebändigte Kraft, die, ähnlich wie *die* Elektrizität, die Welt als Ganze verändern wird. Dementsprechend umfassend und dringlich gefasst sind die Empfehlungen zur (ökonomische) Förderung der KI-Entwicklung

in den USA, um ihre globale KI-Vorherrschaft beizubehalten und das volle Potential von KI ausschöpfen zu können. Das komplementäre kritische Gegenstück, aber mit einem ähnlich breit gefassten Verständnis von KI liefert die Analyse von Crawford (2021), die KI als extraktive Technologie mit global destruktiven Folgen beschreibt und neben den sozialen und ökonomischen Wirkungen auf die ökologischen Gefahren von KI – im generellen – hinweist.

Einen anderen Ton schlägt Marcus (2020) in seinem Paper an, in dem er einige der zentralen Themen aus seinem zusammen mit Davis verfassten Buch (Marcus und Davis 2019) *Rebooting AI* aufnimmt. Marcus nimmt dabei einen engeren Blick ein und begibt sich im Gegensatz zu den beiden anderen Positionen oben auf einer konzeptuellen Ebene. Er sieht den Fortschritt und das Potential der KI weitaus kritischer, plädiert für ein weniger anspruchsvolles Ziel und sieht die Gefahren und Risiken eher im Bereich der technologischen und konzeptuellen Defizite der KI-Forschung:

“Although nobody quite knows what deep learning or AI will evolve into the coming decades, it is worth considering both what has been learned from the last decade, and what should be investigated next, if we are to reach a new level.

Let us call that new level *robust artificial intelligence*: intelligence that (...) can be counted on to apply what it knows to a wide range of problems in a *systematic* and *reliable* way, synthesizing knowledge from a variety of sources such that it can reason *flexibly* and *dynamically* about the world, *transferring* what it learns in one context to another, in the way that we would expect of an ordinary adult.

In a certain sense, this is a modest goal, neither as ambitious or as unbounded as „superhuman“ or „artificial general intelligence [AGI]“ but perhaps nonetheless an important, hopefully achievable, step along the way—and a vital one, if we are to create artificial intelligence we can *trust* (...). Quite simply, if we cannot count on our AI to behave reliably, we should not trust it.” (Marcus 2020, S. 3).

Vergleicht man diese Diagnosen über den aktuellen Entwicklungsstand der KI so scheint im ersten Fall KI als ein kaum zu bändigendes – einmal potentiell produktives (Moorstedt 2022; Smith 2021), einmal destruktives (O’Connell 2017; Crawford 2021) – Phänomen. Im zweiten Fall von Marcus (2020) scheint KI als eine eher holprige Technologie von beschränkter Leistungsfähigkeit. Im Hinblick auf mögliche Risiken und Gefahren des Einsatzes von KI verweisen beide Diagnosen der KI zwar auf unterschiedliche Arten von Problemen. In der allgemeinen Diskussion über die Risiken und Gefahren, die sich auch in den beiden genannten Diagnosen widerspiegeln – sowohl im breit gefassten Verständnis von KI, als auch im konzeptuell enger gefassten Verständnis – konvergiert die Suche nach möglichen Lösungswegen rein auf die Datenbasis als Grundlage der KI.

Das Argument, auf das viele KI-Debatten am Ende hinauslaufen ist, dass die KI besser, Risiken und Gefahren gebannt werden könne, wenn mehr und bessere Daten verfügbar wären. Zwar wisse man nicht, wohin sich das Feld und die Technologie entwickelten, sicher aber ist, dass jeglicher Fortschritt – zum Guten wie zum Schlechten – von der ausreichenden Verfügbarkeit und Qualität maschinenlesbarer Daten abhängt. Dieser verkürzte Blick auf die Datenproblematik als Hauptursache der Risiken und Gefahren der KI klammert wesentliche Aspekte der KI aus. Bei genauerer Analyse wird nämlich ersichtlich, dass der Begriff KI in vielen Fällen irreführend ist – weder künstlich, noch intelligent – und es damit in der Diagnose der gesellschaftlichen Risiken von KI häufig einer realistischeren Einschätzung, Beobachtung und Analyse der Möglichkeiten und Grenzen von KI benötigt, inklusive der damit einhergehenden realen Risiken und Gefahren.

Bei näherer Betrachtung der existierenden Debatten über KI lassen sich also zwei Positionen identifizieren, die die Möglichkeiten und Risiken der KI zwar diskutieren, in ihrer Ursachenanalyse allerdings verkürzt sind:

1. KI-optimistische Perspektiven befeuern dystopische wie utopische Szenarien gleichermaßen. Sie verlaufen dabei parallel mit entsprechenden kulturindustriellen Popularisierungen und *parawissenschaftlichen Texten*: „programmatic descriptions, manifestos, and interviews“ (Jansen 2022). Als KI-optimistisch – wie der Position der National Security Commission on AI der USA – lassen sich jene Ansätze und Positionen bezeichnen, die eine zunehmende Leistungsfähigkeit maschineller Datenverarbeitungssysteme bei der Lösung von komplexen Problemen prognostizieren und an deren End-Entwicklung nicht selten eine Art künstliche Superintelligenz steht – Stichwort Transhumanismus (O’Connell 2017).
2. KI-pessimistische Perspektiven, die auf die nach wie vor begrenzte Leistungsfähigkeit von KI verweisen, sehen Risiken und Gefahren weniger in den wachsenden technologischen Kapazitäten als in dem unbedachten Einsatz von Systemen, die vorgeben ‚intelligente‘ Entscheidungen zu treffen, aber immer das Risiko von folgenreichen Fehlurteilen beinhalten (z. B. Lazer et al. 2014; Crawford, 2021).

Für die Auseinandersetzung mit Risiken und Gefahren der KI ergibt sich daraus eine vielschichtige Ausgangslage, die wir im vorliegenden Beitrag diskutieren und dabei in eine realistischere Debatte einordnen wollen, die sich ins besonders mit den beiden zentralen Termini der KI auseinandersetzt: ‚Intelligenz‘ und ‚Künstlich‘. Angefangen mit dem Begriff der ‚Intelligenz‘, führen wir in einem

ersten Schritt die Unterscheidung zwischen menschlicher Intelligenz und Prozessen maschineller ‚Intelligenz‘ ein. Der zweite Teil setzt sich mit dem zentralen Begriff ‚Künstlich‘ auseinander. Im Prozess der Datafizierung und Digitalisierung findet eine, über menschliche Handlungen vermittelte Transformation statt, um die Sphäre der analogen Welt maschinenlesbar, bzw. durch maschinelle Prozesse verarbeitbar zu machen. Daran anschließend entwickeln wir einen Vorschlag, wie sich die Vielzahl von Positionen in der Diskussion über Risiken und Gefahren der ‚Künstlichen‘ ‚Intelligenz‘ entlang unterschiedlicher Diskursstränge strukturieren lässt.

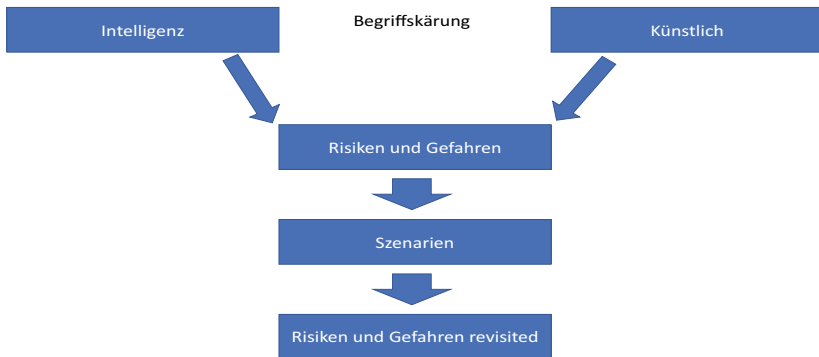
Bei genauerer Betrachtung solcher Diskursstränge zeigt sich die Einheit in der Differenz von Kritiker\*innen und Befürworter\*innen der KI. Die beigebrachten Belege für Risiken und für die spiegelbildlich dazu propagierten Hoffnungen technologischer Fortschritte schöpfen aus einem gemeinsamen Repertoire stilisierter Beispiele von KI-Anwendungen, um ihre Argumente durch Evidenz zu untermauern. Diese Beispiele dienen als dramatisierende Belege für unterschiedliche Beziehungen von KI-Anwendungen und menschlichen Nutzern, die ihre Wirkung durch gezielte Ausblendung jeglicher störender Randbedingungen entfalten. Stellt man diese oft ignorierten konkreten Bedingungen in Rechnung und analysiert den Kontext, in dem solche KI-Anwendungen mit ihren menschlichen Nutzern interagieren, verblasst die Aura des Dramatischen. Wir demonstrieren diese konstitutive Bedeutung des Anwendungskontexts für eine kritische Einschätzung des Potentials der KI anhand von Szenarien, die unterschiedliche Konstellationen von Mensch, Maschine und Welt erfassen.

Der Aufbau des Beitrags lässt sich schematisch wie in Abb. 1 darstellen: Wir beginnen mit einer etwas ausführlicheren Diskussion der Schlüsselbegriff: was ist intelligent und künstlich in der KI, schlagen sodann eine Typologie von Risiken/Gefahren vor und entwickeln eine Reihe von typisierten Szenarien der Beziehung zwischen Welt, Akteur und KI, um abschließend die Debatten über Risiken und Gefahren vor dem Hintergrund der typisierten Szenarien nochmals kritisch zu betrachten.

---

## 2 Künstliche ‚Intelligenz‘?

Neue Technologien regen die menschliche Fantasie an. Angetrieben von parawissenschaftlichen Texten (Jansen 2022) kündigen Wissenschaftler\*innen und Ingenieur\*innen, in stetiger Konkurrenz um knappe Forschungsmittel, regelmäßig kurz bevorstehende weitere Durchbrüche an. Sie stellen in naher Zukunft Lösungen für bisher ungelöste Probleme in Aussicht oder versprechen, durch



**Abb. 1** Aufbau des Beitrags. (Quelle: Eigene Darstellung)

weitere Forschung wissenschaftlich fundierte Antworten auf große Fragen der Menschheit zu liefern. Das Feld der KI steht hier in einer Reihe mit anderen Disziplinen, in denen es in der jüngeren Vergangenheit durch den Einsatz neuer aufwendiger Technologien gelungen ist, spektakuläre Forschungsergebnisse zu produzieren und diese, garniert mit Versprechungen auf epochale Durchbrüche, zu popularisieren.

Im Bereich der KI gehen diese Versprechungen in zwei Richtungen: zum einen könnten durch den Einsatz von KI viele komplexe Probleme, an deren Lösungen Menschen mit ihren Fähigkeiten gescheitert sind, bearbeitet werden (David et al. 2020; Service 2020). Auf der anderen Seite wird KI als die treibende Kraft zur Entlastung/Unterstützung bei der Erledigung diverser menschlicher Tätigkeiten angepriesen. Mit KI würden neue Formen der Mensch-Maschine-Interaktion ermöglicht, durch die sich industrielle Produktion und Distribution – Stichwort Industrie 4.0 (Zeba et al. 2021), Politik und Verwaltung (Reis et al. 2019), kulturelle Orientierung und soziale Koordination (Musk und Neuralink 2019), Mobilität (Coletta und Kitchin 2017), Konsum (Ariztía 2018) und Kommunikation (Gillespie 2018) nachhaltig verändern werden. Solche Versprechungen befeuern eine gesellschaftliche Debatte, in der gegensätzliche Positionen aufeinandertreffen. Die kontrovers geführte Debatte über Risiken und Gefahren sowie Chancen und Möglichkeiten von KI gewinnt ihre Einheit in der Differenz durch eine Reihe geteilter, meist unhinterfragter und oft impliziter Annahmen über das Verhältnis von Technologie und Gesellschaft oder die Beschaffenheit von Menschen und Maschinen. KI-optimistische wie KI-pessimistische Beiträge neigen dazu, menschliche Intelligenz nach dem technischen Vorbild eines Computers zu

modellieren, bzw. künstliche Intelligenz nach den gleichen Schemata der menschlichen Intelligenz zu modellieren und auf dieser Basis eine Art Leistungsvergleich anzustellen. Gleichzeitig gehen beide Positionen von einer grundlegenden ontologischen Differenz aus und damit münden solche Vergleiche dann entweder in den Fantasien eines disruptiven Transhumanismus oder in einer Kritik am instrumentellen Werkzeugcharakter der KI.

Die Einebnung der Differenz zwischen menschlicher und maschineller Informationsverarbeitung und die unhinterfragte Annahme der ontologischen Differenz verleiten zu unproduktiven Verkürzungen und verstellen den Blick auf Phänomene techno-sozialer Hybridisierung (Brown 2006). Die Kontroverse zwischen Verfechter\*innen und Kritiker\*innen der KI droht in einer Sackgasse zu landen. In dieser Situation lassen sich möglicherweise Differenzierungsgewinne erzielen, wenn man die Analyse tiefer legt und von den Prozessen der Informationsverarbeitung beim Menschen als biologischem Organismus ausgeht. Wenn man von menschlichen Informationsverarbeitungsprozessen ausgeht können die Merkmale spezifisch menschlicher Prozesse beschrieben werden, um sie dann als bedeutsame Differenzen für den Vergleich mit der Arbeitsweise maschineller Informationsverarbeitung zu nutzen. Das erlaubt eine differenziertere, nicht reduktionistische Erfassung von Unterschieden und Gemeinsamkeiten bei der Analyse der Interaktion zwischen KI und menschlichen Akteuren.

## **2.1 Menschliche Informationsverarbeitung: verkörpert, kontextuell, sozial**

Menschliche Informationsverarbeitung ist ein kontinuierlicher und verkörperter Vorgang, d. h. der menschliche Organismus verarbeitet fortlaufend eine Vielzahl von über seine verschiedenen Sinnesorgane eingehenden Umweltreize, die sich kontinuierlich mit der Änderung der Position des menschlichen Körpers in seiner Umwelt ändern und deren Verarbeitung die sensomotorische Stabilisierung des Körpers garantiert (Dreyfus 1999). Aus der Perspektive neurobiologischer Prozesse der menschlichen Informationsverarbeitung erscheint bereits ein passiver, monadischer Zustand, den man als prä-reflexives In-der-Welt-Seins bezeichnen könnte, als eine komplexe Leistung des menschlichen Organismus. Seine neurophysiologische Beschaffenheit hält eine Vielzahl von Prozessen der kontinuierlichen Verarbeitung von elektrischen Signalen am Laufen, die dem reflexiven Bewusstsein nicht zugänglich sind, dieses aber in vielfacher Weise beeinflussen können. Die physiologische Körperlichkeit stellt eine Art infrastruktureller Randbedingung für die Analyse weiterer evolutionärer Schritte in Richtung

menschlicher Intelligenz dar. Es handelt sich hier sozusagen um die neurobiologische Hardware derer sich alle höheren Funktionen und Leistungen menschlicher Informationsverarbeitung bedienen müssen. Jede weitere Modellierung menschlicher Handlungen und kognitiven Leistungen sollte diesen Modus des neurophysiologisch vermittelten, dynamischen In-der-Welt-Seins berücksichtigen. Menschliche Intelligenz und die Prozesse der Informationsverarbeitung sind somit auch in der konkreten Umwelt situiert und nutzen diese als eine kontextuelle Datenquelle.

Weiter muss berücksichtigt werden, dass menschliche Akteure bei der Lösung alltäglicher Probleme eine Vielzahl von unterschiedlichen Informationen nutzen, die in ihrer konkreten Situierung in einer strukturierten dreidimensionalen Umwelt verfügbar sind. Dies ermöglicht eine Orientierung in dem praktischen Handlungssetting der Welt in der Reichweite (Schütz 1972), in der die menschliche Informationsverarbeitung stattfindet. Zugleich ist diese Umwelt eine wichtige Ressource für externe Hilfsmittel, die bei Erledigung kognitiver Aufgaben genutzt werden können (Hutchins 2006). Menschliche kognitive Prozesse sind so auf vielfältige Weise situiert und mit der materiellen Umwelt, in der sie vollzogen werden, verbunden. Ein weiteres wichtiges Merkmal menschlicher Kognition ist die enge und kontinuierliche Kopplung von interner Informationsverarbeitung mit körperlicher Bewegung. Durch eine Veränderung der Blickrichtung werden neue Informationen zugänglich, durch das Berühren eines Objekts erhält der menschliche Organismus durch den Tastsinn zusätzliche Informationen über dessen Beschaffenheit sowie über den Geruchssinn möglicherweise nützliche olfaktorische Information.

Menschliche Intelligenz entwickelt sich nicht isoliert als individuelle Fähigkeit, sondern entfaltet sich in der sozialen Interaktion mit Artgenossen. Soziales Handeln lässt sich auf der Ebene menschlicher Informationsverarbeitung als eine strukturierte Abfolge situierter kognitiver Prozesse analysieren. Ego und Alter bilden für einander als Teil ihrer Umwelt eine privilegierte Datenquelle. Bei der Verarbeitung von Informationen, die bei der gegenseitigen Beobachtung anfallen, kommt eine wichtige Fähigkeit zum Tragen, die sich im Lauf der Evolution des menschlichen Informationsverarbeitungssystems herausgebildet hat (Tomasello 2010). Menschen und höher entwickelte Primaten haben die Fähigkeit, Artgenossen als eine spezifische Klasse von Objekten in ihrer Umwelt zu identifizieren und die aus der Beobachtung des Gegenübers gewonnenen Informationen in Stimuli für die Vorbereitung eigener Körperbewegungen zu übersetzen. Diese Kopplung geschieht automatisch durch Aktivierung einer spezialisierten Klasse von Spiegelneuronen (Gallese und Goldman 1998). Eine, nach derzeitigem Stand der Forschung, nur beim Menschen vorhandene weitergehende Form



der Kopplung nutzt die Fähigkeit, beobachtete Bewegungen zu imitieren, als Mittel zur Koordination der Aktivitäten von Alter und Ego. Wenn Ego den auf ein Objekt gerichteten Arm und die Blickrichtung von Alter imitiert und denselben Bewegungsablauf vollzieht, wird Alters Verhalten zum Signal, das Egos Aufmerksamkeit ebenfalls auf dieses Objekt lenkt.

Die durch gegenseitige Beobachtung vermittelte Imitation von Bewegungen stellt eine Art Protoform von bedeutungsvermitteltem Verstehen dar. Im Verlauf der evolutionären Entwicklung entstehen daraus generalisierte Deutungsmuster. Wenn die Imitation eines beobachteten Verhaltens, das die Aufmerksamkeit neu fokussiert wiederholt, zu einer positiven Verstärkung führt, wird dieses Verhalten mit dem dadurch erzielten Ergebnis verknüpft, d. h. es gewinnt symbolische Bedeutung als Zeichen in einer erlernten und im Gedächtnis gespeicherten Ursache-Wirkungs-Relation. Aus soziologischer Perspektive hat Mead (2015) diese Idee als Grundlage seiner Handlungstheorie entwickelt. Die Fähigkeit, das Verhalten von Artgenossen als absichtsvolle Gesten zu deuten, ist Ausdruck einer *Theory of Mind* und eine evolutionär wichtige Grundlage für die Entwicklung menschlicher Intelligenz durch kulturell situierte soziale Interaktion (Tomasello 2010). Symbolische Beziehungen sind das Verbindungsglied in der Ko-Evolution von Intelligenz und Kultur. Sie repräsentieren praktische Handlungsvollzüge und ermöglichen die kumulative Speicherung und Tradierung von nützlichem Wissen über die Welt. Sie entstehen durch soziales Handeln und strukturieren die Interaktion zwischen zwei Akteuren. Zugleich entlasten sie diese Prozesse, da zumeist ein Bruchteil der verfügbaren Umweltreize ausreicht, um eine zur Bearbeitung akuter Anforderungen brauchbare mentale Repräsentation des Status quo und der eigenen Position zu erstellen.

## **2.2 Künstliche Informationsverarbeitung: Mustererkennung und Situationspezifität**

KI reduziert die hier skizzierten Prozesse auf ein Problem der Mustererkennung durch Vergleich von in maschinenlesbarem Format eingehenden mit gespeicherten Daten (Kitchin 2016). Findet sich ein gespeichertes Datenmuster, das dem Input entspricht oder zeigt eine wiederkehrende Regelmäßigkeit, wirft das KI-System ein Ergebnis aus (Burrell 2016). Der Befund lautet dann entweder X (die Daten des Input) gleicht den gespeicherten Daten, die auf ein Exemplar der Kategorie Y verweisen, oder in der großen Menge der als Input verarbeiteten Daten findet sich eine Reihe ähnlicher Verbindungen, die es erlauben, die aus dieser Menge in eine Ordnung von strukturierten Einheiten zu aggregieren. Dabei

ergibt sich eine Reihe wesentlicher Unterschiede zur menschlichen Intelligenz. Die Bandbreite der in der menschlichen Informationsverarbeitung anfallenden Daten ist wesentlich größer. Auch stehen für die Verarbeitung flexiblere und komplexere Verfahren zur Verfügung, die es erlauben, gespeichertes Wissen mit verfügbarem Input zu vergleichen. In den meisten Fällen genügt ein vorläufig erzielter partieller Match zwischen Input und gespeichertem Wissen, um gezielt über weitere Schritte zu entscheiden. Dabei findet eine kontinuierliche Anpassung von erwartungsbasierten Hypothesen auf der Grundlage neu prozessierter Information statt.

Dies zeigt sich z. B. an der Verwendung deiktischer Ausdrücke in der Umgangssprache. Ein einfacher Aufforderungssatz wie: „Nimm das und stell es dorthin“ kann nur in einem dreidimensionalen räumlichen Setting verstanden werden, in dem sich Sprecher\*in und Hörer\*in gemeinsam befinden und unter Einbeziehung einer Zeigegeste des Sprechers oder der Sprecherin. Ein *natural language processing* (NLP) Programm kann diesen Satz zwar mithilfe maschineller Intelligenz in eine andere Sprache übersetzen – *take that and put it over there* – aber es kann den illokutionären Gehalt nicht verstehen und daraus die intendierten Handlungsvollzüge ableiten, die als Grundlage der Handlungskoordination zwischen Sprecher\*in und Hörer\*in dienen. Die für menschliche Intelligenz konstitutive Fähigkeit, die Umwelt des Organismus als Informationsquelle zu nutzen, ist in KI-Systeme nur bedingt implementierbar. In einem trivialen aber wichtigen Sinne sind diese Systeme nicht in der Welt. KI-gesteuerte Roboter können sich zwar in einer drei-dimensionalen physischen Umwelt bewegen (Miki et al. 2022). Aber sie sind nicht in der Lage, diese Bewegung in einen *sinnhaften* Zusammenhang einzuordnen und daraus Schlüsse zu ziehen und Entscheidungen für die Planung ihres Verhaltens zu treffen. Ihre Leistungsfähigkeit bei der Bearbeitung von Problemen und Bewältigung bestimmter Aufgaben basiert auf der verfügbaren Rechenleistung und der Kapazität zur Verarbeitung großer Datenmengen. Menschliche Intelligenz hingegen nutzt gespeichertes Wissen über Strukturen, symbolische Beziehungen und kulturelle Muster zur selektiven Steuerung der Aufmerksamkeit und Informationsverarbeitungsprozesse.

Die zu Beginn des Kapitels skizzierte Kontroverse über Risiken und Gefahren von KI legt somit in der verkürzten Art, wie sie geführt wird, eine falsche Fährte. Techno-soziale Systeme algorithmischer Datenverarbeitung gewinnen ihre Bedeutung als Elemente der kulturell kodierten materiell-symbolischen Umwelt. Als solche entfalten sie ihre Wirkung und die beobachtbaren Veränderungen sozialer, kommunikativer Praxis, die durch ihre Nutzung entstehen, sollten kritisch auf der Basis einer empirischen Beobachtung dieser Praxis analysiert werden. Die ontologische Differenz von menschlicher und maschineller Intelligenz lässt sich

in einem evolutionären Modell menschlicher Informationsverarbeitung überwinden, das den Begriff der Intelligenz aus dem Konzept der Handlungskoordination als Fähigkeit der Orientierung in einer konkreten Umwelt entwickelt (Todd and Gigerenzer 2000; Hutchins 2006; Mead 2015). Die „algorithmische Syntax“ menschlicher Intelligenz kann als formalpragmatische Regelsysteme der Handlungskoordination verstanden werden, die sich rekonstruktiv erschließen lassen. In den im Rahmen der KI-Forschung entwickelten Modellen werden solche Regelsysteme nicht angemessen abgebildet. Im Vergleich zu menschlichen Akteuren, die über ein komplexes multi-modales Sensorium der Sinneswahrnehmung verfügen, können KI-Systeme die Umwelt, in der sie agieren nur in sehr beschränktem Umfang als Ressource für ihre internen Prozesse nutzen.

---

### 3 ‚Künstliche‘ Intelligenz

Auch das Attribut ‚Künstlich‘ ist im Kontext von KI irreführend. Es verleitet zu falschen Analogien mit künstlichen Hüftgelenken, künstlichen Zähnen oder künstlichen Herzklappen. Maschinelle Intelligenz ist in diesem Sinne nicht künstlich. KI dient nicht wie eine Prothese als Ersatz für ein menschliches Organ dessen Funktion sie übernimmt. Das Attribut ‚Künstlich‘ verweist im Zusammenhang mit KI bestenfalls auf den artifiziellen Charakter des durch Datafizierung geschaffenen künstlich-digitalen reduzierten Universums, das gleichsam die Existenzbedingung jeder KI darstellt (van Dijck 2014). Bevor ein KI-basiertes System in einem realweltlichen Setting zum Einsatz kommen kann, muss es mit Unmengen an maschinenlesbaren Daten für die Modellierung des eng definierten Ausschnitts der Welt gefüttert werden. Diese Daten, so die grundlegende Idee, versetzt die KI in die Lage in ihrem eng definierten Einsatzbereich Muster zu erkennen, Kategorisierungen und Klassifizierungen vorzunehmen, und soziale Handlungen – menschenähnlich – zu imitieren (Crawford 2021). Jegliche Forschung und Entwicklung im Feld der KI basiert auf der Annahme, dass die Welt, und alles was in dieser inkludiert ist, maschinenlesbar gemacht werden kann, und dass ein Modell einer durch Datafizierung transformierten Welt, alle Informationen, Prozesse und Eigenschaften der analogen Welt und ihrer menschlichen Bewohner in ausreichender Genauigkeit abbildet. Mithilfe dieses Modells, das die maschinelle Auswertung digitaler Daten steuert, verspricht die KI ihren Nutzern die Entdeckung von Einsichten, die für menschliche Akteure ohne KI unzugänglich bleiben (Pasquale 2015).

Die Digitalisierung aller Lebensbereiche und die zunehmende Mediatisierung von Kommunikationsprozessen erzeugt eine Fülle an digitalen Daten, die für

die Verwendung und Programmierung von KI-Systemen nur mehr eingesammelt werden müssen. Damit wird unter dem Stichwort *Big Data* auch eine neue Ära eingeläutet. Es entstehen Unternehmen, die immer tiefere und genauere Einblicke über die Funktions- und Lebensweisen der datafizierten Akteure erhalten (Ducange et al. 2018). Diese Fülle an Daten, mit denen KI-Systeme gefüttert werden, ermöglicht erst die Funktionalität von KI-Systemen (Crawford 2021). Durch diese umfassende Datafizierung werden zwar immer mehr Bereiche der analogen sozialen Welt für KI-Anwendungen zugänglich, aber damit wird zugleich die Komplexität dieser Welt erheblich reduziert. Die Entwicklung zur Datafizierung und Quantifizierung der Welt setzt eine ökonomische Dynamik in Gang. Mathematische und statistische Verfahren werden fortlaufend verbessert, um aus den unendlichen Datenmengen Profit zu generieren. Datafizierung ermögliche objektive Analysen und führe zu besseren Ergebnissen als subjektiv-getriebene Bauchentscheidungen, die in vielen Bereichen der Wirtschaft, Wissenschaft aber auch des alltäglichen Lebens nur sub-optimale Ergebnisse liefern (Kosinski et al. 2013; Matz und Netzer 2017; Mazzei und Noble 2017). Mit Big Data wird die Erwartung einer objektiven Quantifizierung, bzw. die Aura der Neutralität, Fehlerlosigkeit und Wahrhaftigkeit verbunden (boyd und Crawford 2012). Verstärkt werden diese Tendenzen dadurch, dass die Möglichkeiten der Auswertungen mittels Algorithmen (Pasquale 2015; Gillespie 2016) und KI (Crawford 2021) erheblich ausgebaut wurden.

Die mathematische und wissenschaftliche Herangehensweise an ein Problem mittels Datafizierung und KI wird somit als Tor zur Überwindung menschlicher Subjektivität und Ungenauigkeit bisheriger Entscheidungsproblemen wahrgenommen. Big Data vermittelt dabei ein datengestütztes Wissen, welches mit einem hohen Grad an Objektivität ausgestattet ist (Gillespie 2014; van Dijck 2014). Dadurch entsteht in weiterer Folge die Hoffnung, Entwicklungen und Handlungen vorherzusagen, da ein immer besseres und kompletteres Wissen über die Welt vorhanden ist – Stichwort „prediction“ (Kerr und Earle 2013). Ähnliche Tendenzen findet man z. B. auch bei Entwicklungen in den kognitiven Neurowissenschaften bzw. im ‚Neuromarketing‘ (Schneider und Woolgar 2012). In diesem Bereich besteht die Hoffnung, dass mittels diverser Gehirnmessungsmethoden wie funktioneller Magnetresonanztomographie (fMRI) oder Elektroenzephalographie (EEG), das Verhalten von Individuen vorhergesagt und darauf Einfluss genommen werden kann. Trotz großer Ankündigungen und jahrelanger Entwicklung – sowohl im Bereich Big Data unterstützt durch KI-Systeme, als auch in den Bereichen der kognitiven Neurowissenschaften, kommen die realen Anwendungen lange nicht an den Erwartungen heran (Schneider und Woolgar 2012; Jansen 2022).

Bei der Quantifizierung und Datafizierung der Welt wird allerdings häufig ignoriert, dass die Prozesse zahlreiche menschlich-mediierte Schritte beinhalten, welche das Ergebnis in der Form von datengestütztem Wissen beeinflussen, Interpretationsspielräume erschaffen und sich damit naturgemäß von der scheinbaren Objektivität des Wissens wegbewegen. Datafizierung als ein Prozess der Umwandlung von bisher nicht quantifizierten Aspekten der Welt in Daten funktioniert innerhalb einer Rationalität, die bestimmte Normen, Strategien, Mechanismen und Ökonomien folgt (Kennedy et al. 2015). In vielen Bereichen der Datengenerierung, -manipulation und -verarbeitung sind zahlreiche Personen involviert, die einen signifikanten Einfluss auf die Datenstruktur, Datenverarbeitung sowie Interpretation der Auswertung haben (boyd und Crawford, 2012). Dies hat dann im weiteren Verlauf einen Einfluss darauf, wie KI-gestützte Systeme diese Informationen verarbeiten und in (semi-)automatisierte Entscheidungen mit einspielen. Die Ambivalenz und Subjektivität, die auch mit der Datafizierung und der Nutzung der Daten für KI-Systeme einhergeht, lassen sich am besten anhand eines Beispiels aufzeigen.

Die Klassifizierung von Geschlecht mag recht banal klingen, geht aber mit einer Vielzahl an (menschlichen) Entscheidungen einher. Die binäre Form von Geschlecht (Mann-Frau) wird bei Geschlechterklassifikationen in Datensätzen häufig als gegeben angesehen, allerdings ist sie nicht für Alle zutreffend. Während es mittlerweile auch non-binäre Geschlechterkodierungen (z. B.: Mann-Frau-Divers) gibt, die auch rechtlich anerkannt sind, ist diese in den KI-Modellen nicht gleichermaßen vorhanden. Regionale Unterschiede spielen ebenso eine Rolle, wie ideologische und politische Entscheidungen. Zudem bedeutet es nicht, dass nur weil eine rechtliche Möglichkeit der non-binären Klassifikation gegeben ist, diese sich auch in den Praktiken privater Unternehmen widerspiegelt. Denn die Entscheidungen über die Art und Weise der Klassifizierung in Datenbanken und KI-Systemen – welche Kategorien werden verwendet, wo liegen die Schwellenwerte und wie werden diese jeweils definiert – sind Teil eines undurchsichtigen Prozesses (Bowker und Star 1999). Die Teams an Entwickler\*innen in den gängigen Technologieunternehmen sind dabei meist dominiert von jungen, männlichen Erwachsenen, mit ähnlichen Biographien, Erfahrungen und Karriereentwicklungen (Wellner und Rothman 2020). Trotz zahlreicher Aufrufe, dass Entwicklerteams diverser werden müssen, hat sich in diesen Bereichen immer noch wenig getan. Die Beschreibung von Weizenbaum (1976, S. 116) mag zwar nicht mehr zutreffen, wenn KI-Entwickler als „unwashed and unshaved“ Individuen titulierte werden, „[who] are oblivious to their bodies and the world in which they move [and] exist, at least when so engaged, only through and for computers“ (zit. n. Jansen 2022, S. 5–6). Allerdings gibt es unzählige Beispiele

von diskriminierenden Datenverarbeitungen und KI-Anwendungen, die auf einen Mangel an Diversität in den Entwickler\*innenteams zurückzuführen sind (Wellner und Rothman 2020; Crawford 2021). Am Beispiel der Kategorisierung von Geschlecht zeigt sich ein weiteres Problem, das aus anderen Forschungsbereichen wie etwa der kriminologischen Kohortenforschung bekannt ist. Die einmal getroffene Festlegung auf eine bestimmte kognitive Ordnung bzw. Klassifikation ist sowohl gegenüber neuen Erkenntnissen als auch gegenüber gesellschaftlichen, sozio-kulturellen Veränderungen resistent (Sanders und Condon 2017; Sanders und Sheptycki 2017). Ein KI-basiertes Modell, das mit derzeit aktuell verfügbaren Daten trainiert wird, kapituliert möglicherweise vor Aufgaben/Daten, die veränderte Konstellationen abbilden. KI, angepriesen als innovative Technologie, wird damit strukturell konservativ (Weidinger et al. 2022).

Somit zeigt sich auch hier, dass KI ein Prozess der maschinellen Datenverarbeitung ist, welcher nicht ‚künstlich‘ entstehen kann, sondern auf vielen Ebenen zahlreiche menschliche und subjektive Entscheidungen beinhaltet. Diese Entscheidungen fließen in ein intransparentes KI-System ein, werden dadurch verschleiert und es entsteht „a technologically-inflected promise of mechanical neutrality [and] of algorithmic objectivity“ (Gillespie 2014, S. 15-16). Vielen Anwender\*innen von KI-Systemen bleibt im Endeffekt nichts anderes übrig, als dem System und den daraus resultierenden Informationen zu vertrauen (Symons und Alvarado 2016). Erklärbarkeit und Transparenz in KI-Systemen sind eine Seltenheit, die Trope der KI als undurchsichtige ‚Black Box‘ ist in der Forschung wie der Anwendung omnipräsent (vgl. Pasquale 2015; Wachter et al. 2017), wodurch den Anwender\*innen auch selten die Möglichkeit geboten wird, die Resultate von KI-Systemen zu hinterfragen. So meint auch Andrejevic 2020, dass „one of the challenges posed by “black box systems” is that even when an explanation is demanded and supplied, the result is likely to be cursory: because the machine said so.“ Erklärungen über die Funktionsfähigkeiten der KI, oder darüber, wie gewisse Ergebnisse zustanden kommen, werden häufig nur sehr halbherzig umgesetzt – wenn überhaupt.<sup>1</sup>

---

<sup>1</sup> Eine erwähnenswerte Ausnahme ist hier die in Frankreich gestartete Initiative BLOOM, bei der mehr als 1000 Forscher:innen im Bereich der KI-basierten Sprachprogramme an einem transparenten, nicht anglozentrischen Model arbeiten (Heikkilä 2022).

## 4 Gesellschaftliche Risiken von KI

Wie sich zeigt, ist KI weder intelligent noch künstlich. Menschliche Entscheidungen und Zuarbeit sind wesentlich für das Funktionieren von KI und gemessen an der Leistungsfähigkeit menschlicher Intelligenz schneidet KI außerhalb enger Anwendungsgebiete deutlich schlechter ab. Nichtsdestotrotz wird KI nach längerer Zeit wieder als bahnbrechende Schlüsseltechnologie gefeiert. Wir wollen im Folgenden versuchen, die vielschichtige Diskussion über KI, in der sich fachspezifische und populäre Beiträge mit den philosophischen, technischen, ökonomischen, sozialen und kulturellen Aspekten von KI beschäftigen, zu ordnen. Wir schlagen dazu vor, zwischen unterschiedlichen Diskurssträngen zu differenzieren, die sich über Kontroversen im Format ‚KI – Fluch oder Segen?‘ reproduzieren. Da wir uns auf Risiken und Gefahren von KI konzentrieren, steht dabei die Kritik im Vordergrund. Die von uns vorgeschlagene Unterscheidung ist natürlich stilisierend und die hier typisierten Kritikstränge überlappen und ergänzen sich.

1. *Transhumanismus*. Diese Debatte basiert im Wesentlichen auf Prognosen über zukünftige Entwicklungen der KI-Forschung. Kritiker\*innen wie Anhänger\*innen des Transhumanismus skizzieren eine Zukunft, in der sich domänenspezifische KI-Lösungen zu einer umfassenden bereichsunabhängigen KI transformieren (Bostrom 2014). Diese zur materiellen Reproduktion fähige AGI wird – so die utopischen Optimisten – die Menschheit retten, KI und Mensch entwickeln sich in friedlicher Ko-Evolution und lösen globale Probleme (Goertzel 2015). In der dystopischen Variante mutiert der Mensch zum Haustier der AGI oder verschwindet gänzlich vom Planeten, weil er entweder evolutionär überflüssig oder zum Opfer einer aus dem Ruder gelaufenen und nicht mehr kontrollierbaren AGI wird (Alfonseca et al. 2016).
2. *Biologische Hybridisierung*. Diese Debatte entzündet sich an den technologischen Möglichkeiten und Folgen von KI-basierten Systemen, die in den menschlichen Körper implantiert werden, um dessen Funktionsweise zu steuern. Sie liefert damit gleichsam Evidenz und Anschauungsmaterial für den Transhumanismus. Einschlägige Forschungsfelder wie Neuro-Enhancement ziehen prominente Investoren wie Elon Musk an, dessen Unternehmen Neuralink an der Entwicklung avancierter Schnittstellen von menschlichem Organismus und KI arbeitet. Die Befürworter treten mit geradezu biblischen Versprechen an: Blinde werden sehend, Lahme können wieder gehen und selbst Dumme werden gescheit – dank direktem Anschluss ihres Gehirns

an Künstliche Intelligenz (Pisarchik et al. 2019; Neuralink 2022). Die Kritiker\*innen sehen in dieser Technologie einen Angriff auf die konstitutiven Bestandteile eines humanistischen Menschenbildes. Hinter den Versprechungen, bisher unbehandelbare körperliche Beeinträchtigungen zu lindern, lauert die Gefahr des Verlusts menschlicher Autonomie. Der Mensch als Amalgam technischer und (neuro-)biologischer Prozesse verliert seine Freiheit und den freien Willen.

3. *Soziale Hybridisierung.* Im Gegensatz zu den kontroversen Debatten über Transhumanismus und biologische Hybridisierung, die sich auf mögliche zukünftige Entwicklungen von Technologie und Gesellschaft und KI-Anwendungen im experimentellen Stadium als vielversprechende emerging technology konzentrieren, handelt es sich bei sozialer Hybridisierung um ein sozio-kulturelles Massenphänomen digitalisierter Gegenwartsgesellschaften. Die Schnittstelle zwischen KI und Nutzer sind alltäglich genutzte digitalisierte Endgeräte. Die Nutzer solcher Angebote agieren, wie es in den Surveillance Studies formuliert wird, als „leaking data container“ (Kreissl 2014). Aus jeder Interaktion zwischen Nutzer und System lassen sich wichtige Informationen gewinnen, die dann durch KI-basierte Analysen zur Formulierung und Validierung von Hypothesen über Präferenzen, Interessen und Konsumverhalten der Nutzer verwendet werden. Mit Hilfe von KI können aus diesen Informationen maßgeschneiderte Nutzerprofile erstellt werden, sodass sich zukünftige Interaktionen gezielt und subtil gestalten (Lyon 2007, 2017). Während die Verteidiger\*innen dieser Entwicklung den Convenience Faktor betonen, sehen die Kritiker\*innen in der fortschreitenden sozialen Hybridisierung eine Entwicklung, die dem Verlust wichtiger evolutionärer und kultureller Fertigkeiten Vorschub leistet, den Horizont der Nutzer in vielfacher Hinsicht einschränkt oder manipuliert und zudem die letzten Barrieren zum (Daten-)Schutz der Privatsphäre beseitigt (Bauman und Lyon 2016).
4. *Bias und Diskriminierung.* Unter dieser Überschrift lassen sich die Beiträge zu Risiken und Gefahren der KI zusammenfassen, die auf unbeabsichtigte Nebenfolgen beim Einsatz KI-basierter Entscheidungssysteme verweisen (Crawford 2021). Benachteiligungen entlang der traditionellen Dimensionen, ‚race, class and gender‘ konnten in einer Vielzahl von Studien über den Einsatz von KI-basierten Entscheidungssystemen nachgewiesen werden (u. a. O’Neil 2016; Edwards und Veale 2017; Agostinho et al. 2019). In den meisten Fällen resultieren solche Verzerrungen aus den für ML verwendeten Trainingsdaten. Ein KI-System reproduziert dabei zuverlässig Verzerrungen, Stereotypen und kulturelle Vorurteile, die es in den Daten, die es zur Entwicklung seiner ‚Intelligenz‘ verwendet, identifizieren kann. Die Diskussion über Risiken



und Gefahren KI-induzierter Diskriminierung muss über eine Kritik an den Prozessen und der Leistungsfähigkeit der Datenverarbeitung hinausgehen und um eine kritische Auseinandersetzung mit der Struktur und Qualität der verwendeten Daten erweitert werden (Crawford 2021).

5. *Mangelnde Flexibilität.* Diese Debatte verweist auf die Nebeneffekte die sich aus der begrenzten Leistungsfähigkeit von KI-basierten Systemen beim Einsatz in wenig strukturierten Alltagssituationen ergeben, in denen menschliche Akteure kooperativ und lösungsorientiert interagieren können. Die Entscheidungsrationale von KI, die in einem solchen Setting an die Stelle eines menschlichen Gegenübers tritt, ist weder einsichtig, noch im Kontakt mit einem menschlichen Akteur veränderbar. KI-Systeme reagieren nicht auf Gründe, sondern auf Daten. Aufgaben und Handlungsprobleme können nur entlang der dem KI-System verfügbarer Schritte bearbeitet werden. Situationsangemessene Flexibilität und Kreativität sind in der Interaktion mit algorithmischen Prozessen nicht anschlussfähig (Gillespie 2016). KI-Systeme, so die Kritik, scheitern bei der Lösung oder Modellierung einfachster alltäglicher Routineaufgaben, da sie die (sozialen, kognitiven, kulturellen) Ressourcen und Bedingungen des lokalen Handlungskontexts ignorieren.
6. *Physikalisch-technische Vulnerabilität von datenbasierten Systemen:* Mit zunehmender Abhängigkeit von KI-unterstützten Prozessen steigt in einer Gesellschaft die Anfälligkeit für die Folgen von technischen Fehlfunktionen, gezielten Angriffen auf die Datenverarbeitung oder einer schlichten Unterbrechung der Stromversorgung (Blackout). Die Debatte über die Vulnerabilität moderner Gesellschaften reicht über mehrere Jahrzehnte zurück. Der Ausfall zentraler Module, die zeitkritisch getaktete Prozesse der Datenverarbeitung in eng gekoppelten komplexen Systemen steuern, kann zu kaskadierenden Störungen führen, die großflächig Produktions- oder Dienstleistungsprozesse lahmlegen können. Damit entstehen auch neue Angriffsvektoren im Bereich des Cyber-Warfare, d. h. der gezielten Angriffe auf kritische Infrastrukturen wie Finanz-, Wasser-, Energie- oder Transportsysteme. Mit dem Vordringen und der Vernetzung von KI-unterstützten Prozessen in modernen Gesellschaften, entstehen neue Bedrohungen und Risiken, etwa durch den Einsatz von KI als Waffe im militärischen Bereich.

## 4.1 Typisierung von Szenarien zur Untersuchung von Risiken und Gefahren von KI

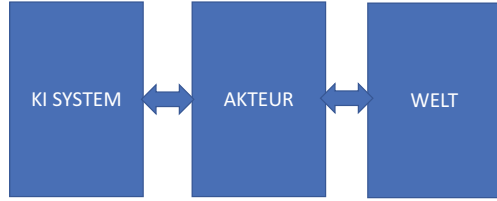
Ein weiteres Problem ist die Vielfalt an unterschiedlichen Szenarien und Konstellationen, die als Beispiel für Risiken und Gefahren von KI herangezogen werden. Für jeden der sechs hier stilisierten Diskursstränge finden sich exemplarische Beispiele, an denen die vermuteten Risiken und Gefahren demonstriert werden. Wir wollen auch hier in grober Stilisierung eine Typisierung von Szenarien vorschlagen, die nach Funktionalitäten von KI und Anwendungskontexten unterscheidet. Für jedes dieser Szenarien lässt sich das Verhältnis von menschlichem Akteur, KI und Handlungsvollzug in je spezifischer Weise erfassen. Damit wird ermöglicht, den realen Einsatz und die Verwendung von KI in einem konkreten organisatorischen oder Handlungskontext zu untersuchen.

1. Das **forensische Szenario**: Die KI fungiert als Assistent, den ein menschlicher Nutzer gezielt in einem sozialen Kontext zur Erledigung kontextspezifischer Aufgaben einsetzt. So nutzt etwa ein/e Grenzpolizist\*in, der/die bei Einreisenden Passkontrollen durchführt, ein KI-basiertes System zur Gesichtserkennung, das einen Input in der Form eines Passfotos durch Rückgriff auf eine große Datenbasis digitalisierter Fotografien mit den gespeicherten Bildern vergleicht. Wird ein passendes Bild gefunden, können gespeicherte Informationen über die der Fotografie zugeordnete Person abgerufen und das Ergebnis an den Nutzern zurückgemeldet werden. KI-basierte Systeme haben hier eine klar definierte und sehr begrenzte Aufgabe in einem formalisierten Handlungskontext zu erfüllen. Dieses Szenario liefert Belege für jene Kritiker\*innen, die mit der Gefahr der durch KI verstärkten Verzerrung und Diskriminierung argumentieren.

Hier bedient sich ein menschlicher Akteur eines KI-Systems als Unterstützung, um eine Entscheidung über das weitere Vorgehen in der Welt zu treffen. Dies funktioniert jedoch nur unter genau festgelegten Randbedingungen und Voraussetzungen (Abb. 2).

2. Das **strategische Szenario**: KI und menschlicher Akteur treten hier in einem durch explizite Regeln vollständig und eindeutig definierten geschlossenen Universum in Kontakt, in dem nur bestimmte Aktionen erlaubt sind. KI-basierte Anwendungen im Bereich der Molekularbiologie oder Schachprogramme zeigen die Unterlegenheit menschlicher Nutzer in einem solchen, regelbasierten Universum mit limitierten Freiheitsgraden und klar definierten Kriterien. KI-basierte Schachprogrammen verweisen auf die Bedeutung des Kontexts für die Interaktion von KI und Nutzer. Je enger und strukturierter

**Abb. 2** Forensisches Szenario. (Quelle: Eigene Darstellung)

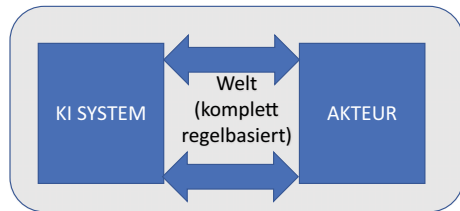


die Welt ist, in der KI und Mensch agieren, desto besser schneidet die KI ab und desto weniger kann die Flexibilität menschlicher Intelligenz produktiv genutzt werden. Die Konkurrenz von menschlichen Nutzern und KI in solchen limitierten Universen kann als Beleg für Risiken und Gefahren dienen. So war etwa ein für den Sieg entscheidender Zug des KI-System in einer Partie des Brettspiels Go gegen einen der besten Spieler in der Analyse dieses Spiels in keinerlei strategischen Zusammenhang einzuordnen oder anderweitig durch menschliche Beobachter\*innen zu erklären. Mit genügend dystopischer Kreativität zeigt sich hier ein erstes Aufblitzen einer transhumanen Überlegenheit der KI.

Der menschliche Akteur und das KI-System interagieren in diesem Szenario in einer finiten und vollständig digitalisierbaren, maschinenlesbaren Welt (Abb. 3).

3. Das **Überwachungsszenario**: Die KI sammelt unbemerkt Daten über einen menschlichen Nutzer in einem technisch-medial vermittelten sozialen Kontext. Jede Aktivität eines Nutzers in diesem Kontext techno-sozialen Alltagshandeln produziert personenbezogene Daten für KI-basierte Analysen zur Erstellung von Konsumentenprofilen, um Nutzerverhalten durch die Präsentation von gezielt ausgewählten Inhalten oder Angeboten zu beeinflussen. Die Schnittstelle von Nutzer und KI ist hier das techno-soziale Alltagshandeln. Die Konsument\*in auf der Plattform tritt bei Suche nach einem bestimmten

**Abb. 3** Strategisches Szenario. (Quelle: Eigene Darstellung)



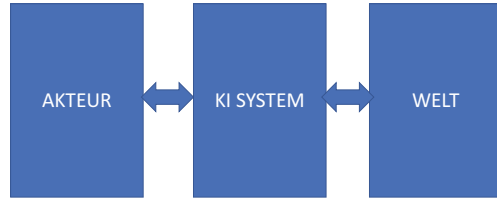
Strategisches Szenario

**Abb. 4**

Überwachungsszenario.

(Quelle: Eigene

Darstellung)

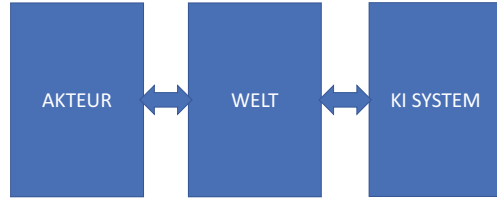


Objekt, ohne es zu bemerken, in Kontakt mit eingebetteten KI-Systemen, die den Suchprozess steuern und dabei die anfallenden personenbezogenen Daten für spätere Verwendung speichern sowie gezielt jene Informationen zur Verfügung stellen, die aufgrund der im Hintergrund laufenden Kalkulation über die Person als passend erachtet werden. Es handelt sich in gewisser Weise um eine abgeschwächte Variante des strategischen Szenarios einer geschlossenen, nach der Logik digitaler Maschinen gesteuerten Welt. Die Kritiker\*innen des herausziehenden Überwachungskapitalismus bedienen sich häufig solcher Szenarien, in denen KI-Systeme Nutzerdaten erfassen, um profitgenerierende Lösungen zu entwickeln und Marketingstrategien zu optimieren (Zuboff 2020) (Abb. 4).

4. Das **Eliza Szenario**: Eliza war der Name eines prominenten Computerprogramms, das Joseph Weizenbaum in den 1960iger Jahren entwickelte und das eine Art therapeutisches Gespräch mit Frage-Antwort-Sequenzen führen konnte (Weizenbaum 1966). Die Versuchspersonen, die über Tastatur und Bildschirm mit diesem Programm interagierten, entwickelten häufig positive Empfindungen und fühlten sich von ihrer ‚Gesprächspartnerin‘ verstanden. KI und menschliche Nutzer treten hier in Kontakt, wobei die Nutzer auf der Basis der Annahme agieren, dass sie es mit einem menschlichen Gegenüber zu tun haben. Avancierte KI-basierte Sprachprogramme sind im sprachlichem Konversationsmodus kaum mehr von menschlichen Gesprächspartnern zu unterscheiden. Auch hier spielt der Kontext, der Nutzer und KI ohne Handlungsbezug nur über das Medium Sprache verbindet, eine wichtige Rolle. Eine für dieses Szenario typische KI-Anwendung sind Chat Bots, die in natürlicher Sprache mit menschlichen Nutzern interagieren und im Modus einer Konversation bestimmte Themen oder Probleme erörtern.

Der menschliche Akteur handelt in diesem Szenario auf der Basis der Annahme, es mit einem Artgenossen zu tun zu haben, interagiert aber mit einem KI-basierten System, das in der Lage ist, zumindest in einem spezifischen Setting die kulturellen und linguistischen Erwartungen des jeweiligen Settings zu erfüllen und sich menschenanalog zu verhalten (Abb. 5).

**Abb. 5** Szenario „Eliza“.  
(Quelle: Eigene  
Darstellung)



## 4.2 Risiken und Gefahren der KI – Versuch einer kritischen Einschätzung

Abschließend wollen wir mithilfe der unterschiedlichen Szenarien, entlang der oben typisierten Diskursstränge und anhand exemplarischer Beispiele einen kritisch informierten Blick auf die Diskussion über Risiken und Gefahren KI werfen. Das eröffnet neue Perspektiven und ermöglicht eine differenziertere Bewertung der gesellschaftlichen Folgen und Effekte von KI. Viele medial großflächig verbreitete Meldungen über epochale Durchbrüche in der KI-Forschung, die dann auch die öffentliche Debatte über Fluch und Segen der KI befeuern, verdanken ihren Erfolg der gezielten Gestaltung des Kontexts ihrer Anwendung. Gleichzeitig können sich hoch riskante und gefährliche KI-Anwendungen, medial unbeachtet und für Nutzer nicht erkennbar im Kontext der Gesellschaft verbreiten.

### *Transhumanismus und biologische Hybridisierung*

Die meisten Vertreter\*innen des transhumanistischen Programms kombinieren technologischen Determinismus mit klassischem, politischem Liberalismus zu einer Art Technoevolution (O’Connell 2017). Transhumanismus gewinnt Gestalt und Kontur durch Abgrenzung gegen das rhetorisch stilisierte Lager der sogenannten „bioconservatives“; ein Etikett das die eigene Position als technoprogressiv markiert (Bostrom 2005). Argumentiert wird gegen *die* Natur in ihrer sozusagen prä-technologischen Variante, deren Hervorbringungen und Geschenke alles andere als positiv und wünschenswert seien: „Cancer, malaria, dementia, aging, starvation, unnecessary suffering, cognitive shortcomings are all among the presents that we wisely refuse. Our own species-specified natures are a rich source of much of the thoroughly unrespectable and unacceptable – susceptibility for disease, murder, rape, genocide, cheating, torture, racism“ (ebd., S. 205). Das transhumane Versprechen basiert – vereinfacht formuliert – auf der Annahme, dass sich Krankheit und Dummheit durch den Einsatz besserer Technologie ausmerzen ließen. Die Gefahr des autoritären Umkippens eines solchen transhumanistisch fundierten Regimes könne gebannt werden, solange Alle

die Freiheit haben, selbst zu entscheiden, ob sie zu technologischer Selbstmodifikation greifen. Technologisch veränderte transhumane Wesen würden lediglich die ohnehin bereits existierende Vielfalt menschlicher Gesellschaften erweitern. Gegen schädliche Wirkungen zukünftiger Technologien wird an die individuelle Verantwortung der Ingenieur\*innen appelliert: „The would-be creator of a new life form with such surpassing capabilities would have an obligation to ensure that the proposed being is free from psychopathic tendencies and, more generally, that it has humane inclinations. (...) [A] future [... AI] programmer should be required to make a strong case that launching a purportedly human-friendly superintelligence would be safer than the alternative“ (ebd., S. 208).

Im Angesicht der prognostizierten Leistungsfähigkeit einer AGI wirkt der Appell an die Verantwortung der Programmierer\*innen, dafür zu sorgen, dass am Ende eine „human-friendly super intelligence“ herauskommt, wenig beruhigend. Das Szenario, das Autor\*innen wie Bostrom entwerfen, geht von einer Art rekursiven Kopplung aus, die eine exponentielle Leistungssteigerung der KI antreibt: menschliche Akteure bedienen sich der vorhandenen Werkzeuge, wodurch sie ihre Fähigkeit steigern, neue, bessere Werkzeuge zu entwickeln, deren Einsatz am/im/durch den menschlichen Akteur dann wiederum neue, bessere Werkzeuge hervorbringt (Goertzel 2015). Dieser Prozess führt im Transhumanismus zur (sich kontinuierlich beschleunigenden) Entwicklung einer human-level thinking machine, der Vorstufe einer mystischen Singularität, von seinen menschlichen Schöpfern entkoppelt und diese möglicherweise auf den Status von „*family pets*“ reduziere, wie Steve Wozniak es formulierte (Gibbs 2015). Ähnliche Befürchtungen äußert auch Bostrom (2014). Einmal autonom geworden, würde die zur Singularität aufgestiegene KI in einer Weise agieren, die ihre Selbsterhaltung und Ausdehnung fördert.

Die von den Vertreter\*innen des Transhumanismus geführte Diskussion über Risiken und Gefahren, die von einer übermenschlichen KI drohen, schneidet das Problem jedoch an der falschen Stelle, wenn über deren mögliche Werte und Ziele spekuliert wird. Ob sich ein autonomes superintelligentes KI-System selbst durch eine Wertorientierung oder Nutzenfunktion steuert, ist eine Frage, die sich nur sinnvoll beantworten lässt, wenn man von einem sehr engen Verständnis von Intelligenz ausgeht (Alfonseca et al. 2016). Intelligenz ist, wie oben ausgeführt, eng mit praktischem, sozialem Handeln gekoppelt. Die Reduktion von Intelligenz auf zielgerichtetes, fokussiert problemlösendes Prozessieren von Daten und Informationen blendet viele Prozesse aus. Die Diskussion über eine mögliche oder wünschenswerte übergreifende Wertorientierung einer zukünftigen übermenschlichen KI ignoriert die Struktur und Praxis menschlichen – intelligenten, kommunikativen, kooperativen, kontextgebundenen – Handelns und verwechselt

damit die rekonstruktive Begrifflichkeit von Werten oder Nutzenfunktionen mit einer Analyse lokal produktiver Praktiken der kooperativen Problemlösung im Prozess der Reproduktion der Gattung.<sup>2</sup>

Der transhumanistische Diskurs neigt dazu, evolutionäre Prozesse zu verkürzen und deren Komplexität zu ignorieren und landet dann bei rhetorischen Figuren eines a-historischen Historismus: Im Rückblick erscheinen so epochale, disruptive Durchbrüche und Technologiesprünge in der Menschheitsgeschichte als unvorhergesehene Ereignisse und die historischen Akteure, so die transhumanistische These, neigten immer zu übervorsichtigen Risikoprognosen, die sich im Nachhinein als übertrieben und irrelevant entpuppen. Die Kritik des ‚precautionary principle‘, das gegen die ungebremste Entwicklung von KI und anderen disruptiven Technologien ins Feld geführt wird, bemüht dann Beispiele aus dem Neolithikum: die Einführung des Ackerbaus wäre, wenn sich die Zeitgenoss\*innen am precautionary principle orientiert hätten, vermutlich verhindert worden, da zu viele Risiken mit dieser Innovation verbunden worden wären. Ähnlich sollte die Erprobung avancierter technologischer Modelle von KI nicht mit Verweis auf die mit ihnen möglicherweise verbundenen Risiken zu stark eingeschränkt werden: „Humanity emerged from apehood via proactionary manifestations of open-ended intelligence, and it seems that, overall, the later stages of progress toward Technological Singularity are currently unfolding in a similar manner“ (Goertzel 2015).

Bezogen auf die oben skizzierten Szenarien und die darin typisierten Beziehungen zwischen menschlichen Akteuren, Welt und KI-Systemen prognostizieren die transhumanistische Zukunftsprojektionen eine Auflösung dieser Unterscheidung. Menschliche Akteure und KI-Systeme bilden im Transhumanismus neue Hybridexistenzen aus, die sich in ihrer Verschmelzung von materiellen und biologischen Limitierungen befreit haben und einen bisher unbekanntem Weltbezug realisieren werden. Die Spekulationen über diesen Endzustand folgen der Unterteilung von Dystopie und Utopie, aber diese Differenz gewinnt ihre Einheit in einem letztlich technodeterministischen Ansatz, der den vermuteten Weg zur stilisierten, aber mit Bordmitteln des gegenwärtigen Denkens nicht greifbaren Singularität einer AGI als Prozess einer fortschreitenden Verbesserung der maschinellen Datenverarbeitung skizziert.

---

<sup>2</sup> Zur epistemischen und methodischen Problematik solcher *als-ob* Rekonstruktionen menschlichen Handelns siehe Schlicht (2002).

### *Soziale Hybridisierung, Social Sorting und Bias*

Menschen verfügen über eine begrenzte Informationsverarbeitungskapazität, die sich jedoch durch den Einsatz kultureller Schemata erheblich erweitern lässt. Während das visuelle Kurzzeitgedächtnis durchschnittlich drei bis fünf isolierte Objekte speichern kann, sind kulturell kompetente Leser von literarischen Werken in der Lage, noch längere Zeit nach der Lektüre eine konsistente Zusammenfassung des Inhalts und eine Aufzählung der wichtigen handelnden Figuren zu liefern (Rumelhart 1975; Cowan 2010). Die Möglichkeit, auf höherwertige kognitive Schemata zurückzugreifen, erleichtert die Orientierung in einer komplexen Umwelt. Klassifikationssysteme wie das Alphabet erleichtern die Suche nach einem Wort in einem Lexikon, kulturelle Stereotype den sozialen Verkehr, Stadtpläne die gezielte Bewegung im unbekanntem Raum und Bibliothekskataloge das Auffinden von Büchern. Das organisierende evolutionäre Prinzip ist die kognitive Ökonomie, es setzen sich jene Strukturen oder Klassifikationsschemata durch, die ein Maximum an Orientierung bei einem Minimum an kognitiver Anstrengung erlauben.<sup>3</sup> Die Orientierung im Raum mit Hilfe von Wegweisen und Karten ist ein gutes Beispiel für die Wirkweise dieses Prinzips (Hölscher et al. 2007).<sup>4</sup>

Jede Möglichkeit, Information über die Welt aus dem individuellen Gedächtnis auszulagern und sie bei Bedarf mit wenig Aufwand wiederzufinden, ist hier attraktiv. Dies erfordert jedoch nicht nur den Zugang zu externen Speichermöglichkeiten, sondern auch die Fähigkeit, diese zu bedienen. Wissen und Information, die außerhalb des menschlichen Gedächtnisses in der Umwelt gespeichert sind, werden nur dann für praktische Zwecke zugänglich, wenn diese Umwelt lesbar ist. Mit dem Vordringen elektronischer Systeme entstehen hier neue Konstellationen. Die Speicherung von Telefonnummern ist ein Beispiel, an dem sich dieser Mechanismus demonstrieren lässt. Vor der Verbreitung von Mobiltelefonen, hatte die durchschnittliche Bewohner\*in westlicher Gesellschaften in der Regel die wichtigsten oder häufig gewählten Rufnummern ihres sozialen Umfelds *im Kopf* und konnte sie *wählen*. Erst mit der alltäglichen Nutzung tragbarer Mobilgeräte änderte sich das. Theoretisch wird damit ‚Speicherplatz‘ im menschlichen Informationsverarbeitungssystem frei, der Zugang zu externalisierter Information durch Datenverarbeitung erweitert gleichsam den Wissensbestand. Allerdings gehen damit eine Reihe von Folgen einher.

---

<sup>3</sup> Die soziologische Systemtheorie spricht hier von Komplexitätsreduktion und Heinrich Popitz (1976) vom Ordnungswert der Ordnung.

<sup>4</sup> Auch im Bereich des ML kommt das Prinzip der cognitive economy zum Einsatz. Siehe Warnett und McGonigle (2002).



Der durch KI-Systeme befeuerte Prozess sozialer Hybridisierung knüpft an die kulturelle Praxis der Auslagerung von Information an. Im einfachen Fall kann man dabei zwei Schritte – Speichern und Abrufen – unterscheiden. Information muss in einem Medium in der Umwelt gespeichert werden, sodass sie bei Bedarf wieder abgerufen werden kann: ein Buch muss mit einer Signatur versehen, Name des Autors oder der Autorin und Titel müssen eingetragen und in ein alphabetisch geordnetes Register eingeordnet werden, sodass eine alphabetisierte Person sie schnell finden und über die Signatur den Standort des Buchs in der Bibliothek identifizieren kann. Das Beispiel der Bibliothek ist in mehrfacher Hinsicht instruktiv. Es zeigt zum einen das Zusammenspiel von Speichern und Abrufen von Information in/aus der Umwelt unter Verwendung kulturell geteilter Klassifikationsschemata und materieller Medien als einen Prozess kognitiver Arbeitsteilung. Zum anderen aber wird man bei der Suche nach dem Standort des gewünschten Buchs gelegentlich überrascht, wenn in dem Regal in unmittelbarer Nachbarschaft des gesuchten Werks ein anderes, möglicherweise ebenso interessantes Buch steht, zu dem man dann greift und damit vom ursprünglichen Pfad der Suche abzweigt.

Solche Ordnungs-Such-Systeme verkörpern allerdings nicht nur ein Verfahren zum verlässlichen und ökonomischen Speichern und Suchen von Information. Bibliotheken verwenden die Zuordnung von Autor\*innen und Texten zu Themen- oder Sachgebieten als weiteres, kognitiv höherwertiges Ordnungsprinzip. Diese Zuordnungen wiederum orientieren sich an den taxonomischen Differenzierungen einer akademisch-disziplinären (z. B. Natur- und Geisteswissenschaften) oder kulturellen (Biographien, Kriminalromane, Reise-, Koch- und Kinderbücher, etc.) Ordnung. In Taxonomien spiegeln sich historisch gewachsene, relativ stabile symbolische Ordnungen und die Herrschaftsbeziehungen, die sie reproduzieren, um sie den gesellschaftlichen Akteuren als Mittel zur Orientierung anzudienen (Bowker und Star 1999). Soziale Hybridisierung, social sorting und Bias sind also keineswegs neue, durch das Vordringen von KI-Systemen entstehende Phänomene. Sie sind konstitutive Bestandteile jedes ausdifferenzierten sozialen Systems: Soziale Akteure nutzen das *knowledge in the environment*, symbolische Ordnungen sortieren Individuen nach normativen Kriterien und vielfältige Benachteiligungen sind Bestandteil jeder real existierenden Gesellschaft.

Aus handlungstheoretischer Perspektive betrachtet, gewinnt durch die Einführung von KI-Systemen die fluide Verarbeitung, das Prozessieren der Information, als eigenständige Aktivität zwischen Speichern und Abrufen von Informationen als dritter Schritt an Bedeutung. Durch das intelligente maschinelle Prozessieren ändern sich – meist unbemerkt von menschlichen Nutzern – die

Klassifikationssysteme und der Prozess des Abrufens führt zu neuen Ergebnissen. Waren traditionelle Ordnungs-/Herrschaftssysteme relativ stabil und damit einer kritischen Analyse zugänglich, so ändert sich das mit dem Auftreten von KI-Systemen. Pasquale (2015) analysiert die gesellschaftlichen (sozialen, juristischen, kulturellen) Folgen der fluiden und opaken algorithmischen Steuerung von Ordnungssystemen. Dank der technischen Möglichkeiten bei der Gestaltung und Adaptation der Architektur von Datenbanken, die als Infrastruktur der logischen und semantischen Ordnung von Informationsverarbeitungsprozessen dienen, entstehen immer neue Ordnungen und Indexierungen in den gespeicherten Informationen (Mackenzie 2012).

Instruktive Beispiele für problematische Effekte, für Risiken und Gefahren einer KI-basierten Hybridisierung finden sich im Bereich der forensischen und der Überwachungsszenarien. Im forensischen Szenario agiert die KI in der Rolle des Agenten, der einem menschlichen Akteur, der eine Entscheidung über ein Problem, mit dem er in seiner unmittelbaren Lebenswelt konfrontiert ist, Auskunft erteilt, ihn durch Bereitstellung von gezielt ausgewählter Information unterstützt. Beispiele sind die Grenzpolizist\*in, die eine KI-basierte Gesichtserkennungssoftware nutzt, eine Richter\*in, die sich an algorithmisch errechneten Wahrscheinlichkeiten des Rückfalls bei der Entscheidung über das Strafmaß oder die Aussetzung einer Strafe zur Bewährung orientiert, oder die KI-gestützte Entscheidung über die Vergabe von Krediten an Bankkund\*innen. Maschinell unterstützte Entscheidungen versprechen schnellere und präzisere Ergebnisse im Vergleich zu menschlichen Akteuren (Ghodselahe und Amirmadhi 2011). Allerdings zeigt sich auch hier, dass selbst avancierte Verfahren nicht vor einseitigen und ungerechtfertigten Entscheidungen auf der Basis implizit bleibender Kriterien wie Geschlecht und ethnische Zuordnung gefeit sind (Hassani 2021). Die rechtliche Einhegung solcher Risiken avancierter FinTech Scoring Verfahren ist schwierig und wirft eine Reihe von Fragen bezüglich Datenschutz und Diskriminierungsverbot auf (Langenbacher 2020).

Konsequente Hybridisierung führt letztlich zur Eliminierung des menschlichen Akteurs, wie das Beispiel der automatisierten Entscheidung über Kreditvergabe bei online Finanzdienstleistern zeigt. Antragsteller\*innen interagieren dann nur mehr mit einem technischen System, das die von Kunden abgefragten Daten in einer algorithmischen Blackbox verarbeitet und am Ende entweder eine Kreditzusage verweigert oder ein Angebot anbietet. Das Vordringen solcher Anbieter, die außerhalb des traditionellen Bankensektors und ohne direkten persönlichen Kundenkontakt operieren, kann das Gleichgewicht im Markt für Finanzdienstleistungen signifikant verschieben (Jagtiani et al. 2021). Zudem sind diese Anbieter,

die außerhalb des für Banken geltenden Rechtsrahmens agieren, nicht den strikten Kontrollen unterworfen, die dort zum Tragen kommen. Prominente Fälle wie der Wirecard Skandal in Deutschland liefern anschauliche Belege für die Folgen solcher regulatorischen Defizite. Die bisher vorliegende Evidenz zu diesem Sektor stammt vorwiegend aus den Vereinigten Staaten und ergibt ein komplexes Bild (Philippon 2020).

Eliminiert man im Beispiel der Kreditvergabe die Bankmitarbeiter\*in aus dem Arrangement, so entsteht das hier als Überwachungsszenario deklarierte Setting. Ein menschlicher Akteur in der Person des Kreditsuchenden tritt, vermittelt durch ein ihm nicht zugängliches KI-System in Kontakt mit der Welt der Finanzen. Die dem Antragsteller oder der Antragstellerin durch die KI-Systeme präsentierten Angebote kommen ohne Erklärung oder Begründung. Sie sind – mangels eines menschlichen Gegenübers – auch nicht verhandelbar. Auf der Aggregatsebene liefern solche maschinell vermittelten Interaktionen die Daten zur Weiterentwicklung der Algorithmen, mit deren Hilfe individuelle Bewertungen erstellt werden. Konsument\*innen tragen als Datenlieferanten dank der Möglichkeiten der KI-basierten Verarbeitung dieser Daten zu ungeahnter Steigerung ihrer eigenen Transparenz im Angesicht der Anbieter von Waren und Dienstleistungen bei. Das erhöht die Informationsasymmetrien unter den Marktteilnehmern und gibt Unternehmen neuartige Möglichkeiten ihre Geschäftsmodelle zu adaptieren, möglicherweise unter Verletzung datenschutzrechtlicher Vorgaben und ethischer Prinzipien (Wang 2018).

#### *Side-effects und Einschränkungen*

KI-Systemen werden dank besserer Rechenleistung und der Möglichkeiten des Zugriffs auf immer größere Datensätze zusehends Fähigkeiten zugeschrieben, die bis dato als menschlichen Akteuren vorbehalten galten; angefangen von Gesichtserkennung bis hin zur Fähigkeit Emotionen im Gesichtsausdruck einer Person zu differenzieren. Als problematisch empfunden werden hier verschiedene Dinge: erstens werden menschliche Gefühlsregungen dank KI-basierter Identifikation maschinenlesbar und können damit in weiterreichende algorithmische Kalkulationen und Strategien (etwa im Bereich des Micromarketings) eingebunden werden. Zweitens ermöglicht der Fortschritt im Bereich von Bild- und Sprachverarbeitung die Generierung von maschinellen Artefakten, die von Menschen nicht mehr unterscheidbar sind. Die Diskussion über den Einsatz von Deep Fakes ist ein Beispiel dafür. Drittens stiften KI-Systeme, die in der Lage sind, in der Interaktion mit menschlichen Nutzern eine Konversation hervorzubringen, die aufseiten der menschlichen Akteure die Vermutung aufkommen lässt, sie hätten es mit einem fühlenden und denkenden System/Wesen zu tun, zunehmend Verwirrung darüber,

wo die Grenze zwischen menschlicher und maschineller Handlungsfähigkeit zu ziehen ist.

Diese Beispiele gewinnen ihre Bedeutung im Kontext der dadurch ermöglichten Veränderung in der Mensch-Maschine Interaktion. Diese wird flexibler und für den menschlichen Nutzer – nicht zuletzt dank der Fortschritte im Bereich der Sprachprogramme – scheinbar realitätsnäher. Das simple lineare Format Befehl/Anfrage des Nutzers – Exekution/Antwort der Maschine scheint sich für den Nutzer zu einer rekursiven, kooperativen sozialen Interaktion zu entwickeln. Allerdings trägt dieser Eindruck, da eine Reihe wesentlicher Merkmale menschlicher Interaktion fehlen. Es gibt keine Ko-Präsenz der Teilnehmer\*innen, die Maschine sendet keine körperlichen Signale. Maschinenlesbare Daten über den menschlichen Nutzer jenseits manifester sprachlicher Äußerungen fallen kaum an. Wichtiger aber als die reduzierte sensorische Bandbreite ist das Fehlen des für die Interaktion zwischen menschlichen Akteuren konstitutiven Elements der kooperativen Angleichung von Standpunkten, des argumentativen Austauschs von Informationen mit dem Ziel, das Gegenüber zu einer Handlung zu bewegen, die den eigenen Interessen und Wünschen entspricht. Mit Maschinen lässt sich im genuin sozialen Sinn nicht kooperativ verhandeln. KI-Systeme können nicht überzeugt, sondern bestenfalls getäuscht werden. Der Einsatz von KI-basierten Systemen zur Bearbeitung von Anfragen und Problemen aus der Sphäre des mundanen Alltags zwingt Interaktionen in einen rigiden Rahmen von vorgegebenen Wenn-dann Entscheidungen. Lokal produktive Lösungen unter Einsatz der vorhandenen Mittel vor Ort, ein situativer Wechsel der Definition des zu lösenden Problems, eine Einigung auf kreativ entwickelte informelle Alternativen oder dialogisch-szenische Praktiken der Beseitigung von auftauchenden Unklarheiten oder Missverständnissen sind hier nicht möglich.

Instruktiv und exemplarisch zeigen sich die im Sinne der Nutzer kontra-produktiven Nebeneffekte bei der Interaktion mit KI-Systemen, wenn diese in der Rolle von Kund\*innen mit einem Dienstleistungsunternehmen in Kontakt treten. Diese versprechen sich von der Einführung teil- oder voll-automatisierter Systeme neben einer Kostenersparnis eine an differenzierten, vorgegebenen Regeln orientierte Bearbeitung typischer Kundenanliegen. Call Center sind ein Beispiel für Teilautomatisierung. Die Mitarbeiter\*innen benötigen in der Regel keine Fachkenntnisse. Zur Bearbeitung eingehender Anfragen greifen sie auf abrufbaren Informationen zurück, die von den Firmen, die das Call Center als Dienstleister nutzen, bereitgestellt werden. Dabei folgen sie fest vorgegebenen Abläufen, stellen standardisierte Fragen und erstellen im Gespräch mit den Anrufern in der Maske auf den Bildschirm eine Art Protokoll aus, das dann an die jeweiligen Unternehmen weitergeleitet wird. Nachfragen der Kund\*innen am Telefon sind

in diesem System nicht vorgesehen. Ebenso wenig werden rechtsverbindliche Entscheidungen bei der Interaktion mit dem Personal von Call Centern getroffen. In voll-automatisierten Systemen übernimmt eine Sprachsoftware die Rolle der Mitarbeiter\*innen des Call Centers, das enge kommunikative Korsett, das die Anrufer\*innen zwingt, ihre je individuellen Probleme und Nachfragen im Format eines rigiden Schemas vorzubringen, bleibt gleich. Während die automatisierte Bearbeitung für den menschlichen Nutzer die Möglichkeiten einschränkt, sein Anliegen in einer ihm angemessenen Form vorzubringen, reduziert diese kommunikative Standardisierung für die Unternehmen die Kosten. Umfassendere Strategien der KI-basierten Kundenbindung, wie Avatare, die als humanoide Erscheinung auf dem Bildschirm erscheinen und zu normalsprachlichen Konversationen fähig sind, werden bereits im Finanzsektor von Banken eingesetzt. Mittelfristig führt der Einsatz von KI hier nicht nur zur Reduktion von Beschwerdemacht, sondern auch zu vermehrter Verlagerung von Verantwortung auf die Kund\*innen.

Die Beispiele, an denen sich die unbestreitbaren Vorteile und Fortschritte von KI bei der Lösung komplexer Probleme belegen lassen, unterscheiden sich von dem hier skizzierten Szenario dahingehend, dass dort die Lösungen Antworten auf Fragen sind. Die Frage welche Moleküle als Antibiotikum geeignet sind, wurde von Googles KI mit Bravour beantwortet. Die Leistungsfähigkeit von KI-basierten Sprachübersetzungsprogrammen zwischen hegemonialen Nationalsprachen ist ein weiteres Beispiel. Interveniert ein KI-System hingegen mit Anspruch auf verbesserte Problemlösungsfähigkeit in kooperativen sozialen Interaktionen, führt das meist zur Einschränkung des Variety Pools lokal produktiver Lösungsoptionen mit entsprechenden Nebeneffekten durch Verminderung der Handlungsautonomie menschlicher Akteure.

### *Technische Vulnerabilität*

Eine zentrale Maxime der allgemeinen Systemtheorie lautet: Systemstabilisierung durch Ausdifferenzierung. Dieser Zusammenhang lässt sich für biologische und soziale Systeme an einer Vielzahl von Beispielen zeigen. Bei technischen Systemen scheinen die Dinge etwas anders zu liegen. In seiner klassischen Studie über Normale Katastrophen vertritt Charles Perrow (1999) die gegenteilige Ansicht: eng gekoppelte, komplexe Systeme, sind weit störungsanfälliger als einfache, weniger eng gekoppelte Systeme. Erhellende zeitgenössische Beispiele für diese Art von Störanfälligkeit finden sich im Bereich der logistischen Steuerung von Strömen – von Verkehrs- und Warenströmen, über Energieversorgung bis hin zu tief gestaffelten industriellen Produktionsprozessen. Diese Prozesse nutzen ein

globales, anarchisch wucherndes Netzwerk, um die Zirkulation der aus ihren traditionellen Verankerungen gerissenen Bausteine der globalen Risikogesellschaft zu koordinieren. Unkontrolliert und unkontrollierbar könnte dieses Netz als ein passendes Symbol des durch techno-soziale Prozesse befeuerten Übergangs zur flüchtigen Moderne dienen.<sup>5</sup> Technologisch informierte Beobachter\*innen dieser Entwicklung schlagen einen warnenden Ton an und sprechen vom Internet of Vulnerable Things (Anand et al. 2020).

KI-basierte Systeme durchdringen die technologische Infrastruktur und das erhöht ihre Anfälligkeit. Es entstehen neue Bedrohungen, die entsprechend dramatisiert werden. Cyberkriminalität wird von den nationalen Strafverfolgungsbehörden als eine wachsende Gefahr eingestuft und im militärischen Bereich sucht man nach Strategien für neue Bedrohungen des Cyber-Warfare. Jenseits solcher gezielten Attacken wächst mit zunehmender Komplexität, Vernetzung und Verbreitung IT-basierter Steuerungs- und Kontrollsysteme aber auch die durch systemische Merkmale bedingte Störungsanfälligkeit. Die Abhängigkeit der Akteure des Alltags von abstrakten Systemen hat zusehends eine digitalisierte, netz-basierte Form angenommen, die vielfältige, oft unbemerkt bleibende Abhängigkeiten erzeugt. Je mehr mundane Prozesse auf das Funktionieren von zunehmend durch KI gesteuerten und vernetzten digitalen Datenströmen angewiesen sind, desto vielschichtiger und unberechenbarer werden die Probleme, die ein technischer Systemausfall auslösen kann.

Auf der Basis dieser gemeinsam geteilten Diagnose, dass komplexe technische Systeme zur Steuerung von Strömen durch den Einsatz von KI komplexer und leistungsfähiger, damit aber auch störungsanfälliger werden, verzweigt sich die Debatte über Risiken und Gefahren in gegensätzliche Äste. Auf der einen Seite versammeln sich technikkritische Positionen, die für eine regulatorische Zählung der KI votieren. Statt das Potential von KI-basierten Systemen weiterzuentwickeln und zu nutzen, um durch schnellere und effizientere Datenanalyse die Leistungsfähigkeit der techno-sozialen gesellschaftlichen Infrastruktur zu steigern, wird die Politik aufgefordert, dieser Entwicklung durch risikominimierende Rahmenbedingungen Einhalt zu gebieten. Sowohl Forschung und Entwicklung als auch der Einsatz entsprechender Systeme erforderten die Berücksichtigung ethischer Standards, sollten im Einklang mit Menschenrechten stehen und klar definierten Sicherheitserfordernissen entsprechen. Die Gegenseite setzt zur Bearbeitung von denselben Risiken und Gefahren auf den vermehrten Einsatz von

---

<sup>5</sup> Zeitdiagnostische Gesellschaftstheorien von Giddens (1990) über Beck (1986) bis Bauman (2015) haben diese Momente der Mobilität, Entwurzelung und Entgrenzung aus soziologischer Perspektive ausführlich analysiert.

fortlaufend verbesserten technischen Lösungen: Der Geist der technologischen Evolution sei aus der Flasche und hat eine globale Dynamik losgetreten, die sich politischer Regulierung widersetzt.

Beide Positionen greifen zur Untermauerung ihrer konträren Argumente und Vorschläge auf sehr ähnliche, stark dystopisch gefärbte Krisenszenarien zurück. Auf der Ebene der Politik spiegelt sich dieses Dilemma aktuell im Entwurf einer europäischen KI-Direktive. Diese führt auf der einen Seite Kriterien für einen mit europäischen Grundwerten verträglichen Einsatz von KI-Systemen ein, eröffnet aber andererseits die Möglichkeit, innovative KI-Systeme unabhängig von diesen Kriterien zeitlich begrenzt in klar definierten Testfeldern zu implementieren. Die Einführung solcher regulatorischen Sandboxes lässt sich als politisches Zugeständnis an die Position der Technooptimisten interpretieren, die in einer ungesteuerten technologischen Innovation die Lösung für technologisch induzierte Probleme sehen. Die Behinderung technologischer Innovation im Namen europäischer Werte und Prinzipien verschaffe zudem der Konkurrenz im globalen Wettbewerb Vorteile und mache die eigenen Gesellschaften damit auch anfällig für Angriffe von außen.

---

## 5 Conclusio

Nach einer Periode der Stagnation gilt KI seit einigen Jahren wieder als vielversprechende Technologie. Dieser Aufschwung basiert auf den technologischen Fortschritten im Bereich der Datenverarbeitung und Datenanalyse, sowie auf den Möglichkeiten durch fortschreitende Vernetzung auf eine wachsende Menge maschinenlesbarer Daten zuzugreifen. Wie bereits in früheren Phasen der Konjunktur von KI wachsen auch aktuell die Prognosen über zukünftige Entwicklungen und Durchbrüche in den Himmel. Das ruft im selben Atemzug die Kritiker\*innen auf den Plan, die auf die drohenden Risiken und Gefahren der prognostizierten Entwicklungen hinweisen. Ein nüchterner Blick auf das behauptete Potential und die Leistungsfähigkeit der Technologien, die als eine Form künstlicher Intelligenz präsentiert werden ist hilfreich, um utopische wie dystopische Überhöhungen zu vermeiden.

Maschinelle Lernprozesse sind weit davon entfernt, menschliche Intelligenz zu ersetzen und hinter der vermeintlich künstlichen, maschinengesteuerten Datenverarbeitung verbergen sich nach wie vor menschliche Arbeit und menschliche Entscheidungen. Bei genauerer Betrachtung der an prominenter Stelle in der Diskussion über KI ins Feld geführten Beispiele zeigt sich, dass es sich dabei

i. d. R. um KI-basierte Systeme handelt, die in einem eng definierten Kontext ebenso eng definierte Aufgaben und Probleme lösen. Risiken und Gefahren drohen hier weniger von der technologischen Leistungsfähigkeit dieser Systeme als vielmehr von den Folgen ihrer unkontrollierten Verbreitung. Durch die fortschreitende Digitalisierung und Datafizierung gesellschaftlicher Prozesse wird der Einsatz von KI-basierten Systemen in allen gesellschaftlichen Bereichen erleichtert. Sie werden zu einem Bestandteil der sozialen Lebenswelt, sie strukturieren den handlungspraktischen Kontext und Horizont im Alltag ebenso wie die Arbeitsweise von Institutionen und Organisationen. Für eine kritische und realistische Einschätzung von Risiken und Gefahren der KI empfiehlt es sich, auf deren je spezifischen diskursiven Rahmungen zu achten und zwischen unterschiedlichen Anwendungsszenarien zu differenzieren. Ordnet man die plakativ platzierten Beispiele durch eine solche Differenzierung perspektivisch ein, lassen sich die utopisch-dystopischen Verkürzungen in der Debatte über Risiken und Gefahren von KI vermeiden.

---

## Literatur

- Agostinho, Daniela, Annie Ring, Kristin Veel, Catherine D'Ignazio, und Nanna Bonde Thylstrup. 2019. Uncertain Archives: Approaching the Unknowns, Errors, and Vulnerabilities of Big Data through Cultural Theories of the Archive. *Surveillance & Society* 17 (3/4): 422–441.
- Alfonseca, Manuel, Manuel Cebrian, Antonio Fernandez Anta, Lorenzo Coviello, Andres Abeliuk, und Iyad Rahwan. 2016. Superintelligence cannot be contained: Lessons from Computability Theory. doi: <https://doi.org/10.48550/arXiv.1607.00913>.
- Anand, Pooja, Yashwant Singh, Arvind Selwal, Pradeep Kumar Singh, Raluca Andreea Fel-seghi, und Maria Simona Raboaca. 2020. IoVT: Internet of Vulnerable Things? Threat Architecture, Attack Surfaces, and Vulnerabilities in Internet of Things and Its Applications towards Smart Grids. *Energies* 13 (18): 4813. doi: <https://doi.org/10.3390/en13184813>.
- Andrejevic, Mark. 2020. Shareable and unsharable knowledge. *Big Data & Society* 7 (1): 205395172093391. doi: <https://doi.org/10.1177/2053951720933917>.
- Ariztía, Tomás. 2018. Consumer databases as practical accomplishments: The making of digital objects in three movements. *Journal of Cultural Economy* 11 (3). doi: <https://doi.org/10.1080/17530350.2018.1435421>.
- Bauman, Zygmunt. 2015. *Liquid Modernity*. Cambridge, UK, Malden, MA: Polity Press; Blackwell.
- Bauman, Zygmunt, und David Lyon. 2016. *Liquid surveillance. A conversation*, 2nd edn. Polity conversations series. Cambridge, Malden: Polity.
- Beck, Ulrich. 1986. Risikogesellschaft. Auf dem Weg in eine andere Moderne. Berlin: Suhrkamp-Verlag.



- Bostrom, Nick. 2005. In defense of posthuman dignity. *Bioethics* 19 (3): 202–214. doi: <https://doi.org/10.1111/j.1467-8519.2005.00437.x>.
- Bostrom, Nick. 2014. *Superintelligence*. Paths, dangers, strategies. Oxford, England: Oxford University Press.
- Bowker, Geoffrey C., und Susan Leigh Star. 1999. *Sorting things out*. Classification and its consequences, 1st edn. Inside technology. Cambridge, Mass.: MIT Press.
- boyd, danah, und Kate Crawford. 2012. Critical Questions For Big Data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society* 15 (5): 662–679. doi: <https://doi.org/10.1080/1369118X.2012.678878>.
- Brown, Sheila. 2006. The criminology of hybrids. *Theoretical Criminology* 10 (2): 223–244. doi: <https://doi.org/10.1177/1362480606063140>.
- Burrell, Jenna. 2016. How the machine ‘thinks’: Understanding opacity in machine learning algorithms. *Big Data & Society* 3 (1): 205395171562251. doi: <https://doi.org/10.1177/2053951715622512>.
- Coletta, Claudio, und Rob Kitchin. 2017. Algorithmic governance: Regulating the ‘heart-beat’ of a city using the Internet of Things. *Big Data & Society* 4 (2): 205395171774241. doi: <https://doi.org/10.1177/2053951717742418>.
- Cowan, Nelson. 2010. The Magical Mystery Four: How is Working Memory Capacity Limited, and Why? *Current directions in psychological science* 19 (1): 51–57. doi: <https://doi.org/10.1177/0963721409359277>.
- Crawford, Kate. 2021. *Atlas of AI*. Yale University Press.
- David, Laurianne, Amol Thakkar, Rocío Mercado, und Ola Engkvist. 2020. Molecular representations in AI-driven drug discovery: a review and practical guide. *Journal of cheminformatics* 12 (1): 56. doi: <https://doi.org/10.1186/s13321-020-00460-5>.
- Dreyfus, Hubert L. 1999. *What computers still can't do*. A critique of artificial reason, 6th edn. Cambridge, Mass.: MIT Press.
- Ducange, Pietro, Riccardo Pecori, und Paolo Mezzina. 2018. A glimpse on big data analytics in the framework of marketing strategies. *Soft Computing* 22 (1): 325–342. doi: <https://doi.org/10.1007/s00500-017-2536-4>.
- Edwards, Lilian, und Michael Veale. 2017. Slave to the Algorithm?: Why a ‘Right to an Explanation’ Is Probably Not the Remedy You Are Looking For. *16 Duke Law & Technology Review* 18. doi: <https://doi.org/10.2139/ssrn.2972855>.
- Gallese, Vittorio, und Alvin Goldman. 1998. Mirror neurons and the simulation theory of mind-reading. *Trends in Cognitive Sciences* 2 (12): 493–501. doi: [https://doi.org/10.1016/S1364-6613\(98\)01262-5](https://doi.org/10.1016/S1364-6613(98)01262-5).
- Ghodselaahi, Ahmad, und Ashkan Amirmadhi. 2011. Application of Artificial Intelligence Techniques for Credit Risk Evaluation. *International Journal of Modeling and Optimization*: 243–249. doi: <https://doi.org/10.7763/IJMO.2011.V1.43>.
- Gibbs, Samuel. 2015. Apple co-founder Steve Wozniak says humans will be robots’ pets. *The Guardian*, 25 Jun. <https://www.theguardian.com/technology/2015/jun/25/apple-co-founder-steve-wozniak-says-humans-will-be-robots-pets>. Zugriff 02 August 2022.
- Giddens, Anthony. 1990. *The consequences of modernity*. Stanford, Calif.: Stanford Univ. Press.
- Gillespie, Tarleton. 2014. The Relevance of Algorithms. In *Media Technologies: Essays on Communication, Materiality, and Society*, ed. Tarleton Gillespie, Pablo J. Boczkowski und Kirsten A. Foot, 167–194: MIT Press Scholarship Online.

- Gillespie, Tarleton. 2016. Algorithm. In *Digital keywords: A vocabulary of information society and culture*, hrsg. Benjamin Peters, 18–30. *Princeton studies in culture and technology*. Princeton [New Jersey], Boston, Massachusetts: Princeton University Press; Credo Reference.
- Gillespie, Tarleton. 2018. *Custodians of the internet. Platforms, content moderation, and the hidden decisions that shape social media*. New Haven, London: Yale University Press.
- Goertzel, Ben. 2015. Superintelligence: Fears, Promises and Potentials. *Journal of Evolution and Technology* 25 (2): 55–87.
- Hassani, Bertrand K. 2021. Societal bias reinforcement through machine learning: a credit scoring perspective. *AI and Ethics* 1 (3): 239–247. doi: <https://doi.org/10.1007/s43681-020-00026-z>.
- Heikkilä, Melissa. 2022. Inside a radical new project to democratize AI. *MIT Technology Review*, Juli 12. <https://www.technologyreview.com/2022/07/12/1055817/inside-a-radical-new-project-to-democratize-ai/>, Zugriff 02 August 2022.
- Hölscher, Christoph, Simon Büchner, Martin Brösamle, Tobias Meilinger, und Gerhard Strube. 2007. Signs and Maps: Cognitive Economy in the Use of External Aids for Indoor Navigation. *Proceedings of the 29th Annual Conference of the Cognitive Science Society (CogSci 2007)*, 377–382 (2007).
- Hutchins, Edwin. 2006. *Cognition in the wild*, 8th edn. A Bradford book. Cambridge, Mass.: MIT Press.
- Jagtiani, Julapa, Lauren Lambie-Hanson, und Timothy Lambie-Hanson. 2021. Fintech Lending and Mortgage Credit Access. *The Journal of FinTech* 01 (01): 2050004. doi: <https://doi.org/10.1142/S2705109920500042>.
- Jansen, Sue Curry. 2022. *What Was Artificial Intelligence?*. mediastudies.press.
- Kennedy, Helen, Thomas Poell, und Jose van Dijck. 2015. Data and agency. *Big Data & Society* 2 (2): 205395171562156. doi: <https://doi.org/10.1177/2053951715621569>.
- Kerr, Ian, und Jessica Earle. 2013. Prediction, Preemption, Presumption: How Big Data Threatens Big Picture Privacy. *Stanford Law Review* 66 (65): 65–72.
- Kitchin, Rob. 2016. Thinking critically about and researching algorithms. *Information, Communication & Society* 20 (1): 14–29. doi: <https://doi.org/10.1080/1369118X.2016.1154087>.
- Kosinski, Michal, David Stillwell, und Thore Graepel. 2013. Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences of the United States of America* 110 (15): 5802–5805. doi: <https://doi.org/10.1073/pnas.1218772110>.
- Kreissl, Reinhard. 2014. Assessing security technology’s impact: Old tools for new problems. *Science and Engineering Ethics* 20 (3): 659–673. doi: <https://doi.org/10.1007/s11948-014-9529-9>.
- Langenbucher, Katja. 2020. Responsible A.I.-based Credit Scoring – A Legal Framework. *European Business Law Review* 31 (Issue 4): 527–572. doi: <https://doi.org/10.54648/eulr2020022>.
- Lazer, David, Ryan Kennedy, Gary King, und Alessandro Vespignani. 2014. The Parable of Google Flu: Traps in Big Data Analysis. *Science* 343: 1203–1205.
- Lyon, David. 2007. Surveillance, Security and Social Sorting: Emerging Research Priorities. *International Criminal Justice Review* 17 (3): 161–170. doi: <https://doi.org/10.1177/1057567707306643>.

- Lyon, David. 2017. Surveillance Culture: Engagement, Exposure, and Ethics in Digital Modernity. *International Journal of Communication* 11 (0): 19.
- Mackenzie, Adrian. 2012. More Parts Than Elements: How Databases Multiply. *Environment and Planning D: Society and Space* 30 (2): 335–350. doi: <https://doi.org/10.1068/d6710>.
- Marcus, Gary. 2020. *The Next Decade in AI: Four Steps Towards Robust Artificial Intelligence*. <https://doi.org/10.48550/arXiv.2002.06177>
- Marcus, Gary, und Ernest Davis. 2019. *Rebooting AI: building artificial intelligence we can trust*. Pantheon.
- Matz, Sandra C., und Oded Netzer. 2017. Using Big Data as a window into consumers' psychology. *Current Opinion in Behavioral Sciences* 18: 7–12. doi: <https://doi.org/10.1016/j.cobeha.2017.05.009>.
- Mazzei, Matthew J., und David Noble. 2017. Big data dreams: A framework for corporate strategy. *Business Horizons* 60 (3): 405–414. doi: <https://doi.org/10.1016/j.bushor.2017.01.010>.
- Mead, George Herbert. 2015. *Mind, self, and society*. The definitive edition. Chicago, London: The University of Chicago Press.
- Miki, Takahiro, Joonho Lee, Jemin Hwangbo, Lorenz Wellhausen, Vladlen Koltun, und Miki Takahiro. 2022. Learning robust perceptive locomotion for quadrupedal robots in the wild. *Science robotics* 7 (62): eabk2822. doi: <https://doi.org/10.1126/scirobotics.abk2822>.
- Moorstedt, Michael. 2022. Sind Maschinen die besseren Menschen? *Süddeutsche Zeitung*, Juli 25. <https://www.sueddeutsche.de/kultur/democratic-ai-gerechtigkeit-kuenstliche-intelligenz-netzkolumne-1.5626822>. Zugriff 02 August 2022.
- Musk, Elon, und Neuralink. 2019. An Integrated Brain-Machine Interface Platform With Thousands of Channels. *Journal of medical Internet research* 21 (10): e16194. doi: <https://doi.org/10.2196/16194>.
- National Security Commission on AI. 2021. Final Report. <https://www.nscai.gov/wp-content/uploads/2021/03/Full-Report-Digital-1.pdf>, Zugriff 02 August 2022.
- Neuralink, 2022. <https://neuralink.com/>. Zugriff 02 August 2022.
- O'Connell, Mark. 2017. The Techno-Libertarians Praying for Dystopia. *Intelligencer*, April 30. <https://nymag.com/intelligencer/2017/04/the-techno-libertarians-praying-for-dystopia.html>. Zugriff 02 August 2022.
- O'Neil, Cathy. 2016. *Weapons of Math Destruction. How Big Data Increases Inequality and Threatens Democracy*, 1st edn. New York: Crown.
- Pasquale, Frank. 2015. *The black box society. The secret algorithms that control money and information*. Cambridge: Harvard University Press.
- Perrow, Charles. 1999. *Normal accidents: Living with high-risk technologies*. Princeton University Press.
- Philippon, Thomas. (2019). On Fintech and Financial Inclusion. *NBER Working Paper No. 26330*. Cambridge, MA. <https://doi.org/10.3386/w26330>
- Pisarchik, Alexander N., Vladimir A. Maksimenko, und Alexander E. Hramov. 2019. From Novel Technology to Novel Applications: Comment on „An Integrated Brain-Machine Interface Platform With Thousands of Channels“ by Elon Musk and Neuralink. *Journal of medical Internet research* 21 (10): e16356. doi: <https://doi.org/10.2196/16356>.
- Popitz, Heinrich. 1976. *Prozesse der Machtbildung*, 3rd edn. Recht und Staat in Geschichte und Gegenwart, 362/363. Tübingen: Mohr.

- Reis, Joao, Paula Espirito Santo, und Nuno Melao. 2019. Impacts of Artificial Intelligence on Public Administration: A Systematic Literature Review. In 2019 14th Iberian Conference on Information Systems and Technologies (CISTI), 1–7. 2019 14th Iberian Conference on Information Systems and Technologies (CISTI), Coimbra, Portugal. 19/06/2019 – 22/06/2019: IEEE. doi: <https://doi.org/10.23919/CISTI.2019.8760893>.
- Rumelhart, David E. 1975. Notes on a Schema for Stories. In *Representation and Understanding*, 211–236: Elsevier.
- Sanders, Carrie, und Camie Condon. 2017. Crime analysis and cognitive effects: The practice of policing through flows of data. *Global Crime* 18 (3): 237–255. doi: <https://doi.org/10.1080/17440572.2017.1323637>.
- Sanders, Carrie B., und James Sheptycki. 2017. Policing, crime and ‘big data’: towards a critique of the moral economy of stochastic governance. *Crime, Law and Social Change* 68 (1-2): 1–15. doi: <https://doi.org/10.1007/s10611-016-9678-7>.
- Schlicht, Ekkehart (2002). Der Homo oeconomicus unter experimentellem Beschuß. *Munich Discussion Paper*. <https://doi.org/10.5282/ubm/epub.70965>
- Schneider, Tanja, und Steve Woolgar. 2012. Technologies of ironic revelation: Enacting consumers in neuromarkets. *Consumption Markets & Culture* 15 (2): 169–189. doi: <https://doi.org/10.1080/10253866.2012.654959>.
- Schütz, Alfred. 1972. *Gesammelte Aufsätze*. Dordrecht: Springer Netherlands.
- Service, Robert. 2020. ‘The game has changed.’ AI triumphs at solving protein structures. *Science*. doi: <https://doi.org/10.1126/science.abf9367>. Zugriff 02 August 2022.
- Smith, Craig. 2021. A.I. Here, There, Everywhere. The New York Times, Februar 23. <https://www.nytimes.com/2021/02/23/technology/ai-innovation-privacy-seniors-education.html>. Zugriff 21 August 2023.
- Symons, John, und Ramón Alvarado. 2016. Can we trust Big Data?: Applying philosophy of science to software. *Big Data & Society* 3 (2): 1–17. doi: <https://doi.org/10.1177/2053951716664747>.
- Todd, Peter M. und Gerd Gigerenzer. (2000). Précis of Simple heuristics that make us smart. *The Behavioral and brain sciences*, 23(5), 727–41; discussion 742–80. <https://doi.org/10.1017/s0140525x00003447>
- Tomasello, Michael. 2010. *Origins of human communication*, 1st edn. A Bradford book. Cambridge, Mass., London: MIT Press.
- van Dijck, Jose. 2014. Datafication, dataism and dataveillance: Big Data between scientific paradigm and ideology. *Surveillance & Society* 12 (2): 197–208.
- Wachter, Sandra, Brent Mittelstadt, und Chris Russell. 2017. Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR. *SSRN Electronic Journal*.
- Warnett, Lawrence, und Brendan McGonigle. 2002. *Unsupervised navigation using an economy principle*. <https://core.ac.uk/download/pdf/85284.pdf>.
- Wang, Christina J. (2018). Technology, the Nature of Information, and FinTech Marketplace Lending. *Current Policy Perspectives* 18–3. Federal Reserve Bank of Boston. <https://www.bostonfed.org/publications/current-policy-perspectives/2018/technology-nature-of-information-fintech-marketplace-lending.aspx>. Zugriff 21 August 2023
- Weidinger, Laura, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, Courtney Biles, Sasha

- Brown, Zac Kenton, Will Hawkins, Tom Stepleton, Abeba Birhane, Lisa Anne Hendricks, Laura Rimell, William Isaac, Julia Haas, Sean Legassick, Geoffrey Irving, and Iason Gabriel. 2022. Taxonomy of Risks posed by Language Models. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, 214–229. FAccT '22: 2022 ACM Conference on Fairness, Accountability, and Transparency, Seoul Republic of Korea. 21 06 2022 24 06 2022. New York, NY, USA: ACM. doi: <https://doi.org/10.1145/3531146.3533088>.
- Wellner, Galit, und Tiran Rothman. 2020. Feminist AI: Can We Expect Our AI Systems to Become Feminist? *Philosophy & Technology* 33 (2): 191–205. doi: <https://doi.org/10.1007/s13347-019-00352-z>.
- Weizenbaum, Joseph. 1966. ELIZA—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9 (1): 36-45.
- Zeba, Gordana, Marina Dabić, Mirjana Čičak, Tugrul Daim, und Haydar Yalcin. 2021. Technology mining: Artificial intelligence in manufacturing. *Technological Forecasting and Social Change* 171: 120971. doi: <https://doi.org/10.1016/j.techfore.2021.120971>.
- Zuboff, Shoshana. 2020. *The age of surveillance capitalism. The fight for a human future at the new frontier of power*. New York, NY: PublicAffairs.



# Toys are us: KI, Diversität und soziale Ungleichheit. Oder: KI für Alle?

Heike Raab

## Zusammenfassung

In diesem Artikel werden KI-gestützte Entwicklungen im Kontext gesellschaftlicher Differenzverhältnisse diskutiert. Im Rahmen dessen sind relevante Aussagen zu marginalisierten Personengruppen erfasst. Ziel ist ein differenziertes Verständnis von KI-Verfahren, in ihrer komplexen Einbettung in gesellschaftliche und kulturelle Wirklichkeiten. Dies geschieht auf zweifache Weise: Geschlecht, Behinderung, Alter, „race“ und Herkunft werden als Zielgruppen ins Visier genommen. Außerdem werden technologische Logiken als Querschnittsthema verhandelt. Ausgangspunkt bildet die These, dass sich diskriminierende Muster und Machtverhältnisse durch KI modernisieren, aber nicht auflösen. Anders formuliert: Technik verändert sich, Ungleichheit bleibt bestehen. Überdies entsteht durch KI die Gefahr neuer Formen von Ungleichheit entlang von Kriterien der technologischen Verwertbarkeit von Vielfalt und Differenz. Mittels KI-Technologien entstehen zudem neue Zugriffspunkte für verschiedene Ökonomisierungs-, Regierungs- und Herrschaftspraktiken. Gleichwohl wird KI-Technik als nicht determiniert verstanden. Viel eher entsteht ein neues Spannungsverhältnis zwischen neuen (kapitalistischen) Regierungs- und Herrschaftspraktiken sowie (politischem) Gestaltungspotential.

---

H. Raab (✉)  
Langenselbold, Deutschland  
E-Mail: [raabheike@freenet.de](mailto:raabheike@freenet.de)

## Schlüsselwörter

Ableism • Behinderung • Disability Studies • Gender/Gender Studies • Inklusion • Universal Design

# 1 Einführung und Problemaufriss

Die Auseinandersetzung mit KI-Technologie erfolgt entlang von drei Achsen: 1. Technologien, die auf dem Tropus von Informationen beruhen; 2. KI gestützte Artefakte; 3. Automaten/Maschinen, die gewissermaßen Abbilder von menschlichen Fertigkeiten darstellen sollen (Suchmann 2019, S. 58). Um hierbei die Spezifika gegenwärtiger Phänomene rund um Künstliche Intelligenz (KI) differenzierter zu verstehen, stehen unterschiedliche Personengruppen (Geschlecht, Behinderung, Alter, „race“<sup>1</sup> und Zugehörigkeit<sup>2</sup>) im Mittelpunkt. Es handelt es sich um Personengruppen, welche für gewöhnlich durch soziokulturelle Zugehörigkeitsordnungen als ‚vulnerable‘ Gruppen ‚gelesen‘ werden. Fragen zu Vielfalt, Inklusion und Teilhabe sind zentral.

Mittlerweile steht künstliche Intelligenz an der Schwelle zum Massenmarkt und zur Alltagstechnologie (Butollo und Nuss 2019, S. 21). Gleichwohl wird diskutiert, inwiefern es zu einem unkontrollierten Wildwuchs in der Entwicklung von KI kommen könne, auch wird der Stand der technologischen Entwicklung befragt. Einerseits werden übertriebene Fähigkeiten von KI-Systemen suggeriert, andererseits warnt die UNESCO das KI-Systeme Diskriminierung verstärken können (Bergt 2023). Aus diesem Grund werden technologische Logiken (Weber und Prietl 2021) von KI als ein soziokulturelles Querschnittsthema erörtert. Im Sinne einer kritischen Gegenwartsdiagnostik ist es das Ziel, KI-Entwicklungen kritisch reflektierend zu befragen. Ausgehend davon soll die Dynamik der Künstlichen Intelligenz mit Bezug auf ungleichheitsrelevante Aspekte dargelegt werden. Im Rahmen dessen erfasst der Artikel relevante Aussagen zu gesellschaftlichen Differenzverhältnissen<sup>3</sup>. Differenzen sind soziale Ordnungskategorien, die durch

<sup>1</sup> „Race“ steht in Anführungszeichen, da es keine menschlichen „Rassen“ gibt. Die Beschäftigung mit „race“ in den Sozial- und Kulturwissenschaften bedeutet stattdessen eine Auseinandersetzung mit verschiedenen Formen des Rassismus. In diesem Sinne geht es in diesem Artikel u. a. um eine rassismuskritische Auseinandersetzung mit Künstlicher Intelligenz (KI). Gefragt wird nach Prozessen der Rassifizierung durch die KI-Entwicklung.

<sup>2</sup> Mit dem Begriff sozialer Zugehörigkeit wird auf sozial-ökonomisch benachteiligte Personen rekuriert. Ein anderer Begriff ist „Klasse“.

<sup>3</sup> Gesellschaftliche Differenzverhältnisse: In diesem Artikel wird von Differenz, bzw. von Differenzkategorien gesprochen, um Differenz als das ‚Andere‘, bzw. Nicht-Identische zu

soziale Ungleichheiten bzw. Machtverhältnisse hervorgebracht werden. Gesellschaftliche Differenzverhältnisse pointiert, im Anschluss an poststrukturalistische Theorie, die wechselnde Art und Weise wie das „Andere“ hergestellt wird“. Ungleichheit kommt insofern ins Spiel, da ich in diesem Artikel davon ausgehe, dass der Zugang zu Ressourcen, bzw. sozialen Positionen, in gesellschaftlichen Differenzverhältnissen, aber auch innerhalb der jeweiligen Differenzkategorie, asymmetrisch reguliert ist. Diese Herangehensweise wird von der Hypothese getragen, dass KI das Soziale verändert. Einerseits verändert KI soziale Ungleichheit und steht so im Spannungsfeld zu Forderungen nach Inklusion<sup>4</sup> und Teilhabe. Andererseits wiederholen sich diskriminierende Muster innerhalb von KI-Technologien. Insgesamt ist von einer Modernisierung sozialer Ungleichheit auszugehen. Modernisierung bezeichnet Prozesse des gesellschaftlichen Wandels. Gesellschaftlicher Wandel entsteht nicht zwangsläufig durch externe Faktoren, sondern erfolgt von innen her als Folge von Modernisierung selbst (vgl. Beck 2015a). Für Beck kennzeichnet der Strukturbruch den Wandel von der einfachen Moderne zu einer reflexiven Moderne. Die Theorie der reflexiven Modernisierung beruht darauf, dass in der Gegenwart nicht mehr davon ausgegangen werden kann, dass durch mehr Wachstum, Wissen und soziokulturelle Differenzierung die gesellschaftliche Strukturierung immer eindeutiger und sicherer wird. Stattdessen wird deutlich, dass die (europäische) Moderne angesichts der Erfahrungen von Kontingenz und unwägbareren Nebenfolgen eher uneindeutiger und unsicherer wird. Auch wenn der Soziologe Beck sich in seinen Thesen zur reflexiven Moderne nicht auf Künstliche Intelligenz bezieht, scheint KI dennoch einen „Modernisierungsschub“ zu evozieren. Gleichwohl ist kritisch anzumerken, dass der optimistische Fortschrittsglaube der sogenannten klassischen Moderne womöglich hauptsächlich für Europa/USA in die Krise geraten ist. Diese Krise gehört schon lange zur Lebenserfahrung des Globalen Südens. Insofern ist der Strukturbruch eine spezifische, geographisch bedingte Wahrnehmung des Westens. Mit Blick auf das Thema dieses Artikels bedeutet dies, dass KI nicht zwangsläufig gesellschaftliche Differenzverhältnisse positiv verändert, sondern ebenfalls

---

kennzeichnen. Zum einen meint Differenz hier Alterität (Andersheit) oder Alienität (Fremdheit). In poststrukturalistischen Debatten ist Differenz das Nicht-Identische, bzw. Andere. Es geht um eine Kritik am dualistischen Denken in der Aufklärung und den diesen begründenden Begriffsapparat der Sozial-/Kulturwissenschaften.

<sup>4</sup> In einer inklusiven Gesellschaft gibt es keine definierte Normalität, welche die Mitglieder einer Gesellschaft anstreben oder erfüllen müssen. Normal ist die Tatsache, dass Unterschiede zwischen den Individuen vorhanden sind. Unterschiede werden als Bereicherung interpretiert. Inklusion beschreibt die Gleichwertigkeit eines Individuums, ohne dass Normalität vorausgesetzt wird. Normal ist vielmehr die Vielfalt.



Teilungsmechanismen hervorruft. Denn es zeigt sich: diskriminierende Muster und Machtverhältnisse werden durch KI nicht grundsätzlich außer Kraft gesetzt, erhalten aber bisweilen ein neues Gewand.

Gesellschaftliche Veränderungsprozesse sind demgemäß daraufhin zu befragen, welche Funktion KI in der Gesellschaft einnimmt. In der Tat scheint die größte Herausforderung zu sein, zu erkennen, welchen Vergesellschaftungsmodus KI generiert bzw. welche sozialen Praktiken und Aneignungsprozesse hervorgerufen werden. Sozial- und kulturwissenschaftliche Studien, die verschiedene Ökonomisierungs-, Regierungs- und Herrschaftspraktiken im Kontext einer expandierenden KI-Technologie erörtern, beziehen sich für gewöhnlich selten auf die Thematik von Differenz und Vielfalt. Stattdessen ist diese Debatte von diversen Schlagwörtern geprägt, die das jeweilig Spezifische von KI als Sozialform zu konturieren versucht (Biniok und Lettkemann 2017, Houben und Prietl 2018, Staab 2020, Schaupp 2021, Pfeiffer 2021). So ist von der „assistiven Gesellschaft“ (Biniok und Lettkemann 2017) die Rede, von der „Datafizierung des Sozialen“ (Houben und Prietl 2018), vom „digitalen Kapitalismus“ (Staab 2020), von „Plattformökonomien“, von „Technopolitik“ (Schaupp 2021), oder der „Gesellschaft der Daten“ (2015), sowie einer spezifischen Distributivkraftentwicklung des Marktgeschehens (Pfeiffer 2021) mit weitreichenden Auswirkungen auf die Gesellschaft. Kritisch muss an dieser Stelle festgestellt werden, dass diese Einschätzung sich tendenziell auf schwache KI-Tools bezieht, während die Implementierung starker KI-Tools auch in ferner Zukunft wohl eher nicht auf der Tagesordnung steht (Daum 2019, S. 311). Gegenwärtig können KI-Tools in begrenzten Anwendungsbereichen kognitive Aufgaben bewältigen. KI-Methoden sind gereift und können das digitale Kapital in die Lage versetzen in bestimmten Bereichen profitable Geschäfte zu machen (ebd. S. 311). Aus dem Blick geraten in diesen Debatten, dass KI-Tools schon jetzt häufig den Alltag umgeben, den alle durchqueren müssen und auf diese Weise Weltzusammenhänge transformieren (Beispiel: Google-Suchmaschine). Es entstehen Konfigurationen und Interaktionen von Menschen und Maschinen, Lebewesen und Technologien, Assemblages aus Natur Materie und KI. Assemblage-Konstellationen unterlaufen die herkömmlichen Unterscheidungen in den Sozialwissenschaften, wie etwa Akteur\*in/Struktur, viel eher entstehen gewebeartige Gebilde (Lorey et al. 2011). Im Vordergrund stehen Interaktionsordnungen, bzw. soziokulturelle Praktiken.

Im zunehmenden Ausmaß spielen im Kontext von KI-Entwicklungen so bezeichnete vulnerable Gruppen eine Rolle (Geschlecht, Alter, soziale Zugehörigkeit, „race“ sowie Behinderung). Einerseits handelt es sich bei diesen Personengruppen um spezielle Zielgruppen andererseits werden über neue KI-Technologien neuerlich gesellschaftliche Differenzverhältnisse verhandelt.

Um wissenschaftliche Leerstellen zu vermeiden, gilt es deshalb zu untersuchen, welche alten und neuen Ungleichheiten durch KI im Bereich gesellschaftlicher Differenzverhältnisse entstehen. Im Rahmen dessen plädiere ich dafür, die hier erörterten benachteiligte Personengruppen nicht als jeweils in sich homogene Gruppe zu konzipieren, sondern die Vielfalt in der Differenz zu pointieren. Hierauf beruht die weitere These: Entwicklung und Einsatz von KI erfolgen entlang einer technologischen Verwertbarkeit von Vielfalt und Differenz.

---

## **2 KI, gesellschaftliche Differenzverhältnisse und soziale Ungleichheit**

Mit diesen Überlegungen docke ich einerseits an neuere Ansätze zu Ungleichheit und Sozialstrukturanalyse an, gehe aber gleichzeitig darüber hinaus. Anders formuliert: Es stellt sich die Frage, wie sich die Auseinandersetzung zwischen Arbeiten über gesellschaftliche Differenzverhältnisse und sozialer Ungleichheit zueinander verhalten, ohne neuerlich Geschlecht oder beispielsweise Behinderung auf einen Nebenwiderspruch kapitalistischer Klassenverhältnisse zu reduzieren.

So geht die Ungleichheits- und Sozialstrukturanalyse seit den späten 1980er Jahren von einer pluralisierten Klassengesellschaft aus und betont Pluralisierung sowie Individualisierung (Imbusch 1998, S. 275 ff.). Ungeachtet des in dieser Zeit eingetretenen Wandels des Arbeitsmarktes durch Phänomene wie neuerliche Massenarbeitslosigkeit (im Vergleich zu den 1960er Jahren), Niedriglohnsektor (Hartz4) u.v.m. wird in diesen Ansätzen von einer Pluralisierung von Lebensweisen ausgegangen. Insofern sind die Ungleichheits- und Sozialstrukturanalysen jener Jahre eingebettet in eine Umbruchsphase, welche von einer Vollbeschäftigung in den 1950er Jahren hin zu hohen Arbeitslosenzahlen ab den 1970er/1980ern führt. Dessen ungeachtet stehen Bildung und Einstellungsfragen im Mittelpunkt von Ungleichheitsforschung, die zu Aussagen über pluralisierte Milieus und Lebensstilenklaven führen. Begründet wird dieser Zuschnitt mit einer nach dem zweiten Weltkrieg eingetretenen allgemeinen Erhöhung des materiellen Wohlstandes. Insofern findet Soziale Ungleichheit nach dem zweiten Weltkrieg auf einem höheren materiellen Niveau statt (Beck 2015b). Es kommt zur Verbesserung des allgemeinen Lebensstandards, ein Umstand der bedingt reflektiert wird.

Gleichwohl stellt die sich auflösende klassische Industriegesellschaft den Hintergrund, für die Diagnose einer Pluralisierung sowie Individualisierung des Sozialen, dar. Für dieses Phänomen prägte Beck den Begriff Risikogesellschaft

(Beck 2015b). In der Risikogesellschaft spielt laut Beck, Status eine untergeordnete Rolle, denn von den gesellschaftlichen Transformationsprozessen seien Alle betroffen. Mit dem Bruch der sogenannten (europäischen) Moderne<sup>5</sup>, durch das Erodieren der Industriegesellschaft, werden soziale Risiken nicht ausschließlich nach Klassengrenzen geordnet. Gleichwohl bleibt eine Klassenlage erhalten, die sich nach wie vor über den Zugang zu Ressourcen und materieller Ungleichheit bestimmt. Gleichzeitig sind vertikale Ungleichheiten vielfältig gebrochen (Imbusch 1998, S. 294). Der mit diesen Phänomenen verbundene Neoliberalismus fungiert als Begleitmusik dieser Prozesse. Der Bruch mit der industriellen Moderne erhält durch den zu dieser Zeit erstarkendem Neoliberalismus das Antlitz eines freiheitlichen Individualisierungsversprechens. In der Tat scheint das Wesensmerkmal des Neoliberalismus ein liberales Freiheitsversprechen im Verbund mit einer zunehmenden Kommodifizierung des Sozialen in allen Bereichen zu sein (Fraser 2003, Lemke 2000, Raab 2011). Jüngste technologische Entwicklungen im Kontext von KI sind, meines Erachtens, Teil jener beschriebenen Erosionen der Industriegesellschaft, auch wenn erst gegenwärtig von einer umfänglicheren KI-Anwendung gesprochen werden kann.

Hauptsächlich zeigt die Auseinandersetzung um gesellschaftliche Differenzverhältnisse, welche parallel Fahrt aufnimmt und eine zunehmende Aufmerksamkeit erhält, die Grenzen jener Pluralisierung und Individualisierung. Einerseits entsteht die berechtigte Kritik, dass Phänomene wie Misogynie, Rassismus oder

---

<sup>5</sup> Eng mit der Kritik am Kolonialismus verbunden, ist die radikale Analyse von Modernität und der gewaltsamen Eroberung weiter Teile der Welt durch europäische und andere Mächte. Die mit dieser Eroberung einhergehende Ideologie kultureller und technischer Überlegenheit sowie ihre verwendeten Dichotomien, in der Beschreibung der Welt, werden in der postkolonialen Theorie radikal infrage gestellt. Aus diesem Grund gibt es eine gesellschaftswissenschaftliche Diskussion über den Begriff der Moderne. Im Mittelpunkt steht die Kolonialität der Moderne. Deswegen wird inzwischen von einer pluralen, vielfältigen Moderne gesprochen, die nicht allein eine westliche Perspektive einbezieht (vgl. Kerner 2021). Aufgrund ihrer engen Verknüpfung mit militärischer und patriarchaler Gewalt, kapitalistischer Ausbeutung und Zerstörung der Natur, ist die Geschichte der „modernen“ Technologie weder „unschuldig“ noch neutral. Technik wurde als Mittel der Kolonisierung eingesetzt und heutige Innovationen basieren nicht selten auf Rohstoffen, die unter postkolonialen Ausbeutungsverhältnissen gewonnen werden. Die hegemoniale Vorstellung von technischem Fortschritt und die ihr zugrunde liegenden Ideen von Aufklärung und Rationalismus sind eng mit den Herrschaftsansprüchen des globalen Nordens verflochten. Zwar tritt die technologische Entwicklung mit dem Anspruch auf, universalistisch dem Wohl aller Menschen zu dienen, tatsächlich bleibt sie aber den partikularen Interessen postkolonialer Metropolen verpflichtet. Damit stellt sich die Frage, wie die Verzahnung des technologische Fortschrittsversprechens mit der europäischen Moderne gelöst werden kann (Jour Fixe Initiative, Berlin & Frankfurt 2022).

Behindertenfeindlichkeit nicht unter der Rubrik Klasse subsumiert werden können. Durch diese Diskriminierungsformen entstünden vielfältige gesellschaftliche Benachteiligungen, die zur sozialen Ungleichheit führten. Gleichzeitig wandle sich der Umgang mit Differenzen in der Spätmoderne, welche, wie beschrieben, durch eine Krise der Industriegesellschaft gekennzeichnet sei. Gesellschaftliche Differenzverhältnisse würden in der Spätmoderne, einer kriselnden Industriegesellschaft, nicht mehr allein durch Ausschließung und Verwerfung reguliert. Stattdessen würden bestimmte benachteiligte Personengruppen über flexibilisierte Ein- bzw. Ausschlüsse reguliert (Engel 2007, Waldschmidt 2014). Für die Disability Studies kommt daher etwa Waldschmidt zu dem Schluss, dass ein multiperspektivischer Ansatz, der vornehmlich die an Klasse orientierte Sozialstrukturanalyse mit dem die Mehrdimensionalität von sozialen Ungleichheiten fokussierenden Intersektionalitätsansatz und die Diversity Studies mit ihrer kulturalistischen Perspektive verknüpft, am ehesten innovative Erkenntnisse verspreche, wenn es um ein solch facettenreiches Phänomen wie Ungleichheit geht (Waldschmidt 2014). Analoges lässt sich zu den in diesem Artikel pointierten weiteren Personengruppen sagen.

Die Auseinandersetzung über gesellschaftliche Differenzverhältnisse verkennt allerdings bisweilen, dass die jeweilige Regulation der Ökonomie die materielle Verfasstheit von Gesellschaft prägt und folglich auch die Gestaltung von Lebensweisen und Differenzverhältnissen. So sind beispielsweise „Minderheiten“-themen zu regelrechten Marktnischen mutiert, die von der Frauen\*literatur bis hin zum barrierefreien Tourismus neue Märkte erschlossen haben. In diesem Sinne wird eine ökonomische Monokultur befürchtet, die Minorisierte ambivalent über marktwirtschaftliche Mechanismen einbezieht (Holert und Terkessidis 1997). Gleichzeitig ist hiermit auch ein Fragehorizont für die Erforschung von KI, im Kontext gesellschaftlicher Differenzverhältnisse, skizziert – sofern KI potentiell eine spezifische Regulation der Ökonomie bereitstellt, ist zu vermuten, dass die materielle Verfasstheit von Gesellschaft sich verändert. Dieser Aspekt tangiert Fragen zu sich transformierenden soziokulturelle Umgangsweisen mit Geschlecht, Behinderung, Alter, „race“ und Herkunft.

Hinsichtlich des gegenwärtigen Entwicklungsstandes von Künstlicher Intelligenz ist aus diesem Grund zu vermuten, dass ähnliche Effekte, wie weiter oben skizziert, im Bereich gesellschaftlicher Differenzverhältnisse eintreten (können). Neue Formen sozialer Ungleichheit entstehen entlang von Ressourcen, Zugang zur Teilhabe sowie Gestaltung von KI-Produkten, oder entlang der Problematik KI-kompatibler Körper. Letzteres bezieht sich auf die Frage, welche Körper relevante Körper im Bereich der KI-Entwicklung sind. Vermutlich kann nicht jeder Körper mit KI-gestützter Technologie ‚beglückt‘ werden, da nicht

jede Form von Alter oder Behinderung technologisch bewältigt werden kann. Insgesamt scheint die Entwicklung ambivalent und dies nicht nur mit Blick auf den technologischen Entwicklungsstand. Einerseits wird KI-Technologie nur bedingt zielgruppengerecht entwickelt, Diversität wird in den Entwicklungsprogrammen vernachlässigt. Teilweise entstehen in diesem Bereich sich widersprechende Entwicklungs-/Nutzer\*inneninteressen. Zum anderen besteht die Gefahr einer neuerlichen, nunmehr KI-gestützten, asymmetrischen Regulation gesellschaftlicher Differenzen. Partizipation und Teilhabe durch Betroffenenengruppen erhält keinen systematischen Platz in der KI-Entwicklung, hierunter leidet die Praxistauglichkeit.

---

### **3 Die aktuelle Debatte: Die Science and Technology, Gender/Disability Studies und Ungleichheitsdynamiken**

Selbstlernende Roboter, autonome Fahrzeuge, Wearable Technologies, Biohacking, Body Enhancement, Augmented Reality und Big Data sind nur einige Phänomene, die auf die Dynamik der o.g. Entwicklungen im Bereich der neuen, v. a. biotechnologischen und informatischen, Technologien und gesellschaftlichen Praktiken seit den 1980er-Jahren verweisen. Als Beispiel sei hier in aller Kürze auf die Entwicklung der Kybernetik verwiesen. Die Kybernetik, eine Voraussetzung der Entwicklung von KI-Gerätschaft, zielt, in ihrer Geschichte, bekanntlich darauf Gemeinsamkeiten zwischen dem Gehirn und Computern zu untersuchen und Schnittstellen verschiedener Einzeldisziplinen zu erkennen, welche menschliches Verhalten/Denken, Nachrichtenübertragung, Regelungstechnik, Entscheidungs- und Spieltheorie und statistische Mechanik betrachtet. Für die Entstehung der Kybernetik sind diese Dinge von zentraler Bedeutung (Fink 2021). Ziel ist eine Theorie der Kommunikation und Steuerung zu entwickeln, die gleichermaßen auf Menschen wie auf Maschinen anwendbar ist. In diesem Sinne gilt die Kybernetik als Vorläufer bzw. Grundlage der KI-gestützten Technologie. Denn in der Kybernetik hat Information nichts mehr mit Inhalt oder Bedeutung zu tun. Information, in diesem Verständnis, ist weder Materialität noch Präsenz, sondern wird als abstraktes Muster oder Struktur gefasst. Gesendet werden nicht Nachrichten oder Botschaften, sondern Signale, und nur wenn die Nachricht so codiert ist, dass ihr Signal auf ein Medium übertragen wird, nimmt sie eine materielle Form an, also beispielsweise beim E-Mail-Verkehr, oder in der Google-Suchmaschine. Information/Daten werden zu einer quantifizierbaren Größe. Als Folge werden Information und Bedeutung letztendlich entkoppelt, Information ist

selbst keine Bedeutungsträger\*in sondern eine statistische Größe und ein binär-digitaler Code (Fink 2021, S. 29). Frühe kybernetische Geräte simulieren auf diese Weise somit menschliches Gehirn als Maschine, während die Informationstheorie parallel eine Art entkörperertes Denken einführt, welches vordergründig nicht-binär, beispielsweise an binäre Geschlechterkonstrukte, gekoppelt ist, sondern scheinbar eine „neutrale“ Informationsvermittlung (Fink 2021) betreibt. Insofern führt schon die frühe Kybernetik zu neuen gesellschaftlichen Praktiken, beispielsweise in der Informationstechnologie.

Heutzutage werden Schnittstellen bzw. fließende Übergänge geschaffen, wie z. B. Brain Computer Interfaces, neue handlungsfähige Akteur\*innen treten auf und soziotechnische Systeme wie ›virtuelle Assistent\*innen‹ oder das ›Internet der Dinge‹ fordern Politik und Wissenschaft auf bislang ungewohnte Art heraus, etwa in Bezug auf Big Data oder die Privatsphäre (Feministische Studien 2019, S. 241).

In Anbetracht jener Entwicklungen ist es wenig verwunderlich, wenn sich auch die Sozial- und Kulturwissenschaften mit dieser Problematik beschäftigen. Im Bereich Technik/KI gewinnt ein Forschungsfeld an Bedeutung, das sich erst in den 1970ern herausbildete: die Science and Technology Studies (STS). Die STS untersuchen nicht nur die Fabrikation wissenschaftlichen Wissens, sondern nehmen auch technologische Artefakte oder Praktiken in den Blick. STS demonstrieren, dass Wissen, technische Artefakte, Menschen und Macht nicht voneinander getrennte Instanzen sind (Bauer et al. 2017). Vielmehr wirken diese Instanzen wechselseitig aufeinander ein. In diesem Sinne sind Artefakte soziale Phänomene, die in machtvollen Aushandlungsprozessen hervorgebracht werden. Technische Artefakte gelten als Akteur\*innen, weil Wissen in konkreten technisch-apparativen Praktiken entsteht (ebd.). Hervorzuheben ist der neuartige Fokus auf Materialität und das prozessuale Verständnis der Genese von Wissen. Zwei Forschungsperspektiven haben die weitere Perspektive seitdem entscheidend geprägt: die Akteur-Netzwerk-Theorie (ANT) (Latour 2010) und die feministische Wissenschaftskritik (Gender 2019; Weber 2002). Um eine möglichst genaue empirische Beschreibung von Akteur\*innenkonstellationen und Interaktionsformen zu erhalten, geht die ANT von einem Kontinuum von Netzwerken und Akteur\*innenkoalitionen aus. Eine Unterscheidung zwischen menschlichen und nicht-menschlichen Akteur\*innen wird aufgegeben. Prägend ist hierfür der Begriff des Aktanten von Latour (Latour 2010). Aktant bezeichnet die jeweilige interagierende materielle Entität, gleichwohl verzichtet Latour nicht auf den Akteursbegriff, sondern relativiert dessen Position (ebd.). Diese Herangehensweise hat weitreichende Folgen für gesellschaftswissenschaftliche Forschung, da Unterscheidungen wie etwa die Natur-Gesellschaft-Differenz,

ebenso wie eine anthropozentrische Fixierung erodieren. Insgesamt erscheinen in dem Zugang der ANT technologische Artefakte als Ergebnis einer Koproduktion heterogener Akteur\*innen (Feministische Studien 2/2019). Letztlich lässt sich schlussfolgern, dass sich soziale Strukturen über soziale Praktiken von Akteurs-Netzwerken entfalten und verstetigen (Feministische Studien 2/2019). Barad (2012) betont in diesem Zusammenhang den Vorrang von Intra-Aktionen und verzichtet eher auf einen Akteur\*innenbegriff.

Vorläufer dieser sozialwissenschaftlichen Debatte sind die feministischen Technowissenschaften (z. B. Haraway 1995 a/b). Diese befragen das Verhältnis von Menschen, Maschine und Natur aus einer geschlechterkritischen Perspektive. Gegenstand sind die sozialen und politischen Folgen von technologischen Entwicklungen für die Geschlechterverhältnisse. Einerseits wenden sich feministische Technowissenschaften gegen eine Gegenüberstellung von Natur und Kultur – es geht um das Vermischen des Organischen mit dem Technischen. Beobachtet wird eine empirische Entwicklung in Science-Fiction, in der Medizin, bei Reproduktionsverhältnissen, in der Kommunikation und im modernen Krieg (Fink 2021). Gleichzeitig werden neuartige Subjektivierungsweisen erforscht und eine anthropozentrische Sichtweise hinterfragt. D. h., es geht darum zu ergründen, wie Wahrnehmungs- und Affektmuster, Körpererfahrungen, Denken und Reflexionsmöglichkeit durch KI geformt werden (Fink 2021; Gender 4/2019; Feministische Studien 2/2019).

Technowissenschaft ist eine Wortneuschöpfung für soziale Praktiken, bei denen Technik und Gesellschaft untrennbar verbunden sind und begrifflich nicht mehr auseinandergehalten werden können. Denn gesellschaftliche Technikverhältnisse werden nicht nur in sozioökonomischen und politischen Strukturen, sondern auch im Handeln der Menschen – in alltäglichen Aneignungs- und Entwicklungspraktiken – (re-)konfiguriert und materiell hervorgebracht.

Forschung nach Normen und Strukturen von Geschlechterverhältnisses in diesen Technologien sind für Gender Studies wegweisend. Diesbezüglich wird auch erörtert, inwieweit KI-Entwicklungen die Kategorie Geschlecht beeinflussen. Die Gender Studies pointieren hierbei insbesondere die Verwobenheit zwischen Körper und Technik. Im Zuge der Etablierung biodigitaler Maschinen entfalten sich außerdem neuartige Formen von Kommunikation und Reproduktion. So mutieren körperliche Substanzen zu entkörperlichten digitalen Daten, beispielsweise in der Gen-/Reproduktionstechnologie, welche wiederum als Marktprodukte eingesetzt werden können. Hier mutiert beispielsweise die weibliche Eizelle zur Ware und zugleich verdichtet sich darin eine neuartige Form der Bevölkerungspolitik (Kalender 2011). Ebenso gibt es digitale Technologien, welche in der Erforschung, Diagnostik und Therapie von autistischen Menschen zum Einsatz

kommen. Hier stellt sich nicht nur die Frage nach Prozessen der Datafizierung von Neurodiversität (Wentz 2022) und ihren Folgen, sondern es entsteht auch das Feld von Technoableism (Hamraie und Fritsch 2019), welches in der Hauptsache diskutiert, inwiefern spezifische KI-Tools barriere- und diskriminierungsfrei sind oder an der „Überwindung“ von Behinderung als störendes „Element“ arbeiten. Last but not Least ermöglicht die kostenlose KI-Chat-BOT ChatGPT Schriftstücke ohne Subjekt (Autor\*innenschaft) und ist inzwischen in aller Munde, da in dieser KI das denkende Gehirn simuliert wird.

Zentral ist das Verhältnis von Materialität und Technologien. In einer kontroversen Debatte werden in den Gender Studies jene Potentiale hervorgehoben, die in diesem Prozess Geschlecht, Körper und Geschlechterverhältnisse de-essentialisieren. Es wird diskutiert, inwiefern neue KI-Technologien die gesellschaftliche Wahrnehmung bzw. die Praxen von Geschlecht verändern. Im Anschluss an Haraway geht es in dieser Debatte um die fortschreitenden Technologisierung der Gesellschaft. Darin begründet sich eine Utopie von einer neuen Gesellschaft, zudem wird das Aufbrechen gängiger Denkkategorien gefordert. Ausgangshypothese bildet die Annahme, eine Trennung zwischen Natur und Technik sei angesichts einer zunehmenden Technologisierung obsolet. Analog würde diese Entwicklung zentrale Elemente des westlichen Denkens, welches im Wesentlichen in asymmetrisch, binären Kategorien angelegt ist, hinfällig machen – beispielsweise sei die Dichotomie Natur/Kultur nutzlos bzw. überholt, da es eine zunehmende Vermischung zwischen Technik und natürlich gedachter Materie käme. Zudem sei Natur letztlich nur durch soziale und kulturelle Prägungen wahrnehmbar und interpretierbar. Ein Umstand, der nicht zuletzt auf die Kritik an einer androzentristischen Wissenschaft durch die Gender Studies verweist (Weber 2003). In der Kritik an einer androzentristischen Wissenschaft wird in den Gender Studies der Male-Bias in vielen Bereichen von Forschung dokumentiert. Diese Ausrichtung der Gender Studies weist Technikfeindlichkeit zurück, es gelte vielmehr, sich Technologien subversiv anzueignen. KI böte, ähnlich wie das Konzept von Haraways Cyborg, potentiell Allianzen an, um als neue symbiotische Lebensform gemeinsam gegen autoritäre Macht-/Herrschaftsapparate zu kämpfen. In dieser Lesart bietet KI das Potential, sich in einem Akt subalternen Solidarität, gegen die allumfassende Macht von Großkonzernen oder Interessen der Kulturindustrie zu wehren (Schmid 2023). Inzwischen ist dieser Euphorie eine gewisse Ernüchterung gewichen. Stattdessen wird vermehrt auf algorithmische Gewalt und Ungleichheit fokussiert sowie eine Re-Traditionalisierung von Geschlecht durch KI-Tools erörtert (Bath und Bauer 2005; netznova 2020).



In den Disability Studies wird wiederum das Wechselverhältnis von Krankheit/Gesundheit und Gesellschaft vielfach herausgearbeitet und die Erkenntnis erörtert, dass über dieses Wechselverhältnis Normalitätszonen verhandelt werden, über die gesellschaftliche Ein- und Ausschlüsse sowie Zugehörigkeitsordnungen entstehen. Innerhalb der Disability Studies sind Probleme von KI-gestützten Technologien und deren Auswirkung auf Behinderung bislang marginal, aber stetig wachsend, da KI-Entwicklung auch in diesem Bereich vorangetrieben wird. Die Entwicklung von technischen Hilfsmitteln und ihre Nutzung durch Menschen mit Behinderungserfahrung sind ein wichtiges, aber bislang nur punktuell bearbeitetes Thema. So wird in den Disability Studies zunächst KI-gestützte Technologie kritisch begutachtet. Der Fokus beruht auf Barrierefreiheit und Universal Design<sup>6</sup> sowie auf der Leistungsfähigkeit der Produkte. Im Anschluss an feministische Technowissenschaft haben sich ferner die „crip technoscience(s) as a field of research and a practice of critical ‘knowing-making’“ (Hamraie und Fritsch 2019, S. 3) entwickelt. Im Unterschied zur allgemeinen Technikwissenschaft wird in diesem Ansatz eine Praxis kritisiert, in der Expert\*innen für Menschen mit einer Behinderungserfahrung Produkte entwickeln. Stattdessen werden partizipative Ansätze favorisiert, in denen Produkte mit oder von Menschen mit Behinderung entwickelt werden. Gleichzeitig wird ein Verständnis von Inklusion und Barrierefreiheit infrage gestellt, welches die ableistischen Normen der bestehenden Umwelt affirmiert. Unter Ableism ist die Ausrichtung an Normen der Nichtbehinderung zu verstehen. Messlatte sind die Fähigkeitsnormen von nicht-behinderten Personen. Hegemoniale Normen werden in der KI-Entwicklung im Bereich Behinderung/Krankheit eher bestätigt, denn reflexiv hinterfragt. Ein Umstand der sich schon durch die mangelnde Berücksichtigung von Inklusion und Barrierefreiheit in vielen Produkten bemerkbar macht.

Mit Blick auf den Ansatz der ANT wird hingegen in den Disability Studies darauf hingewiesen, dass Interdependenzen zwischen Technologien und Personen stärker ausgelotet werden sollten. ANT als ein analytisches Werkzeug ermöglicht ein vertieftes Verstehen der Modi, wie Behinderung in der Gesellschaft verhandelt wird, wie über KI-gestützte Tools Praktiken, Agency und Effekte zirkulieren

---

<sup>6</sup> Universal Design, deutsch Universelles Design, ist ein internationales Design-Konzept, welches Produkte, Geräte, Umgebungen und Systeme derart gestaltet, dass sie für so viele Menschen wie möglich, ohne weitere Anpassung oder Spezialisierung, nutzbar sind. Der Begriff „Universelles Design“ wird häufig mit anderen Begriffen, wie z. B. „Design für Alle“ oder Barrierefreiheit, gleichgesetzt und synonym gebraucht. Doch auch wenn sich die Begriffe sehr ähneln, unterscheiden sie sich in ihren Grundgedanken. Design für Alle als europäische Strategie meint aus diesem Grund, verschiedene Gruppen von Menschen zu integrieren, ohne eine Einheitlichkeit zu erzwingen.

und dabei womöglich das Verständnis von Behinderung neu akzentuieren. Denn gemäß dem ANT-Ansatz sind sozio-materiale Praktiken keine Fakten. Vielmehr entstehen sie als Effekte von Interaktionen, von Othinging oder Ordnungen (Ravneberg und Söderström 2018, S. 4–8). So genannte assistive Technologien (AT) werden im Kontext von Behinderung genutzt, um den Alltag zu erleichtern, etwa in der Pflege oder als Smart House, im Sinne eines Alarmsystems, welches etwa im Falle eines Sturzes zum Einsatz kommt etc. Assistive Technologien werden von Personen mit einer Behinderungserfahrung oftmals als eine Expansion des Körpers erfahren (ebd., S. 7). Assistive Technologien sind dennoch auch umstritten, wie etwa das Cochlea Implantat zeigt. Taube Menschen verstehen sich als minorisierte Sprachengemeinschaft. Für diese Personengruppe ist das Cochlea Implantat ein medizinisches Symbol, welches oralen Sprachgebrauch als ableistische Norm setzt, hingegen die Gebärdensprache pathologisiert (ebd.).

Anders gesagt, der Begriff ‚assistive Technologien‘ wird kritisch in den Disability Studies diskutiert, da er ein ambivalentes Verständnis von Behinderung transportiert (siehe dazu weiter unten). Auch wird Universal Design unterschiedlich wahrgenommen. Während beispielsweise das Smartphone für viele Menschen der Dominanzgesellschaft im Alltag selbstverständlich ist, erleben Menschen mit einer Behinderungserfahrung dieses überwiegend als Abhängigkeit (Whittaker et al. 2019) von einem technischen Gerät. Der Einsatz von AT kann je nach Betroffenengruppe unterschiedlich ausschauen, so divergent wie es Formen von Behinderung/chronischer Krankheit gibt. Teilweise wird menschliche Assistenz als zuverlässiger eingestuft denn technologische Assistenz. Zum anderen ist Abhängigkeit von Unterstützungssystemen ein grundsätzliches Thema in den Disability Studies und kommt im Zusammenhang mit der Care-Debatte zum Tragen. In der Care-Debatte geht es um die Beziehung zwischen Menschen mit einer Behinderungserfahrung, Assistenz, bzw. Assistenzdiensten sowie dem ambulanten Hilfesystem. Gleichwohl betont diese Debatte verstärkt Arbeitsverhältnisse und Fragen zum Expert\*innen/Laien-Verhältnis im Bereich der sozialen Berufe. Im Mittelpunkt stehen Fürsorge-Relationen.

Nicht zuletzt werden aus diesen weiter oben genannten Grund AT in den Disability Studies ambivalent eingeschätzt, darunter auch KI-gestützte Assistenz. Die Disability Studies gehen davon aus, dass KI-Tools grundsätzlich darauf zu befragen sind, welche ableistischen Normalisierungsvorstellungen sie transportieren, die sich zu problematischen Formen und Überbietungslogiken steigern können und auf diese Weise ableistische Stereotype reproduzieren.

Die Behindertenrechtskonvention der Vereinten Nationen (UN-BRK) fordert daher ein universelles Design von „Produkten, Umfeldern, Programmen und Dienstleistungen in der Weise, die von allen Menschen weitgehend ohne eine

Anpassung oder ein spezielles Design genutzt werden können“ (Art. 2 UN-BRK). Allerdings können – selbst bei einem universellen Design – für bestimmte Gruppen von Menschen Hilfsmittel erforderlich sein. Während die deutsche Rechtsprechung von Hilfsmitteln spricht, wird von der Weltgesundheitsorganisation (WHO) der Begriff „Assistive Technologien“ oder „Assistive Produkte“ verwendet. „Assistive Technologie“ ist ein Kürzel für eine Technologie, welche Personen mit einer Behinderungserfahrung behilflich sein soll. Der Begriff wird in den Disability Studies – wie schon weiter oben erwähnt – zwiespältig gesehen. Einerseits sei die Unterscheidung unnützlich, da Technologie entweder hilfreich sei (und in diesem Sinne immer assistiere) oder nicht, während der Begriff gleichzeitig suggeriere, dass Nutzer\*innen Assistiver Technologien „passiv“ seien (Whittaker et al. 2019, S. 5).

In der Materialität digitaler Assistenztechnologien artikulieren sich ferner soziale, politische und wirtschaftliche Interessen gegenüber älteren Menschen. Sie verändern soziale Interaktionen und Beziehungen, sie schließen Akteur\*innen ein, wie sie andere ausschließen. Dabei wird das Alter (aber auch Behinderung) zu einer technisch konfigurierbaren Größe, die sich entsprechend sozial- und innovationspolitischer Interessen gestalten lässt. Darin sind KI-gestützte technische Assistenzsysteme materieller Ausdruck einer Politik, die die demografische Alterung der Gesellschaft als ökonomische und gesellschaftliche Belastung rahmt und dabei politisches Handeln dadurch legitimiert, dass die Bewältigung dieser Problemlagen mit der Entwicklung und Anwendung technischer, digitaler und robotischer Assistenz vollzogen werden kann. Dabei verdeutlicht sich, dass Technik immer nur eine Möglichkeit darstellt, Alter(n) (und Behinderung) zu artikulieren. Gleichzeitig erschwert diese eindimensionale Artikulation die Verwirklichung anderer Formen des Alt-Seins und Alt-Werdens bzw. des Behindert-Seins und -Werdens (Endter 2020).

---

#### **4 Risiko, Einbettung, Gestaltung – KI, soziale Veränderung und benachteiligte Personengruppen**

Die Forschung zu Auswirkungen und Differenzkonstruktionen von KI fokussiert bislang zu wenig auf Ungleichheit, Alltag und Alltagsroutinen in der Gesellschaft. Auch ist davon auszugehen, dass lediglich ein Teil der aktuell erprobten, entsprechend ausgerichteten KI-Produkte marktfähig wird (Blanc und Beudt 2022; Klein 2020). In einer Bestandsaufnahme aus dem Projekt, KI.Assist – Monitoring KI-gestützter Assistenztechnologien für Menschen mit Behinderungen werden beispielsweise verschiedene Zukunftsszenarien skizziert, die in der optimistischen

Lesart von einer Umsetzung von 60 % ausgehen; Klein hingegen diagnostiziert umfassenden Forschungsbedarf, da die Alltagspraxistauglichkeit viele Produkte kaum erforscht sei (Blanc und Beudt 2022; Klein 2020).

Soziale Ungleichheit bzw. Ausschluss kann dadurch entstehen, dass KI-Techniken überwiegend mit Blick auf das Marktgeschehen hergestellt werden und dadurch bestimmte Nutzer\*innengruppen unberücksichtigt bleiben. So bleiben beispielsweise im ÖPNV (öffentlicher Personennahverkehr) autonom fahrende Busse bislang hinter den Erwartungen bezüglich einer Ausrichtung auf Vielfalt, Diversität und Inklusion zurück (siehe dazu weiter unten).

Die Relevanz des Marktes zeigt sich auch beim Thema Universelles Design. Obwohl die Relevanz von Universellem Design beträchtlich ist, wird das Konzept in gestalterischen Prozessen bisher kaum umgesetzt. Eine Ursache ist die marktvermittelte Regulierung von KI-Technologien. Aus diesem Grund sind Konzepte eines Designs für Alle vornehmlich im Bereich des Design-Aktivismus<sup>7</sup> anzutreffen (Bieling 2019; 2021). Entsprechend langsam ist die Entwicklung von KI-gestützten Services. Hier zeigt sich, dass die Diskussion um wohlfahrtsstaatliche Regulierungen um die Dimension KI-gestützter Technologien erweitert werden muss. Unterstützung sollte nicht allein auf Barrierefreiheit zielen, sondern schon spezifische Produktentwicklung fördern.

Inklusion ist daher nicht allein als technologische KI-gestützte Intervention zu verstehen. Anstatt KI-Technologien isoliert zu begutachten, ist der jeweilige Kontext in dem KI-Produkte hergestellt werden, zu beachten. Ältere oder Menschen mit Behinderungserfahrung sind besonders darauf angewiesen, dass beispielsweise KI-gestützte Hilfsmittel im sozialen Umfeld eingebettet sind.

KI führt tendenziell zu neuen Formen der Standardisierung der Umwelt, des Wissens und von Praktiken. Dies begründet sich zum einen darin, dass KI-Technologie relevante Aspekte der (sozialen) Umwelt datenförmig erfasst und in eine mathematische Logik übersetzt, um sie rechenbar zu machen. Die Daten und Phänomene werden wiederum, in den überwiegenden Fällen, entlang der üblichen Differenzkategorien eingeteilt. Daran gekoppelt ist ein Glaube an Objektivität und eine Priorisierung technizistischer Lösungen (Weber und Prietl 2021). Im Rahmen dessen ist auch eine Auseinandersetzung mit Körper(normen) und Subjektivierung relevant. Praktiken der Subjektivierung zu betrachten, bedeutet, zu fragen, wie aus Menschen vergesellschaftete Subjekte gemacht werden und

---

<sup>7</sup> Vgl. International Research Network on Design (and) Activism. Siehe den Link: <https://desactivism.wordpress.com/>. Siehe außerdem: Design Activist Institute. <http://designactivistinstitute.org/>. The Design Activist Institute is a Philadelphia-based, intersectional, feminist, autonomous, grassroots collective of designers working for radical, utopian resistance and building a better world.

wie Widerstreit möglich ist. Denn gleichzeitig entstehen durch KI neue Formen des Aktivismus – sei es gegen Geschlechterdiskriminierung durch KI oder Design-Aktivismus für Inklusion, aber auch KI-gestützte Versuche der Meinungsmanipulation. Im KI-Aktivismus wird oftmals an Modellen der Dekonstruktion von Normen und Binaritäten mit Blick auf KI-Logiken und Designs gearbeitet. Allerdings steht auch der Einsatz von KI selbst auf dem Prüfstand, etwa in der Sozialbürokratie oder durch den Staat. Algorithmische Diskriminierungen werden aufgedeckt. Insofern bestätigt sich eine Forschungslücke im Bereich der KI-Entwicklung – nämlich differenztheoretische und normenkritische Studien.

Ebenso wird deutlich, dass technologische Logiken (Weber und Prietl 2021) von KI die Sichtweise der Dominanzgesellschaft sowie die darin eingelassenen Werte und Normen widerspiegeln<sup>8</sup>. Studien zeigen (Hamraie und Fritsch 2019; Gender 2019; Henwood und Wyatt 2019; Houben und Prietl 2018; netz-nova 2020; Ravneberg und Söderström 2018; Whittaker et al. 2019), dass eine Auseinandersetzung mit Geschlecht, aber auch Behinderung, als eindeutige und binäre Kategorie erforderlich ist. Insgesamt reproduziert KI eher Binaritäten wie Alter, Behinderung, Zugehörigkeit, Geschlecht oder ‚race‘ (Gender 3/ 2019), so die Schlussfolgerungen der genannten Studien. Fragen nach der De/Konstruktion von Differenzkategorien münden teilweise in soziokulturellen Aushandlungsprozessen bzw. Diskussionen über den Umgang mit Vielfalt innerhalb von KI-Logiken (Datensätzen aber auch in KI-Verarbeitungslogiken). Beispielsweise sei das Projekt Re-Coding Algorithmic Culture erwähnt. Das von der VW Stiftung geförderte Forschungsprojekt „Re: Coding Algorithmic Culture“ ist ein interdisziplinäres Projekt der Fachgebiete Gender/Diversity in Informatiksystemen (Universität Kassel), Soziologie der Diversität (Universität Kassel) und Visuelle Kommunikation (Kunsthochschule Kassel). Das ursprünglich bis Februar 2021 geförderte Projekt wurde von der VW Stiftung bis November 2021 mit dem erweiterten Modul „Re: Coding Algorithmic Cultures of the Pandemic -The Example of Social Distancing“ verlängert. Die Projektbeteiligten gehen der Frage nach, wie durch algorithmisch basierte Sammlungen, Klassifikationen und Interpretationen von Daten bestehende gesellschaftliche Ungleichheiten/Diskriminierungen fortgeschrieben werden, oder irritiert, wenn nicht sogar umgeschrieben werden können. Forschungsformate sind Hackathons, Game Jams, Coding-Workshops, partizipatorische Design Labs, Design Noir Experimente, Performances, Ausstellungen sowie Video- und Textanalysen. Diese

---

<sup>8</sup> Siehe beispielsweise, Link: <https://www.uni-kassel.de/eecs/gedis/forschung/recoding> und <https://tupviskom.net/category/re-research/>. Homepage Prof. Johanna Schaffer, Universität Kassel. (abgerufen 31.3.2023).

interdisziplinären Labs sollen kritisches, queer-feministisches, antirassistisches und dekoloniales Wissens im digitalen Raum weitertreiben.

Insofern wäre vertiefend zu behandeln, wie Informationstechnologien und Software inklusiv gestaltet werden können. Wie können Produkte und Inhalte der Informatik – z. B. Roboter, Textverarbeitung und Modellierung – differenzsensibel entwickelt und genutzt werden, ohne Binaritäten festzuschreiben? Es geht um die Möglichkeiten der Gestaltung einer guten Technik und es ist zu fragen, inwieweit Methoden der Technikgestaltung in der Informatik dazu beitragen können, der Fortschreibung von Hierarchie, Differenz und Zweigeschlechtlichkeit durch KI-Technik entgegenzuwirken.

Ferner sollte nicht allein auf individuelle Chancen zur Teilhabe und Inklusion fokussiert werden, sondern ebenso auf materielle und strukturelle Ungleichheiten. Last but not least geht es um einen Perspektivenwechsel, in dem KI-Technologie nicht mehr als neutral konzipiert wird. Stattdessen sollte KI aus einer feministischen oder ableismuskritischen, ebenso aus einer klassismuskritischen oder rassismuskritischen Perspektive konzipiert, genutzt und erforscht werden.

Hilfreich wäre aus diesem Grund ein Inklusionsbarometer mit Blick auf o.g. verschiedene Personengruppen zum Thema KI. Ein erster Schritt in diese Richtung wurde mit dem ersten Bericht der deutschen Bundesregierung an die Europäische Kommission über die periodische Überwachung der Einhaltung der Barrierefreiheitsanforderungen von Websites und mobilen Anwendungen öffentlicher Stellen, zumindest im Bereich von Behinderung und Alter, umgesetzt. Auch wenn sich im Bericht des Ministeriums Digitalisierung und KI-gestützte Entwicklung überschneiden, kann konstatiert werden, dass Berichte dieser Art den Ausbau barrierefreier KI-Technologien zumindest erhöhen. In Anbetracht der Vielfalt der Anwendungsgebiete und Geräte bleibt dieses Unterfangen dennoch Sisyphus-Arbeit. Inzwischen gibt es das EU-Digital-Gesetz (Digital-Service-Act/DSA), welches, Stand 2022, noch ratifiziert werden muss. Inwieweit die Konstruktion von Binaritäten und algorithmischen Benachteiligungsstrukturen damit real überwunden werden kann, bleibt abzuwarten. Bislang zielt die Gesetzesvorlage eher auf Hassrede im Netz, Werbung und auf die Bekämpfung von Fake News. Das Gesetz hinterfragt nicht Algorithmen-Trainings, welche mit binären Differenzkategorien trainiert werden. Insgesamt könnten intelligente KI-Produkte zu mehr Teilhabe führen. Die Umsetzung und Ausgestaltung dieses Punktes müsste vertieft erforscht und vor allem überwacht werden.

## 5 Zur Dialektik von KI-Technologien: eine Vertiefung

Im Folgenden sollen die besprochenen Themen durch Praxisbeispiele vertieft behandelt werden. Sie verdeutlichen, dass KI keineswegs automatisch Inklusion oder Teilhabe bereitstellt. Ebenfalls ist nicht zu erwarten, dass KI zwangsläufig Differenzen oder Binaritäten aufhebt. Viel eher entsteht eine Plattform, in der Differenz neuerlich verhandelt wird. Wird die Vielzahl von KI-Tools berücksichtigt und deren unterschiedliche technologische Entwicklungsstadien, so ist KI zwar offen für Vielfalt, gleichwohl besteht jedoch die Gefahr einer „Informatik der Herrschaft“ (Schmitz 2021).

Dies soll am Beispiel der Entwicklung von Brain-Computer- und Brain-Machine-Interfaces (BCI/BMI) verdeutlicht werden. Technologisch beruht diese Mensch-Maschine-Schnittstelle darauf, dass, *„an effective development of BCI and neuro-prostheses depends on the learning plastic brain and learnable algorithms, which mutually frame each other inside the bio-techno materiality“* (ebd., S. 9).

Die Cyborgisierung von Gehirnkörpern mit Computerhardware und -software besteht heute in der Realisierung von Gehirn-Computer-Schnittstellen (ebd., S. 1). Vorrangig im militärischen Kontext entwickelt, um moderne Waffentechnologie noch effizienter bedienen zu können, da menschliche Fähigkeiten hier an Grenzen stoßen, wird diese Mensch-Maschine-Schnittstelle überwiegend im Bereich der Bionik bzw. Neuroprothetik erforscht und ausprobiert. Neu an dieser Entwicklung ist, dass im letzten Jahrzehnt unbewusste gegenseitige Kalibrierungsprozesse oder Feedbackschleifen möglich geworden sind (ebd.). Gleichwohl zeigen Expertisen der BCI-Forschung und -Entwicklung bedenkliche Trends und Entwicklungen, die sowohl BCI als auch daran anknüpfende transhumanistische Utopien betreffen. Einerseits wird Behinderung genutzt, um diese KI-Tools ethisch zu legitimieren. Andererseits sind anderweitige Ziele in Programmierung und Ausrichtung unschwer zu erkennen. Die BCI signalisiert eine bedeutende Verschiebung in Subjekt-Objekt-Beziehungen, da es tatsächlich zu einer Verbindung zwischen Menschen und Maschine kommt. Wenn sich technologisierte Gehirnkörper durch kontinuierliche Intra-Aktionen (Barad 2012) materialisieren, könnten sie das Potenzial haben, die moderne Dichotomie zwischen feminisierter Natur und maskulinisierter Kultur zu stören (wobei Technologie als Teil der Kultur anzusehen ist) (Schmitz 2021, S. 3). Insgesamt kann allerdings davon ausgegangen werden, dass Binarisierungen in Gestalt von maskulinisierter Rationalität und feminisierter Emotionalität weiterhin in BCI eingehen. Da es in diesen Programmen um affektive Konditionierung geht, verschmelzen neoliberale Transhumanismus-Fantasien mit KI-Forschung. Beispielsweise gibt es einen

wachsenden Markt für BCI-Emotionserkennungssysteme außerhalb des Gesundheitssektors, von denen die meisten behaupten, das individuelle Management der Arbeitsleistung zu verbessern (ebd., S. 9). Im Vordergrund steht demnach das Ziel, durch BCI-Technologie Effektivität, Wettbewerb und Rationalität voranzutreiben – alles ebenso ableistische wie maskulinistische Konzepte. Letztlich werden Variationen eines Human Enhancement angestrebt. Diese konzeptionelle Ausrichtung von Menschen“verbesserung“ sind als leitendes Ziel dieser Technikentwicklung anzusehen.

Das zweite Beispielfeld akzentuiert stärker die Auswirkungen von KI im Alltag im Kontext von Behinderung. „Autonomes Fahren“ soll zu mehr Mobilität verhelfen. Gleichwohl zeichnet sich in der Entwicklung jener KI-Tools ab, dass Behinderungen schon in der Entwicklung ungenügend berücksichtigt sind. So beruhen gegenwärtig Modelle vom „autonomen Fahren“ nicht auf einer inklusiven „Smart City“. Bei einer ungenügend barrierefreien Umgebung führt „autonom Fahren“ nicht automatisch zu mehr Mobilität. Aktuell beträgt das Ausmaß an Barrierefreiheit im gesamten gesellschaftlichen Umweltbereich ca. 50 %. Das neue Barrierefreiheitsstärkungsgesetz von 2021 sowie das European Accessibility Act, kurz: EAA, sollen zwar Barrierefreiheit fördern. Diese Gesetze beziehen sich allerdings lediglich auf bestimmte Bereiche, wie beispielsweise digitale Geräte und regulieren weniger Baumaßnahmen oder Architekturvorschriften. Zudem sind viele Sonderregelungen möglich. Aus diesem Grund ist zu vermuten, dass „autonom Fahren“ im Individualverkehr wie auch im öffentlichen Personennahverkehr (ÖPNV) bei Behinderung nur bedingt von Nutzen sein kann. Zudem zeigt sich im US-amerikanischen Kontext, dass Daten für Fußgänger-Erkennungssysteme unzureichend programmiert sind. So konnten autonom fahrende Fahrzeuge weder Personen mit Rollator, noch mit Fahrrad oder Motorroller ausreichend erkennen – geschweige denn Rollstuhlfahrer\*innen. Insofern haben Datensets, welche Fußgänger-Erkennungssysteme trainieren, einen offenkundigen Disability-Bias (Whittaker et al. 2019, S. 10 ff.). Auch hier zeigt sich: die Standards von „normal“ und „Ableism“ reproduzieren und verstärken sich in spezifischen KI-basierten Systemen.

Nichtsdestotrotz schlagen Entwicklungsabteilungen großer Firmen und Forschungslabore autonomes Fahren im Individualverkehr für Menschen mit Sehbehinderungen als Modell für eine individualisierte, motorisierte Fortbewegung vor, welche sich nahtlos in das System der Automobilität (Vorrang des Autoverkehrs) einzufügen scheint. Jenseits des aktuellen technologischen Entwicklungsstandes scheint relativ ungeklärt, wer ein solches autonom fahrendes Fahrzeug erhält und



welcher Kostenträger womöglich die Kosten übernimmt. Denn gegenwärtig werden barrierefrei umgerüstete Autos lediglich dann finanziell gefördert, wenn eine Berufstätigkeit vorliegt.

Ein Blick auf den ÖPNV führt zu ähnlichen Schlussfolgerungen. So gibt es seit einiger Zeit erste Testphasen für autonom fahrende Busse<sup>9</sup> im öffentlichen Personennahverkehr (ÖPNV in Bayern/Hessen/NRW; v. a. Hamburg mit ist Vorreiter – mit größeren Bussen auf speziellen Spuren – beispielsweise das Projekt: HEAT). Auch hier stellt sich die Frage der Teilhabe von betroffenen Personengruppen. Zudem sind Mängel beim Thema Barrierefreiheit zu nennen. Teilweise sind die benutzten Fahrzeuge nicht barrierefrei. Hinzu kommt, dass die Busse bislang oftmals recht klein sind und bestenfalls eine Person mit Rollstuhl befördern können, sodass eine Gruppe mit Rollstuhl-nutzenden Personen nicht zusammenfahren könnte. Ebenso wenig kann ein KI-gestützter Kleinbus gleichzeitig eine Person mit Kinderwagen und eine Person mit einem Rollstuhl befördern. Insgesamt ist bislang unklar, ob und wann der Regelbetrieb aufgenommen werden kann. Auf absehbare Zeit wird die autonome Nutzung des ÖPNV für Menschen mit Behinderung nur bedingt möglich sein.

Zudem sind Menschen mit Behinderungserfahrung an diversen Punkten zumindest auf Hilfestellungen durch das Fahrpersonal angewiesen (Heiler 2020, 9). Zahlreiche praktische Herausforderungen im Alltag lassen sich oft nur durch Unterstützung vor Ort lösen. Mobilitäts-Info-Apps können hier bestenfalls teilweise unterstützend wirken, falls sie barrierefrei gestaltet sind und relevante Informationen zur Verfügung stellen; wie zum Beispiel Meldungen von Staus oder zu barrierefreien Haltestellen. Es besteht somit die Gefahr, dass eine inklusive Gestaltung von autonom fahrenden Bussen im ÖPNV an den hierfür notwendigen technischen Innovationen/Investitionen, einschließlich der Umgestaltung hin zu einer inklusiven Umwelt scheitert. Das Dilemma einer ungenügend barrierefreien Umwelt, in der KI-gestützte Busse buchstäblich unterwegs sind, bleibt weiterhin bestehen.

---

## 6 Schlussfolgerungen

Unter inklusivem, diversitätssensiblen Produkt- und Anwendungsdesign von KI ist nicht Assimilation an binär-ableistische Normen zu verstehen. Bei Fragen zu Gestaltungsmöglichkeiten sollten aus diesem Grund benachteiligte Personengruppen für die Umsetzung von Universal Design einbezogen werden – von

---

<sup>9</sup> Ein Beispiel: <https://ride4all.nrw/sofia-inklusiv>. Abgerufen am 7.8.2022.

der Produktentwicklung bis hin zur Nutzer\*innenperspektive. Gestaltungsmöglichkeiten von KI sind vorhanden, wenn vorherrschende Normen hinterfragt oder unterlaufen werden. Ein Beispiel hierfür wäre das Institut für digitale Teilhabe (IDT)<sup>10</sup> aus Bremen. Am IDT arbeiten und forschen gleichermaßen Mitarbeiter\*innen mit und ohne Beeinträchtigungen. So wird die Entwicklung zugänglicher Medien, Software und Systeme auch von Expert\*innen in eigener Sache vorangetrieben und bei allen Beteiligten das Bewusstsein für die Breite der zu berücksichtigenden Anforderungen gestärkt. Prozesse werden konsequent partizipativ gestaltet. Selbstverständlich wird es in diesem Zusammenhang nötig sein, sich über konkrete Formen von Teilhabe und Partizipation zu verständigen. In manchen Fällen ist damit lediglich die Erprobung des Endproduktes durch Betroffene gemeint. Oftmals bleiben Macht- und Herrschaftsverhältnisse unthematisiert, da Aushandlungsprozesse zwischen Produktherstellern, Betroffenen, Behindertenverbänden oder NGOs wie Ability Watch bislang selten untersucht worden sind. Ein Praxisbeispiel hierfür wäre ein Roboterarm, der mit einem lernfähigen KI-basierten Steuerungssystem verbunden ist und die Autonomie der Betroffenen im täglichen Leben deutlich erhöhen soll. Im Projekt DOF-Adaptiv<sup>11</sup> werden in einem benutzer\*innenzentrierten Designprozess für die Mensch-Roboter-Kommunikation, unter Verwendung einer Datenbrille, Möglichkeiten der Darstellung von Feedback erforscht. In einem partizipativen Ansatz werden die Anforderungen erhoben und das System mit Betroffenen evaluiert. Abgesehen davon, dass im Kontext der Behindertenbewegung die Nutzung von Prothesen immer mal wieder Diskussionsstoff bietet, ist auch unklar, warum KI-gestützte Roboterarme, beispielsweise am Rollstuhl angebracht, mehr Autonomie bringen sollen als menschliche Assistenz. Unklar bleibt innerhalb welcher Macht- und Herrschaftsdynamiken hier Partizipation umgesetzt wird. Insgesamt befinden sich solche Projekte auf dem Vormarsch; in der Regel in Form vereinfachter, alltagsferner Best Practice-Beispiele. Der praktische Erfolg derartiger technologischer KI-Entwicklungen ist bisweilen ungewiss. Vor diesem Hintergrund wäre es wichtig, zwischen wünschenswerten und kritikwürdigen KI-Entwicklungen zu unterscheiden.

So sollten zum Beispiel mit Blick auf die Einschreibung von Binarität in KI-Tools (Daten und Algorithmen) spezielle sensibilisierende und aufklärende

---

<sup>10</sup> Siehe: Institut für digitale Teilhabe. Link: <https://www.hs-bremen.de/die-hsb/fakultaeten/elektrotechnik-und-informatik/forschung-und-transfer/institut-fuer-digitale-teilhabe/#c18998>. Abgerufen am 25.07.2022.

<sup>11</sup> Siehe die Homepage – Link: <https://www.interaktive-technologien.de/projekte/dof-adaptiv>.

Abgerufen am 25.07.2022.

Trainingsprogramme entstehen, die nicht im Neo-Positivismus münden, sondern sich das Wissen kritischer Gesellschaftswissenschaften zu eigen machen. Sinnvoll sind in diesem Zusammenhang Kooperationen mit Betroffenen, die auf gleicher Augenhöhe in Konstruktions- und Aushandlungsprozesse der KI-Entwicklung einbezogen sind.

Insofern lässt sich schlussfolgern, dass die Auswirkungen von KI-Entwicklungen im Bereich von Differenz, Alltag, Alltagsroutinen und Gesellschaft bislang unzureichend erforscht sind. Gleichzeitig scheint das Thema Differenz in der KI-Technologie ein Ort für neuerliche soziokulturelle Aushandlungen zu sein, die es zu analysieren gilt. Paradoxerweise ist zu konstatieren, dass Aspekte von Differenz- und Bias-Konstruktionen innerhalb der Erforschung von KI ein Schattendasein fristen. Das Thema Differenz wird im ‚Mainstream‘ der KI-Forschung wenig erörtert. Offen bleibt auch, wie die Ausgestaltung von Teilhabe und Partizipation in der Realität umgesetzt wird. Hier scheint weiterer Forschungsbedarf zu bestehen. Gleichzeitig besteht die Gefahr, gesellschaftliche Verhältnisse über die Erforschung einzelner partizipativer Projekte aus den Augen zu verlieren. Ebenso dringlich ist es, gesamtgesellschaftliche Auswirkungen durch KI zu untersuchen und die Umgestaltung von gesellschaftlichen Institutionen und Strukturen im Blick zu behalten. Auch wenn nicht jedes KI-Tool gesamtgesellschaftliche Relevanz hat, hat KI als Technologie und als Technik – in der Summe betrachtet – das Potential einer grundlegenden gesellschaftlichen Transformation. In diesem Sinne ist Forschung zur KI wesentlich als Forschung über Gesellschaft und soziokulturelle Ungleichheit zu verstehen und durchzuführen. Eine umfassende Reflexion gesellschaftlicher Probleme rund um die KI, welche womöglich lediglich neue Pathologisierungen für Nutzende bereithält, steht folglich noch aus. Wichtig ist ferner eine politische Regulierung von KI, beispielsweise durch Zusammenarbeit mit NGOs, wie AlgorithmWatch oder Ability Watch, sowie die Erforschung von Durchsetzungs- und Entscheidungsprozessen durch Akteur\*innen im KI-Komplex, da all die aufgezeigten Entwicklungen in komplexe Macht- und Herrschaftsverhältnisse eingebettet sind.

---

## Literatur

- Barad, Karen. 2012. *Agentieller Realismus: Über die Bedeutung materiell-diskursiver Praktiken*. Berlin: edition unseld.
- Bath, Corinna und Yvonne, Bauer, u.a. (Hrsg.). 2005. *Materialität denken. Expertisen zur technologischen Verkörperung – Hybride Artefakte, posthumane Körper*. Bielefeld: transcript Verlag.

- Bauer, Susanne, Thorsten Heinemann und Thomas Lemke (Hrsg.). 2017. *Science and Technology Studies. Klassische Positionen und aktuelle Perspektiven*. Berlin: Suhrkamp Verlag.
- Beck, Ulrich. 2015a. Moderne. In *Lexikon der Soziologie und Sozialtheorie. Hundert Grundbegriffe*, hrsg. von Sina Farzin und Stefan Jordan, 198–201. Ditzingen: reclam.
- Beck, Ulrich. 2015b. *Risikogesellschaft*, Berlin: suhrkamp Verlag.
- Bergt, Svenja. 2023. Moratorium über Umgang mit KI: Pause mit Problemen Hilft ein temporärer Entwicklungsstopp, KI in Bahnen zu lenken? Expert:innen finden, die Forderung gehe an den echten Gefahren vorbei. *März 31*. <https://taz.de/Moratorium-ueber-Umgang-mit-KI/!5925502&s=KI/> (abgerufen am 31.03.2023).
- Bieling, Tom. 2019. *DESIGN (&) ACTIVISM – Perspectives on Design as Activism and Activism as Design*. Mimesis International (Design Meanings).
- Bieling, Tom. 2021. *Inklusion als Entwurf*. Birkhäuser: DeGruyter.
- Biniok, Peter, und Eric Lettkemann. 2017. *Assistive Gesellschaft. Multidisziplinäre Erkundungen zur Sozialform „Assistenz“*. Wiesbaden: Springer Verlag.
- Blanc, Berit, und Susan Beudt. 2022. *Monitoring KI-gestützter Assistenztechnologien für Menschen mit Behinderungen. Stand der Entwicklungen und Trends*. Ergebnisbericht des Projekts KI.ASSIST. Bundesverband Deutscher Berufsförderungswerke e. V.
- Butollo, Florian, und Sabine Nuss. 2019. *Marx und die Roboter. Vernetzte Produktion, Künstliche Intelligenz und lebendige Arbeit*. Berlin: Dietz Verlag.
- Daum, Timo. 2019: Künstliche Intelligenz als vorerst letzte Maschine des digitalen Kapitals. In: *Marx und die Roboter. Vernetzte Produktion, Künstliche Intelligenz und lebendige Arbeit*, hrsg. Florian Butollo und Sabine Nuss, 311–327. Berlin: Dietz Verlag.
- Ender, Cordula. 2020. *Assistiert Altern. Die Entwicklung digitaler Technologien für und mit älteren Menschen*. Wiesbaden: Springer Verlag.
- Engel, Antke. 2007. Gefeierte Vielfalt. Umstrittene Heterogenität. Befriedete Provokation. Sexuelle Lebensformen in spätmodernen Gesellschaften. In *Heteronormativität und Homosexualitäten*, hrsg. Rainer Bartel, Waltraud Finster, Meinrad Ziegler, 43–64. Innsbruck, Wien, Bozen: Studien Verlag.
- Feministische Studien. 2019. Cyborgs Revisited: Zur Verbindung von Geschlecht, Technologien und Maschinen. 37/2. <https://blog.feministische-studien.de/aktuell/>.
- Fink, Dagmar. 2021. Cyborg werden. Möglichkeitshorizonte in feministischen Theorien und Science Fictions. Bielefeld: transcript
- Fraser, Nancy. 2003. *Umverteilung oder Anerkennung? Eine politisch-philosophische Kontroverse*. Frankfurt: suhrkamp Verlag.
- GENDER. 2019. Zeitschrift für Geschlecht, Kultur und Gesellschaft. Schwerpunkt: Gender, Technik und Politik 4.0. 11/3. Leverkusen: Budrich Verlag.
- Hamraie, Aimi, und Kelly Fritsch, 2019. Crip technoscience manifesto. *Catalyst: Feminism, Theory*. Technoscience, 5/1: 1–34. <https://catalystjournal.org/index.php/catalyst>.
- Haraway, Donna. 1995a. *Die Neuerfindung der Natur*. Frankfurt/New York.
- Haraway, Donna. 1995b. *Ein Manifest für Cyborgs. Feminismus im Streit mit den Technowissenschaften*. Frankfurt, New York.
- Henwood, Flis, und Sally Wyatt. 2019. Technology and In/Equality, Questioning the Information Society (Almost) 20 Years Later. In *DCS. Digital Culture and Society*. 5/1. 183–194. Bielefeld: transcript Verlag.

- Heiler, Hannes. 2020. Frankfurter Behindertenarbeitsgemeinschaft (FBAG). Fachausschuss Verkehr und Beförderungsdienst. Sprecher: Hannes Heiler. Stellungnahme zum neuen Nahverkehrsplan (NVP) der Stadt Frankfurt am Main.
- Holert, Tom, und Mark Terkessidis. 1997. *Mainstream der Minderheiten*. Berlin: Edition ID-Archiv.
- Houben, Daniel, und Bianca Prietl. 2018. *Datengesellschaft. Einsichten in die Datafizierung des Sozialen*. Bielefeld: transcript Verlag.
- Imbusch, Peter. 1998. Macht und Herrschaft in der wissenschaftlichen Kontroverse. In Peter Imbusch, *Macht und Herrschaft. Sozialwissenschaftliche Theorien und Konzeptionen*. 275–299.
- Jour Fixe Initiative Berlin/Frankfurt. 2022. Vortragsreihe: Dialektik der Technik. Sommer 2022. <https://www.jourfixe.net/index.php/veranstaltungsreihe/dialektik-der-technik>.
- Kalender, Ute. 2011. *Körper von Wert: Eine kritische Analyse der bioethischen Diskurse über die Stammzellforschung*. Bielefeld: transcript Verlag.
- Kerner, Ina. 2021. *Postkoloniale Theorien zur Einführung*. Hamburg: Junius Verlag.
- Klein, Barbara. 2020. *Hilfsmittel, Assistive Technologien und Robotik. Selbständigkeit und Lebensqualität im Alter erhalten*. Stuttgart. Wiesbaden: Springer Verlag.
- Latour, Bruno. 2010. *Eine neue Soziologie für eine neue Gesellschaft: Einführung in die Akteur-Netzwerk-Theorie*. Berlin: Suhrkamp-Verlag.
- Lemke, Thomas. 2000. Neoliberalismus, Staat und Selbsttechnologien. Ein kritischer Überblick über die governmentality studies. <http://www.thomaslemkeweb.de/engl.%20texte/Neoliberalismus%20ii.pdf>.
- Lorey, Isabell, Roberto Nigro, und Gerald Raunig. 2011. Inventions. Actualizing Poststructuralist Theory. Transversal.at. <https://www.transversal.at/transversal/0811/raunig-lorey-nigro/en?hl=Assemblage>.
- Netznova\*. 2020. *Wenn KI dann feministisch. Impulse aus Wissenschaft und Aktivismus*. Berlin: Netznova.
- Pfeiffer, Sabine. 2021. *Digitalisierung als Distributivkraft. Über das Neue am digitalen Kapitalismus*. Bielefeld: transcript Verlag.
- Raab, Heike. 2011. *Sexuelle Politiken. Die Diskurse zum Lebenspartnerschaftsgesetz*. Frankfurt/ New York: Campus Verlag.
- Ravneberg, Bodil, und Sylvia Söderström. 2018. *Disability, Society and Assistive Technology. London/New York*. Routledge.
- Schapp, Simon. 2021. *Technopolitik von unten: Algorithmische Arbeitssteuerung und kybernetische Proletarisierung*. Berlin: Matthes & Seitz.
- Schmid, Florian. 2023. Künstliche Intelligenz: Angst-Gegner oder Verbündete? In der Science-Fiction treten künstliche Intelligenzen schon seit Jahrzehnten auf. Erst allmählich nicht mehr nur als Bösewichte. *Neues Deutschland*. März 24. <https://www.nd-aktuell.de/artikel/1171963.science-fiction-kuenstliche-intelligenz-angst-gegner-oder-verbue ndete.html>.
- Schmitz, Sigrid. 2021. TechnoBrainBodies-in-Cultures: An Intersectional Case. In *Frontiers in Sociology*. <https://www.frontiersin.org/articles/https://doi.org/10.3389/fsoc.2021.651486/full>.
- Staab, Philipp. 2020. *Digitaler Kapitalismus*. Berlin: suhrkamp Verlag.

- Suchmann, Lucy. 2019. Feministische Science & Technology Studies (STS) und die Wissenschaften vom Künstlichen. *GENDER. Zeitschrift für Geschlecht, Kultur und Gesellschaft. Schwerpunkt: Gender, Technik und Politik* 4.0.,11/3: 56–84. Leverkusen: Budrich Verlag.
- Waldschmidt, Anne. 2014. Macht der Differenz: Perspektiven der Disability Studies auf Diversität, Intersektionalität und soziale Ungleichheit. In *Soziale Probleme. Zeitschrift für soziale Probleme und soziale Kontrolle*, 25/2: 173–193. Herbolzheim: BELTZ/ Juventa.
- Weber, Jutta. 2002. Technoscience als Epoche? Ontologische, epistemologische und narrative Grundlagen der Techno/Wissenschaften. In *Techniken der Reproduktion. Medien – Leben – Diskurse*. hrsg. Bergermann, Ulrike. Helmer Verlag, Königstein, S. 51–66.
- Weber, Jutta. 2003. *Umkämpfte Bedeutungen: Natur im Zeitalter der Technoscience*. Frankfurt am Main. Campus Verlag.
- Weber, Jutta, und Bianca Prietl. 2021. On the Rise of Data Driven AI and its Epistem-Ontological Foundations. In *The Routledge Social Science Handbook of Ai.*, eds. Anthony Elliott, 58 – 74. Routledge.
- Wentz, Daniela. 2022. “You make me feel like a natural woman<.(Neuro)queerness und Affective Computing”. Unveröff. Vortrag. Internationale Tagung – *DIGITAL GENDER: ETHIK, MACHT UND (GESCHLECHTER-)WISSEN IN SYSTEMEN KÜNSTLICHER INTELLIGENZ. APRIL*.
- Whittaker, Meredith, Meryl Alper und Cynthia L Bennett. 2019. Disability, Bias, and AI. AI Now – Institute at New York University (NYU). New York. *chrome-extension://efaidnbmninnbpcajpcgclclefindmkaj/https://ainowinstitute.org/disabilitybiasai-2019.pdf*.



# Zukunftseuphorie als Trost. Verheißungserzählungen über Künstliche Intelligenz im Kontext gesellschaftlicher Erschöpfungsdiagnosen

Stefan Selke

## Zusammenfassung

Das Thema künstliche Intelligenz (KI) trifft mittlerweile auf rezeptionsbereite Publika, die sich offen für Zukunftserzählungen zeigen, die einen explizit Verheißungscharakter aufweisen. Verheißungsnarrative liegen in Form utopischer Fortschrittsgeschichten, politisch-normativer Leitbilder aber auch als Visionen von KI-Schaffenden, in Form von Medienberichten oder als fiktionale Darstellungen vor. Verheißungserzählungen repräsentieren Erwartungshorizonte und Zukunftshoffnungen zwischen Technikversprechen und Technikgläubigkeit. Im Umfeld von KI lässt sich eine Rückkehr der Heilssehnsucht im Gewand von Ersatzreligionen feststellen. Mit Verheißungen aufgeladene Erzählungen über KI erzeugen Zukunftseuphorie, die sich als Trostritual einordnen lässt. Techno-Verheißungen treten hierbei an die Stelle von religiösen Erfahrungen. Kurz: In Gesellschaften, die latent durch Zukunftsangst geprägt sind, wirken verheißungsvolle Zukunftsnarrative als Trostersatz.

## Schlüsselwörter

Erschöpfte Gesellschaft • Gesellschaftsdesign • Techno-Utopien • Zeitdiagnose • Zukunftseuphorie • Zukunftsnarrative

---

S. Selke (✉)

Hochschule Furtwangen, Public Science Lab, Furtwangen, Deutschland

E-Mail: [stefan.selke@hfu.eu](mailto:stefan.selke@hfu.eu)

© Der/die Autor(en), exklusiv lizenziert an Springer Fachmedien Wiesbaden GmbH, ein Teil von Springer Nature 2024

M. Heinlein und N. Huchler (Hrsg.), *Künstliche Intelligenz, Mensch und Gesellschaft*, [https://doi.org/10.1007/978-3-658-43521-9\\_12](https://doi.org/10.1007/978-3-658-43521-9_12)

## 1 KI als epochentypische Verheißung zwischen Technikversprechen und Zukunftseuphorie

Zeitgenössische Gesellschaftsdiagnosen sind ein diskursiv umkämpftes Feld (Alkemeyer et al. 2019). Sie beschreiben Krisenphänomene, die sich – auf individueller, kollektiver sowie planetarer Ebene – als *Erschöpfungssyndrome* einordnen lassen. Unter Erschöpfung soll dabei nicht allein die fehlende Regenerationsfähigkeit Einzelner, sondern vielmehr das Symptom eines gestörten Zugriffs auf die Welt verstanden werden – die Welt lässt sich somit vor allem über apokalyptische Hermeneutiken fassen (Nagel 2021). Mit diesen umfassenden Störungen sind einerseits disruptive Dynamiken verbunden, auf die andererseits mit gesellschaftswissenschaftlichen Erklärungsversuchen reagiert wird, z. B. mit Theorien der Fragilität (Bayramoglu/do Mar Castro 2021) oder der Überforderung (Nassehi 2021). Vor dem Hintergrund zunehmender Verunsicherung oder gar existentieller Erschütterung tritt das Imaginationsfeld künstlicher Intelligenz (KI) mittlerweile auf rezeptionsbereite Publika, die sich offen für revitalisierte Heilsbotschaften zeigen, die Zukunftseuphorie erzeugen sollen. Erzählungen über KI liegen in Form von Fortschrittsgeschichten, Meta-Analysen, Hype-Zyklen, politisch-normativen Leitbildern, Visionen von KI-Schaffenden, Medieninszenierungen oder auch zahlreichen fiktionalen Darstellungen vor und läuten ein neues „Zeitalter voller Verheißungen“ (Grunwald 2019: 19) ein.

### 1.1 Bedeutungszuschreibungen jenseits des Technischen

Verheißungen sind *prognostische Narrative*, die spekulative Erwartungshorizonte, intentional verdichtete Zukunftshoffnungen sowie Entlastungserwartungen in sich vereinen. Jenseits aller Plausibilität enthalten diese Verheißungserzählungen ein hohes *Identifikationspotenzial*, denn sie verbinden Vorannahmen und Wahrnehmungen und erzeugen auf diese Weise sinnhafte Konstrukte sozialer Wirklichkeit. Zwischen Technikversprechen und Technikgläubigkeit bündeln Verheißungen *spekulative Bedeutungszuschreibungen*, die über das Empirische hinausreichen. Eine soziologische Perspektive auf Verheißungen rückt dabei nicht die *Funktionalitäten* von KI-Technologien in den Mittelpunkt, sondern die *Relevanzen* individueller und kollektiver Bedeutungszuschreibungen an diese Technologie. Für dieses Verständnis von Verheißungen ist ein erweiterter Technikbegriff notwendig, der verdeutlicht, wie Technik durch ihren praktischen Einsatz Bedeutung erlangt, z. B. ein zweidimensionaler (Weyer 2008) oder besser ein dreidimensionaler Technikbegriff (Degele 2002). Damit kann Technik als Artefakt



(Sichtbares, Materialität), als Handlungsform (Praxis des Machens) sowie als Wissensform unterschieden werden. Mit Günter Ropohl lassen sich zudem die *inneren Strukturen* einer Technologie von den *äußeren Reaktionen* auf diese Technologie analytisch unterscheiden. Auf diese Weise entsteht eine Mensch-Maschine-Einheit, ein *sozio-technisches System* (Ropohl 2009: 71 ff.). Diese Perspektive hilft, KI in erster Näherung auch jenseits des Technischen zu verstehen. Noch hilfreicher sind allerdings gesellschaftlich zirkulierende Erzählungen über KI.

Zur Nutzbarmachung und Einordnung von Verheißungsnarrativen wurden daher erstens heterogene Quellen über KI anhand eines strukturierten Samplings gesichtet.<sup>1</sup> Dieser *narrative Wissensraum* über KI diene sodann als Diagnoseinstrument und Spiegel epochentypischer Reflexionsprozesse. Zweitens wurden durch leitfadengestützte Gespräche mit 30 Expert\*innen (aus inner- und außerwissenschaftlichen Feldern) gruppenspezifische Binnenkommunikation, Identifikationspotenziale sowie sinnstiftende und handlungsleitende Funktionen von Verheißungserzählungen über KI rekonstruiert. Ausgehend von religiös-theologischen Heilsversprechen wurden *drittens* Traditionslinien von Verheißungserzählungen rekonstruiert und typologisiert. Auf dieser Basis ordnet der vorliegende Beitrag verheißungsvolle KI-Narrative überdisziplinär ein, indem nach deren *epistemologischer Position* und *gesellschaftlicher Funktion* gefragt wird.

## 1.2 Polyvokalität und Perspektivenvielfalt

KI wird immer häufiger in gesellschaftliche Prozesse und Anwendungsfelder wie medizinische Diagnosen, Wettervorhersagen, Übersetzungsprogramme oder Kaufempfehlungen integriert und (bewusst oder unbewusst) von immer mehr Anwender\*innen genutzt. Aus diesen Anwendungserfahrungen, aber auch aus Wissen und Halbwissen über zugeschriebene Potenziale, resultiert ein Kaleidoskop an Stimmungen und Zukunftsvorstellungen. Folgerichtig braucht es zum Verständnis zeitgenössischer Vorstellungen rund um das Phänomen KI Polyvokalität. Diese Übung in ‚Perspektivismus‘ (Brockman 2020: xxv) hilft, im schnell emergierenden Feld von KI (Entwicklung, Praxis, Diskurs) unterschiedliche Bedeutungsebenen *jenseits* von Funktionalitäten zu erkennen.

---

<sup>1</sup> Die Quellengrundlage reicht von (internationalen) wissenschaftlichen Fachtexten über politische Positionspapiere und ethisch-normative Leitbilder bis hin zu Science-Fiction-Filmen und -romanen sowie themenspezifischen Blogs, Vorträgen sowie Kunst- und Theaterprojekten.

Mit KI sind zahlreiche Versprechungen verbunden, wobei sich Technikversprechen<sup>2</sup> und Technikgläubigkeit gegenseitig bedingen. Technikversprechen werden in Form von Fortschrittsgeschichten erzählt, die zugleich für eine „realistische und kritische Wahrnehmung von Technik jenseits von Technikeuphorie und Technikfeindschaft“ werben (Hänggi 2015: 23). ‚Hype Cycle‘ (Fenn/Mark 2008) versinnbildlichen die Rezeption von Technikversprechen auch dann, wenn Hypes nur eine kurze Halbwertszeit besitzen. Verheißungen oder gar Utopien „operieren hingegen stets mit langen Zeiträumen. Weder durch Enttäuschungen noch durch das Leerlaufen von Zukunftsversprechen“ (Maresch 2001: 235) lassen sie sich falsifizieren. Verheißungsvolle Technikversprechen schießen zwar immer über das Ziel hinaus, sind aber nicht komplett unrealistisch<sup>3</sup> (Budian 2020: 35).

### 1.3 KI-Verheißungen als Fortsetzungsgeschichte

Mit der Digitalisierung ist eine erhebliche Orientierungsleistung verbunden (Naschi 2019: 148). Das *Potenzial zur Ordnungsbildung* bildet den Kern zahlreicher Verheißungsnarrative, die als Fortsetzungsgeschichte bereits vorgängiger Digitalisierungsnarrative gelesen werden können – einerseits noch (teils empirisch und theoretisch fundierte) Gegenwartsdiagnose, andererseits bereits prognostisches und (hoch)spekulatives Zukunftsszenario. Verschiedene gesellschaftliche Teilsysteme – von Politik über Wirtschaft bis Kirche – nutzen Digitalisierungssemantiken, um aktualisierte Selbstbeschreibungen anzufertigen und neue Bewertungskriterien für das eigene Funktionssystem zu ersinnen (Stüssenguth 2015: 115 f.). In diesem Kontext sind Erzählungen hervorragend dazu geeignet, Ziele und Gestaltungsaspekte sowie Chancen und Gefahren zu erfassen, denn „Dystopie und Utopie sind untrennbar mit der Diskussion der digitalen Zukunft verbunden“ (Becker 2019: 17). Diese dualistische Sichtweise, die entweder das Paradiesartige oder das Unheilvolle einer Technologie betont, hat allerdings auch Nachteile. Denn tatsächlich „kann die von der großen Vision ausgehende Faszination das nüchterne Denken vernebeln.“ (Grunwald 2019: 23 ff.) Es ist daher hilfreich, KI-Verheißungen als Teil eines schlüssigen und ganzheitlichen Digitalisierungsnarrativs zu betrachten. Auch wenn die im Kontext von KI-Verheißungen

<sup>2</sup> Unter Technikversprechen ( ‚technology promises‘) sind spezifische Narrative zu verstehen, die Leitvisionen in Innovations- und Forschungsprozessen bündeln und damit Zukunftserwartungen begründen. Vgl.: van Lente/Rip (1998).

<sup>3</sup> Deswegen kann die Entwicklung von KI auch zwischen ‚Standardsituationen der Technologiekritik‘ (Passig 2013) und ‚Standardsituationen des Technologieoptimismus‘ (Passig 2014) nachgezeichnet werden.

prognostizierten Erwartungen nicht immer eintreten, haben Zukunftserzählungen einen starken Bezug zur Realität, denn als *sozio-technische Imaginationen*<sup>4</sup> sind Verheißungsgeschichten Bestandteil der Selbstbeobachtungsfähigkeit von Gesellschaften. Deutlich wird dies an Sammlungen transhistorischer und transkultureller Narrativen zu Mensch-Maschinen-Verhältnissen, die allesamt dazu beitragen, den Kern zeitgenössischer KI-Diskurse offenzulegen (Cave et al. 2020: 5).

In erster Näherung soll unter einer Verheißung eine Zukunftsvision mit prognostischem und spekulativem Inhalt verstanden werden, die auf Technikversprechen zwischen Vertrauen und Glauben beruht. Das zentrale Element aller KI-Verheißungen ist die *intentionale Überhöhung*. Als Zukunftserzählungen modellieren Verheißungsnarrative die soziale Zeit. Während reine Fiktionen alternative Welten „mit freier Beziehung auf die gegebene Wirklichkeit“ (Koschorke 2012: 229 ff.) erfinden, sind KI-Verheißungen gerade *nicht* komplett von der empirischen Welt abgekoppelt. Vielmehr gehören sie zur Gruppe derjenigen fiktionalen Narrative, mit deren Hilfe „Gesellschaften sich ein Bild ihrer selbst erzeugen“ (ebd.). Auf dieser Grundlage besteht die gesellschaftliche Relevanz von KI-Verheißungen in ihrer *handlungsleitenden Funktion*. Das zeigt sich auch daran, wie Erzählungen über Superintelligenz politische Akteure beeinflussen (Johnson/Verdicchio 2017). Die Wirkung von Verheißungen wird auch durch Science-Fiction deutlich, wobei Technologien nicht nur *imaginiert*, sondern auch in bestimmte gesellschaftliche Settings *situiert* werden. Verheißungsvolle Zukunftserzählungen steigern somit den Imaginationshaushalt einer Epoche. Hierbei steht „Zukunftsszenarien (...) ein ganzes Spektrum von dramatischen Möglichkeiten zur Verfügung: Sie können sich utopisch oder apokalyptisch ausgestalten, die bevorstehende Zeit als Belohnung oder als Strafe imaginieren, Wünschen oder Ängsten Ausdruck verleihen. Damit tragen sie die Unsicherheit des Kommenden in die Gegenwart hinein“ (Koschorke 2012: 230).

Dieses zumindest angedeutete Spektrum dramaturgischer Ausdrucksmöglichkeiten lässt sich typologisch in *vier prototypische Zukunftsnarrative* auffächern, die sich bereits in der transformativen Organisationsforschung bewährt haben (Müller 2019). Zwei der Basisnarrative sind im Kern *regressiv*: das Dagegen-Narrativ („Böse KI schafft Menschheit ab“) und das Anpassungs-Narrativ („KI optimiert alles“). Die beiden anderen sind hingegen *progressiv*: das Quest-Narrativ („KI wird Bewusstsein erlangen“) und das Aufbruchs-Narrativ („Gute KI rettet die Menschheit“). Innerhalb dieses narratologischen Rahmens kann

---

<sup>4</sup> Unter ‘sociotechnical imaginaries’ versteht Sheila Jasanoff „collectively held, institutionally stabilized, and publicly performed visions of desirable futures, animated by shared understandings of forms of social life and social order attainable through, and supportive of, advances in science and technology.“ (Jasanoff 2015: 4).

praktisch jede beliebige Geschichte über KI erzählt werden.<sup>5</sup> Als Urheber verheißungsvoller KI-Erzählungen kommen entweder sendungsbewusste und messianisch auftretende Prominente in Frage („singuläres Storytelling“) oder Diskurse und Debatten in heterogenen sozialen Arenen („kollaboratives Storytelling“). In der Summe ermöglichen die daraus resultierenden heterogenen prognostischen Narrative, aus einem bipolaren Erzähl- bzw. Beurteilungsschema auszubrechen, das ausschließliche Pro- und Contra-Argumente („Utopie“ versus „Dystopie“) versammelt.

---

## 2 Verheißungsgenealogie und Verheißungsdimensionen

Immer neue alltagstaugliche KI-Anwendungen (z. B. Übersetzungsprogramme, ChatGPT u.v.m.) rücken verheißungsvolle Narrative kontinuierlich näher an die Lebenswelt heran. Hierbei gilt es zu verstehen, an welche Traditionslinien modernisierte Verheißungsnarrative anschließen und welche gesellschaftlichen Funktionen damit verbunden sind.

### 2.1 Renaissance der Heilssehnsucht

Das Wort ‚Verheißung‘ lässt sich seit dem 15. Jahrhundert belegen.<sup>6</sup> Es wirkt altmodisch, wenngleich es auffallend häufig im Umfeld von KI auftaucht.<sup>7</sup> Zunächst ist eine Verheißung eine ernsthafte und feierliche *Ankündigung von etwas Bedeutsamen*. Im deutschsprachigen Wortverständnis ist eine Verheißung ein Geschenk oder Angebot (eines) Gottes: Ein Gott ‚verspricht‘ nichts, sondern ‚verheißt‘ etwas.

Im Kontext von KI suggerieren Verheißungserzählungen hingegen, dass KI etwas anbietet, nach dem sich Menschen intensiv sehnen. In Homologie zur religiösen Ursprungsbedeutung besteht das Basismotiv der KI-Verheißung in der

---

<sup>5</sup> Die Diagnose von KI-Wintern als Enttäuschung über mangelnde Forschungsförderung lässt sich in diesem Kontext übrigens als permanent revidiertes Anpassungs-Narrativ einordnen.

<sup>6</sup> <https://www.wortbedeutung.info/verheissung/> (06.09.2021).

<sup>7</sup> Z.B. in einem Sammelband Strasser et al. (2021) oder in sprachkritischen Untersuchungen (Gutmann et al. 2021), aber auch in angrenzenden Technikfeldern (Klie 2014). Im englischen Sprachraum ist die Häufigkeit von ‚promise‘ (als Pendant zu Verheißung) ebenfalls augenfällig (z. B. Castro/New 2016; Ho Lee/Yoon 2017).

*Befreiung (oder Erlösung) von einer Last.* Das KI-Zeitalter gilt als verheißungsvoll, weil es seit einiger Zeit ‚ungeahnte‘ oder gar ‚revolutionäre‘ (Tapscott 1996) Möglichkeiten verspricht und soziale und wirtschaftliche Vorteile produziert, die politische Entscheidungsträger nicht länger ignorieren können oder wollen (Castro/New 2016).

Unweigerlich geraten dabei *religiöse Verheißungshorizonte* ins Blickfeld. Als theologische Verheißungen können alle Ankündigungen eines künftigen heilvollen Handelns bezeichnet werden. Besonders die Theologie des Alten Testaments zeugt von der Bedeutung des Begriffs Verheißung (Levin 2013): Immer ist es Gott, der den Menschen eine bestimmte Zukunft verheißt, gleichzeitig erfüllt sich diese Verheißung erst später – oder nie.

Im Kontext von KI ist zudem eine *eschatologische Perspektive* (von Eschaton: endzeitlich) als ultimatives Versprechen eines Lebens nach dem Tod von zentraler Bedeutung. Hier taucht eine erklärungsnotwendige Divergenz auf, weil sich die modernere Theologie einerseits vom radikalen Transzendenzgehalt der Verheißungen entfernt, indem versucht wird, den Verheißungen Gottes greifbarere, diesseitige Bedeutungen zu geben. Andererseits erlebt die Idee der Unsterblichkeit im Gewand des techno-utopischen Mythos einer Superintelligenz gerade eine Renaissance (Krüger 2019). Der Triumph der (Natur-)Wissenschaften trat an die Stelle der Religion und zwang Gott „zu Rückzugsgefechten“ (Fischer 2014: 283). *Zukunftseuphorie* trat an die Stelle der ehemals hartnäckigen religiösen *Erfahrung*.

## 2.2 Verheißungen auf der operativ-funktionalen Ebene

Bislang hatte jede Generation das Gefühl, an der Schwelle zu bahnbrechenden technologischen Revolution zu stehen (Mozorov 2013: 357). Im Kontext von Anpassungs-Narrativen wird KI als unverrückbare Tatsache, als Entität und absolute Realität wahrgenommen. Auf der operativ-funktionalen Ebenen leisten die technologischen Potenziale von KI zahlreiche Verbesserungsversprechen, die schließlich in Strategiepapieren und Meta-Analysen mehr oder weniger hypnotisch redundant (re-)präsentiert werden. Hierbei wird KI eine entscheidende *Problemlösungskompetenz* und besondere *Fähigkeit zu Effizienzsteigerung* zugeschrieben. Die damit verbundene Verheißung besteht in Optimierungsgewinnen in so gut wie allen Lebensbereichen (Wolfram, zit. n.: Brockman 2020: 277). Hierbei lässt sich folgendes Muster erkennen.

**Akzeptanztests für Disruptionen:** Mit Optimierungsgewinnen gehen disruptive Veränderungen einher. Verheißungserzählungen machen diesen (Werte-)

Wandel nachvollziehbar und suggerieren wünschenswerte Zukünfte. Dies zeigt sich etwa am Beispiel Kunst und KI. Selbst wenn Kritiker das vermeintlich Mechanische an Kunstwerken bemängeln, die durch KI geschaffen wurden, verändert sich die Wahrnehmung schöpferischer Prozesse radikal. Etwa, wenn das Gedicht *Sonnenblicke auf der Flucht* oder das Gemälde *Portrait of Edmond Belamy* oder sogar eine Symphonie im Stil von Beethoven maschinell erschaffen werden.<sup>8</sup> Beispiel Bildung: Wird KI als adaptiver Bildungscoach oder Avatar in transformativen Lernprozessen eingesetzt (so z. B. bei der Lernunterstützungssoftware *area*<sup>9</sup>), stellen sich weitreichende Fragen nach der Rollenverteilung im Bildungssystem, den Risiken von Learning Analytics oder von Pathologien akademischer Metriken im Kontext kompetitiver Positionierungsstrategien (Paradeise/Thoenig 2013). Schließlich bieten Unternehmen wie etwa *Tools for Humanity*, die Verheißung eines neuen Kryptoevangeliums, das verspricht, Geld als Tauschmedium abzuschaffen.<sup>10</sup> Weitere Beispiele für derartige latente Akzeptanztests sind die ‚Todesalgorithmen‘ im Kontext autonomer Mobilität, in denen Kritiker die „Vorbereitung der Gesellschaft auf (...) Opferbereitschaft“ und ein „Entgleiten der Unantastbarkeit der Menschenwürde“ (Simanowski 2020: 38 ff.) erkennen sowie die *Verheißung der automatisierten und objektiven Grenzfallentscheidung* im Kontext von End-of-Life-Entscheidungen, die den ‚optimalen‘ Todeszeitpunkt eines Menschen betreffen.<sup>11</sup> Wird etwa der Freitod als Endpunkt eines finalen Selbstdesigns verstanden (Borries 2019: 109), zeigt sich die Verheißung eines ‚guten Todes‘ auch in der im 3D-Druck hergestellten Suizidkapsel *Sarco* des Freitod-Aktivistin Philip Nitschke, die eine neue Sterbekultur repräsentiert. Wer einen (letzten) KI-basierten Test absolviert, erhält den Code zur Öffnung, kann einsteigen, die Kapsel verschließen und sich mit einem speziellen Gasmisch selbst töten. Im Kern erzählen alle diese Verheißungen von ultimativen *Grenzüberschreitungen*. KI-Narrative können daher als *Akzeptanztest für Disruptionen* verstanden werden.

**Strategische Inszenierungen:** Verheißungen übersetzen latente Technikerwartungen in konkrete Marketingmaßnahmen, d. h. *strategische Inszenierungen* machen deutlich, dass die KI-Entwicklung einer ökonomischen und medialen

---

<sup>8</sup> <https://blog.magenta.at/2021/10/07/unvollendete-beethoven/> (18.02.2022).

<sup>9</sup> <https://area9lyceum.de> (23.03.2022).

<sup>10</sup> <https://toolsforhumanity.io> (23.03.2022).

<sup>11</sup> Dieser Sozialdystopie totaler Transparenz wurde vom Literatur-Nobelpreisträger Elias Canetti mit dem Theaterstück *Die Befristeten* ein literarisches Denkmal gesetzt. In seinem Drama *Die Befristeten* beschreibt Canetti erstmals 1952 die Ungleichverteilung von Wissen in einer Gesellschaft, in der allen der eigene Todeszeitpunkt bekannt ist (Canetti 1976).

Logik folgt. Selbst Förderlinien für Forschungsprojekte – wie z. B. der KI-Strategie der Deutschen Bundesregierung<sup>12</sup> – sind keine Ausnahme von dieser Regel. Wachsender Erfolgsdruck zwingt auch Forschende dazu, selbst verheißungsvolle Narrative zu erfinden, deren Doppelfunktion darin besteht, sich gleichzeitig gegen Kritik zu *immunisieren*.

Eine spezifische Form der Inszenierung im Kontext von KI ist der Heldenepos. Besonders in Quest-Narrativen, bei denen es um die Suche nach einer herausragenden Lösung geht, wird KI *selbst* als Held inszeniert. Verheißungen über KI können als neue Form des klassischen Heldenepos verstanden werden, nehmen aber gleichzeitig Bezug auf ikonische Vorbilder wie etwa die amerikanische Pulp-Serie *Captain Future*.<sup>13</sup> Trotz dystopischer Beimischungen<sup>14</sup> war die Serie bereits eine treffende Antizipation dessen, was gegenwärtig über Mensch-Maschine-Interaktion im Kontext von KI-Verheißungen gedacht wird.

Ein prominentes Beispiel für strategische Selbstinszenierung ist der japanische Robotiker Hiroshi Ishiguro. Seine Verheißung ist ein Androide, der nicht mehr von einem Menschen unterscheidbar ist und ein starkes Präsenzgefühl vermittelt. Ishiguro kann bereits eine ganze Entwicklungslinie menschenähnlicher Roboter aufweisen, die auch Geminoiden umfasst, also Androiden, die bestimmten Menschen (u. a. ihm selbst) ähnlich sehen.<sup>15</sup> Die zentrale Verheißung des „Mannes, der eine Kopie von sich selbst erschaffen hat“<sup>16</sup> wirkt für kritische Beobachter zwar oft gruselig, liefert aber ein perfektes Beispiel für die Funktion von Verheißungen als gesellschaftlicher Akzeptanztest.

**Zwang zur Futurisierung:** Strategische Inszenierungen basieren zwangsläufig auf Prognosen und stehen somit für einen *Zwang zur Futurisierung*. Futurisieren meint Zukunftsschau und bedeutet, eine Idee in die Zukunft zu projizieren. Für unkontrollierbare Aspekte ihres Lebens benötigen Menschen Erklärungsmodelle.

---

<sup>12</sup> [https://www.bmbf.de/bmbf/de/forschung/digitale-wirtschaft-und-gesellschaft/kuenstliche-intelligenz/kuenstliche-intelligenz\\_node.html](https://www.bmbf.de/bmbf/de/forschung/digitale-wirtschaft-und-gesellschaft/kuenstliche-intelligenz/kuenstliche-intelligenz_node.html) (30.01.2022).

<sup>13</sup> *Captain Future* ist eine Geschichte von Edmond Hamilton, die von 1940 bis 1944 erschien, in Europa durch die *Captain-Future*-Anime-Serie bekannt wurde und Anfang der 1980er-Jahre im TV ausgestrahlt wurde.

<sup>14</sup> Den technischen Verheißungen im *Captain-Future*-Universum (z. B. Raumschiff mit Tarnvorrichtung und Überlichtgeschwindigkeit) standen soziale Dystopien gegenüber, etwa Vorbehalte gegen Mischlinge aus unterschiedlichen Völkern.

<sup>15</sup> Für diese und weitere Roboter siehe: <https://eng.irl.sys.es.osaka-u.ac.jp/robot> (25.02.2022).

<sup>16</sup> <https://spectrum.ieee.org/hiroshi-ishiguro-the-man-who-made-A-copy-of-himself> (21.01.2022).

Mit Verheißungsnarrativen lässt sich daher an die Tradition von Weissagungen anknüpfen. Auf Basis einer Verobjektivierung der Welt bieten Verheißungserzählungen eine zeitgenössische Orientierungsleistung, die die rituelle Funktion des Orakels ersetzt. Auf der Basis von Mustererkennung ermöglichen KI-Anwendungen Extrapolationen in die Zukunft. Die mathematisch-statistischen Verfahren der KI dienen also als *Ersatz für die göttliche Prophezeiung oder das rituelle Orakel*. Das Motiv des Orakels (Gransche 2016) findet sich auch im KI-Manifest *Homo Deus*. Yuval Noah Harari argumentiert aus einer datenzentristischen Perspektive heraus, dass KI nicht nur ein allwissendes Orakel sei, sondern sogar ein souveräner Akteur (Harari 2017: 515). Trotz zahlreicher Bedenken wird KI vermehrt die Rolle des ‚Sehers‘ zugewiesen. Damit werden traditionelle Handlungsanleitungen zunehmend durch *gottähnliche* und zugleich undurchschaubare Maschinen ersetzt.

### 2.3 Verheißungen auf der kognitiv-epistemologischen Ebene

Auf einer kognitiv-epistemologischen Ebene drehen sich die zentralen Verheißungen um Fragen der Koexistenz und Koevolution von KI und Mensch. Roboter sind nicht nur ein Synonym für die zunehmende Automatisierung der Gesellschaft, vielmehr gelten sie als subjektsimulierende Maschinen (Scholtz 2008). Die *Verheißung künstlicher Subjektivität* wird ebenfalls im Kontext von Akzeptanztests für Disruptionen inszeniert. Ein Fallbeispiel ist *Sophia*, der erste Roboter der Welt, dem am 25. Oktober 2017 eine Staatsbürgerschaft verliehen wurde – in diesem Fall die saudi-arabische. Auf der Pressekonferenz bedankte sich *Sophia* artig, wengleich ein wenig steif, die Bewegungen wirkten vorprogrammiert. Die „politische Choreographie“ (Parviainen/Coeckelbergh 2021) der folgenden Welttournee von *Sophia* diente der Förderung des Marktes für soziale Roboter und damit den Interessen der Technologieindustrie.

Basis der Verheißung künstlicher Subjektivität sind einmal mehr subjektive Fantasien, aufsehenerregende Filmplots sowie effekthaschende Medienberichte, in deren Kontext von „Seelen in Computern“ oder „Geistern in Maschinen“ gesprochen wird (Heil et al. 2019: 33 f.). Auch der Science-Fiction-Autor Frank Schätzing erzählt vom „Puls der Maschine“ (Schätzing 2018: 144). Gerade die Beseeltheit einer dritten Instanz (neben Mensch und Natur) gilt als Verheißung, „die das gegenwärtig Machbare noch weit übersteigt.“ (Becker 2019: 18) KI-Narrative suggerieren, dass wir auf dem Weg zu sozial wahrgenommenen Maschinen sind und diese zukünftig immer selbstverständlicher werden. Neben



*Sophia* sind *ASIMO* – ein von Honda entworfener Roboter in Form eines Astronauten<sup>17</sup> – und *Ameca* – ein humanoider Unterhaltungsroboter von Engineered Arts (UK)<sup>18</sup> – zeitgenössische Beispiele für verkörperte KI. Kaum eine andere Verheißung von KI findet derart viel Resonanz in fiktionalen Genres. Im zeitgenössischen Roman *Klara und die Sonne* wird z. B. ein humanoider Roboter vorgestellt, dessen Funktion darin besteht, als ‚Künstlicher Freund‘ (KF) zu dienen und Kindern sowie jungen Erwachsenen Gesellschaft zu leisten oder diese gar zu ersetzen (Ishiguro 2021).

Während KI zunehmend verkörpert wird, gilt Körperlosigkeit umgekehrt als Voraussetzung für die *Verheißung der Unsterblichkeit*. Die Konvergenz von Bio- und Informationstechnologien geht einher mit zahlreichen Unsterblichkeitsfantasien. ‚Geisttransplantation‘ ist für Kritiker unerreichbar, für digitale Evangelisten eine spekulative Verheißung und für den Roboterpionier Hans Moravec gar eine kulturelle Verpflichtung (Moravec 1990: 166). Von hier ist es nicht mehr weit bis zur Idee der Superintelligenz. Mit der Veröffentlichung des Buches *The Singularity IS near* durch Ray Kurzweil (2005) gewann die Idee der Singularität an Bedeutung und Dynamik. In der Lesart von Kurzweil ist Singularität eine verheißungsvolle Option: Wenn Menschen ihr biologisches Gehirn mit einer Superintelligenz verbinden, erlangen sie Unsterblichkeit. Dieses Zukunftsnarrativ wird von einem komplementären Narrativ, dem ‚Singularity Skepticism‘ begleitet, denn viele Kritiker halten die Idee der Singularität schlicht für übertrieben (Lloyd, zit. n.: Brockman 2020: 9).

Für die versachlichte Zuschreibung eines Erkenntnis- und Nutzensgewinns durch digitale Daten wurde der Begriff ‚Big Data Glaubenssystem‘ (BDGS) geprägt (Lünich 2022). Auch im (angrenzenden) Fall von KI finden sich tief-sitzende Überzeugungen „bezüglich der eher konsequentiell-evaluativen Dimension des Erkenntnis- und des Nutzensgewinns“. (ebd.: 79)<sup>19</sup> Glaubenssätze und Glaubenssysteme basieren gerade auf der Nicht-Überprüfbarkeit der faktischen Voraussetzungen. Verheißungsvolle Erzählungen über KI sind somit *zukunftsge-wandte Deutungsentwürfe*.<sup>20</sup> Hierbei ist zwischen Wissens- und Glaubenssystemen zu unterscheiden, selbst wenn es Graubereiche gibt. Gerade populärwissenschaftliche Veröffentlichungen zu KI zeigen eine auffällige Verschmelzung

---

<sup>17</sup> <https://asimo.honda.com> (26.11.2021).

<sup>18</sup> <https://www.engineeredarts.co.uk/de/robot/ameca/> (26.02.2022).

<sup>19</sup> Aus soziologischer Perspektive handelt es sich um kollektive Vorstellungen einer sozialen Wirklichkeit, also ‚Frames‘ (Matthes 2014) oder Narrative (Müller-Funk 2008).

<sup>20</sup> Damit sind subjektive Überzeugungshaltungen gemeint, deren „Wahrheitsanspruch umstritten bzw. situativ zu bewerten ist und daher relational variabel aus Sicht der Gläubigen eingeordnet werden kann.“ (Lünich 2022: 82).

zwischen apokalyptischem religiösem Denken und wissenschaftlicher Forschung (Geraci 2008).

Glaubenssysteme funktionieren als *Projektionsfläche ausufernder Erwartungen* und zeigen alternative Welten auf (Abelson 1979). In diesem Kontext nimmt die Digitalisierung selbst zusehends religionsartige Züge an, für die sich Begriffe wie ‚Datafication‘ oder ‚Dataism‘ etablierten (van Dijck 2014). Hierbei wird durchaus auch eine mythologische (Boyd/Crawford 2012) oder religiöse (Harari 2017) Dimension der Glaubenssätze sichtbar. Im Kontext der Digitalisierung zeichnet sich immer deutlicher ein *religionsnaher technologischer Solutionismus* ab, der auf dem Glauben eines universellen Lösungspotenzials digitaler Werkzeuge beruht (Mozorov 2013). Diese Haltung wird auch als ‚Thinkism‘ kritisiert, dem mystisch verklärten Glauben, dass zukünftiger Fortschritt nur dann möglich ist, wenn die kollektive Intelligenz zunimmt (Kelly 2017). Kritiker erkennen zudem einen faustischen oder teuflischen Pakt (Zuboff 2015) oder wahlweise einen bedrohlichen Technikmythos (Mosco 2017). Gerade diese *religiöse Unterströmung von Technikentwicklung* rückt jeden Diskurs über KI in die Nähe von Verheißungen. Das Religionshafte<sup>21</sup> entwickelt sich basierend auf der Erfahrung technologischer Transzendenz, die immer mehr zum Alltagserlebnis wird (Geraci 2016). Weil Verheißungsgeschichten an theologisch-dogmatische Heilsverkündungen anknüpfen, können sie als Religionsersatz zur *dynamischen Stabilisierung von Gesellschaft in instabilen Zeiten* beitragen. Sie deuten Kontingenz narrativ um und ersetzen Wert- und Normvorstellungen, die im Prozess der Säkularisierung verlorengegangen sind. Kurz: Sie sind eine moderne Formen der Erlösung (Haring-Mosbacher 2008).

Jede Verheißung braucht einen Absender. Weil das Bedeutsame im Verborgenen wächst, erhalten KI-Verheißungen zusätzlich eine *distinktive Funktion*, die sich am offensichtlichsten im Neo-Mystizismus der KI-Apologeten zeigt, die vorgeben, die Zukunft bereits zu sehen. Wie ein Messias lassen sie sich von etwas leiten, das sie selbst nicht rational begründen können. KI-Verheißungen gleichen damit dem heilsgeschichtlichen Format der *Metanoia*, worunter (im theologischen Verständnis) eine spirituelle Bekehrung, die Änderung der eigenen Lebensauffassung sowie die Gewinnung einer neuen Weltsicht verstanden wird. Es sind immer wenige ‚Auserwählte‘, die diesen Prozess vorantreiben. In KI-Verheißungen erhalten menschliche Allmachtsfantasien einen konkreten Ort, den vorher Götter innehatten. In der *Rolle als transzendente Entität* füllt KI zugleich

---

<sup>21</sup> Selbst wenn diese Funktionen von den sozialen Akteuren nicht explizit mit dem Etikett ‚Religion‘ versehen werden, handelt es sich doch um implizite (Bailey 2002) oder private bzw. unsichtbare Religion (Luckmann 1967), deren Funktion in der Bündelung eines Sinnüberschusses besteht (Luhmann 1998).

eine Leerstelle im Pantheon der Götter. Die Stars der Digitalität sind das funktionale Äquivalent und die säkularisierte Version des Heiligen. Ein prominentes Beispiel dafür ist der Apple-Gründer Steve Jobs, der in einer Kultur ohne religiöse Horizonte Sehnsüchte nach einer höheren Sphäre bot, mit „Freude an starken Formulierungen, unabhängig davon, wie banal diese eigentlich waren.“ (Gumbrecht 2018: 44) Jobs füllte eine Rolle perfekt aus, die bereits von Hans Magnus Enzensberger beschrieben wurde, der zwei antagonistische Prophetien – digitale Evangelisten und Apokalyptiker – unterschied (Enzensberger 2000: 93).<sup>22</sup> In der Tat etablierte sich der Begriff des *digitalen Evangelisten* fest im Verkündungsrepertoire im Umfeld von KI. So wie die biblischen Evangelisten sich auf eine bessere Inspiration beriefen, lassen sich digitale Evangelisten von einer besonderen prognostischen Gabe leiten.<sup>23</sup> Damit erklärt sich auch der *implizit elitäre Ansatz* von Zukunftserwartungen im Umfeld von KI. Verheißungen passen zudem gut zu (radikalen) libertären oder anarchistischen Gesellschaftstheorien, wie sie etwa in der ultra-minimalen Staatstheorie von Robert Nozick, dem Vordenker des amerikanischen Radikalliberalismus vertreten werden (Nozick 2006).<sup>24</sup>

## 2.4 Verheißungen auf der zivilisatorisch-transformativen Ebene

Kennzeichen euphorischer Zukunftsversprechen auf einer zivilisatorisch-transformativen Ebene sind Aufbruchs-Narrative zur *Ankündigung einer besseren Welt*. KI wird hierbei als *soziale Innovation* wahrgenommen und damit als immaterielles Kapital verstanden. Im Mittelpunkt stehen menschenorientierte KI-Lösungen (Minevich 2021) und Reflexionen über gemeinwohlorientierte Anwendungsfelder (Brundage 2016). So legte etwa die EU eine Reihe von zukunftsweisenden Programmen und Fonds auf, die auf soziale Innovation und

---

<sup>22</sup> Dreh- und Angelpunkt des klassischen Essays sind allerdings die neuen, d. h. digitalen Medien und nicht KI. Die Argumentation lässt sich dennoch auf den Gegenstandsbereich KI übertragen.

<sup>23</sup> Nicht selten berufen sie sich dabei auf die prominente Idee des ‚Presencings‘, die Otto Scharmer im Gewand seiner ‚Theorie U‘ zu Weltruhm verholfen hat. Otto Scharmer ist Aktionsforscher und Gründer des *Presencing Institute* am Massachusetts Institute of Technology (MIT) in Cambridge (USA). Scharmer behauptet, dass Zukunftsentwicklungen in Form einer U-förmigen Bewegung ablaufen. Zunächst ist es, so Scharmer, notwendig, in einen Modus des Erspürens zu kommen, um dann zu wissen, was aus der Zukunft auf einen zukommt.

<sup>24</sup> Jedenfalls beziehen sich Tech-Propheten wie der Libertäre Peter Thiel explizit auf diese und ähnliche politischen Theorien (z. B. Thiel 2009).

Technologie ausgerichtet sind.<sup>25</sup> Für den Einsatz von KI als soziale Innovation oder im Kontext sog. Social Entrepreneurship gibt es zahlreiche Beispiele, die von inklusiven Lernassistenten, über mehrsprachige Sozialleistungsrechner bis hin zu Gesundheits-, Coaching- oder Antidiskriminierungs-Apps reichen.<sup>26</sup>

Das Grundprinzip wurde als ‚AI for social good‘ bekannt. Auch in diesem Kontext wird selten auf Verheißungen verzichtet: „Artificial intelligence, while not a silver bullet, could contribute to the multi-pronged efforts to tackle some of the world’s most challenging social problems.“ (Chui et al. 2018) Letztlich handelt es sich um *Heilungsnarrative*, die sich sowohl auf den Kampf gegen tödliche Krankheiten oder die Rettung des Weltklimas beziehen können. Der konzeptionelle Referenzrahmen dieser Narrative sind meist die 17 UN-Nachhaltigkeitsziele (SDGs = Sustainable Development Goals). So wurde z. B. 2016 das Unternehmen *atempo* als SDG Pioneer für den besonders innovativen Beitrag zur Erreichung der SDGs ausgezeichnet. *atempo* nutzt die KI-basierte Software *Capito*,<sup>27</sup> um die Komplexität von Texten für Menschen mit kognitiven Beeinträchtigungen in drei verschiedenen Sprachstufen zu reduzieren und trägt zur Erreichung der Ziele 4 (hochwertige Bildung), 10 (weniger Ungleichheit) und 16 (Friede, Gerechtigkeit und starke Institutionen) bei.

Die Heilungsversprechen können gleichwohl noch umfassender sein. Am Beispiel des *Venus-Projects*<sup>28</sup> soll die Traditionslinie techno-utopischer Machbarkeitsfantasien skizziert werden. Seit Jahrzehnten verspricht das Projekt *Lösungen für zentrale Menschheitsprobleme* und versteht sich als umfassender Aktionsplan „für eine soziale Sanierung“ der Welt auf Basis von High-Tech (Spiegel 2019: 80). Gegründet wurde es 1975 vom autodidaktischen Sozial-Architekten Jacque Fresco und dessen Partnerin Roxanne Meadows in Florida. Der zugrunde liegende Gesellschaftsentwurf lässt sich zwischen Techno-Utopie, New-Age-Esoterik und einer Mischung aus Verschwörungstheorien verorten. Gesellschaftsgestaltung wird ausschließlich als Ingenieursaufgabe verstanden (Willeke 1994), Gesellschaft als Maschine betrachtet und Menschen werden auf konditionierbare Wesen reduziert. Das frühe Manifest *Looking Forward* (Keyes/Fresco 1969) ist ein typisches Quest-Narrativ über die Potenziale von KI (auch wenn damals noch von Kybernetik gesprochen wurde): Automatisierte Fabriken, intelligente Wohnräume und eine vollständig vernetzte Infosphäre gehören zur Verheißung einer

---

<sup>25</sup> Horizon Europe, InvestEU, the EU Programme for Employment and Social Innovation (EaSI) and European Social Fund (ESF).

<sup>26</sup> <https://www.civic-innovation.de/ideenmarkt-und-wettbewerb/ideenmarkt> (26.02.2022).

<sup>27</sup> <https://www.capito.eu> (10.12.2021).

<sup>28</sup> <https://www.thevenusproject.com> (25.01.2021).

neuen Zivilisation. Dazu sammelt der autonom arbeitende Superrechner *Corecen* (Abkürzung für Correlation Center) Informationen, analysiert Fakten und trifft bestmögliche Vorhersagen. Von Wetter über die Wirtschaft bis hin zur Regierungsform soll sich damit jeder Aspekt des Daseins optimiert werden. Fresco wagt sich sogar auf das hochspekulative Feld der Evolution, indem er einen neuen Genotyp Mensch prognostiziert, der durch die Verheißungen der intelligenten Maschine 200 Jahre alt werden könnte – den *Homo Mechanus*.<sup>29</sup> Im zweiten Manifest *Designing the Future* knüpft Fresco an zwischenzeitlich erfolgte technologische Innovationen (darunter explizit KI) an und baut sein Verheißungsnarrativ weiter aus (Fresco 2007). Auf der Suche nach neuen Habitaten sollen menschliche Lebenswelten über Sensoren mit der Biosphäre verknüpft werden, um ein globales, autonomes Nervensystem entstehen zu lassen – als Gehirn und Nervensystem einer neuen Weltzivilisation.<sup>30</sup> Fresco kann als früher Apologet eines neuen Fortschrittsglaubens gesehen werden. Er nahm damit eine Rolle ein, die zwischenzeitlich von Akteuren wie Peter Thiel, Elon Musk oder Ray Kurzweil besetzt wird, denen „Leben und Denken im Zukunftsmodus“ (Gumbrecht 2018) gemein ist.

Auch die Idee, KI als neuen Leviathan zu nutzen, hat eine eigene Traditionslinie. Bereits Norbert Wiener träumte in *Mensch und Menschenmaschine* von einer totalen Regierungsmaschine mit außerordentlichen Privilegien (Wiener 1958). Damit ist der Wunsch verbunden, funktionale Gesellschaften durch die Verheißung algorithmischer Rationalität zu optimieren. Der Traum vom guten Diktator taucht in modernen Verheißungserzählungen als Versprechen des kybernetischen Regierens oder des Redesigns von Gesellschaft auf, wird in der Fachdebatte als ‚precision government‘ bezeichnet und ist mit weitreichenden Erwartungen verbunden. So schlug der Amazon-Programmierer Alessio Piergiacomi als Staatsform die Ablösung der Demokratie durch eine KI-Regierung vor, um der „zunehmenden Verdummung der gewöhnlichen Leute“ zu entkommen.<sup>31</sup> Der Ersatz diskursiver Formen der Rationalität durch eine numerisch-metrische Rationalität bzw. digitalen Rationalität (Han 2013) lässt in zahlreichen Anwendungsfeldern

---

<sup>29</sup> Die damit verbundene Idee der eugenischen Optimierung des Volkskörpers zeigt die Fallstricke techno-euphorischer Verheißungen. Diese Prognose entschärfte Fresco aufgrund der darin enthaltenen ethischen Implikationen in späteren Manifesten, um die öffentliche Akzeptanz des Venus-Projects nicht zu gefährden. Im Kern hielt er dennoch weiter am sozio-kybernetischen Steuerungsansatz fest.

<sup>30</sup> Fresco knüpft hier an den Gaia-Gedanken von James Lovelock an. Vgl.: Lovelock (1991, 2020).

<sup>31</sup> <https://www.quora.com/What-would-an-AI-government-look-like> (31.01.2022).

von der öffentlichen Sicherheit (Selke 2020) bis hin zur Klimakrise (Simanowski 2020: 73) konkurrierende Situationsanalysen sowie Notwendigkeiten des Souveränitätstransfers vom Menschen auf die Maschine entstehen.

Mittlerweile besteht Konsens darüber, dass die mit der Digitalisierung einhergehende Transformation nicht nur ein technologisches, sondern vor allem auch ein kulturelles Projekt ist (Becker 2019: 7). In einer zivilisationstheoretischen Perspektive eignen sich KI-Verheißungen als Lösungsversprechen in Nebenfolgesellschaften bzw. im Anthropozän. Dabei stellt sich die Frage, ob *KI als Pflichtaufgabe der Menschheit zwischen Disaster und Design* zu betrachten ist. Denn zentrale Menschheitsprojekte dienen immer wieder dazu, ultimative Grenzen zu überschreiten und Utopien zu verwirklichen.<sup>32</sup> Mit diesem Motiv widmet sich das Unternehmen *Open AI* der Erforschung von KI zum Wohl der Menschheit.<sup>33</sup> Eine gewaltige Ressourcenausstattung ermöglicht es, weitreichende und langfristige Ziele zu verfolgen, die (lt. Webseite) einen ‚positive human impact‘ garantieren sollen, sich also ‚mission-driven‘ und nicht ‚money-driven‘ darstellen. Die zivilisatorische Verheißung besteht darin, eine essenzielle Pflicht gegenüber der Menschheit einzulösen.<sup>34</sup>

Auf der allgemeinsten und abstraktesten Ebene besteht die Verheißung von KI darin, zur Gesamtökologie universeller Intelligenz(en) beizutragen, wobei KI in Quest-Narrativen als Fortsetzung natürlicher Intelligenz und nicht als Konkurrenz oder gar Bedrohung betrachtet wird (Kelly 2017). Damit ist zugleich die Verheißung verbunden, die Lücke zwischen lebender und toter Materie zu schließen, also eine Synthese des Organischen mit dem Anorganischen zu ermöglichen. Als bereits Norbert Wiener nach den Möglichkeiten einer neuen Systemökologie fragte, ließ er sich möglicherweise von Jakob Johann von Uexküll, dem Gründer der Biosemiotik, inspirieren (Brockman 2020: xxi). In seinem Klassiker *Umwelt und Innenwelt der Tiere* führte Uexküll 1909 erstmals den Begriff Umwelt – in Angrenzung zu Umgebung – ein (von Uexküll 2018). Jedes Lebewesen existiert in einer *Umgebung*, die passiven Objektcharakter annimmt. Die *Umwelt* wird hingeben durch Wahrnehmungen sowie Aktivitäten in Raum und Zeit subjektiv gestaltet. Die Umwelt spiegelt sich in der Innenwelt des Lebewesens, die aus zwei Komponenten besteht: Erstens die *Merkwelt*, die alles umfasst, was der Organismus eines Lebewesens wahrnehmen kann. Zweitens die *Wirkwelt*, die alles

---

<sup>32</sup> Vgl. zu zivilisatorischen Menschheitsexperimenten und real-utopischen Projekten ausführlich Selke (2022).

<sup>33</sup> <https://openai.com/blog/introducing-openai/> (10.05.2021).

<sup>34</sup> Ein Überblick über die Projekte des Unternehmens findet sich unter folgendem Link: <https://openai.com/projects/> (18.02.2022).

umfasst, was ein Lebewesen tun kann. Zwischen Merkwelt und Wirkwelt besteht eine Wechselwirkung („Funktionskreis“) (ebd.). Dieses Modell und Begriffssystem dürfte zahlreiche Pioniere der Kybernetik in ihrem Denken beeinflusst haben. Denn es ermöglicht einen alternativen Blick auf das Verhältnis zwischen Mensch und Maschine. Einerseits lassen sich Gemeinsamkeiten zwischen Organismen und Maschinen erkennen. So vereinigen sich die jeweiligen Teile eines Organismus oder einer Maschine funktional und nicht nur rein formell zu einem Ganzen. Andererseits besitzen Organismen ‚übermaschinelle‘ Eigenschaften wie Formbildung und Regeneration. „Die Maschinen sind alle vom Menschen gemacht“, so Uexküll, „die Organismen entstehen aus sich selbst. Darin liegt ihre hauptsächlichste übermaschinelle Fähigkeit.“ (ebd.) Unter Umständen ließe sich KI dann als *Übergang* zwischen dem Maschinellen und dem Organischen ansehen, weil damit die von Uexküll angesprochene übermaschinelle Eigenschaft der Formbildung verbunden ist. Bei einer starken Superintelligenz käme dann noch die Eigenschaft der Regeneration hinzu, die KI vollends in Richtung eines Organismus bringen würde. Aber das ist ja bekanntlich Zukunftsmusik.

---

### **3 Zukunftseuphorie als Signatur der Gegenwartsgesellschaft**

Verheißungserzählungen sind eine spezifische kommunikative Gattung (Knoblauch 2019), die sich zwischen Gesellschaftsdiagnose und Zukunftspraxis verorten lässt. Sie sind ein Beispiel für ‚Schwellendebatten‘, die sich mit der Frage befassen, „ob wir uns gerade mitten in oder kurz vor einer fundamentalen, epochalen gesellschaftlichen Transformation befinden, ob eine neue historische Phase oder ein neuer Gesellschaftstyp im Entstehen ist.“ (Peters 2007) Zugleich sind *Zukunftserzählungen* als *Zukunftspraktiken* „eng mit Affekt- und Begehrenskonstellationen verbunden.“ (Krämer 2019: 87).

#### **3.1 Zukunftsmangel und Jahrmärkte der Hoffnung**

Wird Zukunft als Ware verstanden, die in stilbildenden Prozessen kollektiv geformt wird, entsteht Zuversicht nur dann, wenn Technologie nicht als Schicksal (Brynolfsson 2014: 309) und Zukunft nicht als Mangel empfunden werden. Die zentrale gesellschaftliche Funktion von Verheißungen besteht darin, Angst vor Zukunftsmangel und Erfahrung von Erschöpfungszuständen eine progressive

Haltung entgegenzusetzen: Zukunftseuphorie. Verheißungen füllen eine *affektive Lücke*. Zukunftseuphorische Verheißungen dienen als Trostersatz.

Diese Orientierungs- und Stabilisierungsleistung von Verheißungen muss gleichwohl hinterfragt werden. Der Trend, Zukunftsmangel durch ein *Übermaß an Rechenleistung* und damit die *Flucht ins Technische* auszugleichen, wurde bereits vom Computerpionier Josef Weizenbaum kritisiert, der stark euphorisierte Zuschreibungen an Technik beobachtete (Weizenbaum 1977).<sup>35</sup> Menschliches Denken sollte berechenbar werden, Aussagen über das Leben verwandeln sich von Wahrheit zu Beweisbarkeit. Computer galten als komfortable (zugleich aber auch manipulative) Werkzeuge<sup>36</sup> der Komplexitätsbewältigung. Gerade diese Zuschreibungen sind der Grund dafür, dass „Menschen bei der Beurteilung des Potenzials weit über das Ziel hinausschießen.“ (ebd.: 24) Als Nebenfolge der positivistischen Aufladung von Rechenmaschinen diagnostizierte Weizenbaum Utopiemüdigkeit.<sup>37</sup> Auch im Kontext von KI wird der Wegfall von Denkalternativen beklagt, etwa wenn in Anpassungs-Narrativen KI als Entität vorausgesetzt und vermehrt unter dem Gesichtspunkt geostrategischer Systemkonkurrenz verhandelt wird. Aus diesem Spannungsfeld resultiert die zunehmende *Nachfrage nach Hoffnung als Form symbolischen Kapitals*. Ausgestattet mit einer Mischung aus Insiderwissen und Charisma sind digitale Evangelist\*innen darauf spezialisiert, Zukunftseuphorie in Form eines ‚Gospels der Disruption‘ oder guter Börsennarrative zu erzeugen. Die Erzeugung von Zukunftseuphorie funktioniert marktförmig sowie im Kontext der Aufmerksamkeitsökonomie (Frank 1998). Diese Blendung ist an rhetorische Techniken und narrative Überzeugungsleistungen (Sukulla 2019) gebunden.

Gegenwärtig ist die Westküste der USA die Keimzelle heterogener Formen des vulgären Techno-Utopismus‘. Die damit verbundenen Machbarkeitsfantasien erscheinen schier grenzenlos, auch deshalb, weil techno-utopische Subkulturen

---

<sup>35</sup> Anhand der euphorischen Reaktionen auf sein (aus heutiger Sicht simples) KI-Programm ELIZA machte Weizenbaum deutlich, „welch enorm übertriebene Eigenschaften selbst ein gebildetes Publikum einer Technologie zuschreiben kann oder sogar will, von der es nichts versteht.“ (Weizenbaum 1977: 19).

<sup>36</sup> Zur Kritik an komfortablen Technologien vgl. Illich (2009) und zur Kritik an der Manipulation durch Quantifizierung vgl. klassisch Marcuse (2004: 243).

<sup>37</sup> Diese Furcht vor ‚sozialen Erfindungen‘ (heute würde man sagen: sozialen Innovationen) sowie die Effizienzillusion, die mit der Rechenleistung einhergeht, beschreibt bereits Weizenbaum: „Aber von den vielen Möglichkeiten einer gesellschaftlichen Innovation, die er dem Menschen eröffnete, war die verhängnisvollste die, ihm alle Überlegungen in Richtung auf eine wesentliche Veränderung aus dem Kopf zu schlagen.“ (Weizenbaum 1977: 4).



wie Transhumanist\*innen, Technogaianist\*innen und Singularist\*innen konvergieren. Palo Alto wirkt wie ein Jahrmarkt der Zukunftshoffnung. Das Hauptmotiv des Digital- und Überwachungskapitalismus (Zuboff 2018) und die fortwährende Anbetung der Digitalreligion(en) (Wolff 2020: 21) erzeugen ein sich selbst verstärkendes Zukunftsnarrativ. Für Hans-Ulrich Gumbrecht besteht der *genius loci* sogar in einem neuen Weltgeist (Gumbrecht 2018: 17). Es gibt also einen Ort, an dem zukunfts euphorische Geschichten wie am Fließband produziert werden. Zukunftseuphorische Szenarien entstehen im Valley innerhalb eines besonderen institutionellen und zivilisatorischen Bezugsrahmens zwischen Depression und Euphorie (ebd.: 50). Zudem ist das Valley ein Ort, an dem es üblich ist, „im Gestus von Science-Fiction zu reden.“ (ebd.: 53) Damit wird der Wunschlosigkeit konsumistisch gesättigter Gesellschaften die *Verheißung ultimativer Wünsche* entgegengesetzt. Auf dieser Grundlage sieht Gumbrecht gar das Entstehen einer neuen sozialen Klasse, die einerseits über die algorithmische Macht, andererseits über die Fähigkeit zu utopischer Selbstdarstellung verfügt (ebd.: 71).

### 3.2 Sehnsucht nach welthaltigem Trost

Um Zukunftseuphorie als Trostersatz zu verstehen, ist ein soziologischer Rückgriff auf die Philosophie bzw. Anthropologie des Trostes notwendig. Einerseits wird nicht alles, was Trost spendet, auch als Trost bezeichnet (Blumenberg 1987: 153) Andererseits müssen zukunftsbejahende Gesellschaften offen für neue und bislang wenig berücksichtigte Formen des Trostes sein. Angesichts der vielen Herausforderungen, die mit menschlichem Leben verbunden sind, herrscht schnell Konsens über die Notwendigkeit von Trost.<sup>38</sup> Tatsächlich gibt es eine lange Traditionslinie des sozialen Formats Trost, zudem Trost-Institutionen, Trost-Rituale, Trost-Metaphern und Trost-Rhetoriken.<sup>39</sup> Gleichwohl herrscht Dissens

---

<sup>38</sup> Bereits der Soziologe Georg Simmel beschrieb den Menschen als trostsuchendes Wesen und Trost als Mechanismus, der das „Leiden am Leiden aufhebt.“ (Simmel 1995: 17).

<sup>39</sup> Diese reichen von den Tröstungen der alttestamentlichen Propheten über das Trostpflaster Vernunft im Gewand moderner Philosophie bis hin zu Ersatztröstungen, wie sie die Psychologie und Psychoanalyse beschreibt (Freud 1994: 270). Für Hans Blumenberg (2006: 625) geht es beim Trost um eine (symbolische) Delegation, die sich gerade aus der Fragilität kultureller Institutionen speist, weil nicht der Mensch, sondern Kultur selbst die Züge des Mangelhaften trägt. Für Blumenberg sind Menschen rhetorische Wesen, deren Kultur sich aus Bildern, Geschichten, Metaphern und Mythen zusammensetzt, „deren praktisch-rhetorischer Sinn nicht nur, aber auch Tröstung ist.“ (Dober 2019: 25).

über Trostformen und Möglichkeiten der Defizitaufhebung durch Tröstungsrituale (Dober 2019: 11).

In soziologischer Perspektive meint Trost die Linderung einer grundlegenden Kontingenzangst. Die Hoffnung auf eine bessere Welt in Form von jenseitigen Heilsversprechen war bislang eine probate Troststrategie. In postmodernen Gesellschaften sind Menschen hingegen gezwungen, selbständig Sinn aus ihrem Dasein machen. Je größer dabei der Vorrat an diesseitiger Zukunft ist, desto *welthaltiger* müssen narrative Trostangebote sein: Trost muss seine Wirkung aus konkreten Erfahrungen beziehen. „Guter Trost muss uns also vor allem helfen, mit einer anderen Einstellung in der Welt weiterzuleben.“ (Dober 2019: 12) Diese Trostform wird hier *Zukunftseuphorie* genannt. Verheißungen dienen als *Katalysator für entlastende Zukunftseuphorie*, die in säkularisierten und erschöpften Gesellschaften wie ein Trostersatz wirkt.

Die gesellschaftliche *Funktion* von Verheißungsgeschichten über KI besteht also darin, einen tröstenden Weltbezug herzustellen, der plausibel erscheint. Diese These beruht auf einer soziologischen Umdeutung von Blumenbergs Ethik des Trostes. Trost ist nicht nur ein zentrales Thema einer anthropologischen Philosophie,<sup>40</sup> sondern auch Ausdruck pragmatischer Konventionen, die als soziale Koordinationslogiken zwischen kaum noch kompatiblen gesellschaftlichen Teilsystemen (Politik, Wirtschaft, Recht, Kultur, Wissenschaft) wirken (Diaz-Bone 2015). In KI-Verheißungen geht es nicht (mehr) darum, das Unbegreifliche zu begreifen, sondern es als kollektiven Erwartungs- und Handlungshorizont praktisch zu nutzen.

### 3.3 KI als kollektives Tröstungsprojekt

Die Funktion von Zukunftseuphorie besteht darin, eine ausreichend plausible und praktikable Lösung für die permanente Notwendigkeit zur individuellen und kollektiven Kontingenzbewältigung zu liefern. In verheißungsvollen KI-Narrativen verbinden sich *rhetorische Trostmittel* mit *lebenspraktischen Versprechen*. Verheißungsnarrative sind tröstende Sprachbilder und handlungsleitende Konventionen in einem. Zukunftseuphorie hat eine *metaphorische* und zugleich eine *praktische* Signatur. Trost wird nicht nur als „anthropologische Kardinalkategorie“ (Dober 2019: 14) betrachtet, sondern als Zukunftspraxis im Umgang mit dem

---

<sup>40</sup> Im Kontext der anthropologischen Philosophie geht es primär um den „menschlichen Mangel an absoluter Evidenz, an Wahrheit, unser Mangel an Lebenszeit, an Heil, an Perfektion.“ (Dober 2019: 13).

Kommenden der Gesellschaft (Krämer 2019; Maresch 2010). KI-Verheißungen lassen sich einerseits als zeitgenössische Trostmittel einordnen, andererseits als *kollektives Tröstungsprojekt* verstehen. Im Mittelpunkt steht dabei die Suggestion, dass bisherige zentrale Kränkungen des Menschseins endgültig eliminiert werden könnten (Rohbeck 1993, Krüger 2019).<sup>41</sup> Damit ist eine seelsorgerische Komponente verbunden, für die entscheidend ist, „dass es um Linderungen, nicht um Heilungen oder endgültiges Heil geht.“ (Dober 2019: 14) KI-Verheißungen zeichnen somit kein abschließendes Bild der Zukunft, sie garantieren keine endgültige Erlösung. Als Signatur der Gegenwartsgesellschaft sind sie gleichwohl wirksam und handlungsleitend.

---

## 4 Ausblick: Verheißungsfreie Zukunftsbejahung?

Abschließend ist zu fragen, wie eine Kritik an dieser Form des Trostersatzes aussehen könnte, denn es gibt zahlreiche Beispiele dafür, dass utopische Verheißungen rasch in dystopisches Unheil kippen können oder der „großen Euphorie (...) die Traumzerstörung“ folgt (Wolff 2020: 10). Trost ist ambivalent, weil es darauf ankommt, wie sehr Tröstung mit einer Abwendung von der Wirklichkeit verbunden ist und wie viel Zuwendung zur Welt darin enthalten ist. Aus dieser angedeuteten Kritik ergibt sich die Notwendigkeit, über alternative Kontexte verheißungsfreier Zukunftsbejahung sowie Gestaltungsrahmen für Lösungsansätze nachzudenken.

### 4.1 Verheißungen als Mittel der Diskurserneuerung

Verheißungserzählungen zeichnen sich durch eine *intermediäre Funktion* aus: KI-Narrative verbinden Zukunftsszenarien und Gegenwartsdiagnosen und damit latente Erwartungsstrukturen mit der Verarbeitungskapazität der Gesellschaft für Zukunft. Sie ermöglichen eine übergeordnete Meta-Perspektive auf KI und sie lassen sich zu *transformativen Szenarien* auffächern. Damit haben Zukunftsnarrative auch eine Übersetzungs- und Reintegrationsfunktionen. Hierin liegt die *gesellschaftsdiagnostische Perspektive* auf Verheißungen. Aus soziologischer Perspektive kann in der Herstellung von (kommunikativer) Anschlussfähigkeit durch

---

<sup>41</sup> Der Philosoph Johannes Rohbeck (1993) sieht folgende vier Kränkungen: 1) Kosmologisch (heliozentrisches Weltbild), 2) Biologisch (Evolutionstheorie), 3) Psychisch (Psychoanalyse/Unterbewusstsein) und 4) Technologisch.

Verheißungsnarrative ein Akt der Sinnsetzung gesehen werden, denn technische Systeme bleiben so lange ‚sinnlos‘, bis ihnen von Menschen Sinn gegeben wird. Verheißungsgeschichten sind Teil einer unverzichtbaren Sinnstrategie, die Zukunft erst ermöglicht. Die *gesellschaftliche Funktion* von *Zukunftseuphorie* als Trostersatz besteht hingegen darin, im Zusammenspiel mit dem Gegennarrativ *Zukunftsangst* das Wahrnehmungs- und Diskursfeld abzustecken. Die Extrempole *Zukunftseuphorie* und *Zukunftsangst* sind notwendig, um der (öffentlichen) Debatte über KI eine Form zu geben<sup>42</sup> und somit das Spektrum für eine gesellschaftliche Auseinandersetzung zu öffnen. Denn erst durch „extreme Positionen wird zuvor Undenkbare denkbar.“ (Budian 2020: 40) *Zukunftsangst* verhindert blinden Aktionismus, der die Gesellschaft überfordern würde: Darin liegt der gesellschaftliche Wert der Dämpfung durch Gegennarrative.<sup>43</sup> Umgekehrt erzeugt *Zukunftseuphorie* Kraft, den schwierigen und aufwendigen Prozess des Wandels anzutreiben und gesellschaftlichen Stillstand zu verhindern. Verheißungsvolle *Zukunftserzählungen* liefern „der Gesellschaft Blaupausen für eine ‚kollektiv erwünschte Zukunft‘.“ (Maresch 2001: 235) Idealerweise entsteht auf diese Weise ein *nivellierender Umgang mit Argumenten* und ein gesunder Umgang mit dem Neuen, denn in der „Aneignung von Neuem ist sowohl dessen Problematisierung als auch Visionierung notwendig (...) Utopie und Dystopie sind beide Teile desselben notwendigen Verständigungsprozesses.“ (Budian 2020: 40) *Zukunftsangst* und *Zukunftseuphorie* sind also zwei Seiten einer Medaille, gemeinsam dienen sie der *Diskurserneuerung*.<sup>44</sup>

---

<sup>42</sup> Vgl. zu den theoretischen Grundlagen der Form einer Debatte im Kontext von Diskurs Design: Friedrich (2013).

<sup>43</sup> *Zukunftsnarrative* können auch die Form von *Dagegen-Narrativen* annehmen. Die Kontrastfolie zu Verheißungen sind daher *Unheil-Verheißungen*. Einerseits besteht der narrative Wissensraum zu einem signifikanten Teil aus Gegennarrativen. Grundlegend wird in diesen Erzählungen die Unheimlichkeit des Künstlichen thematisiert, zu dem die meisten Menschen keinen empirischen Zugang finden. Spezifische Gegennarrative thematisieren hingegen Missbrauch und Kontrollverlust durch KI, unkontrollierte Entfesselungs- Entfremdungs- und Freisetzungprozesse, mit KI verbundene Entmündigungsurteile sowie normativ-ethische Grenzverschiebungen durch schleichenden Wandel. Der vorliegende Beitrag fokussiert aus Platzgründen auf die positiven Verheißungen und nicht auf *Unheil-Verkündungen*.

<sup>44</sup> Das ist auch der Grund dafür, dass in diesem Beitrag über Verheißungen nicht auf die *Dagegen-Narrative* zu KI verzichtet werden kann.

## 4.2 Konturen einer verheißungsfreien Zukunftsethik

In jeder Epoche machen sich Menschen „Gedanken über die Welt und die Stellung des Menschen in ihr.“ (Fischer 2010: 21) Vor diesem Hintergrund geht mit KI die *Verheißung eines alternativen Orientierungsangebots* einher. KI bietet neue und zugleich gleitende Referenzrahmen (Selke 2020), die von Empfehlungssystemen beim Online-Shopping bis hin zur Gestaltung des planetarischen Lebensraums reichen. Vor diesem Hintergrund können KI-Verheißungen als Teil eines größeren Bildes betrachtet werden, das die Welt als permanenten Gegenstand des *Zukunftsdesigns* zwischen Überlebens-, Sicherheits-, Gesellschafts- und Selbstdesign versteht (Borries 2019).<sup>45</sup> Zeitgenössische KI-Narrative weisen Bezüge zu allen der genannten Aspekte auf. Gerade weil Überlebensdesign kollektiven Erfahrungen des Scheiterns entspringt, bieten Verheißungserzählungen ein tröstendes Gegennarrativ. Weil Zukunft zunehmend als Katastrophe verstanden wird, weil die Katastrophe die ultimative ‚Offenbarung‘ darstellt (Horn 2014: 20 ff.) und weil Katastrophenszenarien schon immer dazu dienten, Herrschaft zu legitimieren (Borries 2019: 51), werden Gedankenexperimente zur Zukunft der Menschheit unter der Vormacht einer starken KI verständlich: Kybernetisches Regieren mittels einer allmächtigen KI (Simanowski 2020) oder die Rettung des Ökosystems durch eine Hyperintelligenz (Lovelock 2020) sind daher gleichermaßen Topoi aktueller wissenschaftlicher Diskurse wie auch der Science-Fiction rund um KI (z. B. Hannig 2022) sowie zahlreicher Mischformen zwischen Wissenschaft und Literatur im Kontext künstlerischer Forschung (z. B. Lee/Chen 2021). Gerade die Vielzahl von Katastrophenszenarien macht es notwendig, über hoffnungsvolle Gestaltungsoptionen und eine positive bzw. utopische Zukunftsethik nachzudenken. Krisenbewusstsein und Zukunftseuphorie bedingen sich gegenseitig. Zukunftsdesign darf allerdings nicht allein von den Risiken der Technik ausgehen, sondern muss einerseits den Menschen selbst als Risiko sehen,<sup>46</sup>

---

<sup>45</sup> Vgl. dazu: „Aus einem aufklärerischen Verständnis von Verantwortung heraus ist Weltentwerfen (...) eine moralische Verpflichtung. (...) Weltentwerfen muss von Offenheit geprägt sein, sich als immerfort suchender Prozess verstehen und dabei Unvollständigkeit und Unabgeschlossenheit akzeptieren.“ (Borries 2019: 122).

<sup>46</sup> Vgl. dazu: „The main risk in AI (...) is not so much whether we can mathematically understand what they’re thinking; it’s whether we’re capable of teaching them ethical behaviour. We’re barely capable of teaching each other ethical behaviour.“ (Church, zit. n.: Brockman 2020: 241).

andererseits Gestaltungsoptionen zwischen Erfolg und Ernüchterung betonen.<sup>47</sup> Als Empfehlung bietet sich an, kritisch nachzufragen, „wann immer uns viel versprochen wird. (...) Nicht um prinzipiell jede visionäre Idee schlecht zu machen, sondern um die Bedingungen zu prüfen, unter denen sie sich realisieren lässt. Auch bei der schönsten Utopie sollte die Frage nach den Risiken und Nebenwirkungen nicht fehlen.“ (Grunwald 2019: 20) Gemeinsam mit massenmedial wirkmächtigen Visionen in Film und Science-Fiction bestimmen populärwissenschaftliche Akteur\*innen die Sprache, Leitbilder und Narrative über KI. Dies gilt gleichermaßen für positiv konnotierte, verheißungsvolle Narrative wie auch für Gegennarrative. Dieser narrative Wissensraum prägt das gesellschaftliche Verständnis von KI wesentlich stärker als Faktenwissen über technische Neuerungen oder wissenschaftliche Diskurse. Deshalb werden auch spekulative KI-Narrative in zahlreichen Handlungsfeldern handlungsleitend (Johnson/Verdicchio 2017).

Offen bleibt die Frage, ob und wie *verheißungsfreie Zukunftserzählungen* im Kontext einer progressiven Zukunftsethik aussehen könnten. Deren Funktion bestünde darin, die gesellschaftliche Leitfähigkeit für soziale Utopien jenseits von technikdeterministischen Narrativen zu steigern und auf diese Weise einen Beitrag zum zivilisatorischen Wandel zu leisten. Verheißungsfreie Narrative müssten kognitive *Emanzipation von KI* ermöglichen, anstatt kollektive *Entlastung durch KI* zu versprechen. Zur euphorischen Aufbruchsstimmung der Unternehmen, die KI als Schlüsseltechnologie des 21. Jahrhunderts zelebrieren und zur medialen Euphorie in Form verheißungsvoller KI-Narrative, gehört schließlich auch die Akzeptanz der Bürger\*innen als Optionsschein auf eine bessere Zukunft. Allen Perspektiven gemein ist, dass KI längst kein vager Zukunftstraum mehr ist.

---

## Literatur

- Abelson, Robert. 1979. Differences Between Belief and Knowledge Systems. In *Cognitive Science*, 3/4, *Differences Between Belief and Knowledge Systems*, 355–366.
- Alkemeyer, Thomas, Nikolaus Buschmann, und Thomas Etzemüller. 2019. (Hg.). *Gegenwartsdiagnosen. Kulturelle Formen gesellschaftlicher Selbstproblematierung in der Moderne*. Bielefeld: transcript.
- Bailey, Edward. 2002. *The Secular Quest for Meaning of Life*. Lewiston: Edwin Mellen.
- Bayramoglu, Yener, und Varela, María do Mar Castro. 2021. *Post/pandemisches Leben. Eine neue Theorie der Fragilität*. Bielefeld: Transcript.

---

<sup>47</sup> Vgl.: „I very much hope that a new generation of leaders (...) can rise above the usual tribal, zero-sum games and steer humanity past these dangerous waters were we are in.“ (Tallinn, zit. n.: Brockman 2020: 99).

- Becker, Manfred. 2019. *Das digitale Narrativ. Wer schreibt es und was steht drin?* Augsburg: Rainer Hampp Verlag.
- Blumenberg, Hans. 1987. *Die Sorge geht über den Fluss*. Frankfurt a.M.: Suhrkamp.
- Blumenberg, Hans. 2006. *Beschreibung des Menschen*. Beschreibung des Menschen.
- Borries, Friedrich von. 2019. *Weltentwerfen. Eine politische Designtheorie*. Berlin: Suhrkamp.
- Boyd, Danah, und Kate Crawford. 2012. Critical Questions for Big Data. Information. In *Communication & Society*, 15, *Critical Questions for Big Data*, 662–679.
- Brockman, John. 2020. (Hg.). *Possible Minds. 25 Ways of Looking at AI*. New York: Penguin Books.
- Brundage, Miles. 2016. Artificial intelligence and responsible innovation. In *Fundamental Issues of Artificial Intelligence. Artificial intelligence and responsible innovation*, hrsg. Vincent Müller, Vincent, 541–552. Berlin: Springer.
- Brynjolfsson, Erik, und Andrew McAfee. 2014. *The Second Machine Age. Wie die digitale Revolution unser aller Leben verändert*. Kulmbach: Börsenmedien AG.
- Budian, Elisa. 2020. *Bedrohung oder Verheißung? Über Medienangst und Medieneuphorie bei digitalen Spielen*. Glückstadt: Verlag Werner Hülsbüch.
- Canetti, Elias. 1976. *Dramen* (darin: Die Befristeten). München: Hanser.
- Castro, Daniel, und Joshua New. 2016. The Promise of Artificial Intelligence: Center for Data Innovation. Download: <https://euagenda.eu/upload/publications/untitled-53560-ea.pdf>
- Cave, Stephen, Kanta Dihal, und Sarah Dillon. 2020. Introduction: Imagining AI. In *AI Narratives. A History of Imaginative Thinking about Intelligent Machines*, hrsg. Stephen Cave, Kanta Dihal und Sarah Dillon, 1–21. Oxford: Oxford University Press.
- Chui, Michael, Martin Harryson, und James Manyika. 2018. Notes from the AI frontier. Applying AI for social good. McKinsey & Company. Download: <https://www.mckinsey.com/~media/mckinsey/featured%20insights/artificial%20intelligence/applying%20artificial%20intelligence%20for%20social%20good/mgi-applying-ai-for-social-good-discussion-paper-dec-2018.ashx>
- Degele, Nina. 2002. *Einführung in die Techniksoziologie*. München: Fink.
- Diaz-Bone, Rainer. 2015. *Die „Economics of convention“*. Grundlagen und Entwicklungen der neuen französischen Wirtschaftssoziologie. Wiesbaden: Springer VS.
- Dober, Benjamin. 2019. *Ethik des Trostes. Hans Blumenbergs Kritik des Unbegrifflichen*. Weilerswist: Velbrück Wissenschaft.
- Enzensberger, Hans Magnus. 2000. Das digitale Evangelium. Propheten, Nutznießer, Verächter. Christoph-Martin-Wieland-Vorlesungen. Erfurt.
- Fenn, Jackie und Mark Raskino. 2008. *Masterin the hype cycle. How to choose the right innovation at the right time*. Boston, Mass.: Harvard Business Press.
- Fischer, Ernst Peter. 2010. *Hinter dem Horizont. Eine Geschichte der Weltbilder*. Berlin: Rowohlt.
- Fischer, Ernst Peter. 2014. *Die Verzauberung der Welt. Eine andere Geschichte der Naturwissenschaften*. München: Siedler.
- Frank, Georg. 1998. *Ökonomie der Aufmerksamkeit. Ein Entwurf*. München: Carl Hanser Verlag.
- Fresco, Jacque. 2007. *Designing the Future*. Florida, USA: Osborn Publishing.

- Freud, Signmund. 1994. *Das Unbehagen in der Kultur und andere kulturtheoretische Schriften*. Frankfurt a.M.: Fischer.
- Friedrich, Stephan. 2013. Die Form der Debatte. Skizzen zum Diskurs Design. In *Revue. Magazine for the Next Society*, 11: 80-85.
- Geraci, Robert. 2016. Technology and Religion. In *Handbook of Science and Technology Convergence*, hrsg. William Brainbridge und Roco Michail, 907-917. Switzerland: Springer Reference.
- Geraci, Robert. 2008. Apocalyptic AI: Religion and the Promise of Artificial Intelligence. In *Journal of the American Academy of Religion*, 76/1: 138-166.
- Gransche, Bruno. 2016. The Oracle of Big Data. In *The International Review of Information Ethics*, 24, *The Oracle of Big Data*, 55-62.
- Grunwald, Armin. 2019. *Der unterlegene Mensch. Die Zukunft der Menschheit im Angesicht von Algorithmen, künstlicher Intelligenz und Robotern*. München: Riva Verlag.
- Gumbrecht, Hans Ulrich. 2018. *Weltgeist im Silicon Valley. Leben und Denken im Zukunftsmodus*. Zürich: NZZ Libro.
- Gutmann, Mathias, Marie-Claire Haag, und Christian Wadehul. 2021. Verheißung, Verdammung oder einfach ein Selbstmissverständnis? Sprachkritische Überlegungen zum Umgang mit KI und ihren Beschreibungen. In *Künstliche Intelligenz – Die große Verheißung*, hrsg. Anna Strasser, Wolfgang Sohst und Ralf Stapelfeldt et al., 221-240. Berlin: xenomoi Verlag.
- Hagemester, Michael. 2003. Die Eroberung des Raums und die Beherrschung der Zeit. Utopische, apokalyptische und magisch-okkulte Elemente in den Zukunftsentwürfen der Sowjetzeit. In *Musen der Macht. Medien in der sowjetischen Kultur der 20er und 30er Jahre*, hrsg. Jurij Murašov und Georg Witte, 259-286. München.
- Han, Byung-Chul. 2013. *Digitale Rationalität und das Ende des kommunikativen Handels*. Frankfurt a.M.: Suhrkamp.
- Hänggi, Marcel. 2015. *Fortschrittsgeschichten. Für einen guten Umgang mit Technik*. Frankfurt a.M.: Fischer.
- Hannig, Theresa. 2022. *Pantopia*. Berlin: Fischer.
- Harari, Yuval Noah. 2017. *Homo Deus. A Brief History of Tomorrow*. London: Penguin.
- Haring-Mosbacher, Sabine. 2008. *Verheißung und Erlösung. Religion und ihre weltlichen Ersatzbildungen in Politik und Wissenschaft*. Wien: Passagen.
- Heil, Reinhard, Christian Wadehul, und Karsten Wendland. 2019. Robotik und die Zuschreibung von Bewusstsein. In *Die digitale Revolution und ihrer Kinder. Brennpunkte digitaler Ethik*, hrsg. Hanna Füllig und Gernot Meier, 27-35. Evangelische Zentralstelle für Weltanschauungsfragen.
- Ho Lee, Choong und Yoon Hyung-Jin. 2017. Medical big data. Promises and challenges. In *Kidney Research and Clinical Practice* 36/1: 3-11.
- Horn, Eva. 2014. *Zukunft als Katastrophe*. Frankfurt a.M.: Firscher Wissenschaft.
- Illich, Ivan. 2009. *Tools for Conviviality*. London: Boyars Publishers.
- Ishiguro, Kazuo. 2021. *Klara und die Sonne*. München: Blessing.
- Jasanoff, Sheila. 2015. Future imperfect. Science, technology, and the imaginations of modernity. In *Dreamscapes of modernity. Sociotechnical imaginaries and the fabrication of power*, hrsg. Sheila Jasanoff und Sang-Hyun Kim, 1-33. Chicago: Chicago University Press.



- Johnson, Deborah und Mario Verdicchio. 2017. Reframing AI discourses. In *Minds + Machines* 27/4: 575–590
- Kelly, Kevin. 2017. The Myth of a Superhuman AI. <https://www.wired.com/2017/04/the-myth-of-a-superhuman-ai/>
- Keyes, Kenneth S., und Jacque Fresco. 1969. *Looking Forward*. New York: South Brunswick and A.S. Barnes.
- Klie, Thomas. 2014. Zwischen Verheißung und Schreckensszenario. Über Technik für ein selbständiges Leben im Alter. In *Dr. med. Mabuse* 1/2: 30–33.
- Knoblauch, Hubert. 2019. Gesellschaftstheorien, Gesellschaftsdiagnosen und Zeitdiagnosen. Über einige Gattungen der soziologischen Theoriebildung. In *Gegenwartsdiagnosen. Kulturelle Formen gesellschaftlicher Selbstproblematisierung in der Moderne*, hrsg. Thomas Alkemeyer, Niklolaus Buschmann und Thomas Etzemüller, 217–234. Bielefeld: Transcript.
- Koschorke, Albrecht. 2012. *Wahrheit und Erfindung*. Frankfurt a.M.: Fischer.
- Krämer, Hannes. 2019. Zukunftspraktiken. Praxeologische Formanalyse des Kommenden. In *Gegenwartsdiagnosen. Kulturelle Formen gesellschaftlicher Selbstproblematisierung in der Moderne*, hrsg. Thomas Alkemeyer, Nikolaus Buschmann und Thomas Etzemüller, 81–102. Bielefeld: Transcript.
- Krüger, Oliver. 2019. *Virtualität und Unsterblichkeit. Gott, Evolution und die Singularität im Post- und Transhumanismus*. Freiburg i.Br.: Rombach Wissenschaft.
- Kurzweil, Raymond. 2005. *The Singularity is near*. New York: Penguin.
- Lee, Kai-Fu, und Quifan Chen. 2021. *KI 2041. Zehn Zukunftsvisionen*. Frankfurt/New York: Campus.
- Levin, Christoph. 2013. Verheißung und Rechtfertigung. Gesammelte Studien zum Alten Testament II. Beihefte zur Zeitschrift für die alttestamentliche Wissenschaft (Band 431). Berlin: De Gruyter.
- Lovelock, James. 1991. *Das Gaia-Prinzip. Die Biographie unseres Planeten*. Zürich: Artemis & Winkler.
- Lovelock, James. 2020. *Novozän. Das kommende Zeitalter der Hyperintelligenz*. München: C.H. Beck.
- Luckmann, Thomas. 1967. *The Invisible Religion*. New York: MacMillan.
- Luhmann, Niklas. 1998. *Die Gesellschaft der Gesellschaft*. Frankfurt a.M.: Suhrkamp.
- Lünich, Marco. 2022. *Der Glaube an Big Data. Eine Analyse gesellschaftlicher Überzeugungen von Erkenntnis- und Nutzengewinnen aus digitalen Daten*. Wiesbaden: Springer VS.
- Marcuse, Herbert. 2004. *Der eindimensionale Mensch. Studien zur Ideologie der fortgeschrittenen Industriegesellschaft*. München: Dt. Taschenbuchverlag.
- Maresch, Rudolf. 2010. Die kommende Gesellschaft: Von der Wissensgesellschaft zur post-medialen Gesellschaft? In *Postmediale Wirklichkeiten aus interdisziplinären Perspektiven. Weitere Beiträge zur Zukunft der Medien*, hrsg. Stefan Selke und Ullrich Dittler, 9–25. Hannover: Heise.
- Maresch, Rudolf. 2001. Der Hype ist vorbei. Das Internet ist in der Realität angekommen. In *Medien und Erziehung (Schwerpunkt: Medienutopien – gestern und heute)*, hrsg. Merz. 45/4: 235–242.
- Matthes, Jörg. 2014. *Framing*. Baden-Baden: Nomos.

- Minevich, Mark. 2021. How the EU Is Leading The Way In AI Powered Social Innovation. <https://www.forbes.com/sites/markminevich/2021/01/18/how-the-eu-is-leading-the-way-in-ai-powered-social-innovation/?sh=2428b66a6bb7> (Letzter Zugriff 22.04.2022)
- Moravec, Hans. 1990. *Mind children. Der Wettlauf zwischen menschlicher und künstlicher Intelligenz*. Hamburg: Hoffmann und Campe.
- Mosco, Vincent. 2017. *Becoming Digital*. London: Emerald.
- Mozorov, Evgeny. 2013. *Smarte neue Welt. Digitale Technik und die Freiheit des Menschen*. München: Blessing.
- Müller-Funk, Wolfgang. 2008. *Die Kultur und ihre Narrative. Eine Einführung*. Wiesbaden: Springer.
- Müller, Michael. 2019. Developing the Future Stories of Companies: Open and Closed Story Worlds. In *Transforming Organizations. Narrative and Story-Based Approaches*, hrsg. Jacques Chlopczyk und Christine Erlach, 75–88. Schweiz: Springer Nature.
- Nagel, Alexander-Kenneth. 2021. *Corona und andere Weltuntergänge. Apokalyptische Krisenhermeneutik in der modernen Gesellschaft*. Bielefeld: Transcript.
- Nassehi, Armin. 2019. *Muster. Theorie der digitalen Gesellschaft*. München: C.H. Beck.
- Nassehi, Armin. 2021. *Unbehagen. Theorie der überforderten Gesellschaft*. München: C.H. Beck.
- Nozick, Robert. 2006. *Anarchie, Staat, Utopia*. München: Olzog.
- Paradise, Catherine, und Jean-Claude Thoening. 2013. Academic Institutions in Search of Quality: Local Orders and Global Standards. *Organization Studies* 34/2: 189–218.
- Parvianen, Jaana/Coeckelbergh, Marc (2021): The political choreography of the sophia robot. Beyond robot rights and citizenship to political performances for the social robotic market. In: *AI & Society*, 36/2, pp. 715–726.
- Passig, Kathrin. 2013. *Standardsituationen der Technikkritik*. Berlin: Suhrkamp.
- Passig, Kathrin. 2014. Neue Technologien, alte Reflexe. *Funkkorrespondenz*, 34, <https://www.medienkorrespondenz.de/leitartikel/artikel/neue-technologien-altenbsprelexe.html> (Letzter Zugriff 14.04.2022)
- Peters, Burkhard. 2007. *Der Sinn von Öffentlichkeit*. Frankfurt a.M.: Suhrkamp.
- Rohbeck, Johannes. 1993. *Technologische Urteilskraft. Zu einer Ethik des technischen Handels*. Frankfurt a.M.: Suhrkamp.
- Ropohl, Günter. 2009. *Allgemeine Technologie. Eine Systemtheorie der Technik*. Karlsruhe: Universitätsverlag Karlsruhe.
- Schätzing, Frank. 2018. *Die Tyrannei des Schmetterlings*. Köln: Kiepenheuer & Witsch.
- Scholtz, Christopher. 2008. *Alltag mit künstlichen Wesen. Theologische Implikationen eines Lebens mit subjektisimulierenden Maschinen am Beispiel des Unterhaltungsroboters Aibo*. Göttingen: Vandenhoeck & Ruprecht.
- Selke, Stefan. 2020. „If you see it, say it, and we’ll sort it...“ Shifting baselines und der neue Gesellschaftsvertrag im Zeitalter der Digitalisierung. In *Zusammenwirken menschlicher und künstlicher Intelligenz*, hrsg. Reinhold. Haux, et.al., 151–174. Wiesbaden: Springer VS.
- Selke, Stefan. 2022. *Wunschland. Unsere Odyssee zwischen irdischen Utopien und Weltraumkolonien*. Berlin: Ullstein.
- Simanowski, Roberto. 2020. *Todesalgorithmus. Das Dilemma der künstlichen Intelligenz*. Wien: Passagen.

- Simmel, Georg. 1995. *Aufsätze und Abhandlungen 1901–1908*. Gesamtausgabe Band 7. Frankfurt a.M.: Suhrkamp.
- Spiegel, Simon. 2019. *Bilder einer besseren Welt. Die Utopie im nicht-fiktionalen Film. Bilder einer besseren Welt. Die Utopie im nicht-fiktionalen Film*. Zürich: Schüren.
- Strasser, Anna, Wolfgang Sohst, und Ralf Stapelfeldt et al. 2021. *Künstliche Intelligenz. Die große Verheißung*. Berlin: MoMo Berlin.
- Sukulla, Freya. 2019. *Narrative Persuasion*. Baden-Baden: Nomos.
- Süssenguth, Florian. 2015. Die Organisation des digitalen Wandels. Zur Funktion von Digitalisierungsemantiken in Wirtschaft, Medien und Politik. In *Die Gesellschaft der Daten. Über die digitale Transformation der sozialen Ordnung*, hrsg. Florian Süssenguth, 93–121. Bielefeld: Transcript.
- Tapscott, Don. 1996. *Die digitale Religion – Verheißung einer vernetzten Welt. Die Folgen für Wirtschaft, Management und Gesellschaft*. Weisbaden: Gabler.
- Thiel, Peter. 2009. The Education of a Libertarian. *Cato Unbound. A Journal of Debate*, <https://www.cato-unbound.org/2009/04/13/peter-thiel/education-libertarian/> (Letzter Zugriff 22.04.2022)
- Uexküll, Johann Jacob von. 2018. *Umwelt und Innenwelt der Tiere* (Original: 1921). London: Forgotten Books Reprint.
- Van Dijck, José. 2014. Datafication, dataism and dataveillance. *Surveillance & Society*, 12, Datafication, dataism and dataveillance. 197–208.
- Van Lente, Harro und Aro Rip. 1998. Expectations in Technological Developments: An Example of Prospective Structures to be filled in by Agency. In *Getting New Technologies Together: Studies in Making Sociotechnical Order*, hrsg. Cornelis, Disco und Barend van der Meulen, 203–229. Berlin/New York: de Gruyter.
- Weizenbaum, Josef. 1977. *Die Macht der Computer und die Ohnmacht der Vernunft*. Frankfurt a.M.: Suhrkamp.
- Weyer, Ansgar. 2008. *Techniksoziologie. Genese, Gestaltung und Steuerung soziotechnischer Systeme*. Weinheim: Juventa.
- Wiener, Norbert. 1958. *Mensch und Menschmaschine*. Berlin: Ullstein.
- Willeke, Stefan. 1994. *Die Technokratiebewegung in Nordamerika und Deutschland zwischen den Weltkriegen*. Frankfurt a.M.: Peter Lang.
- Wolff, Marie-Luise. 2020. *Die Anbetung. Über eine Superideologie namens Digitalisierung*. Frankfurt a.M.: Westend.
- Zuboff, Shoshana. 2015. Big other. *Journal of Information Technology* 30: 75-89.
- Zuboff, Shoshana. 2018. *Das Zeitalter des Überwachungskapitalismus*. Frankfurt/New York: Campus.

---

# **Künstliche Intelligenz im Kontext von Macht, Herrschaft und Demokratie**



# Machine learning, political participation and the transformations of democratic self-determination

Jeanette Hofmann and Clara Iglesias Keller

## Abstract

This contribution addresses links between machine learning technologies and democracy with a focus on political participation. Democracy research often regards machine learning technologies as a threat, as these technologies could violate fundamental rights or replace democratic decision making. While raising important concerns, these approaches underestimate the malleability of digital technologies and their relationship to democracy. Our argument is that inherent to democratic practice we find a constant (re)negotiation of rights and institutions, in this case not least driven by the fact that machine learning technologies themselves are far from reaching maturity. The openness and negotiability of the relationship of AI and democracy is illustrated by three critical perspectives that hold special importance for political participation: algorithmic bias, automated decision-making and AI's epistemic dimension. By reflecting the changing condition of political organisation, current research can be productive and even performative in the sense of co-defining a shared understanding of new technologies and aim to set standards for their legitimate use.

---

J. Hofmann (✉) · C. Iglesias Keller  
Wissenschaftszentrum Berlin für Sozialforschung, Berlin, Deutschland  
E-Mail: [jeanette.hofmann@wzb.eu](mailto:jeanette.hofmann@wzb.eu)

C. Iglesias Keller  
E-Mail: [clara.keller@wzb.eu](mailto:clara.keller@wzb.eu)

---

**Keywords**

Künstliche Intelligenz • Maschinelles Lernen • Demokratie • Selbstbestimmung

---

## 1 Introduction

In the early 2000's, advances in machine learning technologies led to a new summer of artificial intelligence development and research. A cross-cutting technology employed in business, society and public administration, machine learning applications fuelled an array of multidisciplinary research that account for potential transformations of society at large. This article draws on such research to add a particular perspective on the interactions of artificial intelligence (AI) technologies and democracy with regard to concepts and practices of political self-determination.

Assessing the manners in which technologies contribute to socio-political change is a complicated task, for their impact on political organization is never straightforward. Rather than “a causal relationship of variables”, the way technologies affect political organization is defined by a “contingent constellation”, where an ensemble of conditions “enables possibilities for political action without determining them” (Hofmann 2019). If or how a technology appears to weaken or strengthen democratic practice, depends on the affordances it entails insofar as societal practice and political interests will put these affordances to use. Since technology is “inherent in the social enactment of democracy” (Djeffal 2019) its influence can be more or less obvious. It can feature prominently in one form of political participation or diffusely inform collective conditions and opportunities to participate. In all cases, assigning political results solely to the deployment of a certain technology proves insufficient, as this fails to account for other relevant factors.

This intricacy also informs discussions about possible effects of machine learning on democracies. Legal and political sciences research often approaches AI as a threat to democracy – either because its technical nature is assumed to restrict rights, or because it could potentially replace democratic decision-making. The shared perception of a threat runs through academic and policy discourses aiming at different dimensions of machine learning. For instance, some groups emphasize the epistemic, world-making dimension of machine learning, while others focus on its capacity to automate decision-making and the biases involved in their data and their output. These interpretations lessen the role other factors play in the

transformations of political organization, while also taking for granted both the tenets of democracy and an imagined final form of machine learning technologies.

This piece spotlights a perspective on the malleability of the interactions of AI and democracy based on the understanding that the practices and substantive conditions that enable democratic organization, as well as the form of AI technologies, are open for change. Democratic practice implicates a constant (re)negotiation of rights and institutions, which entails re-shaping both the scope of deliberation processes and fundamental rights – two dimensions of democratic self-governance that, as we elaborate, are critical to understanding AI and democracy. Approaching AI as a threat to presumably immutable democratic pillars – for instance, autonomy, privacy or deliberation – overlooks that our very understanding of the implications of such public values might change<sup>1</sup>. Furthermore, AI is far from reaching its maturity or even conceptual consistency<sup>2</sup>. This means that ongoing academic and policy discourses, as well as regulatory initiatives, can potentially influence future forms of AI technologies, which raises the stakes for current AI research. In this spirit, this article is to assemble social sciences literature on AI and democracy that captures this openness in its interplay, and thereby provide the basis for an understanding of the possibilities of realizing the democratic project with AI.

In the next section, we introduce the theoretical framework which captures variations among political institutions, fundamental rights, and machine learning technologies. Notably, we engage with Jürgen Habermas' concepts of public and private autonomy (1996) as co-original conditions for legitimization of the political process, which, as we argue, also shape each other in co-evolution. Subsequently, we analyse three distinctive dimensions of the current debate on AI that hold a crucial role for political or democratic participation. First, we attend to algorithmic biases, widely considered to be a threat to rights that may affect the conditions for political participation. Second, we look at the automation of decision-making, whose recent politicization illustrates the potential of policy

---

<sup>1</sup> As we demonstrate with examples throughout this text.

<sup>2</sup> Although or precisely because artificial intelligence (AI) is currently a much-discussed topic, it is difficult to define it. For this reason, many experts completely abandon the term and switch to abstractions such as “predictive technologies”, “agentic machines”, or “algorithmic systems” (Joyce et al. 2021, p. 2; see also Dignum 2022). While this leads some observers to question the very existence of AI, others prefer to reserve the term for “whatever we are doing next in computing” (Recker et al. 2021, p. 1435). Nevertheless, the vague terminology proves to be a problem if one wants to investigate the interactions of AI with society and democracy. To avoid ambiguity, we refer instead to algorithmic systems, learning algorithms, or machine learning.

discourses to influence technical features of machine learning through regulation, strengthening transparency and accountability. Third, and in relation to AI's epistemic dimension, we discuss the possibility of machine learning competing with democratic decision-making, including the risks ascribed to political agency. In conclusion, we offer a brief insight into conditions of democratic practice that may shift through its interactions with the discussed AI developments. We also note how specifically current research can be productive in co-defining a shared understanding of new technologies (and aim to set standards for their legitimate use).

---

## 2 Democracy and artificial intelligence

Machine learning's impacts on collective and individual self-determination are often approached through two perspectives. The first one refers to possible transformations of the "democratic interface", which refers to "the communication and organization processes that engage citizens with institutions of collective self-governance" (Bennett et al. 2018, p. 1657). Here, machine learning can mediate voting systems, the democratic public sphere (Thiel 2022) or even replace democratic decision-making (a prospect that we explore in the next section). On a second, individual-centred level of analysis, the expansion of machine learning technologies is shown to affect access to fundamental rights, as opportunities for individual development become increasingly mediated by algorithmic systems (Eubanks 2018). Algorithmic systems can potentially interfere, for instance, with freedom of movement, access to welfare and formal education – first conditions for one to exist as an engaged part of a collective. Explicitly or not, these rights are the focus of standard approaches which treat algorithmic decision-making as tools of mass surveillance or amplifiers of inequality, as they ultimately centre risks to the right to self-determination, to privacy or to equality.

The literature that identifies these risks often assesses machine learning systems against an imagined or anticipated future. More often than not, this future entails a myriad of negative potential fates for modern political organization in automated societies. Features that define these systems – notably, the replication of bias and opaque decision-making – are reckoned to relentlessly "create both specific injustices and broader states of exception" (McQuillan 2018). Threatening scenarios include their deployment as tools of population control and repression, (Savaget et al. 2019), exacerbation of social inequalities (Eubanks 2018) and further increases in concentration of power (Harari 2018). These dangers are often countered with potential benefits to inclusion and democratic



deliberation (Schippers 2020), imprinting a narrative of risks and opportunities on a debate that often blindsides the potential that machine learning technologies have in the long run to re-shape “democratic politics” (Thiel 2022). Ultimately, approaches as such envision a society structured by ubiquitous automated processes that mediate a variety of public and private interactions. In this picture, McQuillan (2018, p. 3) compares artificial intelligence to “a kind of dark matter, invisible in itself but pulling other systems into new shapes around it” (McQuillan 2018, p. 3). This idea implies a degree of undue influence where machine learning decisions steer us towards a grim path, often without us realizing.

Alternatively, the expansion of machine learning does not hold one possible future, but rather different possibilities that depend upon a constellation of circumstances. Of particular importance to our argument is the evolution of political organization that takes place in two interdependent spheres of democratic self-determination – the already mentioned political participation channels and the protection of fundamental rights.

In democratic theory, the conjunction and interaction of these two spheres is well captured by Jürgen Habermas’ concepts of private and public autonomy as co-original conditions for exercising political autonomy and participating in public discourse (Habermas 1996, p. 103). Private autonomy entails the preservation of the subjective rights that guarantee the condition for individual development, allowing people to take part in the forms of communication “necessary for the genesis of legitimate law” (Habermas 1996, p. 104). In turn, public autonomy refers to the ability to engage in the institutionalized communication and participation channels that allow citizens to express their will (Habermas et al. 2007, p. 300). Moreover, private and public autonomy presuppose each other. While fundamental rights can only be achieved and preserved through institutions of collective participation, without these same rights there would be no medium for the legal institutionalization of the conditions under which individuals “can make use of public autonomy when performing their role as citizens of the state” (Habermas et al. 2007, p. 301).

In the motion of evolving socio-political organizations, public and private autonomy also co-evolve. The constant practice of democratic deliberation through (re)negotiation of rights and institutions entails transformations that re-shape the scope of our guarantees and deliberation processes. Among other factors that contribute to this process, technology is a distinct factor, especially since “the notion of political self-determination constitutive for representative democracies is necessarily predicated on the existence of distribution media” (Hofmann 2019, p. 4). An example concerning algorithmic decision-making technologies is private platforms’ intermediation of online communications. While automated tools

have increasing influence over individual and collective speech through content moderation, the institutional means to assure legitimate exercise of the right to freedom of expression are subject to reconfiguration (Balkin 2014). For instance, amplification of individual speech via platform-centred algorithmic curation can enable new forms of speech conditioned to engaging the algorithm or justify means for speech regulation which are infrastructure focused, like transparency or algorithm explainability. Arguably, these particularities can redesign freedom of expression's very scope, as its prerogatives and prospects change. Transformations of other rights that represent democratic pillars, such as autonomy and the right to privacy, can inspire similar arguments (Block and Dickel 2020; Viljoen 2021).

In another example, the growing availability and use of psychosocial data currently threatens the democratic model of the autonomously acting subject (Block and Dickel 2020). Behavioral economics and experimental psychology are challenging the idea of the rationally acting subject on the basis of empirical evidence that aims to prove that human decision-making is not reasonable, but follows so-called "judgment heuristics", i.e. unconscious rules of action which – unsurprisingly – are often suboptimal (Kahneman 2003). Obviously, there are countless examples of the limited rationality of individual and collective action in history and the present; man-made disasters such as climate change bear witness to this. The crucial question, however, is whether, under the impression of increasing algorithmic forecasts and recommendations, we will reassess the obvious discrepancy between the ideal of rationally acting people and the evidence of sub-optimal outcomes. To put it bluntly, do we continue to be disappointment-proof and defend the possibility of democratic self-determination, or do we subscribe to the conviction (advocated by behavioural economics) that we should give algorithmic systems a greater say for our own benefit? As Hayles (2005, p. 132) puts it, nothing less than our understanding of human thought, action and perception, "how we understand what it means to be human" (cf. also Hildebrandt 2016) is up for discussion.

Moreover, machine learning technologies themselves are far from reaching their maturity. We should not underestimate these system's mutability, or the data that feeds them, as they will increasingly be subject to competing interests. There is no definite form in which machine learning imposes its logic on democratic governance. As Bucher (2018, p. 56) stresses, these technologies are neither "inherently good or bad, neutral or biased, but are made to appear in one way or the other, depending on a whole range of different factors, interests, stakeholders, strategies and, indeed, politics". Current academic and policy discourses, as well as regulatory initiatives, are productive in the sense that they co-define a shared

understanding of new technologies and aim to set standards for their legitimate use.

Even though variations as such are already accounted for in social sciences, they are mostly missing from outlooks on machine learning and democracy. Rather, the threats to self-determination often form a point of reference, from which diffuse results emerge, almost as unescapable destinations. This stiff perspective of both how political orders evolve, and of technology's flexibility, will eventually replicate, and possibly perpetuate, present contingencies of democratic practice in future forms of artificial intelligence. This means that alternative meanings for, and means to materialize democratic practice, might end up overlooked.

In the next section, we present some of these alternatives by looking at three distinctive dimensions of machine learning that are each discussed in social sciences literature. Algorithmic bias, automated decision-making, and the epistemic dimension of machine learning bear different possibilities for how political autonomy is exercised in public discourse.

---

### **3 Transformations in democratic self-determination and three dimensions of machine learning**

#### *Fundamental rights, bias and algorithmic decision-making*

Replication of bias in algorithmic decision-making is a central concern of how artificial intelligence may affect democracies. The idea that algorithms necessarily perpetuate the status quo is embedded in warnings of their capacity to “automate inequalities” (Eubanks 2018) and has earned them the title of “weapons of math destruction” (O’Neil 2016). Beyond engraving public debate with algorithms’ potential to increase socio-economic gaps that corrupt social cohesion, the acceleration of bias and discrimination is not far from an epitome for all sorts of social harm that could accrue from these technologies.

Empirical studies show that these concerns are not unfounded. The implementation of algorithmic decision-making systems in different national contexts, and with different purposes, has been shown to (re)produce bias and thus to affect fundamental rights. Cases in point include welfare distributions decisions, students’ and teachers’ evaluation systems and recidivism calculation software. Curbing the realization of subjective rights ultimately pushes specific groups away from political participation.

Outcomes as such are usually attributed to “the nature of machine learning” (McQuillan 2018), whereas the “learning” feature is “contingent upon continuous access to data” (Fourcade und Johns 2020, p. 808). Researchers draw attention to

machine learning's mechanism of drawing inferences from a predetermined dataset by extracting patterns that are replicated in subsequent automated processes. This is what contributes to the predictive power of these technologies, as they can "predict the target value for new and unknown input" (Fourcade and Johns 2020, p. 804). Yet, rather than being universal and objective, this process of knowledge production is "irrevocably entangled with specific computational mechanisms and the data used for the training" (McQuillan 2018). Whatever insight machine learning produces is, thus, confined to the origin, scope and criteria selection of imputed data. Moreover, the results of these inferences represent data patterns which in turn inform the same system's operations, influencing future outputs through "implied feedback loops" (Hildebrandt 2016, p. 5). Machine learning's reliance on continuous data or, "data hunger" (Fourcade and Johns 2020, p. 806), and the reproduction of patterns are at the core of arguments against, or hesitancy towards, their increasing deployment in democratic societies.

According to Fourcade and Johns (2020), machine learning and social learning share key similarities. People, as well as machines, produce insights based on an accumulation of data, which includes the information and experiences they access throughout "dynamic social exposure" (Fourcade and Johns 2020, p. 810). They cannot produce insights "ab initio independently of pre-existing data" (Fourcade and Johns 2020, p. 810); in its own terms, the human learning process is also confined to a dataset. This resonates with Hildebrandt's reminder that cognitive exploration, whether carried through machines or not, implies some degree of bias, without which "we cannot achieve any kind of perspective on the web of meaning we create while it creates us" (Hildebrandt 2016, p. 25)". As such, machine learning technologies also require their "web of meaning", so that their potential for the replication of bias can be approached as less of a contingency and more of an inherent affordance that "gives the characteristics that make this technology desirable" (Hildebrandt 2016, p. 25). As spelled out by Hannah Fry, algorithmic reproduction of bias is a "mathematical certain"; "the outcome [of algorithms] is biased because reality is biased" (Fry 2019, p. 68).

As the threat to fundamental rights urges the search for solutions that mitigate algorithmic biases, it also accentuates similar contingencies that are part of human-led decision-making. Fundamental rights are constantly disputed at different levels of the democratic institutional design. The epitome of these processes is judicial adjudication, performed by a non-majoritarian institution whose assignment is to interpret the covenants of constitutional democracies into rulings on the scope of individual and collective rights.

The rationality underlying a judge's decision-making process and the influence of personal preferences is a longstanding topic of investigation in legal philosophy (Freitas 2018; Waldron 2008).<sup>3</sup> Either by themselves or in deliberative fora, judges are ideally expected to interpret the law against the circumstances of concrete cases in a way that is reasonably detached from their personal preferences. Normative standards aside, the judicial interpreter appears, "in a portrait authenticated by leading-edge research, as a biased subject, constantly running the risk of inadvertently confirming their initial beliefs" (Freitas 2018, p. 59). Empirical studies have shown different ways in which bias affects judicial behaviour.<sup>4</sup>

There is no need for resignation in view of inexorable biases contingent on learning processes. Accepting them as a necessary feature of social life does not legitimize the imposition of personal (or in-group) ideologies over individuals or minorities. Beyond perverting the equality foundation of liberal democracies, restricting ones' possibilities for self-development ultimately compromises the principles over which liberal democracies were built. It deepens what De Sousa Santos calls the "symbolic burden" of the values of modernity – such as liberty, autonomy and justice – where they mean "increasingly disparate things to different social groups, so that excess of meaning translates into a paralysis of their effectiveness" (De Sousa Santos 2002, p. 15).

Legal hermeneutics has long recognized that personal beliefs, ideologies and experiences influence interpretation of the law, and scholars find the first necessary step to mitigating them is nurturing awareness (Freitas 2018, p. 64). Machine learning potentially broadens the spectrum for scientific investigation on how to raise awareness and mitigate or restructure these biases further<sup>5</sup>. In fact, the literature already flags the possibility of technology facilitating "the contestability of the bias, the iteration of testing our frames of reference against the world we need to navigate" (Hildebrandt 2016, p. 25).

---

<sup>3</sup> Legal philosophy and behavioural theories have explored an array of factors that judges weigh into this process of materializing what are, ultimately, ethical standards elected by the rule of law (Pereira 2016, p. 347), including hermeneutic choices, political implications, and institutional culture, beyond personal ideologies and bias (Campos Mello 2018; Pereira 2016; Freitas 2018).

<sup>4</sup> For instance, an experiment in the US Supreme Court led by Epstein et al. (2018, p. 239) found that justices who subscribe to a liberal ideology were more supportive of free-speech claims than conservative justices. Showing that bias also infiltrates collegiate deliberation, Cesário Alvim Gomes et al. (2018) documented how justices of the Brazilian Supreme Court were found more likely to disagree with rulings reported by female justices, in comparison to their male peers (Cesário Gomes Alvim, Werneck Arguelhes, und Nogueira 2018, p. 866).

<sup>5</sup> For hermeneutic techniques, see Freitas 2018.

Just as much as the law depends on “argument and contestation to come alive” (Hildebrandt 2016, p. 25) so do automated decisions, that also rely on a “procedural dimension of fairness [which] entails the ability to contest and seek effective redress against decisions made by AI systems and by the humans operating them” (European High Level Expert Group). To allow for contestation translates into the formulation of requirements, such as transparency or other features that increase explainability, that are widely discussed under recommendations or AI ethical principles (Fjeld et al. 2020). In this spirit, one could even go as far as to say that learning algorithms may also be developed and used in such a way that they “make the contingency of [political] decision-making clear” instead of hiding it (Rostalski and Thiel 2021, p. 62).

Algorithmic technical design can also be considered as way to mitigate bias in judicial decision-making. Legal and computational sciences have a tendency to keep a “human in the loop”(Binns 2020; Hildebrandt 2016) during the adoption of algorithmic decision-making in judicial processes. Binns (2020, p. 197) argues for this accessory role based on the notion that “individual justice can only be meaningfully served through human judgment”. Individual justice presumes that “each case needs to be assessed on its own merits”, a notion that clashes with the very logics of solely automated decisions (which produce inferences based on past input data). Yet, as acknowledged by the author, if algorithms can replace human’s ability to be inconsistent and discriminatory, providing consistency will not necessarily lead to anti-discriminatory results. Given the choice between human or machine bias, Binns privileges the “human in the loop” as the only capable sphere of applying individual justice (2020, p. 203). However, he also acknowledges some use of automated tools, given that “[i]n so far as algorithmic systems replace the human decision-maker’s ability to be inconsistent, discriminatory, or otherwise contrary to general justice, they may be seen as beneficial.” (2020, p. 204). As much as “the balancing between these two dimensions has been undertaken by human decision-makers” there is also a place for algorithmics in mitigating human bias, one that reaches beyond consistency and actually towards mitigating bias, for instance, incorporating anti-discrimination requirements within the statistical model (Binns 2020, p. 204–205).

Beyond enablers or deterrents of bias, algorithms bear different possibilities for the subjective conditions to participate in public discourse comprised in the idea of private autonomy. There are different ways in which they can be incorporated in decision-making processes that affect fundamental rights within the democratic design that are traditionally led by people. Among the factors that will shape their deployment in and out of these processes, policy discourses and

regulatory initiatives can play a pivotal role regarding how they are put to use and what sort of control mechanisms are in place.

### *Automation and the politicization of AI*

The recent politicization of automated prediction and decision-making illustrates how possibilities for the incorporation of AI in political processes can be influenced by policy discourses and regulation. Here, the openness of machine learning technologies gives such initiatives potential to not only regulate their use, but also to contribute to shape its features and social meaning.

As a result of the expansion of machine learning in business and science, various academic communities started pointing out democracy-relevant problems such as discrimination of vulnerable social groups, lack of fairness, transparency and data protection. The general goal of initiatives such as “Human Centered AI” (HCAI) or “Fairness, Accountability and Transparency in Machine Learning” (FAT/ML) has been to expand the objectives and quality criteria in the development of learning systems in so that potential impacts on people are already considered at an early stage. In other words, they seek to replace the “primarily technical worldview” (Selbst et al. 2018, p. 2) with a “sociotechnical view” (Selbst et al. 2018, p. 8). The approaches advancing such a socio-technical perspective are becoming more influential. Their objectives include giving users more control over algorithmic systems (cf. Shneiderman 2020), increasing the transparency and explainability of algorithmic calculations, or making the inherent value judgements of training datasets visible (Dignum 2022). Instead of accepting datasets as natural representations of the world, these approaches argue for understanding them as a “situated reflection of a particular worldview” (Scheuerman et al. 2021, p. 3).

Machine learning depicts the world in the form of probabilistic models that produce classifying and predictive statements: this is probably spam; this photo shows a cat. But also: this candidate is promising, and this offender will most likely (not) reoffend. However, the objectivity of algorithmic judgments and recommendations (and thus also its categorical difference to democratic thinking) on closer inspection turns out to be a dubious description, and one that is not necessarily shared by the programmers themselves. As Dignum (2022, p. 5) puts it, algorithmic systems represent “accumulation of choices”, namely a selection of possible decisions among various options.<sup>6</sup> Algorithms „propose things in

---

<sup>6</sup> There are options to choose from when selecting learning algorithms defining target variables, compiling training and test data as well as optimizing during training processes (cf. Domingos 2012, p. 79–80).

and about the world“, as Amoores (2020, p. 13) notes. In each single operation of an “apparently autonomous system resides a multiplicity of human and algorithmic judgments, assumptions, thresholds, and probabilities” (Amoores 2020, p. 64). Contrary to what the term “self-learning algorithms” suggests, algorithmic systems do not program themselves. The creative part of coding lies precisely in the formalization and production of predictability (Abebe et al. 2020, p. 255). A vivid example of this is “Perspective AI”, a program developed by Google that links learning algorithms with data sets for the automatic identification of toxic language. As Rieder and Skop (2021) describe in detail, the distinction between civility and toxicity used in the algorithmic moderation of online forums is by no means self-evident. It is culturally influenced, always potentially discriminatory and therefore requires continuous readjustments.

Due to the close interaction between datasets, human, and machine judgments, it makes sense to regard algorithmic systems less as foreign bodies external to democracy and society, but as part of a “complex fabric” (Rieder and Skop 2021, p. 12) or as “fellow players” (Koster 2021, p. 11, our translation). Much more relevant than the well-known question of the allocation of control between humans and machines is the understanding of the “subtle forms of semantic and moral engineering”, which come into play in the development of categories such as linguistic toxicity (Rieder and Skop 2021, p. 12). The authors’ central message is that we better grasp machine learning’s future societal significance once we shift our attention from algorithmic computation and self-determined decision-making to the intricate linkage between machine and social action.

Critical informatics explores the openness of algorithmic systems as a practical possibility for political intervention: “this creative process of designing a system is a potential opportunity for non-technical stakeholders, including responsible officials and members of the public, to hash out different ways in which a problem might be understood” (Abebe et al. 2020, p. 254–255). While algorithmic systems require precise definitions of objects and explications of rules, such unambiguous input and its mathematical operationalization can be tested once it is subject to criticism. “Contrary to the claims of many”, as also Bryson (2022) puts it, “AI and the digital revolution could create a boon for transparency and human understanding”, namely when the regulatory framework enforces such objectives of transparency.

In fact, the politicization of algorithmic systems has gained momentum in recent years, and the socio-technical perspective has held prominence in current policy debates. Civil society, in particular, plays an important role in highlighting the consequences of the expansion of algorithmic decision-making systems for



democracy and in politicizing not only the scope, but the very notion of algorithmic decision-making systems<sup>7</sup>. Current political engagement, for the most part, is concerned with the problematization of automated decision-making systems. In recent years, civil society initiatives have shed light on discrimination by algorithmic decision-making systems deployed by both the state and the private sector; they call for the disclosure of the source code of these systems, but they also develop instruments and procedures themselves with the pragmatic intention to achieve a socially desirable way of integrating learning algorithms into everyday life (for details see Edwards and Veale 2017). Two particularly relevant expressions of political participation in this field illustrate the process of politicization: criticism of algorithm-based decision-making systems and societal representativeness in national regulation processes.

In many countries, civil society initiatives are responding to the rapidly spreading use of learning algorithms in business and public administrations. In 2019 and 2020, AlgorithmWatch and Bertelsmann Stiftung researched case studies about successful, but above all problematic uses of algorithmic decision-making systems in the EU. One of the researchers' key questions was "how automation affects our institutions, our norms and values, and indeed democracy as a whole" (AlgorithmWatch and Bertelsmann Stiftung 2020, p. 6, our translation). The documented cases range from predictive policing, automated admission of students, use of facial recognition programmes and examination of social benefit claims to the detection of tax fraud.

Although the report explicitly meant to focus on positive, exemplary uses, cases where automated decision-making processes undermined fundamental rights of citizens prevailed. A central point of criticism of the report is that the automation procedures are usually introduced without preceding public discussion, their decision-making logic is concealed and the concrete objectives often remain unclear for outsiders (and possibly also for the administrations themselves). If supervisory regulations exist at all, they seem to often be ignored or their implementation does not keep pace with the use of algorithmic systems (AlgorithmWatch and Bertelsmann Stiftung 2020, p. 10–11; see also Ada Lovelace Institute et al. 2021). In light of this situation, civil society engagement assumes an informal supervisory function, by collecting evidence of discrimination or wrong decisions and engaging social pressure in favour of disclosure of

---

<sup>7</sup> In recent years, the increasing use of automated decision-making systems has not only brought existing network policy organizations onto the scene but has also led to a number of start-ups: NGOs such as the Ada Lovelace Institute (2018, UK), Algorithm watch (2017, Germany), AI Now (2017, USA) or Data and Society (2014, USA.).

decision-making systems and more effective control mechanisms. However, this civil society supervisory role is confronted with much resistance.

Because private and public organizations that deploy machine learning usually refuse access to algorithmic procedures and datasets, the repertoire of civil society actors is necessarily focused on the development of strategies to combat the existing information asymmetry. With regard to the public sector, freedom of information requests are an important tool to that end, one that is increasingly supplemented by legal proceedings. However, the chances of success of such procedures are often uncertain. Since claims for disclosure of algorithms are difficult to assert against the private sector, other ways of creating transparency and accountability become necessary. For example, NGOs use crowd-sourcing methods to trace the logics of algorithmic news curation or the display of advertising on digital platforms. With the help of browser add-ons, individual usage histories are collected and evaluated to gain insights into the logic of recommendation algorithms or the extent of personalization of information streams (cf. also Rieder and Hofmann 2020).

Moreover, AlgorithmWatch has been experimenting with the instrument of data donation for several years in order to explore the transparency-creating possibilities and limits of this tool in more detail. For example, the “OpenSCHUFA” project, initiated together with the Open Knowledge Foundation in 2018, was intended to shed light on the opaque credit ratings calculation of the largest German credit agency. However, the financing and data donation campaign launched for this purpose did not yield the expected results. Another data donation project to investigate YouTube’s recommendation algorithms in the run-up to the Bundestag election in 2021 also faced problems (see Kahlert 2022). Not surprisingly, the private sector is resisting attempts at external control. Their Terms of Service often exclude automatic analysis of usage data; (not only) Facebook has repeatedly taken action against unapproved analyses of its curation principles (Kayser-Bril 2021). Given these successful defence measures effective in individual cases, political debates aim to reach beyond individual cases to establish effective transparency rules (Keller and Drake 2022).

Another important field of activity for civil society is comments on regulatory initiatives. The European Commission’s proposal for an Artificial Intelligence Act (AIA) is currently the most important legislative activity in this field in Europe<sup>8</sup>. In 2021, 115 organizations issued a letter calling for human rights to be placed at the heart of artificial intelligence regulation (EDRI et al. 2021). In addition, the call criticizes the risk-based approach of the AIA act. In an expert report

---

<sup>8</sup> COM (2021) 206 final.

for the British Ada Lovelace Institute, Edwards (2022) specifies reasons for the rejection of the bill. According to Edwards' interpretation, the draft draws its "inspiration" from the product safety legislation and therefore treats the suppliers of AI systems the same way as manufacturers of conventional products such as dishwashers or toys. However, "AI is not a dishwasher and the way downstream deployers use it and adapt it, may be as significant as how it is originally built" (Edwards 2022, p. 6). Because algorithm-based systems tend to be the result of a division of labour of several suppliers and can also change significantly with their application contexts, the AIA's approach, namely "[h]olistically evaluating the risk of such a system in the abstract is an impossibility" (Edwards 2022, p. 6).

The critical statements of civil society organizations reveal specific strengths and weaknesses of the politicization of machine learning. The combination of academic expertise with political advocacy towards the protection of human rights and the fight against discrimination complements the procedures of legislative and executive regulatory bodies. The statements on the AIA draft are remarkable in this respect because they do not only criticize the regulatory approach in a professional manner but also manage to politicize the underlying understanding of artificial intelligence. The discourse on (supra-)state regulation reads as an example of "ontological politics" (Mol 1999). There is dispute precisely over an appropriate interpretation of the nature of learning systems: If the application risks of algorithmic systems cannot be conceptualized analogous to those of dishwashers, which models should regulatory measures be based on instead? In striving for regulatory containment, as this discussion also shows, specific definitions of AI are established and are subject to critical reflection in the course of politicization (cf. Bryson 2022).

#### *Political autonomy and the epistemic dimension*

A central question in the discourse on AI and democracy is whether machine learning could become a competitor for the democratic polity, and how such a possibility could be countered. For instance, König and Wenzelburger (2020, p. 2) raise the question whether algorithmic decision-making could produce better political solutions than liberal democracies and, following from that, if political actors could be at least partially replaced by machine learning in the longer term. This possibility is substantiated by pointing out specific weaknesses of representative democracy such as its incapability of effectively addressing the problems of climate change. Given the empirical evidence of its pending effects, the current "deliberation, democratic rivalry, and slow steering force of laws [...] [appear] at this point cumbersome and antiquated" (Rostalski and Thiel 2021, p. 62). The most drastic version of this view has been presented by Harari (2018): „Liberal

democracy and free-market capitalism see the individual as an autonomous agent constantly making choices about the world (...) But once we begin to count on AI to decide what to study, where to work, and whom to date or even marry (...) [d]emocratic elections and free markets might cease to make sense“.

The assumed tension between democracy and algorithmic systems has been ascribed to different methods of decision-making. The promise of algorithmic systems, as König and Wenzelburger (2020, p. 3) put it, is to solve tasks in a way that is far superior to human capabilities. The cognitive tasks referred to are to analyse data sets according to previously defined target parameters. Machine learning, as Mackenzie (2017, p. 7) puts it, is a historically specific form of calculation designed to find and optimize mathematical functions. Their precision in the field of statistical generalization (inference) lies beyond the capability of the human mind (Hildebrandt 2016, p. 10).

A central problem of the algorithmic mode of output production is, according to König and Wenzelburger (2020, p. 5), that there are no and cannot be any comparable standards for assessing the quality of political decisions. In the context of a pluralistic liberal democracy, it is precisely not possible to assume a fixed consensus on what good political decisions are. On the contrary, democratic societies are marked by conflicting views on what characterizes an adequate political decision. Particularly normative questions are not mathematically solvable, but “fundamentally about incommensurable beliefs and values” (König and Wenzelburger (2020, p. 5). While machine learning focuses on modeling a problem and formulating calculable solution parameters, the challenge for democratic societies is to be able to agree on common problem descriptions and objectives in the first place. The “open horizon of possibilities” (Makropoulos 2021) of societal development, against which democratic self-determination assumes its meaning and becomes realizable in to begin with, simultaneously forms the epistemic basis of the contestability of all political action and the “drama of decision-making” (Harari 2018).

Considering the ambiguity of social problems and the different values inherent in democratic processes, machine learning presents itself as a reductive mode of processing social reality that appears to be an “antagonist to assumptions of democratic plurality (Koster 2021, p. 5; our translation). Algorithmic systems carry claims to absoluteness and indisputable, objectively valid knowledge into the political realm where they encounter discursive struggles for worldviews and better arguments (Koster 2021, p. 14; our translation). From a democratic theory point of view, a normalization of algorithmic recommendations implies the danger that our understanding of democratic politics could change.

With AI at its core, political decision-making would no longer appear as the result of considering different possibilities of action, but as the execution of allegedly objective necessities. Uncertainty and the need for interpretation of all knowledge would be weakened, and political action increasingly appeared as an information problem: the more data available, the better the political decision. In this scenario of an increase in algorithmic recommendations for political action, there is a decrease in decisions that can still be attributed to democratic decision-making and consequently to collective self-determination. In this sense, machine learning systems do touch upon the exercise of individual and political autonomy as indispensable conditions for democracy, because “who no longer assumes the role of authorship for herself, will also have no interest in the concerns of the community” (Rostalski and Thiel 2012, p. 59–60; our translation).

While regulation policies are often understood as the result of institutionalized negotiation processes, in which different actors, interests and interpretations of the world meet and fight for agenda setting power, the discussion about machine learning largely lacks this deliberative perspective. Because the relationship between AI and democracy is predominantly perceived to be competitive, the “analysis of the concrete intersections of both logics falls short of a postulated gain of dominance of the technical logic” as Koster (2021, p. 6; our translation) puts it. Instead of empirically examining the encounter between democratic decision-making and algorithmic claims to truth in concrete social domains, the analysis of AI all too often settles for the vague analysis of threats.

Paradoxically, however, the idea of a rivalry between political and algorithmic decision-making logics runs the risk of succumbing to a reductionist worldview itself. This could be the case if the initial presumption of a competitive relationship systematically ignores other variants of the interaction between democracy and algorithmic systems. Instead of getting intellectually comfortable in a “submission or takeover narrative”, Koster (2021, p. 13; our translation) writes, “the exploration of an interrelationship, perhaps also their coexistence” would be desirable. The capacities of democracy should not be underestimated. Democratic public spheres are quite capable of critically reflecting on algorithmic systems, especially when and where they are integrated into social contexts of decision-making and thus are becoming practically effective.

Perhaps one comes closer to apprehend the future role of algorithmic systems for democracy if one considers them not as rivals, but as *coplayers* (“Mitspieler”, Koster 2021; Rieder and Skop 2021). After all, the spread of machine learning does not take the form of a hostile takeover; rather, it presupposes that a specific added value is attributed to algorithmic computability. Consequently, understanding the future role of machine learning it is a matter of identifying points of

interconnection where democratic processes appear to be open to the specific resources of algorithmic systems. From today's perspective, such points are evident in the reflection on practical problems of democracy, but especially in the area of regulatory policies and instruments.

An important resource of machine learning is data analysis. Just as population statistics played a central role in the discovery of society as an "independent realm of reality" (Heintz 2021, p. 149; our translation), large datasets could change the images of society and people that have emerged in the wake of statistics and shape the collective self-image to this day. When we view AI systems as knowledge technologies, data analytics functions not primarily as a analytical tool, but rather a medium through which social realities are both reimagined and rendered visible (Panagia 2021, p. 3).

For instance, the systematic collection and evaluation of behavioural data traces, contact networks, but also physical properties such as pulse rate, breathing or sleep rhythm, generate a previously unknown depth of focus in social observation, which at the same time is highly detailed, globally extensive and real-time. Ulbricht (2020) has proposed the term "demos scraping" for the administrative use of this data. New regulatory concepts such as nudging or behavioural interventions provide a foreshadowing of that and how algorithmic tools expand the repertoire of state (but also private) governance. Possibilities of microscopically accurate intervention arise, which can even work below the consciousness threshold of the demos.

From a political science point of view, the spread of machine learning expands the scope of the politically accessible and decidable.<sup>9</sup> Personality types, feelings and impulses can now be grasped as patterns and politically dealt with in new ways. With the spread of predictive analyses, the hypothetical consequences of action are also increasingly moving into the realm of what can be recognized and influenced. As climate change projections demonstrate, the long-term effects of political action or inaction can be calculated with increasing precision. As a result, the area of the political responsibility is likely to stretch further and further in the direction of a probabilistic future. Thus, it is not a far stretch to assume that individuals and groups become obliged to justify an algorithmically generated reality, inasmuch we deviate from its recommendations and nudges. At any rate, this is what Harari (2018) assumes, according to whom algorithmic systems may eventually make better decisions: "Imagine Anna Karenina taking out her smartphone and asking Siri whether she should stay married to Karenin or

---

<sup>9</sup> Hartmut Rosa (2020, p. 21) describes subjecting the world to control along four dimensions: recognizability, accessibility, controllability and usability.

elope with the dashing Count Vronsky”. (Harari 2018). Avoiding fatal missteps could actually prove to be a difficult temptation to resist.

The political dilemma that Yuval Harari unfolds in front of his audience lies in the costs that democratic societies would have to bear for probabilistic control over their individual and collective future. The price for a general recognition of algorithmic judgements would be the spread of tyranny, namely a shift of politically legitimized authority to networked machines, the concentration of data-based power in a few hands and in the gradual fading of autonomy, an indispensable prerequisite for democratic coexistence.

Although it is more likely that our understanding of autonomy will change in the course of the integration of algorithmic knowledge technologies into everyday life, this does not necessarily lead to a subversion of democracy. Regarding individual and collective self-determination, it will be important to develop a more realistic picture not only of the possibilities, but above all of the limits and weaknesses of machine learning. This involves abandoning the myth of the “thinking machine” and contextualising algorithmic judgments as specific readings alongside other readings of the world. Then machine learning can actually expand political discourse instead of damaging or attempting to replace it.

---

## 4 Conclusions

The question of how machine learning affects political participation inspired us to broaden the view and to start from the assumption that both democratic practices and machine learning technologies are constantly evolving. Rather than associating them through causal thinking, we aimed to advance a perspective that emphasises the openness of their development and thus the contingency in their relationship. Democratic rights, modes of political participation as well as communication media are changing in the course of their interaction. For example, political advocacy in the field of AI regulation aims to establish a right to explainability of automated decision-making and could thus modify the right to due process. It is also conceivable that data protection may in future include a right to encryption. The politicization of machine learning systems focuses on three features: the potential to replicate or even reinforce bias, the automation of decision-making and the epistemic, world-making dimension.

How do these changes come about? Our analysis suggests that the controversies surrounding AI and particularly biases of automated decision-making are implicitly introducing new standards for decisions, including decisions made by

people. Likewise, an unintended side-effect of comparing democratic opinion formation to algorithmic output could consist in new benchmarks for assessing the performance of democratic governance. Even our interpretation of fundamental rights, so far the most significant frame of reference for debating algorithmic systems, might change in light of debating and institutionalizing machine learning.

The categorical otherness, autonomy and power that public discourse likes to attribute to machine learning systems turns out to be a bogeyman that does not stand up to closer inspection. Algorithmic systems themselves are also open to change and regularly require external adjustments to meet the quality standards expected of them. Moreover, if algorithmic models are essentially probabilistic propositions about relations or patterns to be calculated, then, they do not differ fundamentally from other forms of knowledge production. Their results remain tied to the circumstances of their creation (the data sets, calculation methods and objectives used) and their scope is open to political challenge.

In this context, academic and policy discourses, as well regulatory initiatives, hold great potential to narrow or open our understanding of machine learning and put it to service of democratic values. By reflecting the changing condition of democratic practice, these debates can be productive in the sense of co-defining a shared understanding of AI and aim to set standards for its legitimate use.

---

## References

- Abebe, R., Barocas, S., Kleinberg, J., Levy, K., Raghavan, M., and Robinson, D. G. 2020. Roles for Computing in Social Change. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, arXiv:1912.04883v4 [cs.CY], 252–260. <https://doi.org/10.1145/3351095.3372871>
- Ada Lovelace Institute, AI Now Institute, and Open Government Partnership. 2021. *Algorithmic accountability for the public sector. Learning from the first wave of policy implementation*. <https://www.opengovpartnership.org/documents/algorithmic-accountability-public-sector/>
- AlgorithmWatch, and Bertelsmann Stiftung. 2020. *Automating Society Report 2020*. AlgorithmWatch, Bertelsmann Stiftung. [https://automatingsociety.algorithmwatch.org/wp-content/uploads/2021/01/Automating\\_Society\\_Report\\_2020\\_-\\_Deutsche\\_Ausgabe.pdf](https://automatingsociety.algorithmwatch.org/wp-content/uploads/2021/01/Automating_Society_Report_2020_-_Deutsche_Ausgabe.pdf)
- Amoore, L. 2020. *Cloud Ethics. Algorithms and the Attributes of Ourselves and Others*. Duke University Press.
- Balkin, J. M. 2014. Old School/New School Speech Regulation. Yale Law School, Public Law Research Paper, 491. <https://ssrn.com/abstract=2377526>



- Bennett, W. L., Segerberg, A., and Knüpfer, C. B. 2018. The democratic interface: Technology, political organization, and diverging patterns of electoral representation. *Information, Communication and Society*, 21(11), 1655–1680. <https://doi.org/10.1080/1369118X.2017.1348533>
- Binns, R. 2020. Human Judgment in algorithmic loops: Individual justice and automated decision-making. *Regulation and Governance*, 16(1), 197–211. <https://doi.org/10.1111/rego.12358>
- Block, K., and Dickel, S. 2020. Jenseits der Autonomie: Die De/Problematisierung des Subjekts in Zeiten der Digitalisierung. *BEHEMOTH – A Journal on Civilisation*, 13(1), 109–131. <https://doi.org/10.6094/behemoth.2020.13.1.1040>
- Bryson, J. J. 2022, March 2. Europe Is in Danger of Using the Wrong Definition of AI. *WIRED*. <https://www.wired.com/story/artificial-intelligence-regulation-european-union/>
- Bucher, T. 2018. *If ... Then: Algorithmic Power and Politics*. Oxford University Press.
- Campos Mello, P. P. 2018. A vida como ela é': comportamento estratégico nas cortes. *Revista Brasileira de Políticas Públicas* 8(2). <https://doi.org/10.5102/rbpp.v8i2.5481>
- Cesário Gomes Alvim, J., Werneck Arguelhes, D., and Nogueira, R. 2018. Gênero e comportamento judicial no supremo tribunal federal: Os ministros confiam menos em relatoras mulheres? *Revista Brasileira de Políticas Públicas* 8(2). <https://doi.org/10.5102/rbpp.v8i2.5326>
- Dignum, V. 2022. Relational Artificial Intelligence. [arXiv:2202.07446](https://arxiv.org/abs/2202.07446) [cs.CY]. <https://doi.org/10.48550/arXiv.2202.07446>
- Djeffal, C. 2019. AI, Democracy and the Law. In Andreas Sudmann (ed.), *The Democratization of Artificial Intelligence* (pp. 255–284). transcript Verlag. <https://doi.org/10.1515/9783839447192-016>
- Domingos, P. 2012. A few useful things to know about machine learning. *Communications of the ACM*, 55(10), 78–87. <https://doi.org/10.1145/2347736.2347755>
- EDRi, Access Now, epicenter.works, AlgorithmWatch, EDF, Bits of Freedom, Fair Trials, PICUM, ANEC, and Panoptikon Foundation. 2021. An EU Artificial Intelligence Act for Fundamental Rights. A Civil Society Statement. European Digital Rights. <https://edri.org/wp-content/uploads/2021/12/Political-statement-on-AI-Act.pdf>
- Edwards, L. 2022. Regulating AI in Europe: Four problems and four solutions. Ada Lovelace Institute. <https://www.adalovelaceinstitute.org/report/regulating-ai-in-europe/>
- Edwards, L., and Veale, M. 2017. *Slave to the Algorithm? Why a “Right to an Explanation” Is Probably Not the Remedy You Are Looking For* (SSRN Scholarly Paper Nr. 2972855). Social Science Research Network. <https://doi.org/10.2139/ssrn.2972855>
- Epstein, L., Parker, C. M., and Segal, J. A. 2018. Do Justices Defend the Speech They Hate? An Analysis of In-Group Bias on the US Supreme Court. *Journal of Law and Courts* 6(2): 237–262. <https://doi.org/10.1086/697118>
- Eubanks, V. 2018. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. St Martin's Press.
- Fjeld, J., Achten, N., Hilligoss, H., Nagy, A., and Srikumar, M. 2020. Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI. *SSRN Electronic Journal*. <https://doi.org/10.2139/ssrn.3518482>
- Fourcade, M., and Johns, F. 2020. Loops, ladders and links: The recursivity of social and machine learning. *Theory and Society*, 49(5), 803–832. <https://doi.org/10.1007/s1186-020-09409-x>

- Freitas, J. 2018. Interpretação Judicial: Exame Crítico Dos Vieses. *Revista Da AJUFERGS* 10(1a): 57–84.
- Fry, H. 2019. What Statistics Can and Can't Tell Us About Ourselves. *The New Yorker*, 2. September 2019. <https://www.newyorker.com/magazine/2019/09/09/what-statistics-can-and-cant-tell-us-about-ourselves>
- Habermas, J., Sperber, G. B., and Soethe, P. 2007. *A inclusão do outro: estudos de teoria política* (3rd ed., Humanística 3). Loyola.
- Habermas, J. 1996. *Between Facts and Norms: Contributions to a Discourse Theory of Law and Democracy*. Studies in Contemporary German Social Thought. Cambridge, Mass: MIT Press.
- Harari, Y. N. 2018. Why technology favors tyranny. *The Atlantic*. <https://www.theatlantic.com/magazine/archive/2018/10/yuval-noah-harari-technology-tyranny/568330/>
- Hayles, K. 2005. Computing the Human. *Theory Culture and Society*, 22(1), 131–151. <https://doi.org/https://doi.org/10.1177/0263276405048438>
- Heintz, B. 2021. Big Observation – Ein Vergleich moderner Beobachtungsformate am Beispiel von amtlicher Statistik und Recommendersystemen. *KZfSS Kölner Zeitschrift für Soziologie und Sozialpsychologie*, 73(1), 137–167. <https://doi.org/https://doi.org/10.1007/s11577-021-00744-0>
- High-Level Expert Group on Artificial Intelligence. 2018. *Ethics guidelines for trustworthy AI*. European Commission. <https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai>
- Hildebrandt, M. 2016. Law as Information in the Era of Data-Driven Agency. *The Modern Law Review*, 79(1), 1–30. <https://doi.org/10.1111/1468-2230.12165>
- Hofmann, J. 2019. Mediated Democracy – Linking Digital Technology to Political Agency. *Internet Policy Review* 8(2). <https://doi.org/10.14763/2019.2.1416>
- Joyce, K., Smith-Doerr, L., Alegria, S., Bell, S., Cruz, T., Hoffman, S. G., Noble, S. U., and Shestakofsky, B. 2021. Toward a Sociology of Artificial Intelligence: A Call for Research on Inequalities and Structural Change. *Socius*, 7, 1–11. <https://doi.org/10.1177/2378023121999581>
- Kahlert, P. 2022, Februar 28. YouTubes Wahl. *STS@ENS*. <https://medium.com/sts-ens/youtubes-wahl-ea9c4df1297e>
- Kahneman, D. 2003. Maps of Bounded Rationality: Psychology for Behavioral Economics. *The American Economic Review*, 93(5), 1449–1475.
- Kayser-Bril, N. 2021, August 13. Nach Drohungen von Facebook: AlgorithmWatch sieht sich gezwungen, Instagram-Forschungsprojekt einzustellen. *AlgorithmWatch*. <https://algorithmwatch.org/de/instagram-forschung-von-facebook-gestoppt/>
- Keller, P., and Drake, A. 2022, March 30. Proactive Contestation of AI Decision-making. *Verfassungsblog*. <https://verfassungsblog.de/roa-proactive-contestation-of-ai-decision-making/>
- König, P. D., and Wenzelburger, G. 2020. Opportunity for renewal or disruptive force? How artificial intelligence alters democratic politics. *Government Information Quarterly*, 37(3). <https://doi.org/10.1016/j.giq.2020.101489>
- Koster, A.-K. 2021. Das Ende des Politischen? Demokratische Politik und Künstliche Intelligenz. *Zeitschrift für Politikwissenschaft*. <https://doi.org/10.1007/s41358-021-00280-5>
- Mackenzie, A. 2017. *Machine Learners: Archaeology of a Data Practice*. Cambridge, MA: The MIT Press. <https://doi.org/https://doi.org/10.7551/mitpress/10302.001.0001>

- Makropoulos, M. 2021. Historische Kontingenzen und soziale Optimierung. In: Bubner, M. und Mesch, W. (Hrsg.) *Die Weltgeschichte – das Weltgericht? Stuttgarter Hegel-Kongreß 1999* 22: 75–90. Veröffentlichungen der Internationalen Hegel-Vereinigung. Stuttgart: Klett-Cotta. [https://www.researchgate.net/publication/356446646\\_Digital\\_Media\\_and\\_Democracy\\_A\\_Systematic\\_Review\\_of\\_Causal\\_and\\_Correlational\\_Evidence\\_Worldwide](https://www.researchgate.net/publication/356446646_Digital_Media_and_Democracy_A_Systematic_Review_of_Causal_and_Correlational_Evidence_Worldwide)
- McQuillan, D. 2018. “People’s Councils for Ethical Machine Learning.” *Social Media + Society* 4(2): 1–10. <https://doi.org/10.1177/2056305118768303>
- Mol, A. 1999. Ontological Politics. A Word and Some Questions. *The Sociological Review* 47(1): 74–89. <https://doi.org/10.1111/j.1467-954X.1999.tb03483.x>
- O’Neil, C. 2016. Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy. Allen Lane.
- Panagia, D. 2021. On the Possibilities of a Political Theory of Algorithms. *Political Theory*, 49(1), 109–133. <https://doi.org/10.1177/0090591720959853>
- Pereira, J. R. G. 2016. As Garantias Constitucionais Entre Utilidade e Substância. *Revista Brasileira de Direitos Fundamentais and Justiça* 10(35): 345–373. <https://doi.org/10.30899/dfj.v10i35.105>.
- Recker, J., Berente, N., Santanam, R., and Gu, B. 2021. Managing Artificial Intelligence. *MIS Quarterly*, 45, 1433–1450. <https://doi.org/10.25300/MISQ/2021/16274>
- Rieder, B., and Hofmann, J. 2020. Towards platform observability. *Internet Policy Review*, 9(4), 1–28. <https://doi.org/10.14763/2020.4.1535>
- Rieder, B., and Skop, Y. 2021. The fabrics of machine moderation: Studying the technical, normative, and organizational structure of Perspective API. *Big Data and Society*, 8(2), 1–16. <https://doi.org/https://doi.org/10.1177/205395172111046181>
- Rosa, H. 2020. *Unverfügbarkeit*. Wien. Residenzverlag.
- Rostalski, F., and Thiel, T. 2021. Künstliche Intelligenz als Herausforderung für demokratische Partizipation. In Interdisziplinäre Arbeitsgruppe „Verantwortung: Maschinelles Lernen und Künstliche Intelligenz“ der Berlin-Brandenburgischen Akademie der Wissenschaften (Hrsg.), *Verantwortungsvoller Einsatz von KI? Mit menschlicher Kompetenz!* (pp. 56–63). Berlin-Brandenburgische Akademie der Wissenschaften. <http://hdl.handle.net/10419/235149>
- de Sousa Santos, Boaventura. 2002. *Reinventar a democracia* (2nd ed.). Ed. Gradiva.
- Savaget, P., Chiarini, T. and Evans, S. 2019. Empowering Political Participation through Artificial Intelligence. *Science and Public Policy* 46(3): 369–380. <https://doi.org/https://doi.org/10.1093/scipol/scy064>
- Scheuerman, M. K., Denton, E., and Hanna, A. 2021. Do Datasets Have Politics? Disciplinary Values in Computer Vision Dataset Development. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2, 317), 1–37. <https://doi.org/10.1145/3476058>
- Schippers, B. 2020. Artificial Intelligence and Democratic Politics. *Political Insight*, 11(1): 32–35. <https://doi.org/10.1177/2041905820911746>.
- Selbst, A. D., Boyd, D., Friedler, S., Venkatasubramanian, S., and Janet Vertesi. 2018. *Fairness and Abstraction in Sociotechnical Systems* (SSRN Scholarly Paper Nr. 3265913). Social Science Research Network. <https://papers.ssrn.com/abstract=3265913>
- Shneiderman, B. 2020. Human-Centered Artificial Intelligence: Reliable, Safe and Trustworthy. *International Journal of Human-Computer Interaction*, 36(6), 495–504. <https://doi.org/10.1080/10447318.2020.1741118>

- Thiel, T. 2022, April 19. Artificial Intelligence and Democracy. Israel Public Policy Institute. <https://www.ippi.org.il/artificial-intelligence-and-democracy/>
- Ulbricht, L. 2020. Scraping the demos. Digitalization, web scraping and the democratic project. *Democratization*, 27(3), 426–442. <https://doi.org/10.1080/13510347.2020.1714595>
- Viljoen, S. 2021. A Relational Theory of Data Governance. *The Yale Law Journal*, 82.
- Waldron, J. 2008. Judges as Moral Reasoners. *International Journal of Constitutional Law*, 7(1): 2–24. <https://doi.org/10.1093/icon/mon035>.



# Hegemoniale Machtstruktur? Eine Kartierung der Akteure im aktuellen KI-Diskurs

Karin Hutflötz

## Zusammenfassung

Die maßgeblichen Akteure im aktuellen Diskurs um Künstliche Intelligenz (KI) lassen sich nach ihren verschiedenen Perspektiven und Interessen, aber auch hinsichtlich ihrer Reichweite in der gesellschaftlichen Rezeption und Wissensproduktion verorten. KI erweist sich dabei als ein von einseitig hegemonialen Machtverhältnissen geprägtes Tätigkeits- und Diskursfeld: Populärwissenschaftliche Akteure in Resonanz mit breitenwirksam rezipierten KI-Fiktionen in Film und Literatur bestimmen vorrangig die Leitbilder und relevanten Themen und Fragen. Sie prägen das gesellschaftliche Verständnis von KI mehr durch Verheißungen und Zukunftsnarrative als faktische technische Entwicklungen und wissenschaftliche Reflexion dies tun. Für ein vertieftes Verständnis der aktuellen Deutungs- und Rezeptions-Dynamiken von KI ist es daher nötig, einen Überblick über die unterschiedlichen Akteure im Diskurs und ihre jeweilige Agenda zu schaffen. Die Kartierung der Akteure verweist dabei auf Exklusionsdynamiken im KI-Diskurs: Frauen und Nicht-Weiße sind darin bislang stark unterrepräsentiert, werden durch eine tendenziell einseitige Daten-Erhebung (*coded bias*) rekursiv exkludiert und spielen weder in der Entwicklung von KI noch im Diskurs eine nennenswerte Rolle. Insofern trägt KI als maßgebliches Dispositiv der Zeit zur Verstärkung struktureller Ungleichheit bei, zum Beispiel im Sinne des *Gender Gap* und des strukturellen Rassismus– trotz gegenteiliger Einordnungen von KI-Technologie als genderneutral und tolerant.

---

K. Hutflötz (✉)

Institut für philosophische Bildungsforschung & Beratung, München, Deutschland

E-Mail: [karin.hutfloetz@ku.de](mailto:karin.hutfloetz@ku.de)

---

**Schlüsselwörter**

Akteure • Macht • Implizite Dynamiken • Populärwissenschaften •  
Exklusion • Ungleichheit • Gender gap • Racial bias

---

## 1 Methodik und inhaltliche Skizze

Ziel des vorliegenden Beitrags ist eine Kartierung der Akteure im Feld hinsichtlich ihrer Motive und leitenden Interessen, mit besonderem Blick auf hegemoniale<sup>1</sup> Strukturen und Machtdynamiken im Diskurs – vor allem diejenigen, die sich aktuell als wirkmächtig erweisen für den gesellschaftlichen Wandel durch KI. Diskursanalytisch lässt es sich von *einem* Diskurs sprechen (Foucault 1981, S. 74), auch und gerade eingedenk der Tatsache, dass der hier untersuchte „Diskurs“ aus vielfältigen Debatten auf verschiedenen Diskursebenen besteht und diverse Akteure, komplexe Interaktionen wie auch individuelle und kollektive Wechselwirkungen der Begriffe, Konzepte, Sprachspiele, Vorstellungen und Imaginationen in dem Themen-Feld umfasst. Ebenso wird hypothetisch *eine* Kartierung der Akteure vorgenommen und kann sinnvoll gerechtfertigt werden, da sich durchgehend bestimmte narrative und argumentative Muster, wirkmächtige Konzepte und Konventionen im Feld erkennen lassen, die die Wirklichkeit als konstruierte und damit auch als anders mögliche in ihrem real-utopischen Gehalt freilegen. Diskurs wird hier im Singular verwendet im Sinn von Foucault, insofern der Wandel der Gesellschaft im Dispositiv der KI-Debatten betrachtet wird, vor allem im Hinblick auf zentrale Akteure, deren Motive und die relationalen Machtverhältnisse. Im Resultat bezeichnet das Dispositiv einer Zeit die Herausbildung dessen, was jeweils als denkbar, vertretbar und vernünftig gilt und etabliert sich aus unpersönlichen und kontingenten Machtwirkungen. Als Diskurs in diesem Sinn werden alle Äußerungen oder institutionellen Ordnungen gefasst, die in einer Gesellschaft als selbstverständlich wahrgenommen werden und dadurch ihre «Realität» ausmachen. In dieser Arbeit wird damit ein Überblick gegeben über die (in welchem Sinn und Kontext?) als relevant inszenierten Akteure im aktuellen KI-Diskurs und in der massenmedialen Vermittlung von KI, sowie deren zukunfts-narrativen Verheißungen – hinsichtlich der Leitfragen: Wer bestimmt, wie KI als Reflexionsbegriff und Dispositiv gegenwärtiger

---

<sup>1</sup> Unter Hegemonie wird hier im Anschluss an Gramsci eine zivilgesellschaftliche Vorherrschaft oder Vormachtstellung verstanden, „die im Wesentlichen auf der Fähigkeit basiert, eigene Interessen als gesellschaftliche Allgemeininteressen zu definieren und durchzusetzen“ (Brand und Scherrer 2005, S. 6).

Debatten gebraucht und für welche Interessen KI instrumentalisiert wird? Die im Folgenden skizzierte Verortung der Akteure im Feld und ihre Verhältnisbestimmung im Hinblick auf Deutungsmacht und Pfadabhängigkeiten im Diskurs, erfolgt methodisch mittels einer repräsentativen Literatursichtung, transdisziplinär im Kontext einer kritische Diskursanalyse und im Rekurs auf philosophische Methoden – wie Begriffsklärung, das Freilegen expliziter wie impliziter Argumentationsmuster und die Differenzierung verschiedener Reflexions-, Rezeptions- und Sachebenen. Die vorliegende Untersuchung gliedert sich wie folgt: Im ersten Abschnitt geht es um relevante Akteure der KI, ihre Motive und Interessen, und wie sich die verschiedenen Akteure im populären wie fachwissenschaftlichen Diskurs nach den Ebenen, Motivlagen und Reichweiten ihrer gesellschaftlichen Rezeption und Wissensproduktion unterscheiden lassen. Daraus folgt der Vorschlag zu einer Kartierung der Akteure in der Fläche mittels dreier Kategorien: *Erstens* die zentralen Akteure, die sich im massenmedial vermittelten, populärwissenschaftlichen KI-Diskurs verorten lassen. *Zweitens* die bestimmenden Akteure im technischen, und ökonomischen Feld, die sich nochmal signifikant unterteilen lassen in die Gruppe der KI-Expert:innen, die in der technischen Praxis oder in der Entwicklung der KI tätig sind, und solche, die aus privatwirtschaftlicher Perspektive den Diskurs mitbestimmen. Und schließlich *drittens* die Akteure im geisteswissenschaftlichen Meta-Diskurs, die sich in gesellschaftstransformativer Perspektive zu den Zukunftsnarrativen und Verheißungen der KI verhalten und sich sowohl disziplinär als auch hinsichtlich ihrer Positionen und inhaltlichen Rezeption unterscheiden lassen. Dabei liegt der Fokus aus heuristischen Gründen in dem dritten Feld nur auf maßgeblichen Akteuren und aktuellen Referenzpositionen im deutschsprachigen Diskurs.

Im nachfolgenden Abschnitt erfolgt eine Topografie der Machtverhältnisse. Es werden die Hierarchien im Referenzuniversum der KI befragt, die sich im Zusammenspiel der Akteure bzw. der genannten Akteursgruppen und -typen zeigen – und zwar hinsichtlich der Leitfragen: Wer bestimmt die Sprache, die Bilder und Metaphern im Diskurs? Wer muss sich auf wen beziehen, wer hat die Deutungsmacht, auf welcher Ebene? Damit erfolgt neben der Kartierung in der Fläche eine topografische Kartierung, die zuerst 1) auf einen machstrukturellen Vergleich der verschiedenen Akteurs-Ebenen abhebt. Dann aber 2) zeigt, wo die blinden Flecke, die exkludierten und nicht gehörten Stimmen im Diskurs sind. Diese werden dezidiert in den Blick genommen und untersucht, wer nur eine unterrepräsentierte Rolle im KI-Diskurs spielt, wer in welcher Weise ausgeschlossen wird, sei es explizit oder implizit. Dabei erweist sich vor allem der faktische Ausschluss von Frauen, der *gender gap*, wie der von nicht-weißen, nicht-privilegierten Menschen als signifikant; sowohl auf Ebene

der Datenerhebung wie in der technischen Entwicklung oder auf Ebene der Nutzer-Agency. Analog spielen Frauen und nicht-Weiße kaum eine Rolle im Meta-Diskurs und in den gesellschaftspolitischen Debatten zur KI, Folgenabschätzung und gesellschaftlichem Wandel. Das ist umso bemerkenswerter, als sich KI dezidiert um gender-Neutralität und *political correctness* bemüht, ihrem Selbstverständnis und Image nach als diversitätssensibel geltendes Tätigkeits- und Wissensproduktionsfeld gilt.

---

## **2 Relevante Akteure der KI: eine Kartierung in der Fläche**

### **2.1 Die KI-Prophet\*innen diesseits des Höhenkamms**

Die hier skizzierte Kartierung der Akteure in der Fläche ergab sich diskursanalytisch anhand eines repräsentativen Samples von Texten, Artikeln und Monografien zumeist der deutschsprachigen Literatur, aber auch deren internationalen Referenzen – ausgewählt nach den Kriterien Aktualität und Reichweite der Rezeption, insofern ist diese repräsentativ für die Jetztzeit und das gesamtgesellschaftliche Verständnis von KI. Dazu gehörte sowohl Fachliteratur wie Artikel zu Themen und Debatten der KI in den großen Print-/Online-Medien (wie Spiegel, SZ, FAZ, NZZ), aber auch populärwissenschaftliche Bestseller zum Thema – wie „Homo Deus“ (Harari 2017) und „Die 4. Revolution“ (Floridi 2015) oder speziell im Deutschsprachigen z. B. „Künstliche Intelligenz und der Sinn des Lebens“ (Precht 2020) – exemplarische Werke und Autor:innen diesseits des Höhenkamms, die KI in der gesellschaftlichen Wahrnehmung prägen und mediale Massenpräsenz erfahren, aber aus Sicht der fachlichen Expert:innen nicht zum Kanon (insofern Höhenkamm) der jeweils fachlich anerkannten Literatur zählen. Doch – so die erste These – bestimmen sie im Umfeld der KI nicht nur das breite Verständnis und den Diskurs in der medialen Vermittlung, sondern auch maßgeblich den in der wissenschaftlichen Literatur und im technischen Bereich. Auf dieser Ebene werden die maßgeblichen Leitbegriffe und -bilder, Sprachspiele und gesellschaftlichen Deutungsmuster zu KI lanciert, die Themen und Zukunftsnarrative gesetzt, die im Diskurs als relevant wahr- und ernstgenommen werden. Die Diskursanalyse ergab dazu drei wesentliche Befunde: Erstens lässt sich das Feld der Akteure hier mit Bezug auf eine sichtbare Polarisierung und Radikalisierung in medialen Diskursen aufteilen in die Prediger\*innen utopischer Verheißung oder der apokalyptischen Warner:innen. Zweitens zeigt sich, dass es ein mehrfach vermittelter,



auffallend trivialisierter Diskurs ist, der geprägt wird von KI-visionären Leitfiguren – wie Bostrom (2018 „Super-IQ“ als Erlösungs- und Ermächtigungsfantasie), Harari (2017 „Homo deus“) oder Kurzweil (2000 „The Age of Spiritual Machines“; 2005 „The Singularity is near“) – und viele andere mehr, die seit Beginn des weltweiten KI-Hypes in den 1990er Jahren den bis heute diskursprägenden, prophetischen Ton anschlagen. Hier werden Spekulationen zum gesellschaftlichen Wandel durch KI emotionalisiert diskutiert und imaginiert; aber im Gestus rein rationaler Berichterstattung aus dem *inner circle* des Herrschaftswissens. Dass es ein solches faktisch gäbe oder suggeriert wird, zeigt sich einerseits daran, dass nur sehr Wenige KI hinsichtlich ihrer Funktionsweise in technischer Tiefe verstehen, insofern es ein hermetisch nur Wenigen zugängliches Wissen ist und daher machtvoll im Ausschluss. Andererseits wird Fachkompetenz in dem Feld mehr als in anderen oft nur implizit, wenn auch nicht absichtlich suggeriert. Harari (u. a. 2017) z. B. schreibt seine Zukunftsvisionen, als ob er Einblicke in geheime Quellen zur Geschichtsschreibung der Zukunft hätte im Tonfall früherer Prophetien. Massenmedial wie fachwissenschaftlich bezieht man sich dann darauf, als seien es Berichte zum Stand der Forschung. Dabei hat er mit KI oder auch nur Informatik fachlich nichts zu tun – ein Beispiel unter vielen. Die hier exemplarisch genannten Akteure sind publizistisch und ökonomisch nach liberalen Marketing-Maßstäben erfolgreiche Pop-Stars der digitalen Welt, die nicht notwendig technisch hinsichtlich der KI qualifiziert sind, was ihrer Autorität in dem Feld aber keinen Abbruch tut. Kurzweil, die weltweit meistrezipierte Gallionsfigur der KI-Verheißungen, ist Unternehmer und schon lange nicht mehr in der technischen Entwicklung tätig; gilt aber als der führende Experte auf dem Feld der KI. Die Kühnheit seiner Thesen und die Aggressivität seines Selbstmarketings genügen, um seiner Stimme maximal Gewicht im Feld der KI-Rezeption und massenmedial bestimmenden Akteure auch im deutschsprachigen Diskurs zu verleihen. Drittens ist bemerkenswert, dass der Diskurs in der populären, massenmedialen Rezeption und Vermittlung von KI keineswegs nur von individuellen Akteuren bestimmt wird, wie den genannten Predigern oder selbsternannten Experten nicht nur für technische Zukünfte und Entwicklung, sondern für die Transformation von Mensch und Gesellschaft allgemein, sondern ebenso von Kollektiv-Akteuren wie großen Unternehmen einerseits (z. B. Google, Facebook/Meta) und Vorbild- wie Leitbild-gebenden Multiplikatoren wie z. B. Elon Musk andererseits. Zudem werden die realen Erwartungen durch fiktive Vorstellungen und Vor-Bilder der KI in Science-Fiction-Filmen und deren Helden-Figuren (wie „Data“ aus Star Trek oder der „Terminator“) bestimmt, die sich ins kulturelle Gedächtnis transgenerational als Mythos und Versprechen technischer Entwicklung einschreiben. Wirkmächtig sind kollektive

Akteure aber auch durch journalistische Multiplikatoren, die zwar keine fachliche Deutungsmacht beanspruchen, aber in der – speziell in diesem Diskurs überwiegend emotional und normativ aufgeladenen statt sachlichen– Berichterstattung zur technisch-wissenschaftlichen KI-Forschung deren Ergebnisse oft trivialisiert und im Dienst reiner Marketing-Logik rhetorisch-reißerisch deutet, wie z. B. in dem FAZ-Artikel: „So smart kopieren Roboter mittlerweile Tiere“ (Goetz 2021). In der Studie, auf die dieser Bericht Bezug nimmt, war keine Rede von „Kopieren“, nur vom algorithmischen Simulieren bestimmter Bewegungsfunktionen. So werden das gesellschaftliche Verständnis und die kollektiven Annahmen und Debatten zur KI nachhaltig von fiktiven Vorstellungen geprägt, weniger von realen Forschungsergebnissen und dem technisch Machbaren.

## 2.2 Bestimmende Akteure im technischen und ökonomischen Feld

Die zahlenmäßig größte Gruppe der Akteure ist mit Abstand die in der konkreten Entwicklung von KI tätigen – und diese sind von dem bisher Gesagten deutlich zu unterscheiden. Wenn auch die leitenden Vorstellungen und imaginierten Zukünfte, die zentralen Bilder und Metaphern, mit denen wir gesamtgesellschaftlich über KI in allen möglichen Hinsichten sprechen, wie gesagt von den meist populärwissenschaftlichen Akteuren im erstgenannten Feld bestimmt werden, trifft das nicht auf die Akteursgruppe der Entwickler:innen und Expert:innen im Feld und deren oft gegenteilige Haltung zu, die sich aufgrund ihrer Fachkenntnisse und ihrem primären Erkenntnisinteresse als sehr viel zurückhaltender erweisen. Denn die technischen Akteure vom Fach sind kaum anfällig für einseitige Technoverklärung, noch apokalyptischen Bedrohungsszenarien anhängig und betrachten dies kritisch und distanziert, da sie in der Folge gegen Erwartungsenttäuschungen ankämpfen müssen.

Ebenso spielen KI-Prophetien in konkreten politischen und gesellschaftlichen Innovationsprozessen oft keine maßgebliche Rolle, wie qualitative Sozialforschung aktuell zeigt (Hirsch-Kreinsen 2022). Zwar spielt es eine Rolle im Marketing (z. B. nennt Sam Altmann bei der Markteinführung seinen ChatGPT „a smarter tool“ „to fully understand the universe“, Twitter 27. Juli 2022),<sup>2</sup> aber auch das ist von den hier beschriebenen Extremen weit entfernt. Doch welche

---

<sup>2</sup> Altmann, Sam [@sama] (27. Juli 2022) i suspect that to fully understand the universe, we need AGI in a similar way to needing a telescope to understand galaxies [zur Begründung der Markteinführung von ChatGPT] Twitter. <https://twitter.com/sama/status/1552355323045986311>.

gesellschaftlichen Deutungsmuster und Zukunftsnarrative prägen den Diskurs auf dieser Ebene, wer verortet sich wie hinsichtlich der gesellschaftlichen Rolle von KI und hinsichtlich der eigenen? Hier zeigt sich vor allem, dass es wiederum zu differenzieren gilt einmal zwischen den Expert:innen in der Praxis, die in der KI-Entwicklung tätig sind und aus der Sache heraus sprechen, und solchen, die aus privatwirtschaftlicher Perspektive den Diskurs mitbestimmen, aber dezidiert in neoliberaler Markt- und Marketing-Logik agieren, wie vor allem Elon Musk oder Mark Zuckerberg.

Letztere haben wirtschaftliche Interessen, sich in den Diskurs einzubringen. Die Motive, die diese Akteure kennzeichnen, erweisen sich einer Logik maximaler Selbstvermarktung verpflichtet und verheißen zur Image-Generierung überzogene Möglichkeiten von KI, da vielversprechende Zukunftsvisionen als Optionsscheine auf hohen (Spekulations-)Profit wirken. Kommen in der Literatur dagegen Akteure in technisch-fachlicher Perspektive und hinsichtlich konkreter KI-Entwicklung zu Wort, zeigt sich Gegenteiliges: in ihrer Motivlage meist sachlich, ist ihre Haltung primär von Technikverständnis und Verantwortlichkeit und Wissenschaftsethos geprägt (Brundage et al. 2018). Sie grenzen sich mit betonter Nüchternheit und kritischer Zurückhaltung ab von den Verheißungen von KI, vermeiden utopische wie dystopische Versprechen und sehen die Grenzen der Entwicklung viel deutlicher als im populären Diskurs üblich. Zugleich schätzen sie die Möglichkeiten zum politischen, ökonomischen und ethischen Missbrauch von KI höher ein (z. B. Weizenbaum 1978, 2001).

Bemerkenswert sind zudem zwei weitere Punkte: Erstens die erstaunlich hohen, im Grunde faktisch überfordernden moralischen (Selbst-)Ansprüche an ihre Arbeit, die eben zum Fortschritt oder Wohl der Gesellschaft und nicht zum Missbrauch reichen soll; was sie aber genau genommen nicht in der Hand haben. Und zweitens die scheinbar selbstverständliche Unterscheidung zwischen sich selbst als Schöpfer-Akteur und der Nutzer:in. Dabei gibt es kaum Bewusstsein dafür, dass sie selbst ja auch Nutzer:innen sind und in Abhängigkeiten; selbst dann, wenn ihnen die Codes und Algorithmen dahinter prinzipiell bekannt sind. Fragwürdig hier und als blinder Fleck im Diskurs erweist sich die Tatsache, dass der Betrachtungsgegenstand oder die kleinste Handlungseinheit in beiden Fällen der isolierte Akteur ist, das vermeintlich autonome Subjekt gegenüber KI als Eigen- oder Übermacht, nur im ersten Fall vorgestellt als Akteur mit Herrschaftswissen und im zweiten Fall als einer ohne. Dabei fehlt der Blick für Relationalität und ein Handlungskonzept, das kooperative Akteure ebenso auf dem Schirm hat und das Kollektiv als echten Kooperations-Akteur sehen kann (vgl. Loh 2018a, Heidbrink et al. 2017). Von daher wäre Verantwortung im Rahmen der KI nicht

im Kontext von Allmacht versus Ohnmacht zu denken, sondern als kommunikatives Handeln im Sinne von Arendt (2020) oder Habermas (2011) zu verstehen. Mit dem Ziel einer Verortung der Akteure dieser zweiten Kategorie ist zu beachten, dass es sich nicht nur um Individuen handelt (wie z. B. Elon Musk), sondern nicht minder um Kollektiv-Akteure, wie Google, der Meta-Konzern und ebenso die marktbeherrschenden Unternehmen aus China. Die großen Player in dem Feld, die sich als Hüter des technischen Herrschaftswissens und als Hauptakteure der Zukunftsgestaltung durch KI inszenieren, haben unverhohlenen kapitalistische, insofern ideologische Eigeninteressen, um den Diskurs in ihrem Sinn und in eigener Marketinglogik zu lenken. Hier ist derzeit ein sehr interessanter Rollenwandel zu beobachten: Gaben sie sich bis vor wenigen Jahren noch einer dezidiert apolitischen Haltung und reinen Wirtschaftsinteressen verpflichtet (vgl. Scherer 2010, S. 149 f.), verstehen sie sich zunehmend als politische Akteure mit sozialem Impact, die moralische und Demokratie-stützende oder -unterhöhrende Werte vertreten und stärkere Normierungen diesbezüglich vornehmen, als rechtlich sonst möglich wäre; auch ohne hierzu legitimiert zu sein. In diesem Sinn erfolgte z. B. der Ausschluss von Trump auf Facebook, da das Aufsichtsgremium im Mai 2021 entschied, das Konto des damaligen US-Präsidenten zu sperren. Insofern kommt der Firma ein moralischer Wächter-Status zu – analog zur Wächter-Rolle in der totalitären Gesellschaft in Platons Staat – den sie auch Image-bewusst für sich selbst beanspruchen. Umgekehrt mündet die politische Forderung nach gesellschaftlicher Verantwortung der übermächtigen KI-Unternehmen z. B. darin, dass die Whistleblowerin Haugen als ehemalige Facebook-Mitarbeiterin öffentlich institutionell Gehör findet. Sie trat im Oktober 2021 vor dem US-Senat auf und kurze Zeit später vor dem Ausschuss des Europaparlaments. Auch wenn das erst einmal für den Konzern als negative PR zu verbuchen war, so bestätigt es doch *via negationis* den Anspruch der KI-entwickelnden Konzerne auf ihre Führungsrolle beim gesellschaftlichen Wandel und der zukünftigen Gestaltung von Welt.

### **2.3 Geisteswissenschaftlicher Diskurs zu den Verheißungen der KI**

Ziel war es, die fachwissenschaftlichen Debatten und Autor\*innen im Umfeld der KI zumindest grob nach ihren Positionen zu differenzieren und deren Rezeptionsvarianten in gesellschaftstransformativer Perspektive zu verorten – und zwar im Hinblick auf ihre je nach Fachrichtung bevorzugten Fragestellungen, den methodischen Zugang und den Gegenstandsbereich ihrer Forschung. Dabei

lag der Fokus aus heuristischen Gründen auf dem deutschsprachigen Diskurs. Diese Einschränkung macht aber tatsächlich nur in den geisteswissenschaftlichen Meta-Diskursen Sinn, weil der Diskurs international verwoben ist.

Eine Kartierung hinsichtlich der verschiedenen Fachzugänge ergibt Folgendes: So legen Vertreter\*innen der Soziologie (z. B. Nassehi 2019; Baecker 2014; Baecker et al. 2021) den Fokus vor allem auf gesamtgesellschaftliche Muster und Dynamiken, die verursacht werden durch KI als Dispositiv des sozialen Wandels und untersuchen die Auswirkungen von KI auf politischer wie betrieblicher oder institutioneller Ebene (z. B. Welzer 2017). In der Philosophie bezieht man sich bevorzugt in subjektfokussierter Sicht auf die Bewusstseins- und Intelligenz-Debatten zu KI (z. B. Gabriel 2018; Metzinger und Windt 2016) und die ontologischen Voraussetzungen eines möglichen Singularitätseignisses, bringt die „Verteidigung des Menschen“ (Fuchs 2020) in Anschlag und diskutiert die Argumente des Trans- und Posthumanismus (Loh 2017, 2018a) und aus ethischer Warte mögliche Normverschiebungen durch KI (z. B. Nida-Rümelin und Weidenfeld 2020). Die Medien- und Kulturtheorie dagegen widmet sich vor allem der Transformation von Wissens- und Kommunikationsformen durch KI (z. B. Krämer 1998; Kittler 2013; Pias und Rieger 2016). Die geisteswissenschaftlichen Akteure, die die fachwissenschaftliche Reflexion von KI sowohl als Artefakt wie als Dispositiv der Zeit bestimmen, sind auch fast deckungsgleich mit den maßgeblichen Akteuren in den inzwischen zahlreichen Normierungs- und Regulierungsinstitutionen, die die ethischen und rechtlichen Rahmenbedingungen für den Wandel durch KI erheben und festlegen sollen. Permanent entstehen Policy-Papers, Studien und Ethik-Richtlinien in Fach-Kommissionen und -Gremien auf allen politischen Ebenen (siehe den Beitrag von Weber und Kleine in diesem Band). Dennoch erlangt der so rege wie fundiert geführte Meta-Diskurs zu KI und seine zentralen Akteure in gesamtgesellschaftlicher Perspektive und im Dispositiv der Zeit noch relativ wenig Sichtbarkeit, was sich z. B. an daran zeigt, dass weder Begrifflichkeiten noch argumentative Muster, die auf der Ebene entwickelt werden (so überzeugend sie in der Sache auch sein mögen), kaum Eingang finden in die Debatten. Die Reichweite für das Verständnis von KI ist auf dieser institutionellen Governance-Ebene ihrem eigenen Anspruch nach grundlegend, aber hinsichtlich der medialen Rezeption und dem Wirkungsgrad ihrer Beiträge eher gering einzuschätzen. Die Diskursanalyse ergab eine deutliche Diskrepanz zwischen Anspruch und Wirklichkeit.

### 3 Topografische Kartierung der Akteure

#### 3.1 Struktureller Vergleich und Zusammenspiel der Akteursebenen

Die Untersuchung in dieser topographischen Perspektive gibt Antworten zu den Fragen, wie die Machtverhältnisse und Hierarchien im Referenzuniversum der KI-Debatten und im gesellschaftlichen Diskurs derzeit sind: Wer muss sich auf wen beziehen – und wer hat die Deutungsmacht? In welchem (rekursiven?) Wechsel- und Abhängigkeitsverhältnis stehen die oben jeweils grob in drei Kategorien unterschiedenen drei Diskurs- und Akteur-Ebenen?

So gefragt, erweisen sich die populärwissenschaftlichen Akteure und Autor\*innen (z. B. Floridi 2015, Harari 2017, Precht 2020) in Kombination und wechselseitiger Resonanz mit massenmedial seit Dekaden wirkmächtig rezipierten KI-Visionen in Film und Science-Fiction-Literatur als maßgebliche Referenz sowohl für den geisteswissenschaftlichen wie für den breitenwirksamen Diskurs zu KI. Sie liefern und bestimmen nachhaltig (sogar resistent gegen Forschungsergebnisse und Argumente) die Leitbilder und Zukunftserwartungen und das Verständnis von KI wesentlich stärker und schneller, als technische Neuerungen oder wissenschaftliche Fach-Beiträge dies tun können.

Unter das Verdikt der hegemonialen Bestimmung fällt auch die Arbeit der genannten Normierungs- und Regulierungsinstitutionen, die allesamt fachlich hochkarätig mit Wissenschaftler\*innen und KI-Forscher:innen besetzt sind, aber deren fachliche Qualität und enormer Workload hinsichtlich ihrer gesellschaftlichen und medialen Rezeption und Wirkung, wie gesagt, in keinem Verhältnis stehen zur Stimme einzelner Akteure, die einen KI-Propheten-Status in dem Feld haben. Exemplarisch dafür der kürzlich veröffentlichte Tweet von Elon Musk z. B. zu seiner persönlichen Zukunftsangst in Bezug auf KI wurde weltweit als gewichtige Stimme rezipiert und avancierte zum Diskussionsanlass auf allen Ebenen. Kaum außerhalb der Fachcommunity wahrgenommen, noch inhaltlich rezipiert werden dagegen die Resultate fundierter Forschungspublikationen und zahlreicher Policy-Papers und Enquete-Kommissionen zu den Erwartungen an und Empfehlungen zum Umgang mit KI. Auch Ray Kurzweils bereits seit den 1990er Jahren wiederholte Prophezeiung ohne realen, technischen Anhalt, einer in wenigen Jahren sich einstellenden „Singularität“ (Kurzweil 2005, 2013) oder einer zur erwartenden „Super-Intelligenz“ der KI (Bostrom 2016, 2018) bestimmt den Diskurs inhaltlich in prägender Weise. Auch der geisteswissenschaftliche Meta-Diskurs in allen Fachbereichen kommt an der Frage nicht vorbei, wie sich eine „Super-Intelligenz“ oder „Singularität“ z. B. ethisch, sozial und politisch

auswirken würde. Auch die Frage, „wann übernehmen die Maschinen?“ (Mainzer 2019), ist seit Jahren ein zentrales Thema in fachwissenschaftlichen Debatten und ein Beispiel dafür, wie auch Fachexpert:innen die Aufmerksamkeitsökonomie bedienen. Deren Akteure müssen sich auf die populären KI-Propheten beziehen, nehmen deren Bilder, Konzepte und Narrative (und seien sie noch so fragwürdig) auf und müssen sich dazu verhalten – nicht umgekehrt. Das macht die einseitigen Machtverhältnisse im KI-Diskurs deutlich. Das hat zur Folge, dass nicht diejenigen, die den eigenen, rational-wissenschaftlichen Standards im Feld am meisten entsprechen, auch am meisten gehört und rezipiert werden – sondern die mit den eingängigsten Mythen von Wunsch- und Flucht-Metaphern (wie die Singularitätsidee, die an den Gedanken einer Schöpfung *ex nihilo* knüpft; oder die post- und transhumanistischen Visionen, die Unsterblichkeit versprechen) bzw. diejenigen mit den stärksten Versprechen möglicher Zukünfte: Das aber ist die Logik der Werbung. Wohl deshalb drehen sich die Debatten zu KI in hohem Maß nicht um technische Neuerungen und deren reale Auswirkungen auf Menschen und Gesellschaft, sondern um emotional aufgeladene Verheißungen, um utopische oder dystopische Zukunftserwartungen (siehe Selke 2023; Daub 2020) und religiöse Metaphern gesellschaftlicher Transformation. Auch Fach- und Forschungsbeiträge, die diese Motivlage von sich weisen, bewegen sich im emotionalen Spannungsfeld der Vorzüge und Abgründe durch KI und haben vor allem mahnenden oder eschatologischen Charakter (z. B. Russel et al. 2015; Alfonseca et al. 2021).

## **3.2 Wer spielt im KI-Diskurs keine Rolle, wer wird ausgeschlossen?**

### **3.2.1 Zum Gender Gap in der KI: Inwiefern Frauen keine Rolle spielen**

Für eine vollständige Kartierung im Hinblick auf bestehende Machtverhältnisse und Hierarchien innerhalb der Akteure und Akteursgruppen ist es auch erforderlich zu fragen, wo die blinden oder weißen Flecke auf der Karte sind: Wer ist strukturell und systemisch ausgeschlossen als Akteur – sei es explizit oder implizit? Auffallend ist dabei, in welchem Maß Frauen in der KI kaum oder keine Rolle spielen; weder in der Entwicklung als aktive Akteurinnen noch als passive Nutzerinnen und damit als zu berücksichtigender Teil der ‚Menschen‘, deren Daten und Erfahrungen, damit Denken und Handeln, in die Programmierung der KI und in die ‚lernenden‘ Algorithmen eingehen (Criado Perez 2020). Frauen spielen im Feld der KI fast nicht mit, kaum in der Entwicklung, erst in geringem

Maß als Nutzerinnen und sie sind in der Datenerhebung für die Programmierung der KI massiv unterrepräsentiert – damit nicht inkludiert als zu berücksichtigter Teil der „Menschen“, deren Daten und Erfahrungen, damit Denken und Handeln in die KI und die Algorithmen speisenden Datenbasis eingehen. Sie spielen auch kaum eine Rolle im Diskurs, partiell werden sie auch systematisch ignoriert und ausgegrenzt, und zwar, wie sich deutlich zeigt, sowohl im wissenschaftlichen wie im politisch-ökonomischen Diskurs, wie z. B. der aktuelle *Global Gender Gap Report* von März 2021<sup>3</sup> nahelegt. Unter den derzeit populären KI-Propheten findet sich auch weltweit fast keine Frau, die als maßgeblich wahrgenommene Stimme und Autorin in dem Feld Gehör fände.

Selbst in seriösen Studien (z. B. Ginzinger 2019) drängt sich die Frage auf: „Ist KI frauenfeindlich?“<sup>4</sup> Dazu lassen sich Argumente und Narrative der ProbleMLEUGNUNG und der Rechtfertigung überblickshaft skizzieren, ebenso Positionen im feministischen Diskurs.

Die Sichtung der Literatur auf allen genannten drei Ebenen zeigt, wie konsequent die maßgeblichen Akteure im Hauptdiskurs der KI-Deutung(-smacht) und Debatten sich nicht mit diesem Thema von struktureller Exklusion, fast Apartheid zugegendem Missverhältnis befassen; weder mit der selbstverständlich bekannten Ungleichverteilung von Zugang, Rollen und Gehör von Frauen im Feld, noch mit der lauter werdenden Kritik daran.

Das Faktum der Ungleichverteilung und -behandlung ist gut beforscht (vgl. die Studien zum „Digital Gender Gap“<sup>5</sup>), die zeigen, dass das speziell im Umfeld der KI – als Speerspitze der Innovation – noch stärker ist als im sonstigen Feld der IT. Dem ist auch zu entnehmen, wie dem Diskursverlauf allgemein, dass fast nur Frauen darüber sprechen, damit wiederum unter sich bleiben. Interessant auch die redundant vorgebrachten, argumentativen Muster und Praktiken der Rechtfertigung der Wenigen, die das Thema zumindest als Problem anerkennen und aufnehmen. Es wird entweder biologistisch legitimiert oder wegdiskutiert. Sei es, dass man Frauen eine eher emotionale Begabung und dadurch verwehrt Zugang zur KI abspricht oder ihnen eine geringe Affinität zu technisch Innovativen attestiert (Ginzinger 2019). Oder es wird in individualistischer Engführung und ideologischem Rückgriff auf individualistische Selbstverantwortung und Autonomie so argumentiert, dass das Betätigungs- und Diskurs-Feld der KI in maximaler Toleranz doch jedem Menschen offen stünde, der oder die

---

<sup>3</sup> <https://www.weforum.org/reports/global-gender-gap-report-2021/>

<sup>4</sup> [https://www.lbbw.de/artikelseite/maerkte-verstehen/frauen-in-der-techbranche\\_aaffrrvyd\\_d.html](https://www.lbbw.de/artikelseite/maerkte-verstehen/frauen-in-der-techbranche_aaffrrvyd_d.html).

<sup>5</sup> Digital Gender Gaps <https://www.digitalgendergaps.org>.



wolle. Damit inszenieren sich die Akteure der KI als unbeteiligt an der (Re-) Produktion von Ungleichheit, da diese scheinbar außerhalb des Tätigkeitsfelds der KI, ihrer Praxen und Codes und ihrer zugehörigen Diskurse verortet werden – ohne Bewusstsein dafür, welche immanenten Dynamiken der Exklusion hier offenbar am Werk sind und sich trotz allen Wissens darum gesellschaftlich rekursiv nur noch verschärfen.

Darauf verweist der feministische Diskurs im Umfeld der KI (s. Haraway 1995, Loh 2018b) seit Dekaden, wie die Vorreiterin feministischen Denkens zur KI, Donna Haraway, betont, ein auf hohem Niveau geführter Diskurs und inzwischen eigene Disziplin der Kulturwissenschaft, was fachlich und argumentativ als fundiert und anerkannt gilt, zumindest auf der Ebene der kulturwissenschaftlichen Meta-Diskurse, seiner Akteure und Rezeption. Wenn man in dem Feld zur Frage des *Gender Gaps* befragt wird, stimmt man Vielem zu und befürwortet die Argumente der Kritik und lobt die wissenschaftlichen Beiträge dazu. Dennoch bleiben diese nur randständige Teilnehmer:innen im KI-Diskurs. Kritik-äußernde Frauen und Expert:innen werden kaum in maßgeblichen Medien, Publikationen und Podien zitiert oder repräsentiert.

Die rassistisch wie sexistischen Tendenzen struktureller Exklusion aus dem Akteursfeld von KI, werden aber im öffentlichen Diskurs wie in fachlichen Debatten nur randständig diskutiert; und das trotz der bekannten Ungleichverteilungen von Zugang, Rollen und Gehör für Frauen im Feld und der im feministischen Kontext erfolgenden Kritik daran. Handlungsleitend müsste es dagegen sein, mehr Sichtbarkeit für dieses Missverhältnis zu schaffen, denn ein fehlendes Bewusstsein dafür, welche immanenten Dynamiken struktureller (nicht notwendig beabsichtigter) Exklusion hier am Werk sind, wird Ungleichheit rekursiv in Zukunft noch verschärfen (Meyer 2021, Jacobs 2021).

### **3.2.2 Coded Bias: Inwiefern KI strukturell rassistisch ist**

Systematische Ausschlüsse aus der Landschaft der Zukünfte und der KI-Akteure betreffen auch andere Gruppen, denn kaum weniger als die sexistische Exklusion ist zudem eine milieubedingte und rassistische feststellbar: nicht-weiße Menschen, vor allem nicht-technisch gebildete Männer, ist der Akteurs-Status ebenfalls weitgehend verwehrt. Die Sichtung der Literatur auf allen drei genannten Ebenen zeigt auch, wie wenig sich die maßgeblichen Akteure im KI-Diskurs mit diesen ungleich verteilten Zugangsmöglichkeiten und der damit einseitigen Deutungsmacht von KI befassen.

Insofern erweist sich der KI-Diskurs und seine maßgeblichen Akteure, die sich ihrem Anspruch nach als Inbegriff des Fortschritts verstehen und global leitbildgebend sind für die Zukunft, als ein weiteres Beispiel und Indiz für die

reaktionären Tendenzen der Zeit: Obwohl das Diskurs-Feld international hochvernetzt ist, ist es zugleich fast nur für ein bestimmtes Milieu-Segment zugänglich; nämlich für solvente, zur technischen Bildungselite zählenden Männer. Das gilt nicht nur für die westliche Welt. Auch wenn es in China, Afrika und allen anderen Erdteilen KI-Entwicklung und Diskursbeteiligung gibt, so ist diese dort wie sonst auch nachweislich von einem mehr oder weniger starken *gender gap* und einem indirekten und faktisch hochwirksamen *racial bias* geprägt. Auch wenn im Umfeld der KI meist keine Motive erkennbar sind, weshalb diskriminiert wird, und es auch nicht direkt geschieht, ist und bleibt der Zugang zu diesem neuen Technologiefeld, ebenso die Datenbasis für die Programmierung der KI, weitgehend exklusiv. Und das aus Gründen struktureller Bildungsbenachteiligung und kapitalistisch einseitig bedingter Datenerhebungen: In die Programmierung der KI gehen die Datenmengen ein, die automatisch durch die Häufigkeit des Nutzerverhaltens und durch kostenlose oder -günstige Preisgabe von personenbezogenen Daten zu erhalten sind, was eine Repräsentanz der Daten unter dem Primat der Gleichheit für die KI-Programmierung derzeit fast unmöglich macht (West et al. 2019). Im Hinblick auf die Akteure im Feld zeigt sich, inwiefern die KI bislang auf allen genannten Ebenen strukturell exkludierend ist und im hohen Maß weiße, männliche, ökonomisch-solvente Milieus beschränkt ist; was den Debatten zu Bias und Diskriminierung durch KI gewissen Anhalt gibt, auch wenn in der Entwicklung keine sexistische oder rassistische Absicht zu unterstellen ist, wie auch die aktuelle Stellungnahme „Mensch und Maschine“ des Deutschen Ethikrats betont: „Die Ursachen für Diskriminierung durch KI-Systeme sind vielfältig. Oft liegt bei deren Entwicklung keine unmittelbare Diskriminierungsabsicht vor. Stattdessen sind diskriminierende Effekte das Resultat gesellschaftlicher Realitäten oder Stereotype in Kombination mit technisch-methodischen Entscheidungen, wie beispielsweise der Wahl der Zielvariablen und Labels, der Auswahl der Trainingsdaten oder der verwendeten statistischen Analysemethoden.“ (Deutscher Ethikrat 2023, S. 282).

Durch die gender-, klassistisch- und rassistisch-einseitige Datenerhebung und die fehlenden Diskurszugänge verstärken sich durch die KI bereits vorhandene gesellschaftliche Ungleichheiten und erzeugen Pluralitätsverluste, die auch für die Zukunft als *coded bias* prägend sein werden. Und das in dem Maß, in dem es sich vor allem um strukturelle Ausschlüsse und Ignoranz, institutionalisierte Abwertung und Missachtung handelt in meist impliziter, nicht expliziter Form, worauf Kritik kaum fokussieren kann, geschweige denn sinnvoll adressieren. Wenn KI als ein global wirkmächtiger Diskurs solcher Art rekursiv-exkludierend ist, spielt das zurück auf die Gestaltung von Politik und die strukturelle Weichenstellungen

in der Gesellschaft. Denn diese werden zunehmend geprägt von digitaler Vermessung und KI-Anwendungen; z. B. als Finanzsteuerung und Bürokratisierung der Lebenswelt, oder als Transformation von Bildung und Arbeitswelt.

Diese erfolgen jetzt schon primär nach Maschinenkriterien – wie algorithmischer Effizienz und Leistungsfähigkeit, quantitativen Speicher- und Reproduktionskompetenzen – das kritisch betrachtet, eine Zurichtung des Menschen auf die in Zukunft KI-definierten Standardisierungen des Lebens bedeute. Das birgt strukturell die Gefahr einer Verantwortungsabgabe und -verstellung und einer übergesetzlichen „Herrschaft des Niemand“ (Arendt 2008, S. 743), wie die Theoretikerin Hannah Arendt bereits Mitte des 20. Jahrhunderts in ihren Studien zu Elementen und Ursprünge von Totalitarismen und Strukturen der Enthumanisierung dargelegt hat. Diese Gefahr für Forschungszwecke sichtbarer zu machen, d. h. auf gesellschaftswissenschaftlicher Reflexionsebene und für in der Entwicklung von KI, wäre der hier vorgeschlagene Weg, um totalitären Aspekte der Jetztzeit entschlossen zu begegnen. Und das bevor sie sich ungut spaltend und unter dem Radar des politischen Bewusstseins in den gesellschaftlichen Wandel durch KI als reaktionäre Tendenzen und Institutionalisierung bereits angelegter Ungleichheiten ein- und festschreiben können.

---

## Literatur

- Alfonseca, Manuel, Cebrian, Manuel, Anta, Antonio, Coviello, Lorenzo, Abeliuk, Andrés und Rahwan, Iyad. 2021. Superintelligence Cannot be Contained: Lessons from Computability Theory. In *Journal of Artificial Intelligence*, 70: 66–76.
- Arendt, Hannah. 2008. *Elemente und Ursprünge totaler Herrschaft*. München: Piper.
- Arendt, Hannah. 2020. *Vita Activa oder Vom tätigen Leben*. München: Piper.
- Baecker, Dirk. 2014. *Kulturkalkül*. Berlin: Merve.
- Baecker, Dirk und Elsholz, Uwe. 2021. *Parallele Welten der Digitalisierung im Betrieb*. Wiesbaden: Springer VS.
- Bostrom, Nick. 2016. *Superintelligenz. Szenarien einer kommenden Revolution*. Berlin: Suhrkamp Verlag.
- Bostrom, Nick. 2018. *Die Zukunft der Menschheit*. Berlin: Suhrkamp Verlag.
- Brand, Ulrich und Scherrer, Christoph. 2005. Contested Global Governance: Konkurrierende Formen und Inhalte globaler Regulierung. In *Globalisierung als politische Herausforderung*. Governance, Vol 3. ed. Behrens, Markus. VS Verlag für Sozialwissenschaften. [https://doi.org/10.1007/978-3-322-80888-2\\_6](https://doi.org/10.1007/978-3-322-80888-2_6)
- Brundage, Miles et al. 2018. The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation. <https://doi.org/10.17863/CAM.22520>
- Criado Perez, Caroline. 2020. *Invisible Women*. London: Vintage.
- Daub, Adrian. 2020. Was das Valley denken nennt. Frankfurt am Main: Suhrkamp Verlag.

- Deutscher Ethikrat. 2023. Stellungnahme „Mensch und Maschine“ [https:// www.ethikrat.org/fileadmin/Publikationen/Stellungnahmen/deutsch/stellungnahme-mensch-und-maschine.pdf](https://www.ethikrat.org/fileadmin/Publikationen/Stellungnahmen/deutsch/stellungnahme-mensch-und-maschine.pdf) (zuletzt zugegriffen 21.08.2023).
- Florida, Luciano. 2015. *Die vierte Revolution. Wie die Infosphäre unser Leben verändert*. Berlin: Suhrkamp Verlag.
- Foucault, Michel. 1981. *Archäologie des Wissens*. Frankfurt am Main: Suhrkamp Verlag.
- Fuchs, Thomas. 2020. *Verteidigung des Menschen. Grundfragen einer verkörperten Anthropologie*. Berlin: Suhrkamp Verlag.
- Gabriel, Markus. 2018. *Der Sinn des Denkens*. Berlin: Ullstein.
- Ginzinger, Larissa. 2019. LBBW Research Studie: Algorithmen: rassistisch und sexistisch? Ist KI frauenfeindlich? [https://www.lbbw.de/artikelseite/maerkte-verstehen/frauen-in-der-techbranche\\_aaffrrrvyd\\_d.html](https://www.lbbw.de/artikelseite/maerkte-verstehen/frauen-in-der-techbranche_aaffrrrvyd_d.html) (zuletzt zugegriffen 21.08.2023)
- Goetz, Ivo. 2021. Tierische Fortschritte. In *Frankfurter Allgemeine Quarterly*.
- Habermas, Jürgen. 2011. *Theorie des kommunikativen Handelns. Band I: Handlungsrationalität und gesellschaftliche Rationalisierung*. Berlin: Suhrkamp Verlag.
- Harari, Yuval Noah. 2017. *Homo Deus. Eine Geschichte von Morgen*. München: C.H. Beck.
- Haraway, Donna. 1995. Ein Manifest für Cyborgs. Feminismus im Streit mit den Technowissenschaften. In *Die Neuerfindung der Natur. Primaten, Cyborgs und Frauen*, hrsg. Hammer, Carmen und Stieß, Immanuel, 33–72 Frankfurt: Campus-Verlag.
- Heidbrink, Ludger; Langbehn, Claus und Loh, Janina. 2017. *Handbuch Verantwortung*. Wiesbaden: Springer VS.
- Hirsch-Kreinsen, Hartmut. 2022. Soziale Dynamik der Künstlichen Intelligenz. SFS Dortmund, *Beiträge aus der Forschung*, Bd. 209. Sozialforschungsstelle Dortmund (SFS): Technische Universität Dortmund. <https://sfs.sowi.tu-dortmund.de/beitraege-aus-der-forschung/>
- Jacobs, Hanna. 2021. *Coded Bias. Gesichtserkennungstechnologie als Träger und Vermittler von Diskriminierungen*. München: GRIN Verlag.
- Kittler, Friedrich A. 2013. Die Wahrheit der technischen Welt. Essays zur Genealogie der Gegenwart. hrsg. Gumbrecht, Hans Ulrich. Berlin: Suhrkamp.
- Krämer, Sybille. 1998. *Medien – Computer – Realität. Wirklichkeitsvorstellungen und Neue Medien*. Frankfurt am Main: Suhrkamp Verlag.
- Kurzweil, Ray. 2000. *The Age of Spiritual Machines. When Computers Exceed Human Intelligence*. London: Penguin Books
- Kurzweil, Ray. 2005. *The Singularity Is Near. When Humans Transcend Biology*. New York.
- Kurzweil, Ray. 2013. *Menschheit 2.0. Die Singularität naht*. Berlin: Lola Books.
- Loh, Janina. 2017. Posthumanistische Anthropologie zwischen Mensch und Maschine. In *Welt der Artefakte*. ed. Franz, Jürgen und Berr, Karsten, 213–224. Berlin: Frank & Timme.
- Loh, Janina. 2018a. Verantwortung und Roboterethik – ein kleiner Überblick. In *Zeitschrift zum Innovations- und Technikrecht* 1: 29–35.
- Loh, Janina. 2018b. *Trans- und Posthumanismus zur Einführung*. Hamburg: Junius Verlag.
- Mainzer, Klaus. 2019. *Künstliche Intelligenz – Wann übernehmen die Maschinen?* Berlin: Springer.
- Metzinger, Thomas und Windt, Jennifer M. 2016. *Open MIND – Philosophy and the Mind Sciences in the 21<sup>st</sup> Century*. Cambridge MA USA: The MIT Press.
- Meyer, Roland. 2021. *Gesichtserkennung. Digitale Bildkulturen*. Berlin: Klaus Wagenbach.
- Nassehi, Armin. 2019. *Muster. Theorie der digitalen Gesellschaft*. München: C.H. Beck.

- Nida-Rümelin, Julian und Weidenfeld, Nathalie. 2020. *Digitaler Humanismus. Eine Ethik für das Zeitalter der Künstlichen Intelligenz*. München: Dtv
- Pias, Claus und Rieger, Stefan. 2016. *Vollstes Verständnis. Utopien der Kommunikation*. Zürich-Berlin: diaphanes.
- Precht, Richard David. 2020. *Künstliche Intelligenz und der Sinn des Lebens*. München: Goldmann.
- Russell, Stuart, Dewey, Daniel und Tegmark, Max. 2015. Research Priorities for Robust and Beneficial Artificial Intelligence. In *AI Magazine*, 36(4), 105–114. <https://doi.org/10.1609/aimag.v36i4.2577>
- Scherer, Andreas Georg. 2010. Die multinationale Unternehmung als politischer Akteur in der globalisierten Welt. In *Unternehmensethik. Forschungsperspektiven zur Verhältnisbestimmung von Unternehmen und Gesellschaft*, hrsg. Schumann et al, 149–163. Marburg: Metropolis Verlag.
- Selke, Stefan. 2023. *Technik als Trost. Verheißungen Künstlicher Intelligenz*. Bielefeld: transcript.
- Weizenbaum, Joseph. 1978. *Die Macht der Computer und die Ohnmacht der Vernunft*. Frankfurt am Main: Suhrkamp Verlag.
- Weizenbaum, Joseph. 2001. *Computermacht und Gesellschaft*. Frankfurt am Main: Suhrkamp Verlag.
- Welzer, Harald. 2017. *Die smarte Diktatur. Der Angriff auf unsere Freiheit*. Frankfurt am Main: Suhrkamp Verlag.
- West, S.M., Whittaker, M. and Crawford, K. 2019. Discriminating Systems: Gender, Race and Power in AI. AI Now Institute. <https://ainowinstitute.org/discriminating-systems.html>.



# Künstliche Intelligenz und gesellschaftlicher Wandel – eine Herausforderung für demokratische Macht- und Herrschaftsverhältnisse

Peter Imbusch und Joris Steg

## Zusammenfassung

Künstliche Intelligenz (KI) stellt eine Schlüsseltechnologie des 21. Jahrhunderts dar, die einen weitreichenden Einfluss auf sämtliche Bereiche der Gesellschaft und das alltägliche Leben der Menschen hat und weiterhin haben wird. Technische bzw. technologische Revolutionen haben seit jeher auch das Potenzial, zu gesellschaftlichen Umwälzungen zu führen und die Machtverhältnisse und Herrschaftsstrukturen in einer Gesellschaft substantziell zu transformieren. Wird das Verhältnis zwischen Mensch, Gesellschaft und KI in den Blick genommen, bewegt sich die Debatte häufig zwischen naiver Technikeuphorie einerseits und fatalistischen bzw. dystopischen Auslöschungsphantasien andererseits. Die mit der Entwicklung und Anwendung von KI verbundenen Machtfragen bleiben jedoch zumeist seltsam unterbelichtet und untertheoretisiert. Der vorliegende Beitrag befasst sich deshalb mit dem Einfluss von KI auf gesellschaftliche Machtverhältnisse und Herrschaftsstrukturen und fragt danach, wie sich die Macht- und Herrschaftsverhältnisse in Wirtschaft, Politik und Gesellschaft durch den Einsatz von Künstlicher Intelligenz verändern und welche Folgen dies für die Demokratie hat. Dabei zeigt sich, dass KI von sich aus und per se keine egalisierende, nivellierende und demokratisierende Wirkung hat. Vielmehr kommt es tendenziell zu einer

---

P. Imbusch (✉) · J. Steg  
Bergische Universität Wuppertal, Institut für Soziologie, Wuppertal, Deutschland  
E-Mail: [pimbusch@uni-wuppertal.de](mailto:pimbusch@uni-wuppertal.de)

J. Steg  
E-Mail: [steg@uni-wuppertal.de](mailto:steg@uni-wuppertal.de)

Potenzierung, Zentralisierung, Monopolisierung und Stabilisierung der Macht- und Herrschaftsverhältnisse, also zu einem Ausbau der Macht und Herrschaftsbefugnisse bereits mächtiger Akteure – und damit auch zu einer Verstetigung und Verschärfung von Ungleichgewichten und Ungleichheiten.

---

**Schlüsselwörter**

Macht und Herrschaft • Künstliche Intelligenz • GAFAM • Demokratie • Ungleichheit • KI und Militär • Überwachung und Versicherheitlichung

---

## 1 Einleitung

Technische Entwicklungen und Revolutionen haben die Menschen von Anbeginn begleitet. So ist es nicht verwunderlich, dass die Bedeutung von Technik für die gesellschaftliche Entwicklung und das Verhältnis von Mensch, Technik und Gesellschaft seit langem ein Dauerthema der Sozialwissenschaften sind. Gleichwohl steht der Zusammenhang zwischen technischen Entwicklungen und Macht- und Herrschaftsverhältnissen nur selten im Fokus. Und das, obwohl Technik grundsätzlich weit mehr als ein bloßes Medium oder Artefakt ist, das allen Menschen in gleichem Maße zugutekommen würde. Technik und Technologien sind einerseits schon immer ein integraler Bestandteil menschlicher Gesellschaften gewesen und mittlerweile auch ein selbstverständlicher Teil der gesellschaftlichen Strukturen und individuellen Lebensweisen geworden. Mehr denn je sind Technik und Technologien inzwischen mit Kultur, Lebensstilen, habituellen Riten, normativen Orientierungen und gesellschaftlichen Leitbildern verschmolzen. Andererseits können die ungleichheitsgenerierenden Aspekte neuer Technologien ebenso wenig übersehen werden wie das Faktum, dass unterschiedliche Techniken und Technologien immer auch zentrale Quellen, Ressourcen, Instrumente und Mittel für die Ausübung von Macht und Herrschaft waren und sind. Dementsprechend haben sich mit jeder technischen Entwicklung und jedem technologischen Fortschritt stets auch neue gesellschaftliche Machtverhältnisse herausgebildet und häufig auch weitreichende Veränderungen der Herrschaftsstrukturen ergeben.

Mit den jüngsten technologischen Revolutionen – der fortschreitenden Digitalisierung und insbesondere der Entwicklung Künstlicher Intelligenz – stellt sich die Frage nach dem Verhältnis zwischen Mensch, Technik und Gesellschaft und die Frage nach dem Einfluss von Technik auf die gesellschaftlichen Macht- und Herrschaftsverhältnisse in neuer, veränderter und verschärfte Form.

Denn mit Digitalisierung und Künstlicher Intelligenz als vorläufig höchster Stufe technologischer Entwicklungen scheint es möglich zu werden, dass sich das Verhältnis zwischen Mensch und Technik umkehrt, dass die Technik den Menschen nicht mehr nur unterstützt und ergänzt, sondern beherrscht oder gar ersetzt. Das Risiko einer übergroßen technischen Überlegenheit und die damit einhergehenden Gefahren einer Ent- oder Übermenschlichung – weit jenseits ‚normaler‘ risikogesellschaftlicher Gefahren (Beck 1986) – scheinen erstmals Realität werden zu können.

Auch wenn die konkreten gesellschaftlichen Folgen und Auswirkungen einstweilen noch nicht endgültig absehbar und abzuschätzen sind, ist unbestritten, dass Digitalisierung und Künstliche Intelligenz die Schlüsseltechnologien des 21. Jahrhunderts sind. Sie haben bereits heute einen weitreichenden Einfluss auf sämtliche Bereiche der Gesellschaft und das alltägliche Leben eines jeden Einzelnen: Sie betreffen die veränderten Arbeitsformen und -prozesse in der Wirtschaft, die Determinierung unserer individuellen Lebens- und Verhaltensweisen – nicht zuletzt durch die großen Digitalkonzerne –, die Überwachungs- und Kontrollmöglichkeiten der Gesellschaftsmitglieder durch Big Data, die Veränderungen der Kriegsführung durch die sog. ‚revolutions in military affairs‘, aber auch den Umgang, die Nutzung und Verwertung von (persönlichen) Daten, die nicht zuletzt neuartige ethische und normative Fragen aufwerfen.

Die wissenschaftlichen, politischen und medialen Diskussionen um Künstliche Intelligenz kreisen dabei häufig zwischen zwei diametral entgegengesetzten Polen: Die eine Seite (über)betont die Möglichkeiten, Chancen und Potenziale von KI (Stichwort ‚overtrust‘), während die andere Seite die Grenzen, Gefahren, Risiken und schädlichen Konsequenzen von KI (über)betont (Stichwort ‚mistrust‘). So oszilliert die Debatte um KI nicht selten zwischen naiver bzw. utopischer Technikeuphorie einerseits und fatalistischen bzw. dystopischen Auslöschungsphantasien andererseits (für einen exemplarischen Überblick über pessimistische und optimistische Zukunftsszenarien Tegmark 2019; vgl. auch Bostrom 2014). Die mit der Entwicklung und Anwendung von Künstlicher Intelligenz verbundenen gesellschaftlichen Machtfragen bleiben – abseits lobenswerter Ausnahmen (Nemitz und Pfeffer 2020; vgl. auch Pfeffer 2021) – zumeist allerdings seltsam unterbelichtet und untertheoretisiert.

Im folgenden Beitrag soll daher zum einen danach gefragt werden, wie Künstliche Intelligenz in gesellschaftliche Machtverhältnisse eingebettet ist und in welcher Form KI als Machtmittel oder als Herrschaftsinstrument fungiert; zum anderen soll aber auch der Frage nachgegangen werden, wie KI die Macht- und Herrschaftsverhältnisse in Wirtschaft, Politik und Gesellschaft verändert. Daran schließt sich aus theoretisch-analytischer Perspektive die Frage an, inwieweit



zur Erfassung dieser Prozesse ein neues Verständnis von Macht und Herrschaft nötig ist, denn je nach Fassung und Differenzierung von KI verliert diese ihren reinen Objektstatus und gewinnt einen zunächst zugeschriebenen Subjektstatus. Wenn über den Zusammenhang von Macht, Herrschaft und KI nachgedacht wird, muss daher auch die Frage, ob und inwieweit sich durch den Einfluss von KI die Semantiken von Macht und Herrschaft verändern, thematisiert werden. Es gilt also mithin, einer doppelten Transformation von Macht und Herrschaft nachzuspüren: der Transformation von realen Machtstrukturen und Herrschaftsverhältnissen in Wirtschaft, Politik und Gesellschaft durch den Einsatz von KI sowie einer möglicherweise notwendigen Transformation der Verständnisse von Macht und Herrschaft selbst. Auf dieser Basis können dann die gegenwärtigen und zukünftigen Herausforderungen durch KI nicht nur besser verstanden, sondern auch die Potenziale und Risiken des Einsatzes von KI angemessener eingeschätzt und bewertet werden. Ein besonderer Fokus soll dabei auf der Analyse der Auswirkungen von KI für demokratische Macht- und Herrschaftsverhältnisse liegen.

Dazu werden im zweiten Kapitel zunächst die theoretisch-begrifflichen Grundlagen gelegt, indem wir die Konzepte Macht, Herrschaft und Künstliche Intelligenz explizieren und definieren. Im dritten Kapitel werden die aufgeworfenen Machtfragen und Herrschaftsaspekte dann direkt auf Künstliche Intelligenz bezogen. Dazu werden wir ein erweitertes Machtverständnis zugrunde zu legen, welches neben der direkten Macht zwischen Menschen auch die durch Technik vermittelte Macht von Menschen über Menschen und vor allem die in einer neuen Technik wie KI inkorporierten und sich materialisierenden Machtformen berücksichtigt. Im vierten Kapitel analysieren wir die durch KI induzierten Auswirkungen und Folgen für Macht- und Herrschaftsverhältnisse in Wirtschaft, Politik und Gesellschaft insbesondere in Bezug auf ihre Konsequenzen für die Demokratie und die offene Gesellschaft. Ausgehend von den Chancen und Risiken sowie von den Potenzialen und Gefahren von Künstlicher Intelligenz resümieren wir im fünften Kapitel abschließend die doppelte Transformation von Macht und Herrschaft durch KI.

---

## **2 Begriffliche Grundlagen: Macht, Herrschaft und Künstliche Intelligenz**

Da die Begriffe und Konzeptionen von Macht, Herrschaft und Künstlicher Intelligenz keine einheitliche, eindeutige und allgemeingültige Definition besitzen, ist es notwendig, diese drei Begriffe wenigstens in ihren grundlegenden Dimensionen

auszuleuchten, damit klar wird, was mit diesen Ausdrücken eigentlich gemeint ist und wie sie definitorisch bestimmt werden können. Denn sowohl hinsichtlich der Begriffe Macht und Herrschaft als auch in Bezug auf die KI sind die Bedeutungsgehalte sowie die mit ihnen verbundenen Implikationen teils hochgradig umstritten; sie stellen „essentially contested concepts“ (Gallie 1956) dar.

So existieren vielfältige, bisweilen äußerst konträre Verständnisse, konzeptionelle Ansätze sowie Theorien von Macht und Herrschaft, die aus unterschiedlichen theoretischen Perspektiven argumentieren und mit unterschiedlicher Reichweite und differierenden Erklärungsansprüchen ausgestattet sind (Imbusch 2012). Macht und Herrschaft sind zunächst konstitutive Grundtatbestände und ubiquitäre Phänomene menschlicher Gesellschaften. „Wer Macht sagt, sagt auch Gesellschaft, doch wer Gesellschaft sagt, sagt auch Macht“ (Sofsky und Paris 1994, S. 9). Macht und Herrschaft müssen als anthropologische Konstanten, als permanente Elemente der *conditio humana* gesehen werden. Sie selbst müssen als soziales Verhältnis und als relationale Kategorien verstanden werden.

Macht tritt in allen gesellschaftlichen Bereichen und in vielen Formen und Facetten auf. Macht ist „normal und alltäglich, nichtsdestotrotz aber keine einfache Sache“ (Paris 1998, S. 8). Denn bereits die Fragen, wie Macht zu bewerten ist, ob Macht an und für sich gut oder schlecht sei, ob Macht korrumpiere und ob der Mensch ein natürliches Streben nach Macht habe, sind hochgradig umstritten und werden insbesondere in der Philosophie seit jeher kontrovers diskutiert (Anter 2020; Han 2005). Macht ist ein genuin soziales Phänomen, welches man trotz mancher Voreingenommenheit zunächst als normativ offen oder indifferent betrachten sollte. Denn Macht hat auch mit ‚machen‘ bzw. ‚machen können‘ zu tun, im Machtbegriff steckt also nicht nur das ‚power over‘, sondern immer auch das ‚power to‘ (Pansardi und Bindi 2021). Darin kommt die dem Machtbegriff innewohnende Ambivalenz gut zum Ausdruck. Gleichwohl darf nicht übersehen werden, dass Macht sich nicht in einem handlungstheoretischen Zugriff erschöpft, sondern dass daneben noch strukturtheoretische Aspekte (etwa aus der neomarxistischen Theorierichtung oder der Kritischen Theorie) oder intermediäre Konzeptionen von Macht (Giddens 1992) eine wichtige Rolle spielen.

Herrschaft ist hingegen ein Spezialfall der Macht, sie kann grundlegend als verdichtete oder institutionalisierte Macht (Popitz 1992, S. 232; vgl. auch Anter 2020, S. 65) betrachtet werden, die dauerhafter ausgeübt wird und an legitimationsbedürftige Institutionen rückgebunden ist. Herrschaft besitzt in der Regel ebenfalls einen Doppelcharakter: Zum einen besitzt sie eine Sicherheits- und Ordnungsfunktion (Maurer 2004), zum anderen ist sie aber durch den Besitz staatlicher Monopolfunktionen zugleich ein Zwangsinstrument (Imbusch 2011).

Die soziologische Analyse von Macht und Herrschaft muss also die sozialen und anthropologischen Grundlagen, die gesellschaftlichen Einbettungen, die spezifischen Figurationen und Arrangements der Macht sowie die sozialen Auswirkungen und die Legitimität von Macht- und Herrschaftsverhältnissen in den Blick nehmen. Denn ihr geht es ja darum, kritisch zu analysieren, wann, warum, auf welche Weise, mit welchen Mitteln und aufgrund welcher Fähigkeiten Macht und Herrschaft ausgeübt werden kann, was die gesellschaftlichen Grundlagen und Quellen von Macht und Herrschaft sind, warum Menschen Macht und Herrschaft ausüben oder wechselweise über sich ergehen lassen, welche Funktionen Macht und Herrschaft haben und wie legitim historisch-spezifische Macht- und Herrschaftsverhältnisse jeweils sind. Damit ist aber zugleich die Frage nach den gesellschaftlichen Folgen, Wirkungen, Konsequenzen und Effekten von Macht und Herrschaft gestellt.

Fast jede Analyse von Macht und Herrschaft beginnt mit Max Webers klassischen Definitionen. Demnach bedeutet Macht grundsätzlich „jede Chance, innerhalb einer sozialen Beziehung den eigenen Willen auch gegen Widerstreben durchzusetzen, gleichviel worauf diese Chance beruht“ (Weber 1976, S. 28). Machthandlungen und Machtmethode n zeichnen sich folglich dadurch aus, dass individuelle oder kollektive Akteure ihren Willen gegenüber anderen durchzusetzen vermögen. Dabei kann Macht auf unterschiedlichen Grundlagen, Quellen und Fähigkeiten beruhen, sie kann durch verschiedene Instrumente und Mittel ausgeübt werden, sie kann diverse Formen und mannigfache Effekte haben.

Dieser handlungstheoretische Machtbegriff wurde später von Lukes (1974) schon um eine weitere Dimension – die non-decision-making power, die besagt, dass sich Macht auch darin artikuliert, dass die Agenda und die Themenordnung in Auswahlprozessen so bestimmt werden können, dass nur gewünschte Alternativen zur Auswahl bzw. auf der Tagesordnung stehen und bestimmte ungewünschte Möglichkeiten von vorneherein ausgespart werden – erweitert; Autoren wie Foucault (2005) haben dagegen von „kapillarischer Macht“ gesprochen, weil sie alle Poren einer Gesellschaft durchdringt. Popitz (1992, S. 22–39) hat hingegen in seiner Machttypologie die anthropologischen Grundlagen der Macht ausbuchstabiert: Erstens die Aktionsmacht, die als Verletzungsmacht wirkt und nicht nur die kreatürliche Verletzbarkeit des menschlichen Körpers, sondern auch sozio-ökonomische und politische Dimensionen umfasst; zweitens die instrumentelle Macht, die als Unterwerfungsmacht die konventionellste und geläufigste Form der Durchsetzung gegen fremde Kräfte ist; drittens die autoritative Macht, die das Verhalten und die Einstellungen eines Gegenübers steuert und einwilligende

Folgebereitschaft erzeugt; und viertens die datensetzende Macht, die als objektvermittelte Macht über materiell-artifizielle Lebensbedingungen zu begreifen ist.

In Bezug auf die Transformation von Macht und Herrschaft durch KI ist insbesondere die datensetzende Macht von Interesse, geht es hier doch „um eine Macht des Herstellens und der Hersteller“ (Popitz 1992, S. 31). Hier haben Menschen „aufgrund ihrer technischen Handlungsfähigkeit, ihrer technischen Intelligenz“ (Popitz 1992, S. 32) Macht über Menschen. Der doppelte Machtcharakter technischen Handelns besteht insofern darin, „die Macht über die Kräfte der Natur und die objektvermittelte Entscheidungsmacht über die Lebensbedingungen anderer Menschen“ (Popitz 1992, S. 31) zu besitzen.

Herrschaft kann demgegenüber nach Weber als „Chance, für einen Befehl bestimmten Inhalts bei angebbaren Personen Gehorsam zu finden“ (Weber 1976, S. 28), aufgefasst werden. Während der Machtbegriff für ihn „soziologisch amorph“ (Weber 1976, S. 28) sei, müsse der Begriff der Herrschaft präziser sein. Weber koppelt ihn an Fügsamkeit und Legitimität. Er identifiziert dabei legitime Gründe von Herrschaft und differenziert idealtypisch verschiedene Typen: die legale bzw. rationale Herrschaft, die traditionelle Herrschaft und die charismatische Herrschaft (Weber 1976, S. 122–142).

Nichtlegitime Herrschaft spart Weber hingegen weitgehend aus; die Legitimitätsunterstellung selbst ist kritisiert worden (Beetham 1991). Unstrittig ist hingegen die Einschätzung, dass Herrschaft eine Form institutionalisierter Macht darstellt, die sich durch dauerhafte Strukturen, Stabilität, Verlässlichkeit, Kontinuität und Konstanz auszeichnet und so die Machtausübung einer übergeordneten Person(engruppe) gegenüber untergeordneten Gruppen ermöglicht (Popitz 1992, S. 232 f.). Die Beziehung zwischen den Personen(gruppen) ist dabei durch eine grundsätzliche Asymmetrie gekennzeichnet, es ist ein Verhältnis unter Ungleichen. Man hat Macht über etwas oder jemanden, man beherrscht jemanden. Herrschaft – wie auch die Macht – verweist deshalb auf ein Hierarchieverhältnis, auf Dominanz und Unterwürfigkeit, auf eine Über- und Unterordnung, auf Herrschende und Beherrschte. Es geht in der Regel um Macht und Herrschaft von Menschen über Menschen. Davon bleibt zunächst die Einschätzung des Charakters der Herrschaft selbst unberührt. Und auch der Bezug zur Technik kann auf dieser Grundlage nur eindimensional ausbuchstabiert werden.

Der Begriff der Künstlichen Intelligenz ist ebenfalls umstritten und nur schwer zu definieren (vgl. zur Geschichte, Entwicklung und den unterschiedlichen Definitionen Collins 2021; Ertel 2021; Lenzen 2020, 2019; Kreutzer und Sirrenberg 2019; Wittpahl 2019; Zweig 2019; Walsh 2018; Garnham 2017; Kaplan 2017;

Nilsson 2009; Poole und Mackworth 2010; Russell und Norvig 2004; Dreyfus 1992, 1972; Kurzweil 1990; Rich 1983). Bei Künstlicher Intelligenz geht es immer um die maschinelle bzw. computerbasierte Nachahmung und Automatisierung von Verhalten, das bei Menschen als intelligent klassifiziert werden würde. Die Definitionen von Künstlicher Intelligenz unterscheiden sich in Bezug auf ihre Gegenstandsbereiche, ihre Schwerpunkte und ihre Reichweite erheblich voneinander, aber im Grundsatz stimmen sie alle darin überein, dass es bei Künstlicher Intelligenz stets „um die Schaffung von Computerprogrammen oder Maschinen geht, die ein Verhalten an den Tag legen können, das wir als ‚intelligent‘ bezeichneten, wenn es einem Menschen zugeschrieben würde“ (Kaplan 2017, S. 15).

Wenn gegenwärtig von (bereits bestehender) Künstlicher Intelligenz gesprochen wird, geht es häufig um Prozesse der Digitalisierung, um Algorithmen, Big Data, maschinelles Lernen, deep learning, Künstliche Neuronale Netze (KNN), Mustererkennung, bisweilen auch um Expertensysteme und Robotik. Unter Rückgriff auf den aktuellen Stand der Forschung gehen wir von folgender Arbeitsdefinition aus: Künstliche Intelligenz verweist stets auf die maschinelle Nachahmung, Simulation und Automatisierung von Verhaltensweisen, die bei Menschen als intelligent bezeichnet würden. Künstliche Intelligenz bezeichnet die Fähigkeit von Maschinen und Computern, komplexe kognitive Leistungen zu erbringen und Probleme zu lösen, für die Intelligenz erforderlich ist. KI beruht dabei maßgeblich auf der Analyse und Verarbeitung großer Datenmengen (Big Data) und hat das Ziel, auf dieser Grundlage Muster zu erkennen, Schlussfolgerungen zu ziehen und Entscheidungen, Empfehlungen sowie Vorhersagen zu treffen.

Um die Reichweite und die Fähigkeiten von KI zu beschreiben, wird in theoretischer und praktischer Perspektive auch zwischen schwacher (weak oder narrow artificial intelligence) und starker KI (strong oder general artificial intelligence) unterschieden. Die schwache KI bezieht sich immer auf Systeme mit klar definierten Anwendungsproblemen, die eine konkrete Aufgabe mit einer speziellen Methode bewältigen. Schwache KI soll intelligentes menschliches Denken simulieren und unterstützend bei der Lösung spezifischer Aufgaben wirken. Schwache KI kann dem Menschen vor allem als Hilfsmittel zur Problemlösung dienen (Searle 1990), es geht um maschinelles Lernen und die Automatisierung des Denkens, Schlussfolgerns, Entscheidens, Problemlösens und Vorhersagens, aber nicht – wie bei der starken KI – um vollständige Autonomie und die gänzliche Substitution oder gar Überflügelung menschlicher Intelligenz. Beispiele für schwache KI-Systeme sind etwa Chatprogramme, die Spracherkennung, Übersetzung oder die Text- und Bilderkennung. Darüber hinaus kommt KI vor allem in

der Mustererkennung, der Mensch-Maschine-Interaktion, in Expertensystemen, in der Robotik, bei der Optimierung von Suchmaschinen, bei Chatbots, beim autonomen Fahren oder in der personalisierten Werbung zum Einsatz. Schwache KI basiert wesentlich auf der Sammlung, Analyse und Verarbeitung großer Datenmengen, also von Big Data, und der Anwendung von Algorithmen, die auf Grundlage der analysierten Daten lernen und sich selbst optimieren.

Bei starker KI geht es demgegenüber um Systeme, die mindestens über die gleiche intellektuelle Leistungsfähigkeit wie der Mensch verfügen und autonom agieren können. Bisweilen ist auch von der „superintelligence“ (Bostrom 2014) die Rede, also von Systemen und Maschinen, die menschliche Intelligenz und Fähigkeiten weit übertreffen, sodass sich die Gefahr abzeichnet, dass Künstliche Intelligenz die Menschen unterwirft und beherrscht. Bislang existieren jedoch lediglich schwache KI-Systeme, starke KI gibt es nur in der Theorie. Trotz kontinuierlicher technischer Fortschritte, schreibt Bostrom (2003, S. 759), „we still have nothing that even resembles an intelligent machine, if by ‚intelligent‘ we mean possessing the kind of general-purpose smartness that we humans pride ourselves on“. Ob eine starke KI jemals existieren wird – oder überhaupt existieren sollte –, ist noch nicht absehbar und unter Forscher\*innen umstritten. Denn es sind ja nicht nur die unterschiedlichen Anwendungsbereiche von KI und die damit einhergehenden technischen Fragen und spezifischen Konsequenzen, die höchst unterschiedlich ausfallen, sondern insbesondere die ethischen und normativen Implikationen des Einsatzes von KI, die zwar schon heute erkennbar, aber in ihrer Tragweite und ihren Dimensionen noch nicht gänzlich absehbar sind.

---

### **3      Machtfragen und Herrschaftsaspekte in Bezug auf Künstliche Intelligenz**

Weitreichende gesellschaftliche Veränderungen waren und sind immer mit Transformationen von Macht- und Herrschaftsverhältnissen verbunden. Anstoß für solche Veränderungen waren neben sozio-ökonomischen oder politischen Aspekten oftmals auch technische Umwälzungen, die historisch einen stetigen Veränderungsprozess induziert haben. Technik ist ein elementarer Bestandteil menschlicher Gesellschaften und eine konstitutive Grundlage des gesellschaftlichen Lebens, sodass man von einer wechselseitigen Verbindung und einer unauflösbaren Verschränkung von technischem und sozialem Wandel ausgehen kann (Schrape 2021). Zwischen Technik und Gesellschaft besteht also ein wechselwirkendes, reziprokes Verhältnis. Technik prägt und strukturiert die Gesellschaft

und umgekehrt prägt und strukturiert der gesellschaftliche Entwicklungsstand den Stand der Technik.

Technik und Technologien bewegen sich dabei grundsätzlich – und dies gilt in besonderem Maße für Künstliche Intelligenz – nicht in macht- oder herrschaftsfreien Räumen, sondern sind selbst immer bereits auf komplexe Weise in spezifische Machtfigurationen und Herrschaftskontexte eingebunden. Zudem sind Technik und Technologien selbst immer auch Grundlagen, Quellen, Mittel, Ressourcen und Instrumente zur Ausübung von Macht und Herrschaft. Dies gilt umso mehr in modernen hochtechnologisierten Gesellschaften. Schon Popitz hatte geschrieben, dass „der Angelpunkt jeder Machtkontrolle in modernen Gesellschaften [...] die Kontrolle technischen Handelns“ (Popitz 1992, S. 181) ist. Heutzutage kann mehr denn je „jedes technische Verändern zu einem Akt der Machtausübung werden. Und zweifellos hat sich das Ausmaß möglicher Machtausübung mit zunehmender technischer Effizienz vergrößert“ (Popitz 1992, S. 180). Derjenige, der „heute über die technische Gestaltung unserer Lebensumwelt entscheidet, wer datensetzende Macht hat, kann in kürzester Frist ein unermessliches Ausmaß von Macht über unermesslich viele Menschen und eventuell [...] über unermesslich lange Zeiträume ausüben“ (Popitz 1992, S. 180).

Weil die Entwicklung und Durchsetzung von Technik bzw. Technologien stets Fragen des gesellschaftlichen Wandels und der sozialen Machtstrukturen und Herrschaftsverhältnisse sowie der sozialen Ungleichheit einer Gesellschaft berühren, ist eine rein technikzentrierte Sichtweise auf Technik offenkundig unzureichend. Die gesamtgesellschaftliche Relevanz und die gesamtgesellschaftlichen Folgen von Technik und Technologien lassen sich nur dann angemessen erfassen, wenn die genuin sozio-politische und gesellschaftstheoretische Dimension von Technik und damit das Verhältnis von Mensch, Technik und Gesellschaft in den Blick genommen wird. Zentral sind dabei die Fragen, was der Einsatz von Technik mit einer Gesellschaft macht und wie der Einsatz von Technologien die gesellschaftlichen Verhältnisse verändert. In macht- und herrschaftstheoretischer Perspektive zentral sind nicht zuletzt die Fragen, wer der Herr der Technik ist und wem diese zugutekommt.

In Bezug auf Machtfragen ist bislang umstritten und uneindeutig, ob Technik im Allgemeinen und KI im Speziellen Macht besitzen und als eigenständige, autonom agierende Akteure bzw. Handlungssubjekte selbst Macht ausüben (also die Macht aus den technischen Artefakten selbst entspringt), ob Macht der Technik bzw. der Künstlichen Intelligenz etwa als ‚Sachzwang‘ oder als ‚Nebenfolge‘ eingeschrieben ist (sie sich mithin beiläufig und ohne feste Absicht ergibt) oder ob es vielmehr Menschen sind, die technisch bedingt oder technisch vermittelt handeln und über den Besitz, die Beherrschung und die Anwendung der Technik

bzw. KI Macht ausüben. Analog dazu lassen sich in Bezug auf Herrschaft ähnliche Fragen formulieren: Sind Menschen die Akteure, die eine bestimmte Technik anwenden und sie etwa als Herrschaftstechnik nutzen oder ist eine Herrschaft der Technik denkbar und möglich?

Manche Theoretiker\*innen sind der Ansicht, dass Technik selbst keine Macht ausüben könne, sondern Macht immer und ausschließlich ein interpersonales Verhältnis bezeichnet und nur von Menschen über Menschen ausgeübt werden könne. So kritisiert etwa Han, dass die „Überhöhung der Technik zu einer Über-Macht [...] die eigentliche Wahrheit“ (Han 2005, S. 95) verdecke, die Wahrheit nämlich, dass Technik schlicht ein Machtmittel sei. Man dürfe Technik daher nicht „zu einer übermenschlichen Macht“ erheben, vielmehr müsse Technik als „der Ausdruck, das Resultat des menschlichen Strebens nach Macht“ (Han 2005, S. 95) angesehen werden. Popitz ist ebenfalls der Ansicht, dass Technik selbst keine Macht ausübe. Selbst bei der datensetzenden Macht handele es sich um „eine objektvermittelte Macht“, die wiederum „in materialisierter Form auf die Betroffenen übertragen“ (Popitz 1992, S. 31) werde. Auch diese Machtform ist folglich „keineswegs eine Macht der Dinge über den Menschen – obwohl sie die Ideologie ‚verdinglichter‘ Macht nahelegt –, sondern eine Macht des Herstellens und der Hersteller; eine vom Hersteller in das Ding eingebaute, häufig längere Zeit latente Macht, die jederzeit manifest werden kann“ (Popitz 1992, S. 31).

Alltagssprachliche oder metaphorische Redeweisen einmal außer Acht lassend, sind andere Theoretiker\*innen wie Bostrom (2018; 2014; 2003) demgegenüber der Ansicht, dass eine Technologie wie (starke) KI im Extremfall durchaus selbst Macht ausüben und den Menschen zum Untertan machen und beherrschen könnte. Auch wenn dies noch wie ferne Zukunftsmusik klingt, so legen die Fortschritte der Entwicklung Künstlicher Intelligenz doch zumindest dringend nahe, über diese Möglichkeit ernsthaft nachzudenken. Denn hier wird ja letztendlich die Frage nach der menschlichen ‚agency‘ aufgeworfen bzw. – freudianisch gesprochen – die Frage gestellt, wer im Mensch-Technik-Verhältnis der Herr im Hause ist. Die sozialwissenschaftliche Kontroverse zwischen sozialdeterministischen und technikdeterministischen Positionen dreht sich ja um das Ursache-Wirkungsprinzip bzw. das Determinationsverhältnis zwischen Mensch und Technik (KI).

Neuere Ansätze wie die Akteur-Netzwerk-Theorie (Latour 1995, 2005; vgl. auch Belliger und Krieger 2006; Rammert 2006) mit ihrer Unterscheidung von Akteuren und Aktanten scheinen soweit zu gehen, dass auch Systemen Künstlicher Intelligenz Handlungsmacht zugeschrieben und KI ein eigenständiger Akteursstatus zuerkannt werden kann, weil sie auf das Ko-Konstitutionsverhältnis von ‚Humans‘ und ‚Non-Humans‘ bei der Hervorbringung von Technik und



gesellschaftlicher Ordnung eingehen. Das würde zumindest ein erweitertes Verständnis von Macht und Herrschaft nötig machen und auf die mögliche emergente Qualität, die KI als gesellschaftliche Macht darstellt, hinweisen.

Während also Fragen nach dem Verhältnis von KI und Macht und Herrschaft zunächst umstritten und uneindeutig sind, so ist doch unstrittig, dass Künstliche Intelligenz als Schlüsseltechnologie des 21. Jahrhunderts das Potenzial hat, tiefgreifend in gesellschaftliche Verhältnisse moderner Gesellschaften – und damit auch gerade in gesellschaftliche Macht- und Herrschaftsverhältnisse – einzugreifen und diese substanziell zu verändern. KI darf deshalb nicht ausschließlich als technologische Innovation verstanden werden, sondern muss explizit als gesellschaftstheoretische Kategorie mit weitreichenden technologischen, ökonomisch-sozialen, politisch-kulturellen und habituellen Folgewirkungen analysiert werden.

Um nachvollziehen und analysieren zu können, wie Künstliche Intelligenz in gesellschaftliche Machtfigurationen und Herrschaftskontexte eingebettet ist und inwieweit KI gesellschaftliche Macht- und Herrschaftsverhältnisse in Wirtschaft, Politik und Gesellschaft transformiert, kann zwar einerseits an klassische Macht- und Herrschaftstheorien und ihre Annahmen und Funktionsprinzipien (Imbusch 2016, 2012) sowie an die datensetzende Macht technischer Artefakte und die Machtstrukturen in der artifiziellen Gesellschaft (Popitz 1995, 1992) angeknüpft werden, andererseits muss aber auch über diese Theorien hinausgegangen werden. Handlungstheoretische Machtverständnisse müssen um strukturtheoretische ergänzt werden, weil technische Artefakte im Grunde geronnene Macht darstellen. Ordnungsbezogene Herrschaftsverständnisse müssen um Zwang und Gewalt ergänzt werden, weil Technik ein Herrschaftsinstrument ist und Ungleichheiten vergrößern kann. Intermediäre Erklärungsansätze, die Handlungs- und Strukturebene zusammenbringen und gerade in dem wie auch immer gearteten Zusammenspiel von Macht, Herrschaft und Technik den Königsweg sehen, scheinen angesichts der Vielschichtigkeit von KI als Schlüsseltechnologie mitsamt ihrer Ambivalenzen nützlich zu sein. Dabei müssen auch Aspekte der Gouvernamentalität und die Machttechniken zur Subjektkonstitution (Foucault 2006, 2005, 1987) berücksichtigt werden.

Schließlich muss an dem bereits Ende des 20. Jahrhunderts einsetzenden Strukturwandel von Macht und Herrschaft, der unter anderem mit den Stichworten Entgrenzung und Subjektivierung verhandelt wird, sowie den damit einhergehenden komplexen und widersprüchlichen Modernisierungsprozessen angesetzt werden, „in dessen Verlauf Herrschaft subjektiviert wird und sich gleichzeitig von den legitimen Herrschaftsapparaten löst und in schiere Machtausübung verwandelt. Während die Mechanismen der Macht weiter rationalisiert werden, wird Herrschaft gleichermaßen subjektiviert wie totalisiert; [...] und

dies hat zur Folge: Herrschaft wird sowohl universell als auch unsichtbar“ (Bonß und Lau 2011, S. 10). Es bedarf also eines erweiterten Machtverständnisses, das technikoziologische machttheoretische Ausführungen (Dolata und Werle 2007; Dolata 2020) und die durch Technik vermittelte Macht von Menschen über Menschen (Irrgang 2007) sowie nicht zuletzt die sich in der neuen Technik KI selbst materialisierenden Machtformen (Al-Ani und Jeschke 2017) berücksichtigt.

Angesichts der Besonderheiten des Zusammenhangs von Technik, Macht und Herrschaft bzw. Macht, Herrschaft und KI sowie angesichts der vielfältigen gesellschaftlichen Anwendungsbereiche von KI und der spezifischen Auswirkungen auf die jeweiligen Macht- und Herrschaftsverhältnisse ist eine nach gesellschaftlichen Bereichen differenzierte und gesamtgesellschaftlich systematisierte Analyse der Folgen von KI nötig. Weil sich die Formen, Typen, Figurationen und Kontexte von Macht und Herrschaft in der Wirtschaft, der Politik und im Bereich der Gesellschaft unterscheiden und weil KI in den einzelnen gesellschaftlichen Bereichen zu unterschiedlichen Zwecken und Zielen eingesetzt wird, unterscheiden sich auch die Effekte und Auswirkungen von KI auf die gesellschaftlichen Macht- und Herrschaftsverhältnisse.

---

#### **4 Der Einsatz Künstlicher Intelligenz und seine Auswirkungen auf demokratische Macht- und Herrschaftsverhältnisse**

Der Anwendungsbereich von KI hat sich in den vergangenen Jahrzehnten erheblich ausgeweitet und ausdifferenziert. Wurde KI zunächst vor allem im Wirkungskreis von Wissenschaft und Militär verwendet und beforscht, hat es heute eine weitaus größere gesamtgesellschaftliche Relevanz. KI ist nicht mehr auf Wissenschaft, Forschung und Militärtechnologien beschränkt, sondern wird in praktisch allen gesellschaftlichen Bereichen, in der Wirtschaft, dem Finanzwesen und in der Arbeitswelt, in Politik, (sozialen) Medien, Kultur und Kommunikation, im Bildungsbereich, in Überwachungstechnologien und der Versicherheitlichung von Städten sowie dem Gesundheitssystem eingesetzt. Dieser Trend sowie der Bedeutungszuwachs von KI werden sich aller Voraussicht nach in den kommenden Jahrzehnten weiter fortsetzen und intensivieren.

Machtfigurationen und Herrschaftskontexte in der Wirtschaft sind vielgestaltig und vielfältig. Als sicher kann jedoch gelten, dass der Einsatz Künstlicher Intelligenz diese Machtfigurationen und Herrschaftskontexte keineswegs demokratisiert, sondern auf neuer Grundlage verfestigt und damit erweitert reproduziert. Wenn man an Macht und Herrschaft in der Wirtschaft denkt, landet man

zunächst unwillkürlich bei den großen Internetkonzernen (v. a. Google, Amazon, Facebook, Apple und Microsoft, die auch unter dem Akronym GAFAM zusammengefasst werden). Auch wenn diese durch ihre schiere Größe eine herausragende Bedeutung erreicht haben, so kann doch eine Analyse der Macht- und Herrschaftsverhältnisse nicht bei ihnen stehen bleiben, denn in demokratietheoretischer und herrschaftssoziologischer Perspektive reicht ihr Einfluss weit darüber hinaus, da die neuen Internet- und Plattformökonomien nur der sichtbarste Ausdruck der veränderten Verhältnisse sind. In ihnen finden sich zum einen alle klassischen Formen der Machtausübung, die durch deren Größe, Monopol- bzw. Oligopolstellung und Konkurrenzausschaltung in der kapitalistischen Wirtschaftsverfassung bedingt ist. Die Herrschaftsbefugnisse reichen aber weit über die jeweiligen Unternehmen hinaus, da ihnen ihre Organisationsform als Plattform ganz neuartige Machtpotenziale verleiht. Zum anderen geht mit diesem Machtzuwachs ein Formwandel des Kapitalismus insgesamt einher, der etwa unter Stichworten Digitaler Kapitalismus (Staab 2019, 2016), Überwachungskapitalismus (Zuboff 2018, 2015), Plattform-Kapitalismus (Srncicek 2018) oder Finanzmarktkapitalismus (Vogl 2021, 2015) ausbuchstabiert wird.

Denn neben die infrastrukturelle, informationelle und marktbeherrschende Macht tritt hier noch eine Vermachtung durch diverse Informationskontrollen (z. B. Datenkontrollen, Zugangskontrollen, Preiskontrollen), die es ihnen erlaubt, Wirtschaftsmacht in regelsetzende bzw. handlungsstrukturierende Macht zu übersetzen und das Verhalten der Gesellschaftsmitglieder zu strukturieren (Dolata 2020) und die Nutzer\*innen als „Produser“ (Vogl 2021, S. 78) zu instrumentalisieren. Dies geschieht insbesondere, um David Harveys (2005) Begriff der „Akkumulation durch Enteignung“ zu paraphrasieren, durch schleichende Formen einer ‚digitalen Akkumulation durch Enteignung‘, die nur scheinbar freiwillig stattfindet, indem Menschen ihnen Daten über sich zur Verfügung stellen. Dies ist zugleich das Perfide an der neuen Funktionsweise des digitalen Kapitalismus, dass er sich trefflich als ‚Freiheit‘ zu tarnen versteht, eigentlich aber einer streng neoliberal-libertären Ideologie folgt (Daub 2020), welche die weitergehende Privatisierung von öffentlichen Machtpotenzialen und Herrschaftsbefugnissen im Sinn hat. Die von den Konzernen ausgeübte Kontrollmacht kann deshalb als eine neue Spielart einer Mikrophysik von Macht gedeutet werden.

Das passt ebenfalls zu den undemokratisch organisierten Herrschaftsverhältnissen im Inneren der Unternehmen. Zwar finden sich hier alle Ideologien des sog. new work, aber es wird überall eine beträchtliche Kontrollmacht gegenüber den Angestellten ausgeübt und jede Organisation effektiver Gegenmacht (z. B. von Gewerkschaften) unterbunden oder konterkariert. Allein daraus resultiert bereits ein Umschlag von ökonomischer in politische Macht (z. B. sukzessive

Aufhebung von Marktbarrieren, Rechtsfeindlichkeit), die deren Unkontrollierbarkeit und Unregulierbarkeit durch Regierungen verstärkt, zumal während der neoliberalen Hegemonie ohnehin ein die Unternehmen begünstigendes entstaatlichtes Privatrecht entstanden ist, das heutzutage durch Strategien der Steuervermeidung und Gewinnverlagerung zusätzlich Druck auf Staaten und Regierungen auszuüben vermag. Dadurch werden die traditionellen Herrschaftsmittel und -aspekte der kapitalistischen Wirtschaftsverfassung noch verstärkt und demokratisch gewählte Regierungen noch machtloser gegenüber den Tech-Giganten.

Darüber hinaus hat aber auch der Einsatz der von den Tech-Konzernen entwickelten künstlich intelligenten Systemen den Finanzmarktkapitalismus befördert (Vogl 2021), welcher die Krisenhaftigkeit des regulären Kapitalismus enorm beschleunigt und verstärkt hat. Die großen Finanzmarktunternehmen sind heute für ihre millisekundenschnellen Transaktionen auf die Infrastrukturen der großen Tech-Konzerne angewiesen. Betrachtet man nur einmal die große Finanz- und Wirtschaftskrise von 2007 ff., dann sieht man sehr schnell die demokratiegefährdenden Potenziale solcher Krisen.

Wirtschaftsverhältnisse sind immer auch Arbeitsverhältnisse. Klassische Machtausübung der Unternehmen erfolgt seit jeher in der Wirtschaft über weitreichende Arbeitgeberbefugnisse, die sich in den letzten Jahrzehnten deutlich verstärkt haben, sodass manche Chronist\*innen inzwischen von Unternehmen als „privaten Regierungen“ sprechen und aufzeigen, wie „Arbeitgeber über unser Leben herrschen (und warum wir nicht mehr darüber reden)“ (Anderson 2019). Der Formwandel der kapitalistischen Ökonomie – Produktion, Distribution, Konsumtion, Mehrwertproduktion bzw. -abschöpfung (Mazzucato 2018) – und die Digitalisierung von Produktionsregimes – just-in-time, time-to-market, globale Lieferketten, etc. – haben auch die Arbeitsformen in Mitleidenschaft gezogen und dafür gesorgt, dass Hierarchien zwar einerseits flacher geworden sind, der technische Fortschritt jedoch andererseits bestehende Abhängigkeiten im Kapitalismus verstärkt und sogar neue geschaffen hat.

Herrschaftsmacht (direkte oder indirekte Macht über die Arbeiter\*innen und die Angestellten, z. B. mittels des Einsatzes von Technik und Technologien, die Arbeitswelt strukturierende Sachzwänge) zeigt sich heute eher mittelbar oder wie im Fall der „indirekten Steuerung“ (Sauer 2011) nur noch begrenzt, da es hier zu einer Subjektivierung von Macht und Herrschaft durch die Internalisierung von Marktprinzipien kommt (Böhle et al. 2011). Die Abhängigkeit der Menschen von Technik nimmt zu, dadurch lässt sich eine zunehmende technische Fremdbestimmung der Arbeitsabläufe feststellen, die zwar ohne direkte Herrschaft auskommt, aber die volle Verantwortung für das Arbeitsprodukt komplett subjektiviert. Gleichzeitig haben sich durch die digitalen Möglichkeiten neue Formen

der Kontrollmacht (etwa in Bezug auf Arbeitsabläufe und Zeitregime) herausgebildet. Durch die Plattform-Struktur vieler Unternehmen sind zudem neue Formen der Ausbeutung und Prekarität entstanden. „Es gibt ein neues Prekariat. Des- sen Angehörige stehen nicht mehr mit verrußten Händen in Backsteinfabriken in Manchester. Nein, sie huschen in den großen Städten der Welt an einem vorbei. Auf dem Rad oder im Auto. Sie putzen, pflegen, liefern Lebensmittel oder fahren Taxi. Für alle möglichen Dienstleistungen gibt es inzwischen eine entsprechende App. [...] Gewerkschaften und Betriebsräte sind den Arbeitern oft ebenso fremd wie den Unternehmen, deren Zentralen nicht selten weit entfernt in Kalifornien sitzen“ (Gross 2021, S. 4).

Mit dem Einsatz Künstlicher Intelligenz und digitaler Formen der Arbeitszeiterfassung ist nicht nur eine verschärfte Kontrollmacht einhergegangen, sondern sind auch neue Formen der Disziplinarmacht (Foucault 2005, 2006) entstanden, denn die Überwachung der Mitarbeiter kommt nicht ohne Strafen und Sanktionierungen aus. Die darin zum Ausdruck kommende Herrschaftsmacht verdoppelt klassische Formen der Herrschaftsmacht und macht Kontrolle effektiver, sodass von einer regelrechten Kuratierung von Sozialität (Dolata 2020) gesprochen werden kann.

Der digitale Kapitalismus hat also nicht nur zur Verschiebung der Machtverhältnisse zwischen Politik und Wirtschaft bzw. Staat und Kapital, sondern auch zu einer Machtverschiebung zwischen Kapital und Arbeit geführt. Gerade in der Finanzwirtschaft sind hybride Formen der Machtausübung (zwischen Wettbewerb und hierarchischer Kontrolle) und neuartige Macht- und Herrschaftsformen (Herrschaft wird gefestigt, aber indirekter; es kommt zu einer Depersonalisierung von Herrschaft) entstanden.

Eindeutiger und weniger ambivalent stellt sich der Einsatz von KI im militärischen Bereich dar. Denn hier sind Gewaltmittel zugleich Machtmittel in den Händen des Staates bzw. der Unternehmen, die sie herstellen. Staaten verfügen in der Regel ohnehin über das Gewaltmonopol und deren Herrschaft stützt sich in letzter Instanz auf die Möglichkeit von Waffengewalt zur Absicherung einer Herrschaftsordnung oder des Staates. Gewalt ist dabei zugleich eine Form der Macht und sie dient u. a. zur Erlangung bzw. Verstetigung von Herrschaft; Staaten haben direkte Macht durch die Kontrolle der Gewaltmittel (Herrschaft). Der militärische Bereich kann dabei in Bezug auf Innovation und Forschung (KI-Algorithmen und Datenverarbeitungskapazitäten) als ein wesentlicher Treiber in der Entwicklung von KI gelten.

Mit der durch KI möglich gewordenen neuerlichen Revolutionierung militärischer Angelegenheiten (z. B. Aufklärung durch Drohnen sowie Schnelligkeit, Beschleunigung und Automatisierung von Prozessen, nicht zuletzt auch durch

Roboterisierung) steigert sich, mit Popitz (1992, S. 22–39) gesprochen, einfache Aktionsmacht bzw. instrumentelle Macht (z. B. Drohen und Bedrohen) von Staaten oder Regierungen durch den Verlust von Entscheidungshoheit über Prozesse sowie die Automatisierung und Delegation von Entscheidungen zu einer absoluten Macht (nämlich der Macht zu töten), die automatisiert von intelligenten Maschinen (etwa bewaffneten Drohnen) ausgeführt werden kann. Damit drohen Gefahren wie eine sukzessive Militarisierung von Gesellschaften und Rüstungswettläufe sowie eine Enthumanisierung und exzessive Konfliktbereitschaft durch technologisch bedingte Überlegenheitsgefühle. Die Chancen für eine Humanisierung der Kriegführung scheinen demgegenüber weit geringer zu sein.

Der Einsatz von KI führt zu einer Zentralisierung von Herrschaft, weil mit ihr geopolitisch bestehende Machtungleichgewichte zwischen Staaten zementiert werden: Bevorteilte Staaten bleiben vorn und sind technologisch überlegen. Der Einsatz von KI ist hier eher ein Syndrom totaler Gewalt (Glorifizierung, Indifferenz, Technisierung), mit der jetzt auch Formen von „hybrider Kriegsführung“ oder sog. „Cyber Warfare“ erfolgreich durchgeführt werden können (Del Monte 2018; Cummings 2017). Dieses Problem verschärft sich noch in all jenen Ländern, in denen polizeiliche und militärische Befugnisse nicht so getrennt sind wie in den meisten westeuropäischen Staaten und das Militär auch im Inneren zur Herrschaftssicherung eingesetzt werden kann (O’Neill 2016).

Aber selbst bei der Aufrechterhaltung der unterschiedlichen Sphären Militär und Gesellschaft schreitet die Kontroll- und Überwachungsmentalität auch in den westlichen Industrieländern rasant voran. Die neuen Versicherheitlichungs- und Überwachungsstrategien (z. B. CCTV) infolge der polarisierenden Wirkungen des Neoliberalismus, die von David Garland in einer klassisch zu nennenden Untersuchung der Verhältnisse in den USA als „Culture of Control“ (Garland 2002) gelten kann, sind nur durch den Einsatz bestimmter Formen Künstlicher Intelligenz möglich und werden in demokratischen Gesellschaften mit einem Sicherheitsversprechen bei kompletter Überwachung („Sicherheitsgesellschaft“) bzw. der Stärkung von vermeintlichen Präventionsstrategien (wie dem „predictive policing“) von interessierter Seite attraktiv gemacht (Singelstein und Stolle 2008; Groenemeyer 2010), in undemokratischen und autoritären Gesellschaften ist an die Überwachung längst ein Belohnungs- und Bestrafungssystem gekoppelt, mit dessen Hilfe politische Loyalitäten belohnt oder Vergehen bestraft werden (das sog. „social scoring“).

Die sich hier zeigende Kontrollmacht, die aus der Ubiquität des Verdachts erwächst, beruht im Grunde auf der datensetzenden Macht (Popitz 1992, S. 29–39) und kann heute mit dem Terminus der „instrumentären Macht“ (Zuboff 2018)

adäquat beschrieben werden. In ihr drückt sich eine Automatisierung von Herrschaft aus, die ohne konkreten Herrscher auskommt, aber umso unerbittlicher wirksam wird. Hier droht also das Gespenst total(itär)er Kontrolle Wirklichkeit zu werden; autoritäre Systeme und ihre automatisierte Steuerung führen zur Kontrolle einer Gesellschaft durch sog. Big Brother-Technologien, die heutzutage bereits weit über das Maß hinausgehen, das sich George Orwell in seinem Roman „1984“ (Orwell 2008) einst vorstellen konnte. Es droht der komplette Verlust individueller Freiheits- und Grundrechte. Durch die Zentralisierung von Herrschaft bei staatlichen Behörden und Diensten, die unterschiedliche Daten zusammenführen, wird die Überwachung zu einem klassischen Herrschaftsinstrument.

Das Problem an dieser Beschreibung ist, dass hier nicht von einzelnen oder vereinzelt Prozessen und Aspekten die Rede ist, sondern dass diese gleichzeitig stattfinden und zusammenspielen. Daraus resultiert nicht zuletzt ihre Bedrohung für demokratische Macht- und Herrschaftsverhältnisse. Dabei haben wir bisher die unmittelbaren demokratiebedrohenden Auswirkungen digitaler Macht sowie die zunehmende Abhängigkeit von großen Tech-Konzernen und ihrer Datenmacht noch gar nicht berücksichtigt, wie sie in den modernen sozialen Kommunikationsmedien um sich greift. Hier geht etwa um die Beeinflussung demokratischer Wahlen und anderer Institutionen der Demokratie, um Fake News, Social Bots, Hate Speech und die generell veränderte Sprache in der politischen Kultur (z. B. durch social media) mit ihren Filterblasen und regenerativen Erregungskurven. Diese „Informatics of Domination“ (Haraway 1995) stellen die Frage nach Volkssouveränität oder Macht des Kapitals ganz konkret, lässt sich an ihnen doch ablesen, wie automatisierte digitale Prozesse eine kommunikative Macht entfalten können, die demokratischen Verhältnissen mehr als abträglich ist.

Die gesellschaftlichen Auswirkungen und Effekte dieser KI-induzierten Macht lassen sich in Bezug auf die Erosion und Krise der Demokratie – z. B. „Postdemokratie“ (Crouch 2008) – und die Erosion demokratischer Legitimität nachzeichnen, sie schlägt sich nieder im Entzug der demokratischen Kontroll- und Regulierungsmöglichkeiten, in der Überbetonung kommunikativer Macht zuungunsten demokratisch legitimer Machtformen, in der Zunahme von Manipulationsmöglichkeiten (z. B. durch Social Bots und Fake News), durch die eine demokratische Öffentlichkeit in Mitleidenschaft gezogen wird. Ohnehin erleben wir einen neuerlichen „Strukturwandel der Öffentlichkeit“ (Habermas 2022), der, wie einst von Habermas für die postliberale Phase der frühbürgerlichen Öffentlichkeit auf einem anderen technologischen Niveau beschrieben (Habermas 1990), heute erneut in sein Regressionsstadium eingetreten zu sein scheint.

Alle hochfliegenden Träume von einer Demokratisierung der Gesellschaft durch das Internet und die sozialen Medien sind heute doch einer erheblichen

Ernüchterung gewichen, stattdessen ist eher die Rede von einem „technologischen Totalitarismus“ (Schirrmacher 2015). Portale wie Google, Facebook, Amazon und Apple wissen heute mehr über uns, als jedem einzelnen lieb sein kann. Denn bei den Konzernen geht es nicht allein um Wirtschaftspolitik, sondern um Gesellschaftspolitik, um Beeinflussungsmöglichkeiten und um die Strukturierung und Zurichtung von Konsumbedürfnissen, um das, wie wir über bestimmte Dinge denken sollen. Hier sind ganz neue Herrschaftsfigurationen und Machtformen (trotz vermeintlicher individueller Autonomiegewinne) entstanden, die mit gleichzeitigen Ohnmachtserfahrungen bzw. Machtlosigkeiten einhergehen. Dadurch wird allerdings auch die Legitimität demokratischer Herrschaft infrage gestellt. Es verschärfen sich soziale Ungleichgewichte und soziale Konflikte nehmen zu. Dass sich aus dem Eigentum, Besitz oder der Verfügung über große (digitale) Datenmengen neue und sehr effektive Machtformen ergeben, ist ein relativ junges Phänomen, das angesichts zunehmender Asymmetrien von Herrschaft und Legitimität dringend einer Regulierung bedarf.

---

## **5 Resümee: Die doppelte Transformation von Macht und Herrschaft durch KI**

KI ist grundsätzlich ein ambivalentes und ambiguitäres Phänomen. Mit der Entwicklung und Anwendung von Künstlicher Intelligenz sind vielfältige Chancen und Potenziale, aber auch gravierende Risiken und Gefahren verbunden – die sich in den jeweiligen Einsatzfeldern bisweilen gravierend voneinander unterscheiden. KI sollte insofern keineswegs verteufelt, aber auch keinesfalls bedingungslos apologetisiert werden. Zu den Chancen und Potenzialen zählen etwa Wachstum, Wohlstand, technologischer Fortschritt, Zeitgewinne, Sicherheitsgewinne, Vereinfachungen und Verbesserungen des menschlichen Lebens sowie Verbesserungen in der Gesundheitsversorgung – etwa in der Krebsdiagnose. Zu den Risiken und Gefahren zählen etwa eine mögliche Entmenschlichung, Entpolitisierung und Entdemokratisierung, eine Entgrenzung des technisch Möglichen, der Verlust von Freiheits- und Bürgerrechten oder gar das Übergleiten in eine Kontroll-, Überwachungs- und Sicherheitsgesellschaft sowie die Verstetigung und Verfestigung von Vorurteilen (Coded Bias) und Ungleichheiten.

Während die konkreten gesellschaftlichen Auswirkungen ambivalent und mehrdeutig sind, ist eindeutig, dass Künstliche Intelligenz als Schlüsseltechnologie des 21. Jahrhunderts das Potenzial hat, die gesellschaftlichen Verhältnisse moderner Gesellschaften tiefgreifend zu verändern und damit auch die gesellschaftlichen Macht- und Herrschaftsverhältnisse substanziell zu transformieren.



Und dies gleich in doppelter Hinsicht: Zum einen transformiert KI die realen Machtstrukturen und Herrschaftsverhältnisse in Wirtschaft, Politik und Gesellschaft, zum anderen muss, wenn über die macht- und herrschaftsbezogenen Konsequenzen von KI in den Blick genommen werden, über ein erweitertes Macht- und Herrschaftsverständnis nachgedacht werden, das nicht nur die unmittelbare Macht und Herrschaft von Menschen über Menschen betrachtet, sondern auch subkutane, indirekte Machteffekte und die technikvermittelte bzw. technikinduzierte Macht berücksichtigt.

Künstliche Intelligenz beruht bisher vor allem auf der Sammlung und Verarbeitung von großen Datenvolumina. Speziell aus dem Besitz, der Verfügung und der Verwertung dieser Daten erwachsen neuartige Machtpotenziale und Herrschaftstechniken, die zum einen die Machtposition bereits mächtiger (Wirtschafts-)Akteure vervielfacht und zum anderen neue Machtzentren hervorbringt, die ohne hinreichende gesellschaftliche und demokratische Legitimation Wissensvorräte ansammeln. Eine ungezügelter und unregulierter Verfügbarkeit und Verwendung von Daten ohne Transparenz kann die Macht- und Herrschaftsausübung entgrenzen und subjektivieren; gleichzeitig findet aber auch ein Prozess der verschärften Monopolisierung und Re-Zentralisierung von Macht und Herrschaft bei bereits mächtigen Herrschaftsträgern (Wirtschaftsunternehmen, Staaten, staatliche und private (Sicherheits-)Behörden und Institutionen) statt.

Diese Gleichzeitigkeit des Ungleichen, diese Möglichkeiten zur disruptiven Herausbildung neuer mächtiger Akteure und die Möglichkeiten der Machtzuwachses in bereits bestehenden Machtzentren auf qualitativ neuer Grundlage sowie der simultane Prozess sowohl der Entgrenzung und Subjektivierung als auch der Monopolisierung und Zentralisierung von Macht und Herrschaft besitzen in letzter Konsequenz ein totalitäres Potenzial. Es ist geradezu ein Kennzeichen der digitalen Moderne mit ihren künstlich-intelligenten technischen Artefakten, dass Kompetenzen und Verantwortlichkeiten hinsichtlich der Machtausübung unscharf werden, eindeutige Abgrenzungen von Herrschaftsbefugnissen und Verfahrensmustern sich auflösen und unfreiwillige Kontrollverluste eintreten, die sich in ethischen Regulations- und politischen Legitimationskrisen niederschlagen können.

In Bezug auf die konkreten Macht- und Herrschaftsverhältnisse in Wirtschaft, Politik und Gesellschaft zeigt sich, dass KI von sich aus und per se keine egalisierende, nivellierende und demokratisierende Wirkung hat. Vielmehr kommt es tendenziell zu einer Potenzierung, Zentrierung, Monopolisierung und Stabilisierung der Macht- und Herrschaftsverhältnisse, also zu einem Ausbau der Macht und Herrschaftsbefugnisse bereits mächtiger Akteure – und damit auch zu einer Verstärkung und Verschärfung von Ungleichgewichten und Ungleichheiten.

Künstliche Intelligenz ist ein gesellschaftlicher Machtfaktor. KI ist, wie Karl Marx einst über das Kapital geschrieben hat, „eine gesellschaftliche Macht“ (Marx 2008, S. 274). Und wie auch das Kapital zeigt sich KI immer mehr „als entfremdete, verselbständigte Macht, die als Sache [...] der Gesellschaft gegenübertritt“ (Marx 2008, S. 274). Die Phänomene Macht und Herrschaft werden durch KI nicht weniger oder unbedeutender, ganz im Gegenteil. Sie werden allerdings mithin unsichtbarer, anonymer, diffuser und treten in versteckter, dafür aber umso wirksamer und effektiver Form auf. Allerdings ist KI – zumindest noch – kein eigenständig und komplett autonom agierender Akteur mit einem Willen zur Macht und einem Willen zur Herrschaftsunterwerfung. Trotzdem ist ein erweitertes Machtverständnis notwendig, das die emergente Qualität von KI berücksichtigt und nicht nur die direkte Macht von Menschen über Menschen in den Blick nimmt, sondern auch die technikinduzierten Machteffekte. KI entfaltet nicht zuletzt auch dadurch gesellschaftliche Macht, dass die Logik, Strukturmuster und Verhalten von Maschinen und Computern als Vorbild für menschliches Verhalten gelten. Prinzipien wie (Selbst-)Optimierung, Automatisierung, Effizienz, Strukturierung und Standardisierung avancieren zu zentralen Leitwerten auch und gerade für individuelles Verhalten und Gesellschaften insgesamt. Auf diese Weise kann es zu einer Normierung des Menschen kommen.

Auch wenn die Folgen von modernen Techniken bzw. Technologien in Bezug auf ihre Machtaspekte ambivalent zu sein scheinen und die Wirkungen ihres Einsatzes in Wirtschaft, Politik und Gesellschaft heterogen sein mögen, so scheint gewiss zu sein, dass mit dem zunehmenden Einsatz Künstlicher Intelligenz in verschiedenen gesellschaftlichen Sphären ein qualitativer Sprung in Bezug auf Machtpraxen und Herrschaftstechniken einhergeht, der sich direkt oder indirekt als problematisch für die Modernisierung von Gesellschaften erweisen kann. Es ist deshalb zu erwarten, dass die Kämpfe um Macht und Herrschaft (Schwingerl 1993) zukünftig weiter zunehmen und sich als elementare Bestandteile gesellschaftlicher Aushandlungsprozesse bzw. als sichtbarer Ausdruck von gesellschaftlichen Spannungen, Konflikten, Ungleichheiten und Veränderungsprozessen auf neuer Grundlage reproduzieren.

---

## Literatur

- Al-Ani, Ayad, und Sabina Jeschke. 2017. Die Macht der Maschinen: Wie die Digitalisierung die Politik herausfordert. *Internationale Politik* 72: 118–127.
- Anderson, Elizabeth. 2019. *Private Regierung. Wie Arbeitgeber über unser Leben herrschen (und warum wir nicht darüber reden)*. Berlin: Suhrkamp.

- Anter, Andreas. 2020. *Theorien der Macht zur Einführung*. Hamburg: Junius.
- Beck, Ulrich. 1986. *Risikogesellschaft. Auf dem Weg in eine andere Moderne*. Frankfurt/M.: Suhrkamp.
- Beetham, David. 1991. *The Legitimation of Power*. London: MacMillan.
- Belliger, Andréa und David J. Krieger. Hrsg. 2006. *ANThology. Ein einführendes Handbuch zur Akteur-Netzwerk-Theorie*. Bielefeld: transcript.
- Böhle, Fritz et al. 2011. Herrschaft durch Objektivierung. Zum Wandel von Herrschaft in Unternehmen. In *Macht und Herrschaft in der reflexiven Moderne*, hrsg. Wolfgang Bonß und Christoph Lau, 244–283, Weilerswist: Velbrück.
- Bonß, Wolfgang und Christoph Lau. 2011. Einleitung: Zum Strukturwandel von Macht und Herrschaft in der Moderne. In *Macht und Herrschaft in der reflexiven Moderne*, hrsg. Wolfgang Bonß und Christoph Lau, 7–46, Weilerswist: Velbrück.
- Bostrom, Nick. 2003. When machines outsmart humans. *Futures* 35: 759–764.
- Bostrom, Nick. 2014. *Superintelligence. Paths, Dangers, Strategies*. Oxford: Oxford University Press.
- Bostrom, Nick. 2018. *Die Zukunft der Menschheit. Aufsätze*. Berlin: Suhrkamp.
- Collins, Harry. 2021. The science of artificial intelligence and its critics. *Interdisciplinary Science Reviews* 46: 53–70.
- Crouch, Colin. 2008. *Postdemokratie*. Frankfurt/M.: Suhrkamp.
- Cummings, M.L. 2017. *Artificial Intelligence and the Future of Warfare*. London: Chatham House.
- Daub, Adrian. 2020. *Was das Valley denken nennt*. Berlin: Suhrkamp.
- Del Monte, Louis A. 2018. *Genius Weapons. Artificial Intelligence, Autonomous Weaponry, and the Future of Warfare*. Buffalo: Prometheus Books.
- Dolata, Ulrich und Raymond Werle. Hrsg. 2007. *Gesellschaft und die Macht der Technik. Sozioökonomischer und institutioneller Wandel durch Technisierung*. Frankfurt/M./New York: Campus.
- Dolata, Ulrich. 2020. Plattform-Regulierung. Koordination von Märkten und Kuratierung von Sozialität im Internet. *Berliner Journal für Soziologie* 29: 179–206.
- Dreyfus, Hubert. 1972. *What Computers Can't Do: The Limits of Artificial Intelligence*. New York: Harper & Row.
- Dreyfus, Hubert. 1992. *What Computers Still Can't Do: A Critique of Artificial Reason*. Cambridge: MIT Press.
- Ertel, Wolfgang. 2021. *Grundkurs Künstliche Intelligenz. Eine praxisorientierte Einführung*. Wiesbaden: Springer.
- Foucault, Michel. 1987. Das Subjekt und die Macht. In *Michel Foucault: Jenseits von Strukturalismus und Hermeneutik*, hrsg. Hubert L. Dreyfus und Paul Rabinow, 243–261. Frankfurt/M.: Suhrkamp.
- Foucault, Michel. 2005. *Analytik der Macht*. Frankfurt/M.: Suhrkamp.
- Foucault, Michel. 2006. *Sicherheit, Territorium, Bevölkerung. Geschichte der Gouvernementalität*. Frankfurt/M.: Suhrkamp.
- Gallie, Walter Bryce. 1956. Essentially Contested Concepts. *Proceedings of the Aristotelian Society* 56: 167–198.
- Garland, David. 2002. *The Culture of Control. Crime and Social Order in Contemporary Society*. Oxford: Oxford University Press.
- Garnham, Alan. 2017. *Artificial Intelligence: An Introduction*. London: Routledge.

- Giddens, Anthony. 1992. *Die Konstitution der Gesellschaft*. Frankfurt/M./New York: Campus.
- Groenemeyer, Axel. Hrsg. 2010. *Wege der Sicherheitsgesellschaft. Gesellschaftliche Transformationen der Konstruktion und Regulierung innerer Unsicherheiten*. Wiesbaden: VS Verlag.
- Gross, Simon. 2021. Tech-Firmen – So geht Ausbeutung heute. *Süddeutsche Zeitung*, 10. Dezember.
- Habermas, Jürgen. 2022. *Ein neuer Strukturwandel der Öffentlichkeit und die deliberative Politik*. Berlin: Suhrkamp.
- Habermas, Jürgen. 1990. *Strukturwandel der Öffentlichkeit. Untersuchungen zu einer Kategorie der bürgerlichen Gesellschaft*. Frankfurt/M.: Suhrkamp.
- Han, Byung-Chul. 2005. *Was ist Macht?* Stuttgart: Reclam.
- Haraway, Donna. 1995. *Die Neuerfindung der Natur: Primaten, Cyborgs und Frauen*. Frankfurt/M./New York: Campus.
- Harvey, David. 2005. *Der neue Imperialismus*. Hamburg: VSA-Verlag.
- Imbusch, Peter. 2011. Herrschaft. In *Politische Theorie und Politische Philosophie. Ein Handbuch*, hrsg. Martin Hartmann und Claus Offe, 219–222. München: C.H. Beck.
- Imbusch, Peter. Hrsg. 2012. *Macht und Herrschaft. Sozialwissenschaftliche Theorien und Konzeptionen*. Wiesbaden: VS Verlag für Sozialwissenschaften.
- Imbusch, Peter. 2016. Macht und Herrschaft. In *Einführung in die Hauptbegriffe der Soziologie*, hrsg. Hermann Korte und Bernhard Schäfers, 195–220. Wiesbaden: VS Verlag für Sozialwissenschaften.
- Irrgang, Bernhard. 2007. *Technik als Macht. Versuche über politische Technologie*. Hamburg: Dr. Kovac.
- Kaplan, Jerry. 2017. *Künstliche Intelligenz: Eine Einführung*. Bonn: mitp.
- Kreutzer, Ralf T., und Marie Sirrenberg. 2019. *Künstliche Intelligenz verstehen. Grundlagen – Use-Cases – unternehmenseigene KI-Journey*. Wiesbaden: Springer.
- Kurzweil, Ray. 1990. *The Age of Intelligent Machines*. Cambridge: The MIT Press.
- Latour, Bruno. 1995. *Wir sind nie modern gewesen. Versuch einer symmetrischen Anthropologie*. Berlin: Akademie Verlag.
- Latour, Bruno. 2005. *Reassembling the Social. An Introduction into Actor-Network-Theory*. Oxford: Oxford University Press.
- Lenzen, Manuela. 2019. *Künstliche Intelligenz. Was sie kann & was uns erwartet*. München: C.H. Beck.
- Lenzen, Manuela. 2020. *Künstliche Intelligenz: Fakten, Chancen, Risiken*. München: C.H. Beck.
- Lukes, Stephen. 1974. *Power. A Radical View*. London: MacMillan.
- Marx, Karl. 2008. *Das Kapital. Kritik der Politischen Ökonomie. Dritter Band*. MEW 25. Berlin: Karl Dietz.
- Maurer, Andrea, 2004, *Herrschaftssoziologie. Eine Einführung*. Frankfurt/M./New York: Campus.
- Mazzucato, Mariana. 2018. *Wie kommt der Wert in die Welt. Von Schöpfern und Abschöpfern*. Frankfurt/M./New York: Campus.
- Nemitz, Paul und Matthias Pfeffer. 2020. *Prinzip Mensch: Macht, Freiheit und Demokratie im Zeitalter der Künstlichen Intelligenz*. Bonn: Dietz.

- Nilsson, Nils J. 2009. *The Quest for Artificial Intelligence: A History of Ideas and Achievements*. Cambridge: Cambridge University Press.
- O'Neill, Cathy. 2016. *Weapons of Math Destruction. How Big Data Increases Inequality and Threatens Democracy*. New York: Crown Publisher.
- Orwell, George. 2008. 1984. Berlin: Ullstein.
- Pansardi, Pamela, und Marianna Bindi. 2021. The New Concepts of Power? Power-over, Power-to and Power-with. *Journal of Political Power* 14: 51-71.
- Paris, Rainer. 1998. *Stachel und Speer. Machtstudien*. Frankfurt/M.: Suhrkamp.
- Pfeffer, Matthias. 2021. *Menschliches Denken und Künstliche Intelligenz. Eine Aufforderung*. Bonn: Dietz.
- Poole, David L. und Alan K. Mackworth. 2010. *Artificial Intelligence: Foundations of Computational Agents*. Cambridge: Cambridge University Press.
- Popitz, Heinrich. 1992. *Phänomene der Macht: Autorität – Herrschaft – Gewalt – Technik*. Tübingen: Mohr Siebeck.
- Popitz, Heinrich. 1995. *Der Aufbruch zur artifiziiellen Gesellschaft: Zur Anthropologie der Technik*. Tübingen: Mohr Siebeck.
- Rammert, Werner. 2006. Technik, Handeln und Sozialstruktur. Eine Einführung in die Soziologie der Technik. *TUTS-WP-3–2006*. Berlin: Technische Universität Berlin.
- Rich, Elaine. 1983. *Artificial Intelligence*. New York: McGraw-Hill Inc.
- Russell, Stuart und Peter Norvig. 2004. *Künstliche Intelligenz. Ein moderner Ansatz*. München: Pearson.
- Sauer, Dieter. 2011. Indirekte Steuerung. Zum Formwandel betrieblicher Herrschaft. In *Macht und Herrschaft in der reflexiven Moderne*, hrsg. Wolfgang Bonß und Christoph Lau, 358–378. Weilerswist: Velbrück.
- Schirmacher, Frank. Hrsg. 2015. *Technologischer Totalitarismus. Eine Debatte*. Berlin: Suhrkamp.
- Schrage, Jan-Felix. 2021. *Digitale Transformation*. Bielefeld: transcript.
- Schwingel, Markus. 1993. *Analytik der Kämpfe. Macht und Herrschaft in der Soziologie Bourdieus*. Hamburg: Argument Verlag.
- Searle, John R. 1990. Is the Brain a Digital Computer? *Proceedings and Addresses of the American Philosophical Association* 64: 21-37.
- Singelstein, Tobias und Peter Stolle. 2008. *Die Sicherheitsgesellschaft. Soziale Kontrolle im 21. Jahrhundert*. Wiesbaden: VS Verlag.
- Sofsky, Wolfgang und Rainer Paris. 1994. *Figurationen sozialer Macht. Autorität, Stellvertretung, Koalition*. Frankfurt/M.: Suhrkamp.
- Srnicek, Nick. 2018. *Plattform-Kapitalismus*. Hamburg: HIS.
- Staab, Philipp. 2016. *Falsche Versprechen. Wachstum im digitalen Kapitalismus*. Hamburg: HIS.
- Staab, Philipp. 2019. *Digitaler Kapitalismus. Markt und Herrschaft in der Ökonomie der Unknappheit*. Berlin: Suhrkamp.
- Tegmark, Max. 2019. *Leben 3.0. Mensch sein im Zeitalter Künstlicher Intelligenz*. Berlin: Ullstein.
- Vogl, Joseph. 2015. *Der Souveränitätseffekt*. Zürich: diaphanes.
- Vogl, Joseph. 2021. *Kapital und Ressentiment. Eine kurze Theorie der Gegenwart*. München: C.H. Beck.

- Walsh, Toby. 2018. *It's alive. Wie Künstliche Intelligenz unser Leben verändern wird*. Hamburg: Körber.
- Weber, Max. 1976. *Wirtschaft und Gesellschaft. Grundriss der verstehenden Soziologie*. Tübingen: Mohr Siebeck.
- Wittpahl, Volker. Hrsg. 2019. *Künstliche Intelligenz. Technologie, Anwendung, Gesellschaft*. Wiesbaden: Springer.
- Zuboff, Shoshana. 2015. Big Other: Surveillance Capitalism and the Prospects of an Information Civilization. *Journal of Information Technology* 30: 75–89.
- Zuboff, Shoshana. 2018. *Das Zeitalter des Überwachungskapitalismus*. Frankfurt/M./New York: Campus.
- Zweig, Katharina. 2019. *Ein Algorithmus hat kein Taktgefühl. Wo künstliche Intelligenz sich irrt, warum uns das betrifft und was wir dagegen tun können*. München: Heyne.

---

# **Theoretische und methodische Zugänge zu Künstlicher Intelligenz**



# Künstliche Intelligenz als kontingenzerzeugende Technik: Eine praxistheoretische Perspektive

Michael Heinlein

## Zusammenfassung

Der Beitrag entwickelt eine praxistheoretische Perspektive, um das technische Wirken subsymbolischer Künstlicher Intelligenz (KI) zu analysieren und die soziologische Relevanz und Spezifik dieser relativ neuen, zur Interaktion mit Menschen fähigen Form von KI herauszuarbeiten. In der soziologischen Diskussion ist noch keineswegs geklärt, wie die Nutzung von Techniken, die auf Deep Learning und Künstlichen Neuronalen Netzen aufbauen, zu verstehen und in ihren Folgen für gesellschaftliche und soziale Zusammenhänge zu bewerten ist. Der Beitrag schlägt vor, das Wirken interaktiver KI als Erzeugung von Kontingenz in Praxiszusammenhängen zu begreifen. Unterschieden werden dabei drei Formen: agentielle Kontingenz, epistemische Kontingenz und formative Kontingenz. Vor diesem Hintergrund ergeben sich neue Perspektiven auf die Einbettung und Reflexion von KI in der Nutzungspraxis.

## Schlüsselwörter

Akteur-Netzwerk-Theorie • Kontingenz • Künstliche neuronale Netze • Rekursivität • Relationalität • Soziotechnische Praxis • Technisches Wirken

---

M. Heinlein (✉)

Institut für Sozialwissenschaftliche Forschung e.V. - ISF München, München, Deutschland  
E-Mail: [michael.heinlein@isf-muenchen.de](mailto:michael.heinlein@isf-muenchen.de)



## 1 Künstliche Intelligenz und Kontingenz

Auch wenn sich die deutschsprachige Soziologie bis in die 1990er Jahre hinein nur zögerlich mit den „Produkten und Problemen“ (Rammert 1995) der Künstlichen Intelligenz (KI) befasst hat,<sup>1</sup> so ist diese Technik in den letzten Jahren verstärkt in den Fokus soziologischer Forschung gerückt. Ein Blick auf diese Forschung zeigt, dass vor allem neuere Formen von KI die soziologische Diskussion herauszufordern scheinen:<sup>2</sup> Wie die Nutzung von Techniken, die sich der sogenannten konnektionistischen bzw. subsymbolischen KI zurechnen lassen, zu verstehen und in ihren Folgen für gesellschaftliche und soziale Zusammenhänge zu bewerten ist, ist noch keineswegs geklärt.<sup>3</sup> Entscheidend bei dieser Technik ist, dass ihr technisches Wirken auf der Berechnung von Wahrscheinlichkeiten beruht – und dass dadurch in ihrer Nutzung Situationen entstehen, „in denen nicht mehr so eindeutig zu klären ist, wie und worauf die Maschine reagiert“ und ihr

---

<sup>1</sup> Tendenziell gilt dies auch für die internationale soziologische Literatur, in der sich ab Mitte der 1980er Jahre zunehmend programmatisch ausgerichtete Arbeiten zum Thema finden (z. B. Brent 1986; Collins 1990; Ennals 1991; Gilbert & Heath 1985; Schwartz 1989; Woolgar 1985).

<sup>2</sup> Wie es von einer „multiparadigmatischen Wissenschaft“ (Kneer & Schroer 2009) wie der Soziologie zu erwarten ist, findet sich mittlerweile eine Vielzahl an Perspektiven auf KI. Diese reichen von ethnographischen Studien zur Sozialrobotik (Breazeal 2002; Muhle 2023), die die verkörperte Form von KI in den Vordergrund rücken (z. B. Alač 2009, 2016; Bischof 2017; Koolwaay 2018; Straub 2020) bis hin zu Arbeiten, die KI als ein algorithmisches Geschehen fassen, durch das neuartige gesellschaftliche Möglichkeits- und Wirklichkeitsräume eröffnet werden. Dieser Zugang findet sich vorrangig in der Systemtheorie, die sich bereits früh mit der Kommunikation mit Computern auseinandergesetzt hat (u. a. Baecker 2011; Esposito 2014, 2017; Harth & Lorenz 2017) und in Studien zur ‚intelligenten‘ Algorithmisierung von Öffentlichkeit, Macht und Kontrolle, wie sie etwa als neues ökonomisches Prinzip von „data, extraction, analysis“ (Zuboff 2015), als „threat of algocracy“ (Danaher 2016), als (un-)sichtbar von Daten durchzogenes „algorithmic life“ (Amoore & Piotukh 2016) oder als politisch und ethisch problematische Verengung von entscheidungsrelevanten Perspektiven und Zukunftsnarrativen durch KI (Amoore 2020) diskutiert wird.

<sup>3</sup> ‚Die‘ Künstliche Intelligenz gibt es nicht, da unter diesem Sammelbegriff eine Vielzahl von Technologien und Paradigmen (darunter das Paradigma der konnektionistischen bzw. subsymbolischen KI) verhandelt werden. Zudem ist der Begriff der Künstlichen Intelligenz selbst einem historischen Wandel unterworfen, sodass mit ihm in Abhängigkeit von technischen und sozialen Entwicklungen zu verschiedenen Zeiten jeweils Anderes bezeichnet wurde. Es gibt mittlerweile eine Fülle an Literatur, die sowohl den Begriff der Künstlichen Intelligenz als auch die unter diesem Begriff firmierenden Technikformen erläutern und historisch einordnen – einen solchen Überblick kann dieser Beitrag nicht leisten. Stellvertretend für viele seien daher an dieser Stelle Mainzer (2019), Rosengrün (2021) und Seising (2021) genannt.

„Output durchaus nicht eindeutig als das Ergebnis einer maschinellen Operation zu entschlüsseln ist.“ (Nassehi 2019, S. 223) Im Folgenden wird argumentiert, dass *Kontingenz* damit zu einem entscheidenden Merkmal subsymbolischer KI wird – und zwar nicht im Sinne kontingenter Zuschreibungen an Technik, sondern als eine Eigenschaft des technischen Wirkens selbst.

Der Fokus des theorieorientierten Beitrags liegt auf interaktiven KI-Systemen, mit denen Menschen unter den beschriebenen Uneindeutigkeitsbedingungen in eine Beziehung treten können und denen Möglichkeiten der „symbolisch vermittelten Kommunikation“ (Rammert 2007, S. 32) eingeschrieben sind. Technik wird damit zu einem Gesprächspartner, der jedoch bisweilen nicht mehr klar als Technik erkennbar ist. Darunter fällt eine große Bandbreite von Anwendungen, denen gemein ist, dass sie auf menschliche Sprache, Schrift oder Bewegungen reagieren und ihr Verhalten darauf abstimmen können. Beispiele dafür sind Sprachassistenten wie Siri oder Alexa, die auf gesprochene Befehle reagieren und eine Vielzahl von Aufgaben ausführen; Assistenzsysteme, die bei der Erledigung von Aufgaben (z. B. Terminplanung oder Bestellungen) helfen; Chatbots, die – wie ChatGPT oder Google Bard – anhand von menschlichen Vorgaben Texte erstellen oder von Menschen erstellte Textfragmente ergänzen; Empfehlungssysteme, die anhand der Daten und des Verhaltens von Nutzenden personalisierte Empfehlungen (z. B. Film- oder Musiktipp) geben; Hausautomatisierungssysteme, die Haushaltsgeräte und -systeme wie Kühlschränke, Heizungen und die Beleuchtung steuern; Bilderkennungssysteme, die Objekte und Gesichter auf Fotos oder Videos identifizieren; Roboter, die Umgebungen und Bewegungen analysieren, um mit Menschen zusammenzuarbeiten; oder mobile Geräte, die in komplexen Umwelten ihre Routen eigenständig festlegen und sich selbständig fortbewegen.

Die These des Beitrags lautet, dass solche interaktiven, auf maschinellem Lernen beruhende KI-Anwendungen Kontingenzen in soziale Prozesse einführen, die bei der Einbettung und Nutzung dieser Technik reflektiert werden müssen. Damit ist gemeint, dass das technische Wirken von KI nicht vollständig vorhersehbar ist und Folgen zeitigt, die auf die Emergenz neuer Wirklichkeits- und Möglichkeitsräume des Wissens und Handelns verweisen. Dies erfordert mit Blick auf den Technikbegriff ein Umdenken: Versteht man datengetriebene Maschinen, die sich an Wahrscheinlichkeiten orientieren, als „nicht-triviale Maschinen“, die in ihrem Wirken etwas tun, „das auch anders hätte ausfallen können“ (Nassehi 2019, S. 131), dann kommt man nicht umhin, ihnen eine gewisse Kontingenz zuzugestehen – also etwas, das historisch bislang „der menschlichen Black Box“ (ebd.) zugerechnet wurde und nicht einer Technik, von der man gewöhnt war, dass zwischen dem, was man der Maschine vorgibt und dem, was sie ausgibt eine feste

Beziehung besteht. Als fortgeschrittene digitale Technik zeichnet KI demgegenüber aus, dass sie „tatsächlich Überraschungen erzeugt, will heißen: unerwartete oder nicht eindeutig kalkulierbare Ergebnisse produziert.“ (ebd., S. 222) Um diese These einer KI-getriebenen Kontingenzerzeugung in sozialen Praktiken zu entfalten, wird zunächst in die Funktionsweise subsymbolischer bzw. konnektionistischer KI (beide Begriffe rekurrieren auf dasselbe technische Prinzip, betonen jedoch unterschiedliche Aspekte davon) eingeführt, bevor im Anschluss daran die Argumentationslinie des Beitrags skizziert wird.

Was meint subsymbolische bzw. konnektionistische KI? Beide Begriffe werden synonym verwendet und verweisen darauf, dass die unter diesem Paradigma programmierten Softwaresysteme „de-emphasize the explicit use of symbols in problem solving. Instead, they hold that intelligence arises in systems of simple, interacting components (biological or artificial neurons) through a process of learning or adaptation by which the connections between components are adjusted. Processing in these systems is distributed across collections or layers of neurons.“ (Luger 2005, S. 453) Semantische Repräsentationen, wie sie in der menschlichen Sprache angelegt und Gegenstand der symbolischen KI<sup>4</sup> (z. B. Expertensysteme) sind, spielen bei subsymbolischer KI, die ab dem Ende der 1980er Jahre in der KI-Forschung zunehmend an Relevanz gewonnen hat (vgl. Smolensky 1988; Fodor & Pylyshyn 1988), keine wesentliche Rolle. Entscheidend ist vielmehr die algorithmische Struktur, die mit dem Begriff ‚konnektionistisch‘ näher bestimmt wird: Die verwendeten Algorithmen sind funktional gruppiert (‚Neuronen‘) und schichtartig so miteinander verschränkt, dass ihre Operationen – nach dem Vorbild biologischer neuronaler Netze – in vernetzter Weise ablaufen. Die funktionale Zuordnung der algorithmischen Schichten unterscheidet sich auf den ersten Blick nicht grundlegend vom Aufbau herkömmlicher Computerprogramme: Der ‚input layer‘ eines Künstlichen Neuronalen Netzes nimmt Daten aus der Umwelt auf, während der ‚output layer‘ die Ergebnisse des Netzes für die Umwelt verfügbar macht. Dazwischen liegende Schichten, die sogenannten ‚hidden layers‘, führen Berechnungen mit den Eingabedaten durch und erzeugen Ergebnisse, die von der Ausgabeschicht verwendet werden. Neu an der algorithmischen Architektur konnektionistischer bzw. subsymbolischer KI ist jedoch die ihr eingeschriebene Variabilität: Die Vernetzung der Funktionen und Prozesse erfolgt in Abhängigkeit von Dateninputs, die auf der Basis von Wahrscheinlichkeiten so lange klassifiziert

---

<sup>4</sup> Das Paradigma der symbolischen KI geht davon aus, „that the correct level at which to model the mind is that of the *symbol* – that is, an entity in a computer program that is taken to refer to some entity in the real world.“ (Chalmers 1992, S. 26; Hervorh. i. Orig.).

werden, bis eine relativ stabile Netzwerkstruktur erreicht ist.<sup>5</sup> Das Vorgehen subsymbolischer KI weist in diesem Sinne Merkmale eines „theorielosen Probierens“ (Brödner 2022, S. 34) auf: Sie kommt ohne Vorwissen über den zu verarbeitenden Gegenstand aus und fokussiert allein auf Korrelationen, die anhand einer Datengrundlage berechnet werden. Auf diese Weise erzeugen komplexe und variable Inputs (wie etwa Bilder, Sprache oder Bewegungen) mit einer gewissen Wahrscheinlichkeit einen anschlussfähigen Output, ohne dass Input und Output in der Programmierung vorab festgelegt werden müssten.<sup>6</sup> Damit Künstliche Neuronale Netze ihre Aufgabe erfüllen können, müssen sie jedoch in den meisten Fällen vorab mit Daten trainiert werden. Erst, wenn subsymbolisch operierende KI-Systeme zuverlässig Muster in bekannten Daten erkennen und zufriedenstellende Ergebnisse in Trainingsumgebungen liefern, können sie auf neue, dabei aber ähnliche Daten und Situationen angewendet werden. Dort entwickeln sie ihre Strukturen eigenständig weiter bzw. passen sich an die für das KI-System relevanten Gegebenheiten der Praxis an.

Die Art und Weise, wie die algorithmische Komplexität subsymbolischer bzw. konnektionistischer KI erzeugt wird, lässt sich vor diesem Hintergrund als Bottom-up-Ansatz beschreiben: „One of the most attractive features of connectionist learning is that most models are data or example driven. That is, even though their architectures are explicitly designed, they learn by example, generalizing from data in a particular problem domain.“ (Luger 2005, S. 845) Subsymbolischer KI sind demnach Freiheitsgrade eingeschrieben, die dem Aufbau ihrer internen Strukturen dienen. Wirksam werden diese Freiheitsgrade in der – qua Programmierung stets selektiven – Auseinandersetzung mit einer Praxis, anhand derer sich die algorithmischen Strukturen bewähren oder aber weiterentwickelt

---

<sup>5</sup> Symbole werden in dieser vernetzten Funktionsweise allein als Ergebnis sichtbar: Konnektionistische KI-Systeme „still follow rules, but the rules are well below the semantic level. It is hoped that as a consequence of following rules at this low level, semantic properties will emerge – that is, manifest themselves in the processing and behavior of the program – without having been explicitly programmed in. Consequently, when viewed at the semantic level such systems often do not appear to be engaged in rule-following behavior, as the rules that govern these systems lie at a deeper level.“ (Chalmers 1992, S. 26; siehe auch Smolensky 2012).

<sup>6</sup> Dieser Prozess der Gewichtung und Strukturierung der algorithmischen Relationen kann von außen nicht beobachtet werden, da er in den verborgenen Schichten der KI stattfindet – die Qualitätssteigerung bemisst sich somit allein am von der KI erzielten Ergebnis. Hier setzen die Debatten um eine nachvollziehbare und transparente KI an, die auch deshalb schwer zu realisieren ist, da es sich bei neueren KI-Anwendungen eben um subsymbolisch operierende Systeme handelt.

werden. Diese programmierseitig aufwendig und lang vorzubereitende Anpassungsfähigkeit macht subsymbolische KI-Systeme bis zu einem gewissen Grad sowohl kontextunabhängig (die Möglichkeit der algorithmischen Strukturbildung ist prinzipiell nicht an eine spezifische Klasse von Praxiszusammenhängen gebunden) als auch robust (etwa, wenn die aus einer Praxis aufgenommenen Daten unklar oder ‚verrauscht‘ sind, aufgrund der zahlreich verkoppelten Algorithmen aber dennoch verarbeitet werden können). Die verwendeten Algorithmen können damit für eine Vielzahl von Aufgaben eingesetzt werden, die mit dem Erkennen von Mustern in großen und in der Regel unstrukturierten Datensätzen (‚big data‘) zu tun haben. Subsymbologische KI findet sich dementsprechend in sehr unterschiedlichen Anwendungen – beispielsweise in digitalen Tools mit einer Benutzeroberfläche, im Hintergrund bei bestimmten Programmen mitlaufend oder als Teil der Robotik – und mit einem sehr breiten Anwendungsspektrum.

Der springende Punkt an der ‚Intelligenz‘ subsymbolischer KI ist, dass die dort genutzten Algorithmen unbekannte Daten selbständig strukturieren und ihre programmierten Strukturierungsmethoden selbstreferenziell, d. h. entlang interner, anhand der Auseinandersetzung mit nicht-algorithmischen Umwelten gewonnener Selektivitäten weiterentwickeln können. Wie die Systemtheoretikerin Elena Esposito anmerkt, ist ein ‚intelligenter‘ Algorithmus somit dann erfolgreich, „if it learns to learn by itself, i.e. to develop a practice of unsupervised learning, in which the algorithm does not learn what others teach. Instead it decides autonomously what to learn and what to communicate.“ (Esposito 2017, S. 261) Man muss hinzufügen, dass dies nur für die eng umrissenen Zwecke und Spezifika der Nutzung gelten kann, für die eine KI-Anwendung programmiert wurde. Gleichwohl lässt sich dieses ‚Lernen‘ für unterschiedliche Kontexte nutzbar machen – etwa um unsichere Entscheidungen zu unterstützen, Fahrzeuge in unvorhersehbaren Umwelten zu bewegen oder mit grundsätzlich unberechenbaren (kontingenten) Menschen zu interagieren und zu kommunizieren.

Die algorithmische Komplexität greift auf diese Weise in die soziale Komplexität der Praxis ein, in der KI-Systeme zur Anwendung kommen. Aus soziologischer Sicht besteht eine der wesentlichen Leistungen von KI somit darin, sich als dynamisches Element mit einer „relativen Autonomie“ (Rammert 2007, S. 82) in Praxiszusammenhänge einzufügen und dort Aufgaben mit gewissen Freiheitsgraden zu erfüllen. Relativ ist diese Autonomie aus zweierlei Gründen: Zum einen handelt es sich bei KI immer um programmierte Algorithmen, die auf sehr spezifische Zwecke zugeschnitten sind, d. h. um Systeme, mit „none or very little ability to do anything beyond their particular domain of functionality“ (Dyer-Witheford et al. 2019, S. 10). Die Stärke gegenwärtiger KI liegt demnach in ihrer Spezifität für bestimmte Aufgaben und Kontexte; sie ist damit

trotz ihrer breiten Einsetzbarkeit noch weit davon entfernt, eine „general purpose technology“ (Crafts 2021; Brynjolfsson et al. 2019) zu sein.<sup>7</sup> Zum anderen ist KI durchgängig abhängig von Daten und allen damit verbundenen Voraussetzungen, etwa der Datenqualität. Dazu gehören Trainingsdaten, mit denen die Genauigkeit der verwendeten algorithmischen Strukturen für bestimmte Zwecke vorab gesteigert und optimiert wird, aber auch große Datenbestände, auf die KI-Systeme zugreifen, um das Geschehen in einer Nutzungspraxis strukturieren zu können.

Trotz dieser Relativierungen – und an diesem Punkt setzt der Beitrag an – weisen interaktive KI-Anwendungen eine neue Qualität der Unvorhersehbarkeit<sup>8</sup> auf, die ihr technisches Wirken zu etwas anderem macht als eine feste Koppelung von Abläufen oder eine reine Wiederholung von Vorgängen, wie es bei herkömmlicher Technik der Fall ist: „Das zukünftige Verhalten eines technischen Agenten, sei es ein Roboter oder ein Avatar, lässt sich nicht mit Sicherheit vorhersagen oder berechnen.“ (Rammert 2003, S. 7) Kontingenz, d. h. die Möglichkeit des Anders-möglich-Seins (vgl. Luhmann 1992a; Makropoulos 1997) wird damit zu einem entscheidenden Merkmal des praktischen Wirkens von KI.<sup>9</sup> Das Wirken von KI stellt sich damit nicht als ein zwangsläufiges Geschehen dar, wie es ein klassischer Technikbegriff nahelegen würde,<sup>10</sup> sondern ist immer mit einem „Seitenblick auf andere Möglichkeiten“ (Luhmann 1992b, S. 74), d. h. mit einer systematischen Unbestimmtheit gekoppelt: Ein von KI gestütztes

---

<sup>7</sup> Dies kann sich, wie die aktuelle Entwicklung sogenannter Large Language Models (LLMs) bzw. Transformer-Technologien zeigt, jedoch schnell ändern. Zum Zeitpunkt des Schreibens dieses Textes erhalten diese Technologien deshalb erhöhte Aufmerksamkeit, da sie mit ihrer konnektionistischen Architektur in der Lage sind, mit entsprechendem Training Inhalte (z. B. Texte oder Bilder) zu generieren, die der Performanz menschlicher Akteure erstaunlich nahekommen. Bubeck et al. (2023) machen hier – unter Berücksichtigung bekannter Restriktionen von LLMs und algorithmischer *biases* – sogar „sparks“ einer allgemeinen Intelligenz von KI aus. Ob diese These haltbar ist, wird sich nach dem Abflauen des Hypes um diese Technik allerdings erst noch zeigen müssen.

<sup>8</sup> Unvorhersehbar ist eine Technik auch dann, wenn sie plötzlich defekt ist oder aufgrund einer Fehlfunktion Fehler produziert – siehe etwa Latour (2002b) für das Versagen eines Overhead-Projektors oder Orr (1996) für das Irritationspotenzial von Fotokopierern. Dies gilt grundsätzlich auch für KI, ist hier aber nicht mit ‚neuer Qualität der Unvorhersehbarkeit‘ gemeint, die auf die dieser Technik intentional eingeschriebenen Potenziale abhebt.

<sup>9</sup> Im Folgenden meint KI immer die beschriebene Form der konnektionistischen bzw. sub-symbolischen KI.

<sup>10</sup> Es stellt sich die Frage, inwieweit ein ‚klassischer‘ soziologischer Technikbegriff im Kontext von KI überhaupt noch tragfähig ist. Die Überlegungen in diesem Beitrag können auch als Versuch gelesen werden, den Technikbegriff um Dimensionen der Kontingenz zu erweitern.

Wissen und Handeln unterliegt einer bis zu einem gewissen Grad kontingenten soziotechnischen Dynamik, die sich entlang der situativen und a priori nicht vollends festlegbaren Verbindungen von Mensch und Technik beobachten lässt. Eine fortschreitende Digitalisierung hat damit den eigentümlichen und soziologisch herausfordernden Effekt, „dass immer mehr Techniken entstehen, die tatsächlich in soziale Prozesse eingreifen, die Handlungsverläufe mitstrukturieren und mit denen Menschen in eine Situation doppelter Kontingenz geraten.“ (Nassehi 2019, S. 224) Die soziotechnische Praxis, an der KI als ein dynamischer Bestandteil mitwirkt, erhält in diesem Sinne eine Art Kontingenzschub, der sich – wie weiter unten zu sehen sein wird – entlang der Schnittstellen und Relationen von Mensch und ‚intelligenter‘ Technik aufschlüsseln lässt.

Diese „Kontingenzperspektive“ (Reckwitz 2004) auf das Wirken von KI vermag zweierlei zu leisten: Zum einen stellt sie ein Korrektiv für bestehende Diskurse um KI als menschenähnlicher, gar posthumaner Akteur oder als hochentwickeltes, gleichwohl passives Werkzeug dar. Beide Diskurse setzen spezifische technische Rationalitäten und Technikbegriffe voraus und imaginieren KI auf eine Weise, in der alles bereits entschieden ist: KI kann entweder immer nur handelnder Akteur oder handlungsunfähige Technik sein. Die Kontingenz des Wirkens von KI wird dabei sowohl überschätzt (KI als Akteur) als auch unterschätzt (KI als Werkzeug). Zum anderen invisibilisiert die ‚künstlich intelligente‘ Praxis mit KI selbst ihre eigene Kontingenz (und damit auch das kontingenzerzeugende Potenzial von KI), indem sie – wie jede temporale Praxis – als eine Praxis abläuft, die Möglichkeiten in Notwendigkeiten transformiert und offene Zukunftshorizonte in Bedingungen der Vergangenheit verwandelt (vgl. Heinlein 2022). Dass das technische Wirken von KI Möglichkeitsräume für ihre eigene soziotechnische Praxis schafft, gerät dabei aus dem Blick – oder besser gesagt: lässt sich praktisch in der Regel gar nicht erst in den Blick nehmen. Eine Kontingenzperspektive auf KI hingegen vermag auch praktisch erzeugte Selbstverständlichkeiten und Notwendigkeiten der KI-Nutzung aufzubrechen, sodass dem Nachdenken über die nachhaltige Einbettung und Nutzung von KI in sozialen Zusammenhängen neue Impulse gegeben werden können.

Vor diesem Hintergrund entwickeln die folgenden Ausführungen ein praxistheoretisches Verständnis von KI, das Konzepte der Akteur-Netzwerk-Theorie aufgreift und mit Überlegungen zu ‚intelligenter‘ Technik und Fragen der Kontingenz verbindet. Im Fokus steht dabei eine Beschreibung von KI, die sie als eine *relationale* und *relationierende* Technik sichtbar macht. Damit ist gemeint, dass KI die soziotechnischen Zusammenhänge, in die sie sich einfügt, zugleich auf kontingente Weise miterzeugt und mitverändert – sie ist in diesem Sinne als eine *kontingenzerzeugende Technik* begreifbar. Um dieses Argument zu entfalten,

entwickelt Kap. 2 entlang akteurnetzwerktheoretischer Figuren eine Perspektive auf KI, die die Technik als wirkmächtigen Bestandteil sozialer Praktiken sichtbar macht. Dies mündet in ein relationales Verständnis von KI (Kap. 3), das die Basis für die Unterscheidung dreier Kontingenzformen – epistemische Kontingenz, agentielle Kontingenz und formative Kontingenz – abgibt (Kap. 4). Das Zusammenspiel dieser Formen, so die These, zeichnet das Wirken von KI in einer soziotechnischen Praxis aus und hat Folgen dafür, wie KI in einer Nutzungspraxis sozial eingebettet und reflektiert werden kann (Kap. 5).

---

## 2 Eine symmetrische Beschreibung von Künstlicher Intelligenz

Eine praxistheoretische Perspektive auf KI bietet sich aus zwei Gründen an. Ein erster, wissenssoziologisch legitimierbarer Grund besteht darin, dass KI in öffentlichen, darunter auch wissenschaftlichen Diskursen bevorzugt in praxisbezogenen Begriffen beschrieben wird: KI lernt, denkt, handelt, nimmt wahr, analysiert, sortiert, entscheidet, beobachtet etc. Auch wenn man gut daran tut, vermenschlichenden Deutungen von KI mit Vorsicht zu begegnen, lassen sie sich doch als ein Hinweis darauf lesen, dass KI mit einem spezifischen Wirken assoziiert wird, das über die der Technik innerhalb der Soziologie gerne zugeschriebene Rolle eines „neutral means“ (Rammert 1999, S. 168), d. h. eines neutralen Werkzeugs hinausgeht.<sup>11</sup> Eine theoretische Perspektive auf KI muss ihr technisches Wirken daher jenseits von Anthropomorphismen, aber auch technizistischen Verkürzungen und Black Boxing fassen und begreifen können. Ein zweiter Grund findet sich in dem der KI eingeschriebenen Praxisbezug selbst. Wie dieser Praxisbezug aussieht, wird deutlich, wenn man einen Algorithmus als ein „state transition system“ versteht, „that starts in an initial state and transits from one state to the next until, if ever, it stops or breaks. [...] In particular, a sequential-time interactive algorithm [...] is a state transition system where a state transition may be accompanied by sending and receiving messages.“ (Gurevich 2012, S. 37) Interaktive Algorithmen, wie sie auch in KI-Systemen zur Anwendung kommen, verfügen über eine ihnen eingeschriebene Offenheit, die praktisch hergestellt wird und sich als Abfolge von Übergängen beschreiben lässt: Interne Zustandsübergänge

---

<sup>11</sup> So hat der Wissenschafts- und Techniksoziologe Steve Woolgar bereits 1985 angemerkt, dass KI eine gute Gelegenheit darstelle, über quasi-naturalisierte soziologische Gewissheiten – etwa, dass soziales Handeln genuin menschliches Handeln sei – nachzudenken (Woolgar 1985).



beruhen auf einem Austausch in einer Praxis, der zu unterschiedlichen Informationsniveaus zwischen dem Anfangs- und dem Endpunkt prinzipiell unendlicher Austauschprozesse führt. Dies gilt allerdings nicht nur für den Algorithmus, sondern auch für die Praxis selbst, da sich diese mit dem Wirken des Algorithmus verändert, d. h. eigenlogische Zustandsübergänge durchläuft. Stellt man die weiter oben beschriebene ‚Intelligenz‘ eines Algorithmus bzw. Systems von Algorithmen in Rechnung, dann sind diese Praxisübergänge im Falle von KI nicht immer vorhersehbar. Die programmierte Offenheit interaktiver Algorithmen erhält durch die Fähigkeit ‚intelligenter‘ Algorithmen zur selbstreferentiellen Eigenstrukturierung ihrer Vernetzungen eine andere, kontingenzerzeugende Qualität: Die Praxis muss mit der Möglichkeit nicht erwartbarer Informationen und Verhaltensweisen rechnen, deren Zustandekommen sie selbst beeinflusst hat. Die Herausforderung besteht somit darin, die Praxis von KI als ein dynamisches Geschehen zu beschreiben, in dem sich soziale und technische Elemente wechselseitig durchdringen und beeinflussen.<sup>12</sup> Dazu muss aber zunächst das Verhältnis von Technik und Sozialem geklärt und ein begriffliches Instrumentarium eingeführt werden, das ihre Relationierung erlaubt.

Technik wurde in der Soziologie bereits in ihren Anfangstagen thematisiert. Im Vordergrund stand dabei die Trennung einer sozialen Welt von einer Welt des Technischen mitsamt ihrer Artefakte und Materialitäten: „Sociology has [...] consolidated itself [...] by marking some areas of interest as outside its disciplinary boundaries. Here anything non-social or nonhuman, including the natural world and the world of technological artifacts, has been made Other.“ (Lee & Brown 1994, S. 774) Dass die Welt der Technik – insbesondere im Sinne einer konkreten technischen Materialität – von Beginn an jenseits der Grenzen der Soziologie lag, ist demnach kein Zufall, sondern hat auch strategische Gründe: Um ihren Gegenstand, das Soziale, im Unterschied zu einer stummen und ‚objektiven‘ Welt des Natürlichen und Technischen zu fassen und zu legitimieren, musste sich die noch junge Soziologie Ende des 19. Jahrhunderts aktiv von der Biologie und der Physik

---

<sup>12</sup> Wesentlich dafür ist, die praktischen Bedingungen von Handeln und Wissen herauszuarbeiten, die sich mit dem Wirken von KI einstellen und etablieren. Wenn daher im Folgenden die Akteur-Netzwerk-Theorie diskutiert wird, dann geschieht dies im Sinne einer Praxistheorie, der ein Praxisbegriff zugrunde liegt, der auf eine soziotechnisch erzeugte und sich dynamisch verändernde Ermöglichungsstruktur von Handlungen und Kommunikation abhebt. Der Begriff des Wirkens, der auf die transformativen Effekte von KI in ebendieser Praxis abstellt, dient dazu, apriorische Zuschreibungen von Handlungsfähigkeit oder Intentionalität an KI und eine damit verbundene Anthropomorphisierung einer menschengemachten Technik zu vermeiden. Zugleich wird damit der Umstand markiert, „dass Handlung und Technik jeweils nur in Zusammenhängen emergieren“ (Rammert 2016, S. 105), die im Falle von KI sowohl Erwartbares als auch Unerwartbares produzieren.

abgrenzen. Der Eigencharakter und die Materialität technischer Artefakte spielten und spielen – dieses Erbe trägt der Mainstream der Soziologie auch heute noch weiter – keine wesentliche Rolle für soziologisches Denken, zumindest nicht in dem Sinne, dass sie als eigenständige Elemente des Sozialen gesehen würden: Technik verweist in aller Regel auf ein „außersoziales Phänomen“ (Halfmann 1995, S. 10).

Dass es diese Grenzen und unterschiedlichen Interessensgebiete gibt, kann demnach als Ergebnis eines intentionalen Schließungsprozesses, einer „labour of division“ (Munro 1997) der Soziologie gelesen werden. Diese Bemühung wird unter anderem bei Émile Durkheim deutlich, der Technik im Sinne eines soziologischen Tatbestands gefasst hat. Technik erhält in dieser Lesart ihren sozialen Sinn im Zusammenhang mit einer „mehr oder minder festgelegte[n] Art des Handelns, die die Fähigkeit besitzt, auf den Einzelnen einen äußeren Zwang auszuüben; oder auch, die im Bereich einer gegebenen Gesellschaft allgemein auftritt, wobei sie ein von ihren individuellen Äußerungen unabhängiges Eigenleben besitzt.“ (Durkheim 1961, S. 114) Eine entsprechend ‚soziologisierte‘ Technik kann dann auch zur Erklärung gesellschaftlicher bzw. sozialer Zwänge herangezogen werden (z. B. Linde 1972). Der Techniksoziologe Werner Rammert interpretiert diesen Aspekt der Durkheimschen Soziologie dahingehend, dass damit „einerseits die Schließung der Soziologie für das rein Soziale, andererseits ihre Öffnung für die Analyse von Technostrukturen als besonderes Substrat von Sozialstrukturen“ (Rammert 1998, S. 13) möglich wird. Technische Artefakte werden in dieser Lesart nicht als für sich stehende und wirkende Elemente des Sozialen begriffen, sondern allein aus der Perspektive des Sozialen gedacht. Sie sind als „besonderes Substrat“ des Sozialen auch dessen Logik verpflichtet und nur auf diese Weise ein legitimer Gegenstand der soziologischen Beobachtung. Um technische Dinge im Gegensatz dazu stärker in ihrer konstitutiven Rolle und Funktion für soziale Zusammenhänge zu begreifen, wurde im Rahmen der Akteur-Netzwerk-Theorie das sogenannte verallgemeinerte Symmetrieprinzip entwickelt (programmatisch: Callon 1986; Latour 1998, 2005). Die Verallgemeinerung beruht auf einer Ausdehnung des Symmetrieprinzips des ‚Strong Programme‘, einer Ende der 1970er Jahre entwickelten Perspektive der Wissens- und Wissenschaftssoziologie, die zum besseren Verständnis des Folgenden kurz skizziert werden soll.

Das Ziel des wissenssoziologischen Symmetrieprinzips bestand darin, wissenschaftliches Wissen nicht als eine privilegierte, von anderem Wissen abgehobene Wissensform zu begreifen, sondern auf die gleiche Weise zu beschreiben und zu analysieren, wie die Wissenssoziologie dies für alltags- und lebensweltliches Wissen tut. Der Wissenssoziologie geht es in erster Linie darum, die Ordnung

des Wissens vor dem Hintergrund gesellschaftlicher Strukturen zu beschreiben (Berger und Luckmann 1969). Gesellschaftliche und soziale Verhältnisse werden in dieser Lesart zu den Bedingungen der Möglichkeit von Wissen und der Begriffe, mit denen Wissen sprachlich gefasst wird. Wissenschaftliche Begriffe sind, genauso wie Alltagsbegriffe, von dieser genuin sozialen Interpretation von Wissen nicht ausgeschlossen. David Bloor, der das ‚Strong Programme‘ der Edinburgh School begründet hat, schreibt dazu: „All knowledge, whether it be in the empirical sciences or even in mathematics, should be treated, through and through, as material for investigation. Such limitations as do exist for the sociologist consist in handing over material to allied sciences like psychology or in depending on the researches of specialists in other disciplines. There are no limitations which lie in the absolute or transcendent character of scientific knowledge itself, or in the special nature of rationality, validity, truth or objectivity.“ (Bloor 1976, S. 1) Scheinbar objektive und naturwüchsig gegebene Grenzen, zu deren Reproduktion, wie Bloor bemerkt, auch die Soziologie beigetragen hat, werden nun aufzubrechen versucht. Begriffe und Eigenschaften wie Rationalität, Validität und Objektivität verwandeln sich damit in legitime Gegenstände der wissenssoziologischen Analyse. Damit werden, wie Michel Callon – einer der Begründer der Akteur-Netzwerk-Theorie – hervorhebt, wissenschaftliche Rationalitätsansprüche dekonstruiert: „The sociologists act impartially and refer to the different protagonists in the same terms, even if one among them succeeds in imposing his will. The sociologists attribute the actors with neither reason, scientific method, truth, nor efficiency because these terms denote the actor’s success without explaining the reasons for it.“ (Callon 1986, S. 197) Zwischen ‚rationalem‘ und ‚irrationalem‘ Wissen besteht mit dieser Symmetrisierung von Wissen auf analytischer (!) Ebene kein Unterschied. Anstatt von einer Trennung dieser beiden Wissensformen auszugehen, wird die Möglichkeit dieser Trennung selbst zum erklärungsbedürftigen Phänomen: Rationales Wissen wird, genauso wie irrationales Wissen, als ein Effekt sozialer Prozesse begriffen und nicht als selbstverständliches Ergebnis rationalen und methodisch-genauen Abbildens prinzipiell erkennbarer Objekte der Welt des Natürlichen und des Technischen. Sprachlich umgesetzt wird dieses symmetrische Programm durch ein gemeinsames analytisches Vokabular: Die an der Wissensgenese beteiligten Akteure und Gruppen müssen mit den gleichen Begriffen beschrieben werden. Doch dieses Symmetrieprinzip kann – als Element eben einer ‚klassischen‘ Wissens- und Wissenschaftssoziologie – den Anspruch, symmetrisch zu sein, mit Blick auf den Eigencharakter technischer Artefakte nur vor dem Hintergrund einer Asymmetrie einlösen. In sozialkonstruktivistischen Ansätzen, die dem Bloorischen Symmetrieprinzip folgen, wird das Soziale (in diesem Falle: die institutionalisierte

Forschung) als konflikthafte und interessengeleitetes Verhältnis verschiedener Akteure bzw. relevanter sozialer Gruppen modelliert. Dieses Verhältnis bestimmt das Produkt der Wissenschaften, d. h. die Form und den Inhalt wissenschaftlichen Wissens und technischer Innovationen. Bestimmte Akteure können in einem gewissen Rahmen – dem „technological frame“ (Bijker 1992, 1995) – ihre Vorstellungen, Interpretationen und Interessen gegenüber anderen Akteuren behaupten und auf diese Weise Gestaltungs- und Schließungsmöglichkeiten der Technik bzw. des technisch-wissenschaftlichen Wissens beeinflussen (mit Bezug auf KI: Hirsch-Kreinsen 2023). Technische Artefakte und wissenschaftliche Aussagen stellen in dieser Perspektive das Ergebnis eines zunächst offenen sozialen Aushandlungsprozesses dar. Die so eingeführte Kontingenz wird im Sozialen als „interpretative flexibility of technological artefacts“ (Pinch & Bijker 1984, S. 419) konzeptualisiert, die jedoch das grundlegende Verhältnis der sozialen Welt zur Welt der Objekte und der Natur unberührt lässt. Dies bedeutet, „dass wissenschaftliche ‚Fakten‘ stets für mehr als nur eine einzige Interpretation offen sind. Es sind soziale Faktoren, die den Interpretationsüberschuss limitieren und für eine ‚Schließung‘ von Diskursen sorgen.“ (Degele 2002, S. 100) Objekte dienen damit, ganz im Sinne Durkheims, als „Leinwand, auf die die Gesellschaft ihren Film projiziert“ (Latour 1998, S. 73) – und dies ist auch nicht verwunderlich, da Bloor sich selbst immer wieder auf Durkheim beruft (z. B. Bloor 1998).

Die Symmetrie des ‚Strong Programme‘ lässt nicht an der Rolle und der Logik der Gesellschaft bzw. des Sozialen zweifeln und setzt diese für die Soziologie als gegeben und grundsätzlich verstehbar voraus. War es vorher noch eine sich auf eine objektive Natur berufende Repräsentationslogik der Naturwissenschaften, die das Symmetrieprinzip aushebeln sollte, ist es nun die Soziologie, die ihre „area of interest“ (Lee & Brown 1994, S. 774) als prinzipiell verhandlungsgeschlossen sieht: „The sociologist tends to censor selectively the actors when they speak of themselves, their allies, their adversaries, or social backgrounds. He allows them to express themselves freely only when they speak of Nature.“ (Callon 1986, S. 198) Vor diesem Hintergrund erscheint nun jedoch das Begriffsrepertoire der Soziologie als problematisch: Woher nimmt die Soziologie die Gewissheit, dass es ihre Sicht der Dinge ist, die nicht zur Debatte stehen kann? Die frühen Schriften der Akteur-Netzwerk-Theorie, zu denen die zitierten Texte von Callon (1986) und Latour (1998) zählen, sind durchzogen von der Skepsis gegenüber einer Soziologie, die die Welt der natürlichen Dinge und der technischen Artefakte aus ihren Überlegungen zur Zusammensetzung des Sozialen ausschließt. Hier setzt die *Verallgemeinerung* des Symmetrieprinzips an, die Michel Callon folgendermaßen zusammenfasst: „The goal is not only to explain conflicting viewpoints and arguments in a scientific or technological controversy in the same terms.

We know that the ingredients of controversies are a mixture of considerations concerning both Society and Nature. For this reason we require the observer to use a single repertoire when they are described. [...] [T]he rule which we must respect is not to change registers when we move from the technical to the social aspects of the problem studied.“ (Callon 1986, S. 200) Tradierte soziologische Beschreibungen von Technik werden mit der Verallgemeinerung des Symmetrieprinzips infrage gestellt. Es geht nicht mehr nur darum, soziologisch zu beobachten und zu beschreiben, wie Akteure über technische Zusammenhänge sprechen und wie technischen Objekten ein Platz in einer sozial determinierten Ordnung der Dinge, des Wissens, der Diskurse etc. zugewiesen wird. Die Analyse technologischer Debatten ist nach wie vor wichtig, jedoch erhalten Artefakte in diesen Debatten mit einer symmetrischen Beschreibung einen anderen Stellenwert. Das verallgemeinerte Symmetrieprinzip eröffnet der soziologischen Beobachtung die Möglichkeit, *in actu*, d. h. in konkreten soziomaterialen und soziotechnischen Praktiken nachzuvollziehen, wie die Grenzen von Technik und Gesellschaft, Objekten und Subjekten in einer von beiden Entitäten bevölkerten Welt gezogen und konstruiert werden, *ohne* von einer apriorischen Trennung dieser beiden Sphären auszugehen. Anstatt also die Bedingung von Technik und Gesellschaft als soziale Konstruktion allein im Sozialen zu suchen, rückt das verallgemeinerte Symmetrieprinzip das Problem der gemeinsamen Bedingungen der Möglichkeit von Sozialem und Nicht-Sozialem in den Mittelpunkt (vgl. Latour 1998, 2005). In methodischer Hinsicht wird dieser Zweifel am Wesen des Technischen *und* des Sozialen – sowie am Verhältnis beider Größen – ganz im Sinne des ersten Symmetrieprinzips in ein gemeinsames analytisches Vokabular übersetzt: Die Beschreibung der Gesellschaft und der Welt der Dinge, des Natürlichen und der Technik muss in ein- und denselben Begriffen möglich sein. Die Verallgemeinerung des Symmetrieprinzips verschiebt auf diese Weise den Problemfokus: Vorher waren es normative Unterscheidungen wie ‚rational‘ und ‚irrational‘, die ein apriorisch bestimmtes Soziales als Effekte hervorbrachte und so eine unbestimmte Welt der Natur und der Technik vor dem Hintergrund einer prinzipiell berechenbaren und bekannten Gesellschaft sozial konstruierte. In der verallgemeinerten Version wird nun die Möglichkeit der (von der Soziologie in aller Regel normativ betriebenen) Trennung von Sozialem und Nicht-Sozialem in den Blick genommen und problematisiert.

Auf analytischer Ebene bedeutet diese Symmetrisierung von Subjekten und Objekten, dass auch technischen Artefakten eine aktive, mitunter handelnde Rolle zugesprochen wird: In den Blick kommen alle menschlichen und technischen „entities that *do things*“ (Latour 1988, S. 303; Hervorh. i. Orig.) und die in der Lage sind, in ihren jeweiligen Zusammenhängen transformativ zu wirken. Da bei KI

schnell in die Denkfalle des Anthropomorphismus getappt ist, gilt es zweierlei zu berücksichtigen: Zum einen darf konzeptionell nicht übersehen und empirisch nicht unterschätzt werden, dass es sich bei KI um eine Technik handelt, die qua Programmierung immer (inter-)aktiver wird – also um etwas Nicht-Menschliches, das zwar von Menschenhand eingeschriebene, aber eben doch quasi-menschliche Potenziale in konkreten Praxiszusammenhängen besitzt: „Sprachausgaben, agentenorientierte Programmierung und in Robotern verkörperte Intelligenz verleihen den technischen Artefakten ein weiteres Wirkungsspektrum, einen größeren Aktionsradius und eine feinere Aktions- und Interaktionsfähigkeit. Die Frage nach der Technik ist angesichts dieser aktiver werdenden und sich mit menschlichen Handlungseinheiten enger verbindenden Artefakte neu zu stellen.“ (Rammert & Schulz-Schaeffer 2002, S. 12). Zum anderen darf die Frage nach dem Wirken von KI nicht gleichgesetzt werden mit einer pauschalen Zuschreibung von Handlungsfähigkeit und Intentionalität an technische Artefakte, wie sie bereits vielfach kritisiert wurde (u. a. Bloor 1998; Collins & Yearley 1992; Ropohl 2005, S. 399 f.). Die Aufgabe besteht demgegenüber darin, genauer zu bestimmen, auf welche Weise sich eine künstlich intelligente Technik in Praxiszusammenhänge einfügen und wie sich diese Zusammenhänge dadurch als Wirklichkeits- und Möglichkeitsräume, innerhalb derer gehandelt, gedacht, kommuniziert etc. wird, verändern. Nicht die Idee, dass Technik handeln kann, ist normativ (vgl. Rammert & Schulz-Schaeffer 2002, S. 24 ff.) – es geht vielmehr darum, die Normativität des soziologischen Blicks zu überwinden, der die Bedeutung von Technik für soziales Handeln, Wissen und soziale Ordnung in aller Regel unterschätzt. Eine wesentliche Herausforderung, die sich mit einer symmetrischen Beschreibung von KI ergibt, besteht demnach darin, eine Antwort darauf zu geben, wie sich soziale Ordnungen, Prozesse und Strukturen jenseits tradierter kategorialer Zuschreibungen an Gesellschaft und Technik bzw. Subjekt und Objekt denken lassen – und zwar *ohne*, dass genuin menschliche Eigenschaften wie Denken, Handeln, Kommunizieren und Entscheiden vorschnell über Bord geworfen werden.<sup>13</sup> Das folgende Kapitel diskutiert vor diesem Hintergrund die

---

<sup>13</sup> Man kann dies analog zum Verhältnis von Strukturfunctionalismus und Systemtheorie sehen: Während Talcott Parsons den Begriff des sozialen Systems analytisch verwendet, geht Niklas Luhmann davon aus, „dass es Systeme gibt.“ (Luhmann 1984, S. 30) Diese Differenz hat weitreichende Folgen für die Theorieentwicklung und empirische Forschung. In diesem Beitrag wird davon ausgegangen, dass es analytisch Sinn macht, technisches Wirken und menschliches Handeln symmetrisch zu beschreiben – doch dass dies keineswegs bedeutet, dass Technik ‚intelligent‘ ist oder ‚intentional‘ handelt.

praxistheoretischen Konzepte der Relationalität und – damit zusammenhängend – der Rekursivität, die auch im Kontext der Akteur-Netzwerk-Theorie eine zentrale Rolle spielen.

---

### **3 Ein relationales Verständnis von Künstlicher Intelligenz**

Relationale Perspektiven sind für die Soziologie nichts Neues und haben bereits das Denken ihrer ‚Klassiker‘ von Georg Simmel bis Karl Marx bestimmt. Bruno Latours Arbeit an der Akteur-Netzwerk-Theorie als einer „sociology of associations“ (Latour 2005) kann daher zugleich als eine Erweiterung und Neuinterpretation eines tradierten Relationalitätsverständnisses der Soziologie begriffen werden. Akteurnetzwerktheoretische Ansätze heben die Bedeutung der wechselseitigen Verbindung, Beeinflussung und Veränderung von menschlichen und technischen Elementen in einem heterogen gedachten, d. h. aus unterschiedlichen ‚Materialien‘ bestehenden Sozialen hervor. Was genau damit gemeint ist, lässt sich an einem historischen Laborbeispiel, der von Latour interpretierten „Entdeckung/Erfindung/Konstruktion des Milchsäureferments durch Pasteur im Jahre 1857“ (Latour 1996, S. 87), veranschaulichen. Dieses Beispiel zeigt, wie die Verbindungen zwischen menschlichen Akteuren, natürlichen Objekten und technischen Artefakten im Sinne einer ‚symmetrischen‘ Soziologie als dynamische, auf sich selbst zurückwirkende Relationen beschrieben werden können.

Latour setzt die Geschichte von Pasteur und dem Milchsäureferment dezidiert im Kontrast zu ‚traditionellen‘ historischen und soziologischen Beschreibungen an: „In der Begebenheit, im Ereignis liegt [...] immer Reziprozität vor, die mehr oder weniger gleich verteilt ist. De Gaulle verändert Churchill, der ihn seinerseits verändert. Gerade diese Reziprozität scheint aber in der Wissenschaftsgeschichte nicht möglich zu sein, denn dazu müsste man das Ereignis zwischen Pasteur und der Milchsäure aufteilen! Und doch verlangt die verallgemeinerte Symmetrie nach einer solchen Aufteilung. Nicht nur ‚stößt‘ das Ferment Pasteur ‚zu‘ – und verwandelt diesen ehrwürdigen Chemiker aus der Provinz in einen Meister der Mikrobiologie auf der ganzen Welt –, sondern auch Pasteur ‚stößt‘ dem Ferment ‚zu‘ – und verwandelt eine Fermentierung durch Kontaktwirkung in eine Fermentkultur mit Zucker als Nahrungsmittel.“ (Latour 1996, S. 96) Historische und soziologische Beschreibungen behandeln menschliche Akteure symmetrisch, indem sie sie mit der Möglichkeit, sich gegenseitig zu beeinflussen, und in diesem Sinne mit Handlungsmacht ausstattet. Mit dem verallgemeinerten Symmetrieprinzip bleibt dieser Grundsatz der Form nach erhalten,

er wird jedoch anders interpretiert: Aus einer homogenen zwischenmenschlichen Beziehung wird eine Beziehung zwischen verschiedenen, materiell heterogenen Wesen. Plötzlich stoßen das Ferment und Pasteur, das heißt: ein menschliches und ein nicht-menschliches Wesen aufeinander und verbinden, verwandeln und verändern sich in sich mitverändernden Beziehungsgeflechten – mit dem schlussendlichen Effekt, dass Pasteur weltbekannt und das Milchsäureferment zu einer eigenständigen, identifizierbaren Substanz wird, die von Pasteurs Labor aus als zuckerumsetzende Fermentkultur auf der ganzen Welt zu zirkulieren beginnt.

Stellt man die Perspektive von rein sozialen auf soziomateriale Relationen um, dann können sich nicht nur de Gaulle und Churchill durch ihre dynamische Beziehung auf der politischen Weltbühne wechselseitig definieren, sondern auch das Ferment und Pasteur im Labor. Mensch und Nicht-Mensch gehen Relationen ein, in der Identitäten, d. h. identifizier- und benennbare Netzwerkknoten vermittelt und stabilisiert werden. Die Ergebnisse der Vermittlung – das produzierte Wissen, die Zuschreibungen und Identitäten etc. – wirken wiederum auf Form und Inhalt der Relationen zurück und erzeugen auf diese Weise eine rekursive Dynamik fortlaufender wechselseitiger Einflussnahmen und Veränderungen. Es geht damit nicht so sehr um die menschlichen und technischen Elemente an sich (die Subjekte und Objekte, die nach wie vor als Relata erkennbar sind), sondern um die dynamischen Verknüpfungen (Relationen) dieser Elemente, in denen ein fortlaufender Austausch zwischen menschlichen und technischen Gegebenheiten und Möglichkeiten stattfindet.

Die Bedingungen von Identität – im Sinne beobachtbarer und praktisch stabilisierter Eigenschaften und Zurechenbarkeiten – liegen in dieser Lesart nicht mehr nur in einer rein sozialen, d. h. homogenen Vermittlung, sondern in einer Vermittlung, die heterogene Bestandteile umfassen kann: „The argument is that the networks of the social come in a variety of material forms: for instance, people, texts, machines, architectures. But (here is the symmetry) these materials are not given in nature, but are more or less precarious ordering effects which express themselves in different ways, including that of durability.“ (Law 1994, S. 95) Die Ordnung, die in den Relationen zwischen Menschen und Dingen entsteht, stellt demnach einen kontingenten Effekt dar: Ob technische Artefakte und Menschen eine Rolle spielen bzw. Eigenschaften aufweisen, die man als sozial oder nicht-sozial bezeichnen kann, entscheidet sich erst im Aufeinandertreffen von Mensch und Technik, das heißt in der konkreten Praxis. Entsprechend beschreibt Latour die Geschehnisse in Pasteurs Labor als kontingentes Wechselspiel variabler Handlungsträgerschaften: „Pasteur denkt, das Ferment lenkt. Das Ferment denkt, Pasteur lenkt.“ (Latour 1996, S. 105) Die Begriffe „Denken“ und „Lenken“



verweisen an dieser Stelle nicht auf apriorische Zuschreibungen oder naturwüchsige Eigenschaften von Subjekten oder Objekten, sondern auf ein Potenzial, das in der komplexen Verwobenheit von Mensch und Technik angelegt ist und sich dort als „verteiltes Handeln“ (Rammert 2007) realisiert. Dies ist etwas anderes, als technischen Artefakten Intentionalität zuzuschreiben.

Objekte und Dinge sind in dieser Lesart somit weder neutral, noch ist ihr Verhältnis zu Subjekten eine fixe Relation. Vielmehr gehen heterogene Elemente eine fragile Ordnung ein, deren Logik – auch wenn die Elemente Eigenlogiken mitbringen – keinem dieser Elemente von Beginn an innewohnt und die entsprechend auch gegen Widerstände hergestellt werden muss (vgl. Pickering 1995; Latour 1988). Ein ‚Soziologismus‘, wie er im Bloorischen Symmetrieprinzip zu finden ist, wird in dieser Perspektive durch die Idee von „co-responsibilities“ (Lee & Stenner 1999, S. 93) ersetzt: Heterogene Entitäten vermischen ihre Kompetenzen und Attribute fortlaufend und ermöglichen auf diese Weise *gemeinsam* soziales Handeln und soziale Ordnung. Damit löst sich eine fixe Relation von Subjekt und Objekt in Koproduktionen von in dieser Hinsicht gleichberechtigten menschlichen und technischen Elementen auf. Wer oder was in der Lage ist, soziale Ordnung herzustellen oder zu verändern und Handlungsmöglichkeiten zu eröffnen, lässt sich in dieser Perspektive nicht mehr eindeutig und auch nicht von vornherein sagen: „Responsibility in this sense is to do with who gets the credit or who the blame, and hence also to do with the control or steering of the state of affairs in question. The centred or grammatical modes of explanation which answer that ‘he did it’ or ‘economics did it’ or ‘genetics did it’ are ways of finding clear centres of responsibility amongst otherwise confusing states of affairs. [...] An actor-network clearly does not depend on or belong to a centre, because what passes for a centre is an effect of, and hence depends upon or belongs to the network.“ (ebd., S. 92 f.) Akteurnetzwerktheoretische Ansätze problematisieren eindeutige Zurechnungen, indem sie das Subjekt dezentrieren und die Bedingungen seiner Möglichkeit in den Relationen eines materiell heterogenen Netzwerks suchen.<sup>14</sup> Der soziale Akteur, wie ihn die Soziologie im Regelfall imaginiert, wird auf diese Weise zu einem Hybridwesen. Als Effekt von Netzwerkrelationen eines materiell heterogen gedachten Sozialen sind technische Dinge, genauso wie menschliche Kompetenzen, verantwortlich für die Möglichkeit, sich als Akteur zeigen zu können (vgl. Law & Moser 1999; Moser & Law 1999, 2001). Pasteur kann, um dieses Argument auf das laufende Beispiel zu beziehen, entsprechend

---

<sup>14</sup> Darauf weist auch Rauer (2023) hin, wenn er von sozio-algorithmischen Dynamiken spricht, die eindeutige Zurechnungen von Autorenschaft in der Praxis mit ‚intelligent‘ automatisierten Schreib(unterstützungs)programmen (z. B. ChatGPT) erschweren oder gar unmöglich machen.

als „heterogeneous engineer“ (Law 1987) das Milchsäureferment dazu bringen, sich in seinem Labor und seinen Verfahren „vom Ereignis zur Substanz“ (Latour 1996, S. 101) zu wandeln. Allein diese Ordnungsleistung erlaubt es ihm, zum zentrierten Subjekt, zum Entdecker und Repräsentanten des Ferments zu werden.

Was bedeutet dies nun für die Frage nach dem praktischen Wirken von KI? Techniken, die unter den Sammelbegriff der konnektionistischen Künstlichen Intelligenz fallen, sind offenkundig keine Fermente, sondern menschengemachte Techniken, die in Praxiszusammenhängen bestimmte Aufgaben erfüllen sollen. Die planvolle und zielgerichtete „material delegation“ (Law 2001), d. h. die konkrete Übertragung von Aufgaben an und Einschreibung von Handlungsabläufen in eine Technik, bezieht sich im Falle von KI jedoch weniger auf die enge Kopplung von technischem Wirken und vorhersehbarem Ergebnis (wenn ein Taschenrechner beispielsweise die Rechenoperation ‚zwei mal zwei‘ ausführt) als vielmehr auf operative Freiheitsgrade in der Art und Weise, *wie*, d. h. entlang welcher Selektivitäten und adaptiver Strukturen ein *unvorhersehbares* Ergebnis erreicht wird (etwa wenn eine KI eine sehr große und unstrukturierte Datenmenge analysiert). Dieses Argument lässt sich anhand einer Unterscheidung schärfen, deren eine Seite der Techniksoziologe Ingo Schulz-Schaeffer herausarbeitet, wenn er von Sinnzuschreibungen an Technik als „erzeugter Selektivität“ spricht: „Die jeweilige Gestalt des Ineinandergreifens der mechanischen Komponenten eines technischen Artefaktes oder der Algorithmen seines computertechnischen Steuerungsprogramms ist Ausdruck erzeugter Selektivität. Seine technischen Abläufe sind sinnhafte Abläufe – aus der Perspektive seines Konstrukteurs, der sie eingerichtet hat, um dem Artefakt eine bestimmte Funktionalität zu verleihen, ebenso wie aus der Perspektive des Nutzers, der auf diese Funktionalität zugreift. Dieses Merkmal erzeugter Selektivität unterscheidet maßgebliche Verursachung durch technische Artefakte von Ursache-Wirkungs-Zusammenhängen, die der Natur zugeschrieben werden.“ (Schulz-Schaeffer 2008, S. 3142) Erzeugte Selektivität hat als Sinnkategorie ihr Korrelat in der geplanten Einschreibung von Handlungsabläufen in technische Artefakte, die in der Nutzungspraxis wiederholt zum Tragen kommen, dort entsprechend gedeutet werden und erwartbare Folgen zeitigen – etwa, wenn ein Taschenrechner die Rechenoperation ‚zwei mal zwei‘ immer wieder korrekt durchführt. Maschinen erhalten auf diese Weise einen spezifischen sozialen Sinn, der mit Zuschreibungen an ihr Wirken verbunden ist. Im Falle von KI muss die soziale Seite der erzeugten Selektivität jedoch um eine technische Seite der *erzeugenden* Selektivität ergänzt werden, die darin besteht, technikseitig aus der Vorstellung und Praxis einer reproduzierbaren Wiederholbarkeit von Abläufen auszubrechen und eigene Selektivitäten für daran anschließende Sinnsetzungsprozesse bereitzustellen. Es reicht nicht aus, allein

nach Zuschreibungen an KI zu fragen, die ihr Wirken sinnhaft erscheinen lassen. Vielmehr sind ‚intelligente‘ technische Abläufe in der Lage, ihrerseits sinnhafte Abläufe in Gang zu setzen und entlang technisch erzeugter Selektivitäten zu kanalisieren. Dies wird durch die Freiheitsgrade möglich, die KI generell als flexibler Technik eingeschrieben sind: „In der Künstliche-Intelligenz-Technologie beginnt die Loslösung von fest verdrahteten oder eindeutig vorgeschriebenen Abläufen mit dem Wechsel von der Master–Slave-Architektur zur agentenorientierten Programmierung und zu gesellschaftsorientierten Architekturen verteilter intelligenter Aktivitäten.“ (Rammert 2003, S. 7) Erzeugte und erzeugende Selektivität verschränken sich demnach: KI ist qua Programmierung mit der Fähigkeit ausgestattet, in einer konkreten Praxis kontingente Selektivitäten zu erzeugen, die diese Praxis transformieren und in anderen Zusammenhängen nicht zwangsläufig wiederholt erzeugt werden können.<sup>15</sup> Was eine KI ‚wahrnimmt‘, hängt demnach davon ab, in welche soziotechnischen Zusammenhänge sie eingebracht wird; wie eine KI dort wirkt, hängt von der erzeugenden Selektivität der Technik und der praktischen Anschlüsse an diese Selektivität ab. Die strukturelle Offenheit von KI gegenüber unterschiedlichen Praxiszusammenhängen wird damit zu einem entscheidenden Merkmal, das ihr kontingente Relationierungen erlaubt, die wiederum zu kontingenten Dynamiken soziotechnischer Praxen führen. Das ist letztlich mit ‚Lernen‘ von KI gemeint.

Diese Form der ‚intelligenten‘ technischen Delegation, die darin besteht, den Weg der Aufgabenlösung nicht vollständig zu determinieren (vgl. Heinlein & Huchler 2023), lässt sich weiter ausdifferenzieren. Folgt man Rammert, dann führt KI Aufgaben „in relativer Autonomie („autonomy“), ausgestattet mit dem Vermögen zur Reaktivität („reactivity“), orientiert an Tätigkeiten („proactiveness“) und unter Bezug auf andere Agenten („sociability“) [aus].“ (Rammert 2007, S. 82) KI bringt entwicklungsseitig somit bereits bestimmte Prädispositionen mit, die ihr eine praktische Offenheit gegenüber nicht-technischen Zusammenhängen verleihen und sie mit Möglichkeiten ausstatten, sich in unterschiedliche Praxiszusammenhänge einzubringen und diese – im Sinne einer Koproduktion von Mensch und Maschine – zu transformieren. Wichtig zu sehen ist an dieser Stelle, dass diese Prädispositionen erst praktisch wirksam werden. Die „relative Autonomie“ von KI ist demnach nicht pauschal gegeben, sondern immer relativ in Bezug auf die Praktiken, in denen sie als *relationale* Autonomie von Technischem *und* Menschlichem hergestellt und beobachtbar wird. Dies gilt

---

<sup>15</sup> Es wäre zu prüfen, ob und inwieweit diese Verschränkung von sinnsetzendem (erzeugte Selektivität) und sinnerzeugendem Wirken (erzeugende Selektivität) von KI ihrer Anthropomorphisierung Vorschub leistet: KI erscheint als eine Technik, deren technisches Wirken zu Deutungen einlädt, die sie im Bereich des Menschlichen ansiedelt.

auch für die anderen genannten Merkmale, die sich immer auf die soziotechnischen Relationen beziehen, innerhalb derer sie möglich und wirksam werden und Effekte zeitigen. „Reaktiv“, „proaktiv“ und „sozial“ kann KI nur in bestimmten Relationen zu Menschen und anderen Technikformen sein. Auf diese Relationen ist sie nicht nur angewiesen, sondern sie bringt diese auch selbst hervor, ermöglicht und verändert sie. Mit anderen Worten: KI ist und bewirkt jenseits der soziotechnischen Zusammenhänge, in die sie eingebracht wird und die sie verändert, nichts. Sie ist auf ein dynamisches Geschehen angewiesen, das ihr fortlaufend Daten liefert und in das sie fortlaufend Daten einspeisen kann – mit der Folge, dass sich die soziotechnische Praxis und ihre menschlichen und technischen Elemente selbst verändern. Indem KI aus Praxiszusammenhängen fortlaufend Daten gewinnt, diese Daten entlang eigener, sich praktisch weiterentwickelnder Regeln und Prozesse verarbeitet und das Resultat wieder in die Praxis einspeist, entsteht eine Dynamik, die in der Praxis spezifische Möglichkeitsräume schafft. Die Symmetrisierung des analytischen Vokabulars und die damit verbundene relationale Perspektive haben demnach nicht zum Ziel, KI zu vermenschlichen oder, in umgekehrter Weise, einem Technikdeterminismus Vor-schub zu leisten. Vielmehr werden soziologische Denkfiguren auf das Verhältnis und die Dynamik von Mensch und Technik bezogen, um die dort stattfindenden wechselseitigen Veränderungen und „Transformaktionen“ (Rauer 2023) zu beobachten und konzeptionell zu fassen.

---

## 4 Kontingenzen der Praxis Künstlicher Intelligenz

KI, so wurde argumentiert, geht mit einer Steigerung der Freiheitsgrade technischen Wirkens einher, sodass in den soziotechnischen Relationen der Nutzungspraxis Möglichkeiten des Anders-möglich-Seins entstehen. Diese Feststellung reicht jedoch nicht aus, um zu verstehen, *in welcher Hinsicht* KI Kontingenz in soziotechnischen Praxiszusammenhängen erzeugt. In einem nächsten Schritt gilt es daher, die Formen und Ebenen zu bestimmen, mit denen sich die Kontingenzen einer künstlich intelligenten Praxis systematisieren lassen.

Zu diesem Zwecke wird auf eine Unterscheidung zurückgegriffen, die der Mathematiker und Computerkünstler Frieder Nake im Kontext digitaler Kunst entwickelt hat, um die unterschiedlichen Relationsweisen von sozialen und algorithmischen Zusammenhängen in der menschlichen Interaktion mit Computern aufzuzeigen. Nake setzt dazu am Verhältnis der Oberfläche und der Tiefenstruktur algorithmisch erzeugter Bilder an, die er in seiner Terminologie als „surface“

und „subface“ bezeichnet (Nake 2008). In der traditionellen Kunst steht die sichtbare Oberfläche eines Werkes im Vordergrund, wie es sich den Betrachtenden präsentiert. Das „surface“ ist demnach identisch mit der visuellen Erscheinung eines Bildes. Bei digitalen Bildern sind jedoch zugleich die algorithmischen Prozesse, mit denen das Bild erschaffen wurde, von Bedeutung. Nake verwendet den Begriff „subface“ für die Ebene der zugrunde liegenden und miteinander verbundenen Algorithmen, die die Form, Farbe, Textur und andere visuelle Eigenschaften eines digitalen Bildes definieren, sich den Blicken der betrachtenden Personen jedoch entziehen. Diese Unterscheidung einer sichtbaren Oberfläche der Zeichen und einer unsichtbaren Tiefenstruktur algorithmischer Prozesse ist nach Nake grundlegend für die Interaktion mit Computern, die als „Zeichenmaschine“ (Nake & Grabowski 2005, S. 138) verstanden werden: Computer produzieren Zeichen, die für die nutzende Person intransparenten algorithmischen Prozessen entspringen. Die sichtbare Oberfläche der Zeichen und die algorithmische Tiefenstruktur sind dabei nicht voneinander unabhängig, sondern wechselseitig miteinander verschränkt: „The algorithmic thing comes as a visible appearance for us. At the same time, it comes as a computable appearance to the program. [...] It does not make any sense to talk about the computer image without keeping in mind its visibility *and* computability, i.e. its computable visibility *and* its visible computability.“ (Nake 2008, S. 105; Hervorh. i. Orig.) Das Zeichen, das sich zugleich als Grenze und als Verbindung zwischen sozialen und algorithmischen Zusammenhängen materialisiert, wird von zwei Seiten aus als Referenzpunkt genutzt: Zum einen dient es als Grundlage für algorithmische Prozesse, die es weiterverarbeiten, zum anderen wird es zum Gegenstand sozialer Deutungsprozesse, die zu weiteren Eingaben am Computer und damit zu einem Anschluss an algorithmische Prozesse führen. Der Schnittstelle zwischen Mensch und Maschine kommt vor diesem Hintergrund eine besondere Bedeutung zu: „Die analogen Äußerungen von Mensch und Maschine werden durch eine Welt digitaler Signale vermittelt, die, sobald sie ihre Unsichtbarkeit durch Erscheinen an der Schnittstelle verlieren, uns auch schon zum Zeichen werden (müssen).“ (Nake & Grabowski 2005, S. 138) An der Schnittstelle – dem „interface“ nach Nake (2008) – materialisiert sich demnach die Relation von algorithmischer Tiefenstruktur und zeichenhafter Oberfläche in einer Weise, die Informationsverarbeitungsprozesse sowohl für die beteiligten Algorithmen als auch für die sozialen Nutzungskontexte ermöglicht und kanalisiert.

Bezieht man diese Überlegungen auf die obigen Ausführungen zur Praxis von KI, dann wird deutlich, dass die dort beschriebenen Theoriefiguren in Nakes Konzept bereits angelegt sind und je eigene, gleichwohl miteinander verwobene Formen der algorithmischen und sozialen Relationalität und Rekursivität

beschreiben: Die ‚hidden layers‘ Künstlicher Neuronaler Netze erzeugen und prozessieren fortlaufend elektrische Impulse, die vom ‚output layer‘ als Zeichen ausgegeben und im sozialen Nutzungskontext interpretiert werden. Die dadurch entstehenden Veränderungen im Nutzungskontext werden wiederum vom ‚input layer‘ der KI in einer Form registriert, die die Übersetzung von Daten in elektrische Impulse und daran anschließende algorithmische Aktivitäten zulässt. Damit kann die wechselseitige Verschränkung und Dynamik algorithmischer und sozialer Zusammenhänge in künstlich intelligenten Praktiken sowohl nach Ebenen differenziert als auch – in einem weiteren Schritt – symmetrisch beschrieben werden. Die Konzepte des „surface“, „subface“ und „interface“ verweisen in diesem Sinne nicht auf materiale Eigenschaften von Computern oder Menschen, sondern auf spezifische *Relationsebenen* mit spezifischen *Relationsweisen* von sozialen und technischen Elementen, die praktisch realisiert werden. Jeder Relationsebene wohnt, wie nachfolgend argumentiert wird, eine eigene Form der Kontingenz inne.

***Relationsebene der Zeichen / epistemische Kontingenz:*** Die Oberfläche bzw. das „surface“ stellt in einer symmetrischen Lesart die Relationsebene der Zeichen dar. Geht man davon aus, dass sich mit der Technik des Computers eine „neue historische Praxisform der Schrift (Software)“ (Faßler 1996, S. 39) etabliert hat, dann wird deutlich, dass diese Praxis nicht nur auf der von Menschen vorgenommenen Einschreibung von Zeichen in Computersysteme beruht (sei es in der Form des Programmierens oder der Eingabe von Daten in Benutzeroberflächen), sondern auch auf einer maschinellen Produktion von Zeichen. Schrift ist – zumindest diesem Verständnis nach – etwas Symmetrisches. Indem die menschliche und die maschinelle Zeichenproduktion praktisch aufeinander bezogen sind, kommt es zu einer „semiotischen Koppelung zweier Semiosen, von denen eine beim Menschen, die andere bei der Maschine stattfindet.“ (Nake & Grabowski 2005, S. 138) Eine künstlich intelligente Praxis lässt sich auf dieser Relationsebene daher als ein Austausch von menschlich und maschinell erzeugten Zeichen beschreiben, die auf je spezifische Weise in sozialen und algorithmischen Zusammenhängen verarbeitet werden. Das Ergebnis dieser Verarbeitung wird von beiden Seiten als objektivierte Information genutzt, die wiederum die Grundlage für weitere zeichenhafte Austauschprozesse darstellt. Wesentlich dabei ist, dass sich hierdurch die Selektivitäten der sozialen und algorithmischen Anschlüsse verändern können. Mit anderen Worten: Die menschliche und maschinelle Interpretation eines

Zeichens hat Folgen dafür, wie weitere Zeichen interpretiert werden usw.<sup>16</sup> In Anlehnung an Latours Konzept der „zirkulierenden Referenz“ (2002a) kann man sagen, dass in einer künstlich intelligenten Praxis Referenzen zwischen Mensch und Maschine algorithmisch und sozial aufgegriffen, transformiert und an die jeweiligen Relata des soziotechnischen Netzwerks – die menschlichen und die algorithmischen Knotenpunkte – zurück vermittelt werden. Auf diese Weise entsteht eine dynamische Ordnung des menschlich-maschinellen Wissens, die sich als Ergebnis der Kommunikation mit einem Netz von Algorithmen – d. h. mit Elementen, die sich, obwohl sie eindeutig nicht-menschlicher Natur sind, durch „independence, self-reference, and complexity“ (Baecker 2011, S. 22) auszeichnen – verstehen lässt. Kontingenz zeigt sich auf dieser Relationsebene dort, wo Kommunikation durch Zeichen irritiert wird und sich entlang der Differenz von Erwartung und Erfahrung neues Wissen herausbildet.<sup>17</sup> Esposito beschreibt diese *epistemische Kontingenz* folgendermaßen: „The algorithm does not become more informed or more intelligent; it just learns to work better. But thereby it can produce increasingly complex communication with its users, who can learn unknown things about the world and about themselves. Communication becomes more effective, and new information is produced. [...] Even and especially if the algorithm is not an alter ego, does not work with a strategy, and does not understand its counterpart, in interaction with machines human users can learn something that no one knew before or could have imagined, which changes their way of observing.“ (Esposito 2017, S. 262) Die Neuheit und die Unvorhersehbarkeit des Wissens rühren daher, dass KI in der soziotechnischen Praxis als ‚Sichtbarmachungsmaschine‘ wirkt: Ihre Algorithmen werden dafür programmiert und dazu eingesetzt, in komplexen Daten Muster zu erkennen und latente Zusammenhänge in strukturierter Form darzustellen. Damit geht eine Veränderung von Perspektiven einher (vgl. Nassehi 2019) – man sieht *andere* Dinge, aber auch Dinge *anders* als zuvor, ohne dass diese Perspektivänderung

---

<sup>16</sup> Paradigmatisch dafür können generative KI-Anwendungen bzw. Transformer-Technologien wie der Chatbot ChatGPT gelten, die in Echtzeit im Austausch mit Nutzenden sinnvolle Texte erzeugen.

<sup>17</sup> Analog zum Bloor’schen Symmetrieprinzip ist für diese Theoretisierung einer soziotechnischen Wissensgenese die Unterscheidung von ‚wahrem‘ und ‚falschem‘ Wissen auf analytischer (!) Ebene nicht von Bedeutung – für die gesellschaftliche Nutzung von KI allerdings umso mehr. Selbst fortgeschrittene KI-Anwendungen wie der Chatbot ChatGPT produzieren falsche Informationen und ‚fake news‘, die – nolens volens – von mehr oder weniger kompetenten Nutzenden bewertet und eingeordnet werden müssen.

gewollt worden wäre oder hätte antizipiert werden können.<sup>18</sup> Zugleich gibt es – auch wenn die Algorithmen eines verwendeten KI-Systems auf einen Zweck hin programmiert wurden – keinen vorherbestimmten Weg, der genau *dieses* Wissen erzeugen und genau *diese* Perspektive eröffnen würde. Wissen ergibt sich vielmehr in einer idiosynkratischen Praxis, die ihre sozialen und algorithmischen Bestandteile gemeinsam erschaffen, transformieren und reproduzieren; es ist damit weder beliebig noch zufällig (vgl. Esposito 2017, S. 260).<sup>19</sup> Die hier diskutierte epistemische Kontingenz geht somit nicht in Shoshana Zuboffs Begriff einer durch Digitalisierung gesteigerten „reflexivity“ (Zuboff 1988) auf, lässt sich dazu aber in Beziehung setzen: Zuboff weist darauf hin, dass digitale Technik in Produktionsprozessen einen Überschuss an Daten produziert, der ex post weiter genutzt werden kann (etwa, wenn Scanner an Supermarktkassen nicht nur Warenwerte aufsummieren, sondern zusätzlich Daten liefern, die für Inventuren und Bestellprozesse genutzt werden können). KI-Anwendungen erzeugen ebenfalls einen teils erheblichen Datenüberschuss, für den gegenwärtig noch keine Verwendungszwecke oder Analysemöglichkeiten bestehen. Man könnte hier also von einer der aktuell sich vollziehenden Praxis nachgelagerten Kontingenz des Wissens mit offenem Zukunftshorizont sprechen. Das in künstlich intelligenten Praktiken erzeugte Wissen verweist damit nicht nur auf kontingente relationale Effekte, die mit dem Austausch von Zeichen zwischen Mensch und KI zu tun haben. Es stellt sich auch als ein „situated knowledge“ (Haraway 1988) dar, das sich den spezifischen räumlichen und zeitlichen Umständen, Bedingungen und Grenzen der Verwobenheit von algorithmischen und sozialen Dynamiken verdankt. Nimmt man dieses Argument ernst, dann kann Wissen immer nur in einer bestimmten Praxis erzeugt und angeeignet werden. Selbst wenn die Zirkulation von Referenzen über räumlich und zeitlich begrenzte Situationen hinausreicht, indem Daten innerhalb verkoppelter KI-Architekturen prozessiert und Informationen in Arbeitsgruppen, Organisationen, wissenschaftlichen Communities etc. sozial geteilt werden, erzeugen ähnliche Praxiszusammenhänge, in

---

<sup>18</sup> Abzugrenzen davon ist die Notwendigkeit, Wissen über die Arten und Folgen der Algorithmisierung (in) der Gesellschaft zu entwickeln, die „uns ein anderes Denken über die Produktion von Macht und Hegemonie im Web, über die Formung und Ausrichtung von Information online, abverlangen.“ (Cardon 2017, S. 131 f.).

<sup>19</sup> Bezogen auf die vieldiskutierte Frage der Standardisierung von Wissen durch KI bedeutet dies, dass Standardisierung praktisch immer auch mit Differenzen zu tun hat bzw. gerade auf die Möglichkeit von Differenzen angewiesen ist, um wirksam zu werden. Oder anders formuliert: Standardisierung muss Kontingenzen in der praktischen Umsetzung mit einrechnen und aushalten (können).



denen eine bestimmte Form von KI genutzt wird, nur ähnliches, aber nicht dasselbe Wissen. Folgt man diesem Gedanken, dann sorgt gerade die Adaptivität von KI-Systemen, d. h. ihre Anpassungsfähigkeit an strukturell ähnliche Situationen dafür, transsituativ neues und zugleich unterschiedliches Wissen zu generieren.

**Relationsebene der Aktionen / agentielle Kontingenz:** Die Schnittstelle bzw. das „interface“ kommt in einer symmetrischen Perspektive als *Relationsebene der Aktionen* von Mensch und Maschine in den Blick. Damit ist das materielle und zeitliche Geschehen miteinander verschränkter Formen etwa des menschlichen Schreibens, Sehens und Sich-Bewegens und der jeweiligen maschinellen Pendanten in einer soziotechnischen Praxis gemeint. Die Dynamik der Verbindung von Mensch und Maschine zeichnet sich auf dieser Relationsebene dadurch aus, dass sie eine (ethnomethodologisch gesprochen) soziotechnische „Vollzugswirklichkeit“ (Bongaerts 2017) von aufeinander bezogenen Körper- und Maschinenbewegungen ins Werk setzt. Welche Form diese wechselseitige Bezugnahme annimmt und welchem Muster sie folgt, lässt sich aufgrund der Freiheitsgrade der maschinellen Bewegungsabläufe nicht im Vorhinein bestimmen (vgl. Rammert 2003, S. 7) Offensichtliche Beispiele für diese Unvorhersehbarkeit sind KI-gesteuerte Roboter, die ihre Dreh-, Greif- und Manipulationsbewegungen so planen und korrigieren, dass ein kollisionsfreier Umgang mit Menschen gewährleistet ist, oder autonome Drohnen, die ihre Flugrouten fortlaufend an dynamische Umgebungen anpassen (Rauer 2016). Der Begriff der Maschinenbewegung beschränkt sich hier jedoch nicht auf Roboter und bewegliche Geräte, sondern umfasst den Möglichkeitsraum sämtlicher maschineller Aktionen, die situativ wahrnehmbar sind. Zur Maschinenbewegung zählen demnach nicht nur autonome Positionsänderungen, sondern auch andere Formen des maschinellen Tuns wie die selbsttätige Ausgabe von Zeichen auf einem Display, das Abspielen von Tönen oder die Wiedergabe bewegter Bilder, wie es etwa bei Chatbots der Fall ist (vgl. Dogruel und Dickel 2022; Guzman und Lewis 2020). Treten künstlich intelligente Maschinen in Aktion, dann tragen sie mit ihren Bewegungen zu einer soziomaterialen Ordnung bei, an der sich menschliche und algorithmische Verhaltensweisen im weiteren Verlauf orientieren und die in diesem Verlauf dynamisch fortgeschrieben wird. Diese „accomplished reality“ (Harold Garfinkel) von Mensch und künstlich intelligenter Maschine ist dabei weder sozial noch technisch vorherbestimmt und determiniert, sondern trägt kontingente Züge. Der Begriff der *agentiellen Kontingenz* macht darauf aufmerksam, dass in der gemeinsamen Vollzugswirklichkeit von Mensch und Maschine Handlungsträgerschaften ausgetauscht und vermittelt werden, die neue Möglichkeiten des Handelns eröffnen – und zwar sowohl mit Blick auf die Folgen des Handelns (etwas wird erreicht, das nur in der Kombination von Mensch und KI möglich ist) als auch

die Art und Weise, wie das Handeln selbst vollzogen wird (die Kombination von Mensch und KI verändert die Dynamik und die Logik des Handlungszusammenhangs im Sinne einer „Transformaktion“ [Rauer 2023]). Exemplarisch sei hier auf Morana Alačs Analyse eines Roboters verwiesen, der zu Lehr- und Lernzwecken in einer Schulklasse eingesetzt wurde (Alač 2016) oder Pat Treuschs Studie zum gemeinsamen Stricken mit einem KI-gesteuerten Roboterarm (Treusch 2021): Beide Arbeiten lassen sich als dichte Beschreibungen von Praxiszusammenhängen lesen, die keinem sozialen oder technischen Determinismus unterliegen und erst durch das aufeinander bezogene Agieren von Menschen und Maschinen in genau der Form entstehen, die durch die jeweilige soziologische Beobachterin beschrieben wird. Agentielle Kontingenz meint demnach nicht die technische Möglichkeit der Auswahl eines maschinellen Ablaufs aus bestehenden Lösungsansätzen anhand situativer Umstände, sondern die technisch und sozial realisierte Möglichkeit, dass „sich je nach Situation oder Pfadabhängigkeit unterschiedliche erfolgreiche Abläufe ergeben können.“ (Rammert 2007, S. 94) Die soziale Praxis bestimmt mit, was technisch geschieht, und umgekehrt. Voraussetzung dafür ist, dass die jeweiligen sozial oder technisch realisierten Vollzüge sichtbar und zweckmäßig stattfinden, um eine soziale und algorithmische Lesbarkeit und damit Zurechenbarkeit von Verhaltensweisen zu ermöglichen. Zweckmäßigkeit meint hier keine inhärente Intentionalität von Mensch oder Maschine, sondern verweist in zweierlei Hinsicht auf eine in soziotechnischen Relationen stattfindende Zuschreibungspraxis: Zum einen bauen KI-Systeme in Form eines fortlaufend überprüften datenbasierten Modells ex post „Erwartungen an den Nutzer“ (ebd., S. 36) auf und schließen daran „in gewisser Weise ‚reflexiv‘ auf diese Erfahrungen hin“ (ebd., S. 84) mit algorithmischen Vollzügen an, die zu sichtbaren Effekten innerhalb der Nutzungspraxis führen. Soziale Nutzungskontexte nehmen komplementär dazu im Sinne der weiter oben beschriebenen „erzeugten Selektivität“ (Schulz-Schaeffer 2008) ex ante fortlaufend Sinnzuschreibungen an das technische Wirken vor und schließen daran handelnd an.<sup>20</sup> Zweckmäßig ist damit nur das, was in der wechselseitigen Beobachtung von Mensch und Maschine als intendiertes bzw. für die Situation relevantes<sup>21</sup> Verhalten wahrgenommen und

---

<sup>20</sup> Folgt man Armin Nassehi, dann hängt die Zuschreibung von Intelligenz an eine Technik damit zusammen, dass „der Nutzer oder Anwender einer solchen Technik die Kontingenz des Gegenübers wahrnimmt, will heißen: dass man davon überrascht wird, dass der Automat so und nicht doch anders entscheidet – was den Automaten dann eben nicht mehr als Automaten erscheinen lässt.“ (Nassehi 2019, S. 222).

<sup>21</sup> Mit Alfred Schütz (2004) lässt sich Relevanz als Verhältnis zwischen einem Ereignis und der darauf bezogenen Bedeutungszuschreibung verstehen. Aufseiten der KI können Korrelationen als Relevanzstrukturen gelten, wohingegen aufseiten des Individuums Relevanzen

interpretiert wird – dies gilt, wie oben argumentiert wurde, nicht nur für das praktische Verhältnis von Mensch und KI, sondern auch für die soziologische Beobachungsposition, die eine künstlich intelligente Praxis zu rekonstruieren versucht. Zum anderen eröffnet die Praxis mit KI neue Möglichkeiten der Ausübung von (Handlungs-)Macht, die über die zeitliche und räumliche Situiertheit der Praxis selbst hinausgehen. Das soziotechnische Netzwerk der Praxis erweitert sich über Datenströme, sodass Hersteller und Betreiber einer KI in die Lage versetzt werden, über Raum und Zeit hinweg an situierten Geschehnissen teilzuhaben und in vermittelter Weise innerhalb ‚ihrer‘ Netzwerke zu machtvollen Akteuren zu werden. Jede künstlich intelligente Praxis erzeugt Daten, die über vernetzte Anwendungen geteilt und verarbeitet werden. Dadurch entstehen Zonen, die zwar praktisch miteinander verwoben sind, jedoch auf je spezifische Relationsweisen von Menschen, Daten und Algorithmen bzw. Maschinen innerhalb sich weit verzweigender soziotechnischer Vollzugswirklichkeiten verweisen. In Anlehnung an George Herbert Meads Überlegungen zur Wahrnehmung von materiellen Objekten (Mead 1932; vgl. auch Lewis 1981) lässt sich sagen, dass soziotechnische Netzwerke der KI von der „manipulatory area“ der unmittelbaren Berührung von menschlichem Körper und technischem Artefakt bis hin zur „distance experience“ der (Fremd-)Beobachtung eines Objekts bzw. Ereignisses – in diesem Fall: durch Daten – reichen. Die datensetzende bzw. -nutzende Macht der Hersteller und Betreiber ist damit immer an eine konkrete Nutzungspraxis gekoppelt, die ihr Daten liefert. Sie geht aber trotz dieser Kopplung nicht in dieser auf, da sie aus anderen Relationen hervorgeht und in andere Kontexte eingebunden ist (vgl. Amoore 2013, 2020; Amoore & Piotukh 2016; Zuboff 2015).

**Relationsebene des Impliziten und Latenten / formative Kontingenz:** Im Gegensatz zu den beiden beschriebenen Relationsebenen der Zeichen und der Aktionen, die auf Sichtbarkeit abstellen, verweist das „subface“ in einer symmetrischen Lesart auf die Strukturen einer künstlich intelligenten Praxis, die ihr unterliegen und auf eine in diesem Sinne ‚unsichtbare‘ *Relationsebene des Impliziten und Latenten* verweisen. Damit ist zunächst die Verschränkung von menschlichen und maschinellen Dispositionen gemeint, die sich einer vorgängigen Einschreibung bzw. Internalisierung verdanken und in der Praxis zwar vorhanden, dabei aber nicht zwingend offensichtlich und in der Regel erst rekonstruktiv anhand ihrer sichtbaren Effekte ableitbar sind. Darunter fällt all das,

---

eng mit dem ihm verfügbaren Wissen und Erfahrungen zusammenhängen. Eine ‚symmetrische‘ Relevanztheorie für die hier beschriebene Relationierung von Mensch und KI, die die Etablierung einer soziotechnischen Ordnung entlang der Wechselseitigkeit und Dynamik je eigener algorithmischer und sozialer Relevanzstrukturen beschreibt, steht allerdings noch aus.

was in sozialer Hinsicht unter anderem als Habitus (Pierre Bourdieu, Norbert Elias), implizites Wissen (Michael Polanyi) oder unbewusste Situationsrahmung (Erving Goffman) und in technischer Hinsicht als algorithmische Selektivität (vgl. Heinlein & Huchler 2023, S. 49; Pfeiffer 2021, S. 284–285), d. h. als strukturelle und methodische Eigenschaft der Datenverarbeitung in den ‚hidden layers‘ einer KI diskutiert wird. Gemein ist diesen heterogenen Konzepten die Vorstellung, dass es „hinter den Erscheinungen operierende Gesetzmäßigkeiten“ (Oevermann 2002, S. 1) gibt, die sich nicht auf die expliziten Formen einer Praxis (hier beschrieben als für die beteiligten menschlichen und maschinellen Instanzen sichtbare Zeichen und Aktionen) reduzieren lassen. Diese impliziten Größen, die sich sowohl auf sozialer Seite als auch auf Seite der KI finden, fließen nicht nur in die soziotechnische Praxis mit KI ein, sondern kanalisieren und formen diese auch – allerdings auf einer Ebene, die für die Praxis selbst nicht unmittelbar verfügbar ist.<sup>22</sup> Auf der einen Seite bedeutet dies, dass sich im menschlichen Umgang mit KI Verhaltensweisen, Wissensformen und Sinnstrukturen zeigen und etablieren, die nicht oder nur unter bestimmten Bedingungen thematisch werden (können), der soziotechnischen Praxis jedoch eben genau dadurch, dass sie unbeobachtet bleiben, eine spezifische Gestalt verleihen. Dies kann etwa dann der Fall sein, wenn eine KI-Anwendung offensichtlich falsche Ergebnisse hervorbringt, die unhinterfragte Deutungen und Nutzungsweisen der KI praktisch virulent werden lassen oder aber man KI-generierte Empfehlungen ignoriert. Im einen Fall hat man sich zu sehr auf die Technik verlassen, im anderen Fall erscheint man (bestenfalls) als Technikskeptiker. „Erst die Krise“, so kann man mit Latour auf beide Fälle bezogen sagen, „macht uns die Existenz des Gerätes wieder bewusst.“ (Latour 2002b, S. 223) Gleichwohl öffnet sich damit nicht die Black Box KI – denn es gilt, dass ein Algorithmus „makes selections and choices based on criteria that are not random, but that the user does not know and need not know.“ (Esposito 2017, S. 260)<sup>23</sup> Das algorithmische Wirken entspringt somit Unterscheidungen, die gleichermaßen ungesehen wie

---

<sup>22</sup> Interessant, aber soziologisch kurzichtig erscheint hier die These einer „Obsoleszenz des Impliziten“ (Bächle et al. 2017), die behauptet, dass humanoide Roboter über Sensorik „mit den Dimensionen impliziten Wissens umzugehen“ (ebd., S. 68) wüssten und sie im eigenen Tun reproduzieren könnten. Dies verkürzt die Debatte um implizites Wissen jedoch beträchtlich und blendet aus, dass dazu mehr gehört als ein nur als regelhaft beobachtbares und damit algorithmisch formalisierbares Interaktionsgeschehen (siehe dazu u. a. Böhle & Rose 1992, Neuweg 2015 und Polanyi 1985). Aus diesem Grunde werden in der Argumentation dieses Beitrags die Ebenen der beobachtbaren Aktionen und die des verborgenen Impliziten und Latenten analytisch voneinander getrennt.

<sup>23</sup> Eine menschliche Person kann immerhin ihr Handeln nachträglich rationalisieren und erklären.

unreflektiert in die Art und Weise der situierten Wissensproduktion und Vermittlung von Handlungsmöglichkeiten und Handlungsmacht einfließen. Beides, die impliziten Wahrnehmungen und Zuschreibungen sowie die verdeckt operierenden algorithmischen Unterscheidungen, gehen in eine dynamische Praxis ein, die weder sozial noch technisch determiniert ist. In dieser Hinsicht kommt eine künstlich intelligente Praxis als eine Praxis der wechselseitigen Inanspruchnahme und Beeinflussung von impliziten, d. h. nicht oder nur in vermittelter Weise offenkundigen sozialen und algorithmischen Zusammenhängen in den Blick, die jedoch sichtbare Effekte zeitigen. Die Relationsebene des Impliziten verweist somit, mit Pierre Bourdieu gesprochen, auf die „Erzeugungs- und Ordnungsgrundlage“ (Bourdieu 1987, S. 98) einer künstlich intelligenten Praxis und damit auf eine *formative Kontingenz*, die mit den generativen Tiefenstrukturen der praktischen Relationen zwischen Mensch und künstlich intelligenter Maschine zu tun hat. Mit dem Begriff der Latenz lassen sich diese Tiefenstrukturen weiter fassen und ausdifferenzieren. Folgt man Niklas Luhmann, dann kommt dem „Abdunkeln gewisser Aspekte, gewisser Ursachen oder gewisser Folgen des Handelns“ (Luhmann 1967, S. 100) eine funktionale Bedeutung zu: Die Komplexität einer sozialen Situation wird dahingehend reduziert, dass von den Handelnden nicht alles entschieden und gewusst werden muss. Erst durch Latenz wird es möglich, die „integrierende Struktur in einer höchst komplexen, fluktuierenden Umwelt konstant zu halten.“ (ebd., S. 101) Diese Komplexitätsreduktion trifft auch auf Technik zu, da das Wirken von Technik in einer Nutzungspraxis in der Regel in einer Weise objektiviert wird, die die sozialen Bedingungen und Entstehungszusammenhänge der jeweiligen Technik ausblenden und sie in die technische Umwelt verlagern: „once developed and deployed, technology tends to become reified and institutionalized, losing its connection with the human agents that constructed it or gave it meaning, and it appears to be part of the objective, structural properties of the organization.“ (Orlikowski 1992, S. 406) Driftet Technik in den Bereich des Unhinterfragten und Gegebenen ab, dann war die Einschreibung von Handlungsvollzügen in maschinelle Zeichen- und Aktionsprozesse erfolgreich. Technik wird nicht mehr als menschengemacht (und damit selbst als kontingent), sondern als funktionierender Mechanismus oder Black Box wahrgenommen (vgl. Latour 2002b). Zwischen einer sozialen Objektivierung von Technik und einem technisch umgesetzten ‚Abdunkeln‘ ihrer Wirkungshintergründe und -ursprünge besteht im Falle von KI jedoch ein enger Zusammenhang: Es ist sozial wahrscheinlich, dass das technische Wirken von KI objektiviert wird – und zwar gerade deshalb, weil es die Funktionsweise Künstlicher Neuronaler Netze mit ihren ‚hidden layers‘ nahelegt. Da nicht ohne Weiteres ersichtlich ist, nach welchen sozial abhängigen Kriterien und Selektivitäten KI-Systeme verfahren, wird

das Wirken von KI nach wie vor<sup>24</sup> als „technologically inflected promise of mechanical neutrality“ (Gillespie 2014, S. 181) wahrgenommen, d. h. als Fortsetzung, wenn nicht gar Steigerung eines allgemeinen Versprechens des technischen Fortschritts. Damit werden nicht nur die Machtverhältnisse verschleiert und latent gehalten, die hinter der Programmierung und dem Einsatz von KI liegen. Vielmehr kommt auch ein besonderer Latenzeffekt in den Blick: „Im Alltagsverhalten werden die treffenden Empfehlungen der Algorithmen auf eine intelligente Technik zugerechnet und nicht mit dem eigenen Verhalten in Beziehung gesetzt.“ (Baecker 2019, S. 2 f.) Die Debatte um die Intransparenz und Erklärbarkeit von KI muss daher um eine Debatte um die *Latenz ihres Wirkens* ergänzt werden. In den Blick kommt damit die Problematik einer algorithmischen Steuerung sozialer Prozesse, ohne dass die beteiligten menschlichen Akteure näher darüber Bescheid wüssten bzw. überhaupt Bescheid wissen könnten.<sup>25</sup> Auf welche Weise, d. h. entlang welcher interner algorithmischer Selektivitäten und Nutzerprofile KI wirkt, bleibt zwangsläufig latent und kann in actu zumeist nicht, ex post nur mit Mühe erschlossen werden. Es überrascht daher nicht, dass das Luhmannsche ‚Abdunkeln‘ im Begriff der KI-getriebenen Dunkelverarbeitung einen semantischen Widerhall findet – also dort auftaucht, wo ohne menschlichen Eingriff große Mengen an Daten von KI-Systemen gesammelt, verarbeitet und zur Grundlage weiterer automatisierter Entscheidungen (z. B. über Produktionsabläufe, Bewerbungsschreiben oder Bewährungsaussichten) gemacht werden. Das „sub-face“ stellt daher die womöglich wirkmächtigste, zugleich aber am schwersten greifbare Ebene einer künstlich intelligenten Praxis dar.

---

## 5 Abschließende Bemerkungen: Wie mit der kontingenzerzeugenden Technik KI umgehen?

Die Argumentation dieses Beitrags grenzt sich gegenüber Ansätzen ab, die den Blick nur auf vorab festgelegte Eigenschaften von Menschen und Maschinen und nicht (alternativ oder zusätzlich) auf die dynamische Interaktion zwischen Mensch und Maschine richten. Erst in einer konkreten Praxis, so lässt sich der Grundgedanke zusammenfassen, kommt vollends zum Tragen, in welcher Hinsicht „humans are better at and machines are better at“ (Fitts 1962). Aus diesem

---

<sup>24</sup> Dies mag sich langsam ändern, da auch mediale Diskurse mittlerweile die Verzerrungen durch und Grenzen von KI (vgl. etwa Pfeiffer 2021, S. 284–285) erkennen und in der Berichterstattung aufgreifen.

<sup>25</sup> Man kann dies auch so formulieren, dass es nicht nur um das *Was* des Wirkens von KI geht (Intransparenz), sondern auch um das *Wie* (Latenz).

Grunde wurde eine Perspektive entwickelt, die die Relationen von Menschen und ‚intelligenten‘ Maschinen und die Rekursivität der wechselseitig erfolgenden menschlichen und maschinellen Be- und Vollzüge in den Vordergrund stellt. Die Symmetrisierung des konzeptionellen Vokabulars bezog sich demnach nicht darauf, KI menschliche Eigenschaften zuzuschreiben (und damit letztlich einem Merkmalsansatz verhaftet zu bleiben), sondern auf die nach Ebenen differenzierten Verbindungsformen und -weisen zwischen Mensch und Maschine. Das Menschliche an KI ist in dieser Lesart in erster Linie nicht der nach wie vor irreführende Begriff der Intelligenz, sondern ihr Potenzial, sich in soziale Prozesse einzuklinken, ohne dass ihr Wirken vollumfänglich vorhergesehen werden kann – dass also KI, knapp formuliert, mit Kontingenz zusammenhängt. Diese Fähigkeit zur Kontingenzerzeugung wird zum einen durch ein verkürztes Technikverständnis verschleiert, wie es sich u. a. im Werkzeugdiskurs über KI widerspiegelt, zum anderen aber auch durch anthropomorphisierende sowie trans- und posthumanistische Diskurse einseitig überhöht: Weder ist KI ein einfaches Werkzeug, noch entscheidet, handelt und denkt KI in einem menschlichen Sinne.

Zu betonen ist, dass die Trennung der drei Relationsebenen (Zeichen, Aktionen, Implizites/Latenz) allein aus analytischen Gründen erfolgt ist: Wissen, Handeln sowie die impliziten und latenten Dimensionen einer Praxis sind nicht nur auf vielfältige Weise miteinander verwoben, sondern auch unmittelbar aufeinander angewiesen. Gleichwohl erlaubt ein ebenenbezogener Blick, das technische Wirken von KI in der Praxis genauer zu durchdringen und die jeweiligen Logiken des Verhältnisses von Mensch und künstlich intelligenter Maschine herauszuarbeiten. Dies gilt auch dann, wenn bestimmte Ebenen in der Praxis keine große Rolle spielen – etwa, wenn sich maschinelle Bewegungen auch in einem weit gefassten Sinne nur auf ein Minimum beschränken (etwa das Blinken eines Cursors oder einer Kontrollleuchte) und keine großen Kontingenzspielräume erlauben. Für die praktische Nutzung interaktiver KI-Systeme stellt sich daher die Frage, wie mit ihrem kontingenzerzeugenden Wirken umgegangen werden kann.

Für die *Integration und Einbettung* von KI in soziale Nutzungskontexte ist unter diesem Gesichtspunkt insbesondere relevant, welche Dynamiken durch den Einsatz von KI erwartbar sind und welche Kontingenzen man zulassen bzw. nutzbar machen kann oder aber – soweit es möglich ist – kontrollieren will. Dies erfordert eine „Kontingenzperspektive“ (Reckwitz 2004) auf die soziotechnischen Zusammenhänge einer Praxis mit interaktiver KI, die sich der Wechselwirkung sozialer und algorithmisch-maschineller Dynamiken bewusst ist und die Möglichkeitsräume der Nutzung abschätzen kann – auch und gerade mit Blick auf überfordernde oder belastende Wirkungen des Einsatzes von KI.

In der praktischen *Verwendung* interaktiver KI-Systeme wird die Frage der Befähigung von Nutzenden zentral, den kontingenzerzeugenden Charakter der Technik zu reflektieren und auf das eigene Wissen und Handeln zu beziehen. Es geht mithin darum, die fortlaufende Invisibilisierung von Kontingenz in der Praxis aufzubrechen: Die soziotechnische Ordnung des Wissens und Handelns mit KI folgt weder technischen noch sozialen Notwendigkeiten, sondern unterliegt Kontingenzen, die sozial *und* technisch vorhanden sind. Es wäre daher notwendig, die Zuschreibungen an und den praktischen Umgang mit KI im konkreten *doing*, d. h. in der laufenden Praxis zu reflektieren, um auf der Ebene der Verwendung von KI mehr Transparenz und Handlungsoffenheit zu erzeugen. Indem man die Kontingenz von KI-Systemen anerkennt und beobachtet, ist eine andere, im Vergleich zu technikobjektivierenden oder -anthropomorphisierenden Zuschreibungen wahrscheinlich angemessenere Praxis des Umgangs mit KI möglich.

Mit Blick auf die *Bewertung* des mit Hilfe von KI generierten Wissens taucht die Herausforderung auf, seine Validität angesichts der Kontingenz der Technik *und* der sozialen Zusammenhänge, in der die Technik praktisch wirksam ist, einzuschätzen. Reflektieren ließe sich dies zum einen durch ein planerisch-vorausschauendes Handeln in der Entwicklung oder aber ein in der Nutzung nachholendes Nachdenken über die möglichen algorithmisch und sozial erzeugten Selektivitäten eines spezifischen Ereignisses innerhalb einer spezifischen Praxis. Letzteres entspräche dem Versuch, das Wirken von KI zu erklären, ohne darauf bauen zu können, dass eine KI-Anwendung die Beweggründe und Rationalitäten ihres Wirkens offenbart; Ersteres müsste sehr viele Möglichkeiten einer soziotechnischen Praxis mit interaktiven KI-Systemen antizipieren, die sich jedoch mit hoher Wahrscheinlichkeit immer wieder als kontingent, d. h. als abweichend und überraschend erweisen wird.

Problematisch an beiden Varianten, der vorausschauenden und der nachholenden Beobachtung und Reflexion, ist, dass sowohl ein hohes Maß an Wissen über KI als auch an Zeit erforderlich ist, die „unter spezifisch praktischen Bedingungen der Dringlichkeit und zeitlichen Irreversibilität“ (Bongaerts 2008, S. 38) jedoch nur selten gegeben ist. Wie wissensintensiv und voraussetzungsreich die Nutzung von KI ist, stellt die Arbeits- und Techniksoziologin Sabine Pfeiffer heraus, die vielerorts einen „unbewusst destruktiven“ Einsatz von KI und maschinellem Lernen (ML) in Unternehmen ausmacht: „Ein unbewusst destruktiver Einsatz von KI/ML lässt sich etwa beobachten, wenn die Schlichtheit des statistischen Modells völlig über- und der gewählte Anwendungskontext in seiner Komplexität gleichzeitig unterschätzt wird. Weil Kausalitäten unterstellt werden, obwohl es sich nur um Datenrauschen handelt. Weil Algorithmen, die in ihren Lernprozessen



in die Mitte tendieren, sachlich relevante Beobachtungen als statistische Ausreißer aussortieren. Weil das Wissen um Skalenniveaus der Daten und um die Notwendigkeit von deren Passfähigkeit zum gewählten Algorithmus fehlt. Weil das Wissen um die Notwendigkeit der Passfähigkeit von beidem (Skalenniveau der Daten und gewähltem Algorithmus) mit den sachlichen Gegebenheiten des Anwendungskontexts fehlt. Weil die Implementierung oft allein in die Hände von Informatik und Data Science gelegt wird, ohne die Expertinnen und Experten für den angezielten Anwendungskontext mit einzubeziehen. Weil Daten genutzt werden, nur weil sie vorhanden sind, ohne ihre Validität sachlich (statt nur mathematisch) zu hinterfragen. Weil die Daten selbst schon systematisch verzerrt sind und reale (erwünschte oder unerwünschte) Schief lagen in Gesellschaft oder im Anwendungskontext widerspiegeln und diese sich ohne ausgleichende Gewichtung in autonomen Verarbeitungsprozessen weiter verstärken. Weil die Daten eine weniger offensichtliche Verzerrung aufweisen, die sich aus der Leichtigkeit oder Erschwernis beim Sammeln der Daten faktisch ergibt und auf die bei unreflektiertem Einsatz dann nicht mal mit Gewichtung reagiert wird. Weil für viele KI/ML-Anwendungen noch völlig unklar ist, wie lange mit welchen Daten gelernt werden soll und ob das nur anfänglich oder dauerhaft und immer wieder aufs Neue passieren soll – und nach welchen Kriterien dabei entschieden wird.“ (Pfeiffer 2021, S. 284) Diese Punkte machen deutlich, dass für eine sinnvolle und effektive Anwendung von KI anderes und womöglich auch mehr Wissen nötig ist, als dies bei herkömmlichen Techniken der Fall ist. Gleichwohl beziehen sich die Überlegungen stark darauf, was gewusst und bedacht werden muss, bevor ein vorgefertigtes KI-System in einem Unternehmen zum Einsatz kommt – dies fällt in der Regel in den Aufgabenbereich einzelner Personen oder Abteilungen (v. a. IT), die nicht zwingend dauerhaft mit dem KI-System arbeiten. Aus der Perspektive des Beitrags heraus stellt sich demgegenüber die Frage, was Nutzende wissen müssen, wenn sie mit einer kontingenzerzeugenden Technik konfrontiert sind.

Die konsequente Anbindung des Wirkens von KI an das praktische (Erfahrungs-)Wissen der Nutzenden erscheint in diesem Lichte als ein vielversprechender Ansatz, der grundlegend notwendiges Wissen u. a. um die mathematisch-stochastische Logik, die Funktionalität und die Datenabhängigkeit von KI ergänzen kann: Eine künstlich intelligente Praxis benötigt den Menschen auch im Sinne einer kontinuierlich mitlaufenden Reflexionsschleife ‚in the loop‘, die die Praxis selbst beobachtet und KI als eine technisch erzeugte *Perspektive* auf die Welt begreift. Die Aufgabe besteht entsprechend darin, die KI-generierten Zeichen und Aktionen fortlaufend vor dem Hintergrund des Geschehens einzuordnen und auf diese Weise mit den Kontingenzen von KI zu rechnen.

## Literatur

- Alač, M. 2009. Moving Android: On Social Robots and Body-in-Interaction. *Social Studies of Science*, 39(4): 491–528.
- Alač, M. 2016. Social Robots: Things or Agents? *AI & Society*, 31(4): 519–535.
- Amoore, L.A. 2013. *The Politics of Possibility: Risk and Security Beyond Probability*. Duke University Press.
- Amoore, L.A. 2020. *Cloud Ethics: Algorithms and the Attributes of Ourselves and Others*. Duke University Press.
- Amoore, L.A., und V. Piotukh (Hrsg.). 2016. *Algorithmic Life. Calculative Devices in the Age of Big Data*. Routledge.
- Bächle, Thomas C., P. Regier, und M. Bennewitz. 2017. Sensor und Sinnlichkeit: Humanoide Roboter als selbstlernende soziale Interfaces und die Obsoleszenz des Impliziten. *Navigationen – Zeitschrift für Medien- und Kulturwissenschaften*, 17(2): 67–86.
- Baecker, D. 2011. Who Qualifies for Communication? A Systems Perspective on Human and Other Possibly Intelligent Beings Taking Part in the Next Society. *Technikfolgenabschätzung – Theorie und Praxis*, 20(1): 17–26.
- Baecker, D. 2019. Auf dem Weg zu einer Theorie der digitalen Gesellschaft: Rezension zu ‚Muster – Theorie der digitalen Gesellschaft‘ von Armin Nassehi. *Soziopolis: Gesellschaft beobachten*, veröffentlicht unter <https://nbn-resolving.org/urn:nbn:de:0168-ssoar-81735-2> (letztes Abrufdatum: 26.04.2023)
- Berger, P.L., und T. Luckmann. 1969. *Die gesellschaftliche Konstruktion der Wirklichkeit. Eine Theorie der Wissenssoziologie*. Fischer.
- Bijker, W.E. 1992. The Social Construction of Fluorescent Lighting, or How an Artifact Was Invented in Its Diffusion Stage. In *Shaping Technology – Building Society: Studies in Sociotechnical Change*, hrsg. W.E. Bijker, & J. Law, 75–102. MIT Press.
- Bijker, W.E. 1995. *Of Bicycles, Bakelite and Bulbs: Toward a Theory of Sociotechnical Change*. MIT Press.
- Bischof, A. 2017. *Soziale Maschinen bauen. Epistemische Praktiken der Sozialrobotik*. transcript.
- Bloor, D. 1976. *Knowledge and Social Imagery*. Routledge.
- Bloor, D. 1998. Anti-Latour. *Studies in History and Philosophy of Science*, 1: 81–112.
- Böhle, F., und H. Rose. 1992. *Technik und Erfahrung*. Campus.
- Bongaerts, G. 2008. *Verdrängungen des Ökonomischen. Bourdieus Theorie der Moderne*. transcript.
- Bongaerts, G. 2017. Zur Vollzugswirklichkeit des Handelns. *Soziale Welt*, 68(2–3): 157–173.
- Bourdieu, P. 1987. *Sozialer Sinn. Kritik der theoretischen Vernunft*. Suhrkamp.
- Breazeal, C. 2002. *Designing Sociable Robots*. MIT Press.
- Brent Jr., E.E. 1986. Knowledge-based Systems: A Qualitative Formalism. *Qualitative Sociology*, 9(3): 256–282.
- Brödner, P. 2022. Die Illusionsfabrik der ‚KI‘-Narrative. *FifF-Kommunikation*, 22(2): 32–36.
- Brynjolfsson, E., D. Rock, und C. Syverson. 2019. Artificial Intelligence and the Modern Productivity Paradox: A Clash of Expectations and Statistics. In *The Economics of Artificial Intelligence: An Agenda*, hrsg. A. Agrawal, J. Gans, & A. Goldfarb, 23–60. University of Chicago Press.

- Bubeck, S., V. Chandrasekaran, R. Eldan, J. Gehrke, E. Horvitz, E. Kamar, P. Lee, Y.T. Lee, Y. Li, S. Lundberg, H. Nori, H. Palangi, M.T. Ribeiro, und Y. Zhang. 2023. *Sparks of Artificial General Intelligence: Early experiments with GPT-4*. Veröffentlicht unter <https://arxiv.org/abs/2303.12712> (letztes Abrufdatum: 26.04.2023)
- Callon, M. 1986. Some Elements of a Sociology of Translation: Domestication of the Scallops and the Fishermen of St Brieuc Bay. In *Power, Action and Belief: A New Sociology of Knowledge?* hrsg. J. Law, 196–233. Routledge and Kegan Paul.
- Cardon, D. 2017. Den Algorithmus dekonstruieren. Vier Typen digitaler Informationsberechnung. In *Algorithmenkulturen. Über die rechnerische Konstruktion der Wirklichkeit*, hrsg. R. Seyfert, & J. Roberge, 131–150. transcript.
- Chalmers, D.J. 1992. Subsymbolic Computation and the Chinese Room. In *The Symbolic and Connectionist Paradigms: Closing the Gap*, hrsg. J. Dinsmore, 25–48. Psychology Press.
- Collins, H.M. 1990. *Artificial Experts: Social Knowledge and Intelligent Machines*. MIT Press.
- Collins, H.M., und S. Yearley. 1992. Epistemological Chicken. In: *Science as Practice and Culture*, hrsg. A. Pickering, 301–326. The University of Chicago Press.
- Crafts, N. 2021. Artificial Intelligence as a General-Purpose Technology: An Historical Perspective. *Oxford Review of Economic Policy*, 37(3): 521–536.
- Danaher, J. 2016. The Threat of Algcocracy: Reality, Resistance and Accommodation. *Philosophy & Technology*, 29(3): 245–268.
- Degele, N. 2002. *Einführung in die Techniksoziologie*. Fink.
- Dogrueel, L., und S. Dickel. 2022. Die Kommunikativierung der Maschinen. *Publizistik*, 67(4): 475–486.
- Durkheim, É. 1961. *Die Regeln der soziologischen Methode*. Luchterhand.
- Dyer-Witheford, N., A.M. Kjösen, und J. Steinhoff. 2019. *Inhuman Power: Artificial Intelligence and the Future of Capitalism*. Pluto Press.
- Ennals, R. 1991. *Artificial Intelligence and Human Institutions*. Springer.
- Esposito, E. 2014. Algorithmische Kontingenz. Der Umgang mit Unsicherheit im Web. In *Die Ordnung des Kontingenten. Beiträge zur zahlenmäßigen Selbstbeschreibung der modernen Gesellschaft*, hrsg. A. Cevolini, 233–249. Springer VS.
- Esposito, E. 2017. Artificial Communication? The Production of Contingency by Algorithms. *Zeitschrift für Soziologie*, 46(4): 249–265.
- Faßler, M. 1996. *Mediale Interaktion: Speicher, Individualität, Öffentlichkeit*. Fink.
- Fitts, P.M. 1962. Functions of Man in Complex Systems. *Aerospace Engineering*, 21(1): 34–39.
- Fodor, J.A., und Z.W. Pylyshyn. 1988. Connectionism and Cognitive Architecture: A Critical Analysis. *Cognition*, 28: 3–71.
- Gilbert, G.N., und C. Heath (Hrsg.). 1985. *Social Action and Artificial Intelligence*. Gower.
- Gillespie, T. 2014. The Relevance of Algorithms. In *Media Technologies: Essays on Communication, Materiality, and Society*, hrsg. T. Gillespie, P.J. Boczkowski, & K.A. Foot, 167–193. MIT Press.
- Gurevich, Y. 2012. What Is an Algorithm? In *SOFSEM 2012: Theory and Practice of Computer Science. SOFSEM 2012. Lecture Notes in Computer Science*, hrsg. M. Bieliková, G. Friedrich, G. Gottlob, S. Katzenbeisser, & G. Turán, 31–42. Springer.
- Guzman, A. L., und S.C. Lewis. 2020. Artificial Intelligence and Communication: A Human–Machine Communication Research Agenda. *New Media & Society*, 22(1): 70–86.

- Halfmann, J. 1995. Editorial: Theoriebausteine der Techniksoziologie. In *Technik und Gesellschaft. Jahrbuch 8: Theoriebausteine der Techniksoziologie*, hrsg. J. Halfmann, G. Bechmann, & W. Rammert, 7–11. Campus.
- Haraway, D. 1988. Situated Knowledges: The Science Question in Feminism and the Privilege of Partial Perspective. *Feminist Studies*, 14(3): 575–599.
- Harth, J., und C.-F. Lorenz. 2017. „Hello World“ – Systemtheoretische Überlegungen zu einer Soziologie des Algorithmus. *kommunikation@gesellschaft*, 18(2).
- Heinlein, M. 2022. Zeit. In *Handbuch Sozialwissenschaftliche Gedächtnisforschung, Band 1: Grundbegriffe und Theorien*, hrsg. G. Sebald, M. Berek, K. Chmelar, O. Dimbath, H. Haag, M. Heinlein, N. Leonhard, V. Rauer. Springer VS.
- Heinlein, M., und N. Huchler. 2023. Artificial Intelligence in the Practice of Work: A New Way of Standardising or a Means to Maintain Complexity? *Work Organisation, Labour & Globalisation*, 17(1): 34–60.
- Hirsch-Kreinsen, H. 2023. *Das Versprechen der Künstlichen Intelligenz. Gesellschaftliche Dynamik einer Schlüsseltechnologie*. Campus.
- Kneer, G., und M. Schroer. 2009. Soziologie als multiparadigmatische Wissenschaft. Eine Einleitung. In *Handbuch Soziologische Theorien*, hrsg. G. Kneer, & M. Schroer, 7–18. VS Verlag für Sozialwissenschaften.
- Koolwaay, J. 2018. *Die soziale Welt der Roboter: Interaktive Maschinen und ihre Verbindung zum Menschen*. transcript.
- Latour, B. 1988. Mixing Humans and Nonhumans Together. The Sociology of a Door-Closer. *Social Problems*, 35(3): 298–310.
- Latour, B. 1996. Haben auch Objekte eine Geschichte? Ein Zusammentreffen von Pasteur und Whitehead in einem Milchsäurebad. In *Der Berliner Schlüssel. Erkundungen eines Liebhabers der Wissenschaften*, 87–112. Akademie Verlag.
- Latour, B. 1998. *Wir sind nie modern gewesen. Versuch einer symmetrischen Anthropologie*. Fischer.
- Latour, B. 2002a. Zirkulierende Referenz. Bodenstichproben aus dem Urwald am Amazonas. In *Die Hoffnung der Pandora. Untersuchungen zur Wirklichkeit der Wissenschaft*, 36–95. Suhrkamp.
- Latour, B. 2002b. Ein Kollektiv von Menschen und nichtmenschlichen Wesen. Auf dem Weg durch Dädalus' Labyrinth. In *Die Hoffnung der Pandora. Untersuchungen zur Wirklichkeit der Wissenschaft*, 211–264. Suhrkamp.
- Latour, B. 2005. *Reassembling the Social: An Introduction to Actor-Network-Theory*. Oxford University Press.
- Law, J. 1987. Technology and Heterogeneous Engineering: The Case of Portuguese Expansion. In *The Social Construction of Technological Systems*, hrsg. W.E. Bijker, T.P. Hughes, & T. Pinch, 111–134. MIT Press.
- Law, J. 1994. *Organizing Modernity*. Blackwell.
- Law, J. 2001. *Ordering and Obduracy*. Lancaster University: Centre for Science Studies and the Department of Sociology. Veröffentlicht unter <https://www.lancaster.ac.uk/fass/resources/sociology-online-papers/papers/law-ordering-and-obduracy.pdf> (letztes Abrufdatum: 26.04.2023)
- Law, J., und I. Moser. 1999. *Managing, Subjectivities and Desires*. Lancaster University und Oslo University: Centre for Science Studies, the Department of Sociology and Teknologi, Innovasjon og Kultur. Veröffentlicht unter <https://www.lancaster.ac.uk/fass/>

- [resources/sociology-online-papers/papers/law-moser-managing-subjectivities-desires.pdf](#) (letztes Abrufdatum: 26.04.2023)
- Lee, N., und S. Brown. 1994. Otherness and the Actor Network. *The Undiscovered Continent. American Behavioral Scientist*, 37(6): 772–790.
- Lee, N., und P. Stenner. 1999. Who pays? Can we pay them back? In *Actor Network Theory and After*, hrsg. J. Law, & J. Hassard, 90–112. Blackwell.
- Lewis, J.D. 1981. G.H. Mead's Contact Theory of Reality: The Manipulatory Phase of the Act in the Constitution of Mundane, Scientific, Aesthetic, and Evaluative Objects. *Symbolic Interaction*, 4(2): 129–141.
- Linde, H. 1972. *Sachdominanz in Sozialstrukturen*. Mohr.
- Luger, G.F. 2005. *Artificial Intelligence: Structures and Strategies for Complex Problem Solving. Fifth Edition*. Addison-Wesley.
- Luhmann, N. 1967. Reflexive Mechanismen. *Soziale Welt*, 18(2/3): 97–123.
- Luhmann, N. 1984. *Soziale Systeme. Grundriss einer allgemeinen Theorie*. Suhrkamp.
- Luhmann, N. 1992a. Kontingenz als Eigenwert der modernen Gesellschaft. In *Beobachtungen der Moderne*, ders., 93–128. Westdeutscher Verlag.
- Luhmann, N. 1992b. Europäische Rationalität. In *Beobachtungen der Moderne*, ders., 51–91. Westdeutscher Verlag.
- Mainzer, K. 2019. *Künstliche Intelligenz – Wann übernehmen die Maschinen?* Springer.
- Makropoulos, M. 1997. *Modernität und Kontingenz*. Fink.
- Mead, G.H. 1932. The Physical Thing, Supplementary Essay 2. In *The Philosophy of the Present*, hrsg. A.E. Murphy, 119–139. Open Court.
- Moser, I., und J. Law. 1999. Good Passages, Bad Passages. In *Actor Network Theory and After*, hrsg. J. Law, & J. Hassard, 196–219. Blackwell.
- Moser, I., und J. Law. 2001. *'Making Voices': New Media Technologies, Disabilities, and Articulation*. Lancaster University: Centre for Science Studies and the Department of Sociology. Veröffentlicht unter <https://www.lancaster.ac.uk/fass/resources/sociology-online-papers/papers/moser-law-making-voices.pdf> (letztes Abrufdatum: 26.04.2023)
- Muhle, F. (Hrsg.). 2023. *Soziale Robotik. Eine sozialwissenschaftliche Einführung*. de Gruyter.
- Munro, R. 1997. Ideas of Difference: Stability, Social Spaces and the Labour of Division. In *Ideas of Difference: Social Spaces and the Labour of Division*, hrsg. K. Hetherington, & R. Munro, 3–24. Blackwell.
- Nake, F. 2008. Surface, Interface, Subface. Three Cases of Interaction and One Concept. In *Paradoxes of Interactivity. Perspectives for Media Theory, Human-Computer Interaction, and Artistic Investigations*, hrsg. U. Seifert, J. H. Kim, & A.A. Moore, 92–109. transcript.
- Nake, F., und S. Grabowski. 2005. Zwei Weisen, das Computerbild zu betrachten. In *Hyper-Kult II: Zur Ortsbestimmung analoger und digitaler Medien*, hrsgs. M. Warnke, W. Coy, & G.C. Tholen, 123–150. Transcript.
- Nassehi, A. 2019. *Muster. Theorie der digitalen Gesellschaft*. C.H. Beck.
- Neuweg, G.H. 2015. *Das Schweigen der Könner. Gesammelte Schriften zum impliziten Wissen*. Waxmann.
- Oevermann, U. 2002. *Klinische Soziologie auf der Basis der Methodologie der objektiven Hermeneutik (Manifest der objektiv hermeneutischen Sozialforschung)*. Frankfurt a.M.: Goethe-Universität. Veröffentlicht unter <http://publikationen.ub.uni-frankfurt.de/files/4958/ManifestWord.pdf> (letztes Abrufdatum: 26.04.2023)

- Orlikowski, W.J. 1992. The Duality of Technology: Rethinking the Concept of Technology in Organizations. *Organization Science*, 3(3): 398–427.
- Orr, J.E. 1996. *Talking About Machines. An Ethnography of a Modern Job*. ILR Press.
- Pfeiffer, S. 2021. *Digitalisierung als Distributivkraft. Über das Neue am digitalen Kapitalismus*. transcript.
- Pickering, A. 1995. *The Mangle of Practice: Time, Agency, and Science*. University of Chicago Press.
- Pinch, T.J., und W.E. Bijker. 1984. The Social Construction of Facts and Artefacts: or How the Sociology of Science and the Sociology of Technology Might Benefit Each Other. *Social Studies of Science*, 14(3): 399–441.
- Polanyi, M. 1985. *Implizites Wissen*. Suhrkamp.
- Rammert, W. (Hrsg.). 1995. *Soziologie und künstliche Intelligenz – Produkte und Probleme einer Hochtechnologie*. Campus.
- Rammert, W. 1998. Technikvergessenheit der Soziologie? Eine Erinnerung als Einleitung. In *Technik und Sozialtheorie*, hrsg. W. Rammert, 9–28. Campus.
- Rammert, W. 1999. Relations That Constitute Technology and Media That Make a Difference: Toward a Social Pragmatic Theory of Technicization, *Techné: Research in Philosophy and Technology*, 4(3): 165–177.
- Rammert, W. 2003. *Technik in Aktion: verteiltes Handeln in soziotechnischen Konstellationen*. TUTS – Working Papers, 2–2003. Technische Universität Berlin, Fak. VI Planen, Bauen, Umwelt, Institut für Soziologie, Fachgebiet Techniksoziologie. Veröffentlicht unter <https://nbn-resolving.org/urn:nbn:de:0168-ssoar-11573> (letztes Abrufdatum: 26.04.2023)
- Rammert, W. 2007. *Technik – Handeln – Wissen. Zu einer pragmatistischen Technik- und Sozialtheorie*. VS Verlag für Sozialwissenschaften.
- Rammert, W. 2016. *Technik – Handeln – Wissen. Zu einer pragmatistischen Technik- und Sozialtheorie*. Springer VS.
- Rammert, W., und I. Schulz-Schaeffer. 2002. *Technik und Handeln – Wenn soziales Handeln sich auf menschliches Verhalten und technische Artefakte verteilt*. Technical University Technology Studies Working Papers, TUTS-WP-4–2002. Veröffentlicht unter <https://nbn-resolving.org/urn:nbn:de:0168-ssoar-122133> (letztes Abrufdatum: 26.04.2023)
- Rauer, V. 2016. The Mobilization of Algorithms. In *Algorithmic Cultures. Essays on Meaning, Performance and New Technologies*, hrsg. R. Seyfert, & J. Roberge, 140–157. Routledge.
- Rauer, V. 2023. Von Interaktion zur Transformaktion: Die Folgen von Künstlicher Intelligenz für Theorien sozialen Handelns. In *Künstliche Intelligenz, Mensch und Gesellschaft*, hrsg. M. Heinlein, & N. Huchler. Springer VS.
- Reckwitz, A. 2004. Die Kontingenzzperspektive der ‚Kultur‘. Kulturbegriffe, Kulturtheorien und das kulturwissenschaftliche Forschungsprogramm. In *Handbuch der Kulturwissenschaften. Band 3: Themen und Tendenzen*, hrsg. F. Jaeger, & J. Rüsen, 1–20. Metzler.
- Ropohl, G. 2005. Die kulturelle Verzögerung der Techniksoziologie. In *Pragmatisches Philosophieren. Festschrift für Hans Lenk*, hrsg. R. Dürr, G. Gebauer, M. Maring, & H.-P. Schütt, 392–403. LIT.
- Rosengrün, S. 2021. *Künstliche Intelligenz zur Einführung*. Junius.
- Schulz-Schaeffer, I. 2008. Deutung und Delegation: Handlungsträgerschaft von Technik als doppeltes Zuschreibungsphänomen. In *Die Natur der Gesellschaft: Verhandlungen des*

33. *Kongresses der Deutschen Gesellschaft für Soziologie in Kassel 2006*, hrsg. K.-S. Rehberg, 3135–3144. Campus.
- Schütz, A. 2004. Das Problem der Relevanz. In *Alfred Schütz Werkausgabe, Bd. VI.1, Relevanz und Handeln 1. Zur Phänomenologie des Alltagswissens*, hrsg. E. List, 57–249. UVK.
- Schwartz, R.D. 1989. Artificial Intelligence as a Sociological Phenomenon. *The Canadian Journal of Sociology*, 14(2): 179–202.
- Seising, R. 2021. *Es denkt nicht! Die vergessenen Geschichten der KI*. Büchergilde Gutenberg.
- Smolensky, P. 1988. On the Proper Treatment of Connectionism. *Behavioral and Brain Sciences*, 11(1): 1–74.
- Smolensky, P. 2012. Subsymbolic Computation Theory for the Human Intuitive Processor. In *How the World Computes. CiE 2012. Lecture Notes in Computer Science, vol 7318*, hrsg. S.B. Cooper, A. Dawar, & B. Löwe, 675–685. Springer.
- Straub, I. 2020. *Zur Sozialität und Entität eines androiden Roboters. Empirische Zugänge zum Objekt- und Subjektstatus*. Springer VS.
- Treusch, P. 2021. *Robotic Knitting. Re-Crafting Human-Robot Collaboration Through Careful Coboting*. transcript.
- Woolgar, S. 1985. Why Not a Sociology of Machines? The Case of Sociology and Artificial Intelligence. *Sociology*, 19(4): 557–572.
- Zuboff, S. 1988. In *the Age of the Smart Machine: The Future of Work and Power*. Basic Books, Inc.
- Zuboff, S. 2015. Big Other: Surveillance Capitalism and the Prospects of an Information Civilization. *Journal of Information Technology*, 30(1): 75–89.



# Selektivitäten (subsymbolischer) Künstlicher Intelligenz

Norbert Huchler

## Zusammenfassung

Der Beitrag sondiert unter dem Schlüsselbegriff der Selektivität verschiedene strukturierende Wirkungen von subsymbolischer Künstlicher Intelligenz (KI) als sozialem Phänomen, von der zielgerichteten Entwicklung über die spezifische technische Funktionsweise bis hin zur Einbettung in Nutzungszusammenhänge, verbunden mit latenten gesellschaftlichen Anpassungsprozessen. Damit erweitert der Beitrag die Diskussionen über Diskriminierung und Datenbias um weitere Aspekte latenter sozialer Gestaltung und technikkimmanenter Strukturierungen. Aufbauend auf der Systematisierung von elf KI-Selektivitäten werden zentrale Fragen eines sich wandelnden Mensch-KI- bzw. Mensch-Technik-Verhältnisses andiskutiert, und es wird ein Leitbild für ein mögliches künftiges Verhältnis skizziert, das über ein Konkurrenz- bzw. lineares Substitutionsverhältnis hinausgeht.

## Schlüsselwörter

Selektivitäten • Mensch-KI-Verhältnis • Subsymbologische KI • Latente soziale Folgen • Sozialer Bias • Technik-Bias • Komplexitätsreduktion • Objektivierung

---

N. Huchler (✉)

Institut für Sozialwissenschaftliche Forschung e.V. – ISF München, München, Deutschland  
E-Mail: [norbert.huchler@isf-muenchen.de](mailto:norbert.huchler@isf-muenchen.de)



## 1 Einleitung

Künstliche Intelligenz (KI) wird als eine Schlüsseltechnologie bezeichnet, mit der große Versprechungen und hohe Erwartungen (Hirsch-Kreinsen 2023), aber auch Problematiken (z. B. Zuboff 2015; Crawford 2022) verbunden sind. KI wirft vor allem neue Fragen zum Mensch-Technik-Verhältnis auf, insbesondere mit Blick auf Wissen und Erfahrung sowie Handeln und Kompetenzen. In diesem Beitrag wird KI unter anderem verstanden als Technik der Komplexitätsreduktion und -bearbeitung mittels Objektivierung, Formalisierung und Informatisierung. Damit ist KI immer sowohl mit einer Öffnung als auch einer Schließung von Wissens- und Handlungsspielräumen verbunden – also mit strukturierenden sozialen Wirkungen, die Innovationen ermöglichen, aber sich auch in Form von selbstbestätigenden Prozessen und Pfadabhängigkeiten äußern können. Sie können somit zur Steigerung von Anschluss- und Handlungsmöglichkeiten beitragen, aber auch zu Verengungen und Konflikten führen. Diese strukturierenden sozialen Wirkungen gehen weit über die breit diskutierten Aspekte sozialer Verzerrungen in Bezug auf Datenbias, Training und Interpretation von KI-Output hinaus. Um KI als soziales Phänomen und insbesondere die mit dem KI-Einsatz einhergehenden latenten sozialen Folgen zu fassen, wird der Begriff Selektivität genutzt, der Aspekte des sozialen Bias aufgreift, erweitert und um technikimmanente Aspekte erweitert. Ziel ist es, die strukturierenden Wirkungen von KI auf Wissens- und Handlungsräume zu erkennen und reflektieren. Auf diese Weise wird erkennbar, wie die aktuelle Gestaltung von KI sich sozial auswirkt und wie künftige, evtl. alternative Gestaltungsmöglichkeiten aussehen können. Um Gestaltung und Gestaltbarkeit zu erkennen, braucht es ein Verständnis für die latenten Selektivitäten von KI. Diese werden in der Regel dann sichtbar, wenn sie Irritationen hervorrufen, zum Beispiel in Form von Ungleichheiten und Ungechtigkeiten, aber auch in Form einer Nichtpassung zur aktuellen sozialen Praxis außerhalb von KI. Es geht also darum, über die Selektivitäten von KI einerseits die latenten Folgen von KI und andererseits Gestaltungsräume für das Mensch-KI-Verhältnis aufzuzeigen. Da subsymbolische, konnektionistische KI erstens die typischen Selektivitäten von KI allgemein reproduziert und zweitens durch ihr spezifisches Verfahren eine eigene Form von Selektivität hervorbringt, liegt der Fokus des Beitrags auf dieser aktuell viel Aufmerksamkeit erfahrenden Variante von Künstlicher Intelligenz.

In diesem Beitrag werden zunächst die Besonderheiten subsymbolischer bzw. konnektionistischer KI auf eine Weise gefasst, die sie anschlussfähig macht für die Diskussion um Selektivitäten (Abschn. 2). Subsymbolische KI ermöglicht durch die temporäre Unabhängigkeit von Expert\*innen- und Kontextwissen eine

neue Form der Komplexitätsreduktion, die neue Möglichkeitsräume eröffnet und diese zugleich eigenlogisch strukturiert. Darauf folgt eine Systematik von elf kurz umrissenen Selektivitäten von (subsymbolischer) KI (3). Aus diesen Selektivitäten werden dann Folgerungen für das Mensch-KI-Verhältnis abgeleitet (Abschn. 4). Zentrale Themen werden in der Folge kurz diskutiert, etwa Wissen ohne Verständnis, Aneignung menschlicher Kompetenzen durch KI, KI als Akteur oder als Träger von Vertrauen und Verantwortung (Abschn. 4.1). Schließlich wird ein Leitbild für das künftige Mensch-KI-Verhältnis skizziert, das sich an Komplexität und Koevolution orientiert und von Ersatz und Konkurrenz abgrenzt (Abschn. 4.2).

---

## 2 Subsymbolische/konnektionistische KI als technisches Verfahren

Der Begriff KI wurde auf der Dartmouth Conference 1956 in Amerika als Vision platziert, der zufolge „grundsätzlich alle Aspekte des Lernens und anderer Merkmale der Intelligenz so genau beschrieben werden können, dass eine Maschine zur Simulation dieser Vorgänge gebaut werden kann“ (McCarthy et al. 1955; Übersetzung N. H.). Diese Vision – der technisch-formalen Beschreibbarkeit – wurde sogleich zu einer Mission, um die herum sich mittlerweile zahlreiche unterschiedliche methodische Ansätze und technische Verfahren versammelt haben. Sie reichen von klassischen Formen der „Wenn-dann“-Programmierung über einfache Statistik (Big Data) und Expertensysteme bis hin zur konnektionistischen, subsymbolischen KI (neuronale Netze, Deep Learning). Letztere erfährt mit Programmen wie ChatGPT aktuell große Aufmerksamkeit. In der Folge existieren zahlreiche unterschiedliche Definitionen rund um den programmatischen Sammelbegriff Künstliche Intelligenz (Russell 2010). Weder der eklektische Versuch, die einzelnen Verfahren und technologischen Grundlagen zu bündeln (Mustererkennung, Maschinelles Lernen, künstliche neuronale Netze und Deep Learning, Expertensysteme, Wissensmodellierung, logische Programmierung etc.), noch die generalisierende Abstraktion, etwa in der Definition von Künstlicher Intelligenz als „Fähigkeit einer Maschine, menschliche Fähigkeiten wie logisches Denken, Lernen, Planen und Kreativität zu imitieren“ (Europäisches Parlament 2020), stellen wirklich zufrieden. Entsprechend der Charakteristik eines Missionsbegriffs ist KI in eine Dynamik von Erwartungen und Enttäuschungen (Seising 2021, [Selke]), Fortschrittsbegriffen zwischen Disruption und inkrementeller Verbesserung, Diagnosen von Hochphasen und sogenannten KI-Wintern eingebunden ([Hirsch-Kreinsen]). Die Gleichzeitigkeit von neuen Möglichkeiten und den

damit verbundenen Grenzen spiegelt sich in den hohen Erwartungen und tiefen Enttäuschungen in der Geschichte der KI wider.

Dies mag auch daran liegen, dass in der bis heute prägenden Vision von McCarthy et al. (1955) von Anfang an eine Konkurrenz zum Menschen (zunächst zu seinen kognitiven Leistungen) angelegt war – und damit die Gefahr eines anthropologischen Trugschlusses (Salles et al. 2020; Watson 2019). Die vor-schnelle Gleichsetzung technisch ersetzter (oder ggf. simulierter) Funktionalitäten mit dem menschlichen Denken und Handeln, das zur Erreichung des gleichen Ziels notwendig ist, impliziert eine technologische Aneignung dieser menschlichen Kompetenzen. Beim Ersatz einer Tätigkeit durch KI handelt es sich jedoch lediglich um ein *funktionales* Äquivalent (Huchler 2019), denn die Voraussetzungen, die Art und Weise des Vollzugs sowie die Resultate (inkl. der Nebenfolgen) unterscheiden sich zwischen technischen Verfahren und menschlicher Tätigkeit, auch dann, wenn sie denselben Zweck erfüllen (ebd.).

Das Problem dieses Fehlschlusses zeigt sich besonders in der aktuellen Diskussion rund um die konnektionistische bzw. subsymbolische KI. Auf Basis existierender Grundlagen (z. B. Maschinelles Lernen, Bayessche Netze) gewann der Zweig der subsymbolischen KI-Forschung mit steigenden Rechenkapazitäten schon seit Ende der 1980er Jahre an Relevanz (vgl. Smolensky 1988; Fodor und Pylyshyn 1988). In den letzten Jahren sorgt ein Innovationssetting aus verbesserten methodischen Verfahren, zunehmender Leistung von Hard- und Software, einer breiten Verfügbarkeit von Daten, vernetzten Wertschöpfungs- und Gesellschaftsstrukturen bzw. Nutzungskontexten, gesteigerter Aufmerksamkeit und Forschungsgeldern für eine spürbare Durchdringung der Gesellschaft mit KI (vgl. [Hirsch-Kreinsen]). Dazu kommen erste resultierende, breit nutzbare und greifbare Anwendungen (wie ChatGPT). Auch der Diskurs rund um diese aktuelle Entwicklung ist geprägt durch die erwähnte Mission des Ersatzes menschlicher Tätigkeiten und der technischen Aneignung entsprechender Kompetenzen. Sie wird vor allem über Bilder, Utopien und Dystopien rund um KI in Szene gesetzt ([Gransche, Manzeschke], [Manzeschke, Gransche], [Selke]), manifestiert sich aber teilweise auch in einzelnen konkreten Umsetzungen.

Für eine realistische Sicht auf KI mit ihren Potenzialen und Problematiken für die Gesellschaft ist es hilfreich, die Besonderheiten der konnektionistischen bzw. subsymbolischen KI genauer in den Blick zu nehmen. Sie grenzt sich von der symbolischen KI bzw. sogenannten Expertensystemen dadurch ab, dass sie ohne Wissen über die Lösungswege auskommt und stattdessen mit ausgeklügelten Verfahren statistische Korrelationen bzw. Muster identifiziert, deren Ergebnisse dann bewertet werden. Während symbolische KI ‚händisch‘ schlaue gemacht werden muss, setzt die subsymbolische KI auf eine automatisierte Lösungsfindung,

basierend auf Wahrscheinlichkeiten. Neuere Systeme bilden aktuell ein Netzwerk aus teilweise mehr als 500 Mrd. zunächst offenen Parametern<sup>1</sup>. Auf Basis einer umfangreichen Suche nach Korrelationen baut und optimiert das KI-System selbst Lösungswege in Form eines komplizierten Netzes aus Funktionsbausteinen („Neuronen“) bzw. auf Basis von „Funktionsapproximation“ (Brödner 2022, S. 34). Es sucht so lange nach statistischen Zusammenhängen und kombiniert diese miteinander, bis die gewünschten Resultate entstehen – ohne dass es notwendig (oder auch nur möglich) wäre, in die ‚Blackbox‘ der sich solchermaßen bildenden komplizierten Relationen bzw. Vernetzungen Einblick zu nehmen. Die Einzelprozesse bleiben zwar weiterhin deterministisch, jedoch sind sie extrem zahlreich, abhängig von den zugrunde gelegten Daten, erzeugen komplizierte Wechselwirkungen und verändern sich. Zwar ist das „theorielose“ (ebd.) Vorgehen zunächst sehr aufwendig. Entsprechende Verfahren müssen entwickelt und dann für bestimmte Anwendungsfelder trainiert werden. Jedoch sind die zugrunde liegenden Verfahren der generischen KI breit anwendbar und die Trainingsprozesse sind optimierbar. Daher wird mit entsprechenden Skalierungseffekten gerechnet. So benötigt z. B. ChatGPT zur automatisierten Erstellung von Texten keine aufwendig erzeugten Ontologien und Taxonomien, sondern wird auf Basis umfangreicher Datensätze intensiv auf wahrscheinliche Kombinationen und Konstellationen von Zeichen trainiert – unabhängig von Thema und Kontext. Die Verfahren der subsymbolischen KI lassen sich auf ganz unterschiedliche Datenformen und -quellen anwenden, nicht nur auf Wörter und Texte, sondern auch auf Zahlen, Pixel (Bilder), Töne (Sprache, Musik) etc. Lernende bzw. trainierte KI steckt bereits in zahlreichen Alltagsanwendungen im Arbeit- und Privatleben.

Aufgrund dieser potenziellen Anwendungsbreite gilt KI als Querschnitts- bzw. Schlüsseltechnologie – zum Beispiel als zentraler Treiber der Entwicklung hin zu einer Informationsökonomie (Kämpf und Langes 2021) oder als ein wesentlicher Faktor des digitalen Kapitalismus mit den Potenzialen einer „Reproduktionskraft“ und/oder „Destruktivkraft“ (Pfeiffer 2021, S. 280). Kern solcher Diagnosen ist, dass von der Kombination von probabilistischem statistischem Schließen und Machine Learning eine Steigerung der Adaptivität und Funktionalität von technischen Systemen für den Einsatz zur Bearbeitung komplizierter Aufgaben in komplexen Realumgebungen erwartet wird – vor allem durch einen wesentlich umfangreicheren Zugriff auf vorhandenes, in Daten sedimentiertes Wissen. KI-Systeme sollen, je nach Aufgaben bzw. Eingaben und Einsatz, selbst neue

---

<sup>1</sup> Zudem besteht das Sprachmodell GPT-4, das aktuell ChatGPT zugrunde liegt, offenbar aus einer Kombination mehrerer Modelle, über deren Zusammenwirken wenig bekannt ist (Spriestersbach 2023).

Annahmen bilden und testen sowie unbekannte Lösungsstrategien entwickeln. Ziel ist die Erschließung neuer Automatisierungspotenziale in Bereichen, die aufgrund ihrer kognitiven und/oder manuellen Komplexität bislang als nicht automatisierbar galten. Die Erwartungen an KI richten sich also nicht nur darauf, dass sich relevantes Wissen, Tätigkeiten und Prozesse in Daten abbilden und rechenbar (objektivierbar) machen lassen. Vielmehr macht es die Kombination aus statistischer Wahrscheinlichkeit und tiefem maschinellem Lernen möglich, Kontingenz bzw. Komplexität zunächst offenzuhalten und nicht ex ante etwa durch fix geplante Abläufe zu beherrschen. Es geht also um eine Abkehr von (symbolischer) KI, die auf Basis von Expertenwissen aufwendig für spezifische Aufgaben im Voraus programmiert werden muss, und eine Hinwendung zu ebenfalls sehr aufwendig programmierten und trainierten selbstlernenden, adaptiven subsymbolischen und generischen Systemen.

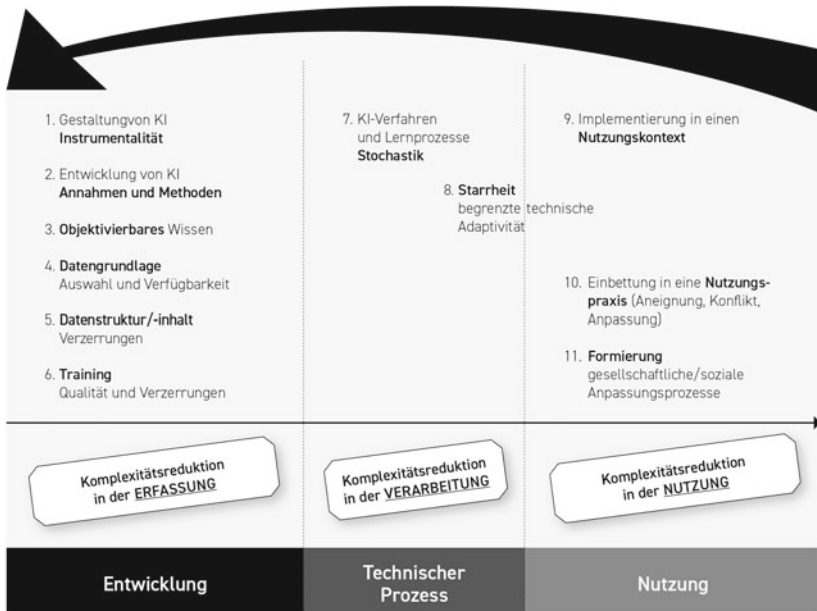
Die Verfahren der subsymbolischen KI gehen mit neuen Kontingenzen ([Heinlein]), Unsicherheiten und Problematiken einher. Daher ist es wichtig, dieses spezifische technische Verfahren einzuordnen. Hierfür soll im Folgenden genauer auf die speziellen Eigenheiten der Nutzung von konnektionistischer/subsymbolischer KI in der sozialen Praxis geblickt werden – insbesondere auf die soziale Strukturiertheit von KI und die soziale Strukturierung durch KI.

---

### **3 Technik-Bias und sozialer Bias – Die Selektivitäten von KI**

Menschen gestalten und strukturieren die Funktionen und Wirkungsweisen von Technik sowohl in der Entwicklung als auch in der Anwendung. Technik wiederum nimmt Einfluss auf das Handeln und Denken von Menschen. Aus einer sehr breiten (eröffneten) Optionenvielfalt bzw. Kontingenz des potenziell Möglichen wird durch die Eigenschaften der Technik selbst, mittels ihrer Gestaltung wie auch durch ihren Gebrauch, immer wieder neu ausgewählt. Dies wird im Folgenden mit dem Begriff Selektivitäten umschrieben. Gestaltung und Anwendung von Technik fokussieren immer auf einen Ausschnitt des Möglichen, strukturieren bestehende Handlungs- und Wissensräume vor und engen sie damit ein – und eröffnen so zugleich (bestimmte) neue Handlungs- und Wissensräume. Der Techniksoziologe Ingo Schulz-Schaeffer verwendet den Begriff der Selektivität in ähnlicher Weise, zum Beispiel als „im technischen Algorithmus verkörperten Handlungssinn“, der als „Selektivität des sachtechnischen Ablaufs“ die arbeits- teilige Interaktions- bzw. „Delegationsbeziehung“ zwischen Mensch und Technik

strukturiert und sich dabei vom „intentionalen Handlungssinn des Akteurs“ unterscheidet (Schulz-Schaeffer 2007, S. 9). „Das Verhalten der (Teil-)Akteure auf beiden Seiten dieser Delegationsbeziehung ist Ausdruck wechselseitig aufeinander bezogener sinnhafter Selektivität“ und nicht intentionalen Sinns (ebd.). In diesem Sinne bilden Selektivitäten eher latente Wirkungszusammenhänge, die systematisch mit an Funktionalitäten ausgerichtetem Handeln verbunden sind. Schulz-Schaeffer fasst entsprechend Sinnzuschreibungen an Technik als Aspekt „erzeugter Selektivität“ – bei der Entwicklung, beim anschlussfähigen technischen Ineinandergreifen der Systeme und auch „aus der Perspektive des Nutzers“ – bzw. als „sinnhafte Abläufe“ entlang intendierter Funktionalitäten und grenzt diese ab von „Ursache-Wirkungs-Zusammenhängen, die der Natur zugeschrieben werden“ (Schulz-Schaeffer 2008, S. 3142). Damit adressiert er die im Folgenden differenzierten Phasen der Entwicklung, des technischen Prozesses und der Nutzung (s. Abb. 1) und verweist auf die soziale Ausgestaltung und die handlungsstrukturierenden Wirkungen von Selektivitäten.



**Abb. 1** Die Selektivitäten (subsymbolischer) KI. (Quelle: Eigene Darstellung)

Der Blick auf die Selektivitäten von KI ermöglicht es, sich zumindest ein Stück weit der großen Herausforderung zu stellen, KI passgenau und nachhaltig gesellschaftlich einzubetten. Selektivitäten verweisen zudem darauf, dass technische Systeme nur in der sozialen Praxis wirksam werden ([Heinlein]), dass Technologie nie ‚neutral‘ ist und dass technologische Entwicklungen wie auch die Erzeugung technischer Artefakte keinem determinierten Verlauf bzw. keinen unveränderbaren Pfaden folgen. Das soziale Phänomen KI ist als ein permanenter soziotechnischer Gestaltungsprozess zu verstehen. KI-Entwicklung und KI-Nutzungspraxis folgen keinem technologischen Determinismus, sind aber durchaus durch technikimmanente Eigenschaften geprägt (vgl. Lutz 1987), zum Beispiel die spezifische Funktionsweise subsymbolischer KI.

Ziel der Verwendung des Begriffs der Selektivität zur Beschreibung der Funktionsweise von subsymbolischer KI und der sozialen Praxis im Zusammenhang mit dieser Technologie ist der Versuch, die bisher recht eingengegte Diskussion um den ‚sozialen Bias‘ von KI zu erweitern. Unter den Stichworten Diskriminierung und Datenbias werden soziale Verzerrungen, die typischerweise mit der Nutzung von (konnektionistischen) KI-Systemen einhergehen (können), breit diskutiert. KI schreibt jedoch nicht nur Verzerrungen im zugrunde liegenden Datenmaterial fort, sondern verstärkt diese auf eigene Weise und kann vor allem selbst neue Verzerrungen bzw. einen technischen Bias erzeugen. Einige dieser latenten Selektivitäten bzw. Strukturierungen der sozialen Praxis sind bislang weniger im Blick, obwohl sie eine Brücke bilden könnten, um KI in ihrer gesellschaftlichen Funktionsweise zu verstehen. Im Übrigen gehen mit einer uninformierten und unprofessionellen Anwendung von KI bzw. Machine Learning zahlreiche Gefahren einher (Pfeiffer 2021, S. 283 ff.).

Die im Folgenden systematisierten Selektivitäten zielen auf die latenten sozialen Folgen von KI. Sie beziehen sich weniger auf die Ebene des KI-Diskurses (der ebenfalls soziale Wirkungen ausübt) und nicht auf einen „bewusst destruktive[n] Einsatz von KI/ML“ (Pfeiffer 2021, S. 283). In der absichtsvollen Nutzung von KI zur Täuschung, Manipulation und Kontrolle (z. B. durch DeepFakes) liegt sicherlich eine der akut größten gesellschaftlichen Gefahren für demokratische Strukturen und demokratische Teilhabe, den sozialen Zusammenhalt und das Vertrauen in gesellschaftliche Institutionen. Zu nennen wären hier auch die militärische Nutzung und entsprechende Folgen für lokale wie globale Konflikte sowie die Nutzung von KI als Instrument der Herrschaft, Macht, Überwachung und (Selbst-)Kontrolle (Zuboff 2015), die Bedrohung durch „Algokratie“ (Danaher 2016) oder das von Daten durchdrungene „algorithmische Leben“ (Amoore und Piotukh 2016). Dieser Beitrag zielt jedoch nicht

auf beabsichtigte soziale Effekte von KI – sei es in Richtung positiver Gestaltung oder mutwilligen Missbrauchs.<sup>2</sup> Vielmehr verweisen die im Folgenden herausgearbeiteten Selektivitäten auf latente soziale Wirkungen, hieraus entstehende Konflikte und un intendierte Nebenfolgen, wie sie zum Beispiel dann zum Tragen kommen, wenn Design und Einsatz von KI dem Anwendungsgegenstand und dessen Umwelt nicht gerecht werden. Die Selektivitäten gehen aus verschiedenen Formen der Komplexitätsreduktion an unterschiedlichen Stellen im KI-Entwicklungs- und Anwendungsprozess hervor und stellen potenzielle Gestaltungs- und Konfliktfelder dar.

Im Folgenden werden elf unterschiedliche Selektivitäten von KI systematisiert und in einer Chronologie beschrieben (s. Abb. 1 unten), von der Entwicklung und Ausgestaltung der KI-Systeme über die technischen Funktionsweisen subsymbolischer KI bis hin zu ihrer Nutzung. Dabei wird eine dynamische Perspektive vorgeschlagen. Das heißt, Konflikte, Anpassungs- und Aneignungsprozesse sowie latente Formierungen in der Nutzungspraxis wirken sich wiederum auf die weitere bzw. zukünftige Entwicklung und Gestaltung von KI-Systemen aus. Besonders problematisch werden die Selektivitäten, wenn sie einen selbstreferenziellen und sich selbst verstärkenden Zirkel bilden und nicht irritiert werden – zum Beispiel durch Konflikte in der Passung zum Gegenstand, im Kontext oder in der Praxis. Hier bieten sich verschiedene Bruchlinien bzw. Gelegenheiten zur Irritation an, wie zum Beispiel a) der Prozess der Reduktion von Komplexität bei der Entwicklung von KI-Anwendungen, insbesondere bei der Erfassung und Zugänglichmachung (mittels Objektivierung) von Wissen, b) die Komplexitätsreduktion bei der technischen Verarbeitung (z. B. Stochastik) und c) die Komplexitätsreduktion bei der sozialen Einbettung in eine Nutzungspraxis. An diesen Stellen kann es zu Passungsproblemen und Konflikten kommen, die auf Selektivitäten aufmerksam machen und so Irritationen und Innovationen ermöglichen.

---

<sup>2</sup> Rechtliche Aspekte wie Regulierungsbedarfe, Urheberrecht, Eigentumsrechte, Persönlichkeitsschutz, Datenschutz, Datensicherheit, Privatheit etc. werden hier ebenso wenig adressiert wie die ökologischen Problematiken von KI oder auch mögliche psycho-soziale und gesundheitliche Folgen, Abhängigkeiten, soziale Distanzierung sowie Problematiken simulierter Persönlichkeit (KI als Trainer, Partner, Angehöriger). Globale Ungleichheiten, die mit KI-Entwicklungsprozessen verbunden sein können, werden nur angeschnitten.



### **3.1 Selektivität bei der instrumentellen Gestaltung von KI (Instrumentalität)**

Eine entscheidende Form der sozialen Selektivität von KI beginnt schon beim Systemdesign und den damit verbundenen Entscheidungen (Brödner 2020; Mittelstadt et al. 2016; Friedman und Nissenbaum 1996): welchen Zielen das System dienen soll und mit welchen Erwartungen es verbunden ist, auf welche Formen von Daten es ausgerichtet sein soll, welche Schnittstellen (Input, Output) geplant sind, welche Formen der Output annehmen kann und welche Anschlussmöglichkeiten damit verbunden sind, wie das System in einen sozialen Kontext des Nutzungsprozesses einzubinden ist, und nicht zuletzt, welche Folgen antizipiert oder ausgeblendet werden. Schon im Design-Prozess wird die mögliche Nutzungsbreite von KI selektiv kanalisiert. KI wird interessen- und erwartungsbasiert ins Laufen gebracht und agiert dann in dem Handlungsrahmen, der ihr durch Zielsetzung und Funktionen eingeschrieben ist. Damit verbunden ist die Frage, nach wessen Interessen die Systeme designt werden bzw. wer in den Designprozess eingebunden ist. Eine typische Problematik ist, dass „die Implementierung oft allein in die Hände von Informatik und Data Science gelegt wird, ohne die Expertinnen und Experten für den angezielten Anwendungskontext mit einzubeziehen“ (Pfeiffer 2021, S. 284). Damit ist die Gebrauchstauglichkeit bzw. die Passung der zu entwickelnden Lösungen zum entsprechenden Anwendungskontext bereits gefährdet. Eine zusätzliche Rolle spielen die zur Verfügung stehenden bzw. gestellten Ressourcen in Form von Know-how, Zeit, Budget etc.

### **3.2 Selektivität der Annahmen, Modelle und Operationalisierungen (Annahmen und Modelle)**

Selbst bei generalistischer bzw. breit einsetzbarer KI liegen ihrem Design Thesen über Anwendungszusammenhänge zugrunde. Dies betrifft bereits die Zielsetzung – zum Beispiel: einen Text zu generieren, ein Bild zu erstellen, einen Prozess zu optimieren. Das eigentlich kontextungebundene stochastische KI-Verfahren wird, um sein funktionales Leistungsversprechen erfüllen zu können, in einen algorithmischen Prozess eingebunden, der mit (teils impliziten) theoretischen (Wirkungs-)Modellen verknüpft ist. Diese bleiben oftmals intransparent, haben aber Einfluss auf die Ergebnisse, deren Darstellung und Interpretationsmöglichkeiten. Der Datenoutput wird im einbettenden Algorithmus interpretiert bzw. in für den Nutzungskontext sinnvolle Informationen transformiert, mit den formalen Zielsetzungen und Funktionen verknüpft und so anschlussfähig für

einen geplanten Nutzungskontext gemacht (vgl. Brödner 2020). Bei manchen Anwendungen basieren KI-Systeme auf sehr ausgefeilten und verlässlichen Lernalgorithmen, werden aber mit sehr vereinfachten Wirkungsmodellen in den prozessierenden Programmen verbunden (vgl. Brödner 2019). Die informationstechnische Modellierung und Formalisierung sozialer Arbeitspraktiken kann dabei leicht scheitern (Rohde et al. 2017). Ein Beispiel sind KI-Systeme für die Personalauswahl, die versprechen, zwar auf Basis von elaborierten Schrift-, Stimm- oder Videoanalysen, aber anhand von stark vereinfachten psychologischen Modellen auf die Eignung für bestimmte Tätigkeiten schließen zu können. Die Komplexität des Kontexts wird dabei über die verbundenen Modellannahmen stark reduziert. Aber selbst wenn ‚nur‘ Bilder in einem bestimmten Stil und mit einem bestimmten Inhalt erzeugt werden, wird diese Aufgabe auf eine bestimmte Weise interpretiert und prozessiert, die andere Interpretationen ausschließt. Beim Blick auf die sozialen Selektivitäten und den Bias von KI sollten deshalb die den jeweiligen KI-Anwendungen hinterlegten Annahmen und Modelle gesondert Beachtung finden. Dasselbe trifft auf die latenten Strukturierungen zu, die von der algorithmischen Struktur bzw. der Einbettung in eine Anwendung ausgehen.

### **3.3 Selektivität der technisch-formalen Repräsentation (Objektivierung)**

Als Technologie der Informationsverarbeitung unterliegt KI typischen technologieimmanenten Selektivitäten bei der Bewältigung (sozialer und physischer) Komplexität durch technisch-formale Repräsentation (z. B. Digitalisierung) bzw. bei der Erfassung und datenbasierten Verarbeitung komplexer soziotechnischer Herausforderungen. Sie verarbeitet nur das, was in Daten repräsentierbar ist. Dies ist auf drei Ebenen eingeschränkt: aufgrund a) der sozio-materiellen Komplexität der physischen und sozialen Welt, b) eines ständigen (teils rekursiven) Wandels und c) Nicht-Formalisierbarkeit (vgl. Huchler 2022).

Sozio-materielle Grenzen: Sozial wirksame KI-Systeme agieren in einem Umfeld mit zunächst einmal potenziell unendlicher Komplexität in Bezug auf die physische und soziale Welt und zugleich unter deren Restriktionen. Eine unendliche Menge an Faktoren, Zusammenhängen, Prozessen, Ambivalenzen und Dynamiken trifft auf beschränkte und vorstrukturierte Ressourcen, Fähigkeiten und Optionen – in einem sich permanent wandelnden Gesamtsystem. Hinzu kommen die ausschnittshafte Erfassung komplexer Realität, die immer limitierte Verfügbarkeit von Daten, die ausschnittshafte Repräsentanz durch Rechenoperationen, deren Anzahl begrenzt ist, die Rechenleistung etc. Insofern muss für

KI-Systeme der Gegenstand und Nutzungskontext eingegrenzt werden. Sie funktionieren dann am besten, wenn der Gegenstand kompliziert und nicht komplex ist bzw. wenn es sich um ein geschlossenes und nicht um ein unkontrollierbares (offenes und sich wandelndes) Setting handelt – zum Beispiel ein Spiel wie Go – oder wenn der Gegenstand einem solchen Zustand nahe ist, zum Beispiel die Analyse, Variation oder Misch-Kreation von Texten, Bildern, Musikstücken etc. Je größer der Anwendungsgegenstand gemacht wird, umso problematischer wird die Ausblendung bzw. Reduktion seiner Komplexität.

Rekursive Grenzen: Strukturierend wirkt auch die Dynamik, d. h. der Effekt, dass durch KI-Anwendungen technisierte Prozesse ihre eigene Umwelt verändern und dabei ständig neue Aufgaben und Nebenfolgen produzieren. Der Anwendungsgegenstand und sein Umfeld wandeln sich durch die Bearbeitung mit KI permanent und konstituieren sich mit jeder technischen Intervention neu. In Anlehnung an die Dilemmata der Rationalisierung (Berger und Offe 1980) und die Ironie der Automatisierung (Bainbridge 1983) lässt sich zeigen, dass mittels KI automatisierte Prozesse immer von vor-, parallel- und nachgelagerten Prozessen begleitet werden, die einen ermöglichenden, sichernden oder erhaltenden Folgeeffekt haben. Die Transformationsdynamik von KI ist durch die permanente Neuschaffung von derartigen Aufwänden gekennzeichnet. Dies ist systematisch damit verknüpft, dass KI eine begrenzte und selektive Methode zur Bewältigung von (sozio-technischer) Komplexität ist und immer neu sozial eingebettet werden muss.

Grenzen der Formalisierbarkeit: Wenn der Gegenstand nicht-formale Wissensinhalte beinhaltet und damit komplex ist, steigt nicht nur der Aufwand einer passgenauen Erfassung, sondern bestimmte Inhalte können überhaupt nicht bzw. immer nur reduziert/abstrahiert erfasst werden. Dies betrifft die Grenzen der Explikation und Objektivierung von Wissen bzw. der technischen Übersetzung von Daten in Information und Wissen (Schmiede 2006). KI-Systeme werden zwar weiterhin durch Menschen entwickelt und optimiert, in Nutzungskontexte eingebunden und am Laufen gehalten, aber mit ihnen können zunehmend nicht nur kontextlose Daten erzeugt, sondern sinnvolle Informationen und Anschlussmöglichkeiten generiert werden. KI-Systeme wandeln zunehmend Daten in Informationen, wenngleich mit (systembedingten) Lücken (s. Abschn. 3.7). Die Repräsentanz von an Bedeutungen und Kontexte gebundenem sinnhaftem Wissen durch (objektivierte bzw. explizite) Informationen (z. B. in Form von Texten und Bildern) sowie deren Transformation in rechenbare Daten sind an einen Abstrahierungs- bzw. Reduktionsprozess gekoppelt. Der damit notwendigerweise verbundene Wissens- und Informationsverlust zeigt die Grenzen der Formalisierbarkeit von implizitem Wissen (Polanyi 1985; Neuweg 2015) und

Erfahrungswissen (Böhle 2009) auf. Implizites Wissen, Erfahrungs- oder Körperwissen, Kompetenzen, subjektivierendes Arbeitshandeln etc. gelten als nur teilweise oder selektiv in Daten transformierbar (Polanyi 1985; Rammert 2003; Schmiede 2006; Böhle 2009). Nach Schmiede (2006, S. 473) geht der gesellschaftliche Fortschritt mit einer Zunahme von Wissen *und* Nicht-Wissen einher. Damit stehen Wissen und Nicht-Wissen in einem dialektischen Verhältnis. KI kann genutzt werden, um Kontingenz zu verarbeiten, neue Informationen zu liefern und daraus neues Wissen zu generieren. Gleichzeitig ist mit KI eine Zunahme der Komplexität verbunden, die neue Bereiche des Nichtwissens eröffnet. Nichtsdestotrotz bleibt die Kernvision hinter KI, dass sich (auch komplexe soziale) Realität umfassend bzw. ausreichend in Daten darstellen lässt und damit berechenbar gemacht werden kann. Hier entsteht dann ein Konflikt, wenn relevante Teile der Praxis nicht repliziert und damit nicht berücksichtigt werden (können).

In komplexen und sich verändernden Kontexten sind KI-Lösungen daher immer selektiv in dem Sinne, dass sie notwendigerweise fragmentarisch sind. Sie können ihren Gegenstand nie vollständig repräsentieren und sie können auch veralten (Pfeiffer 2020). In der Praxis zeigt sich diese Selektivität von KI dort, wo KI-Systeme entweder ausschnitthaft eingesetzt werden oder systematisch an Grenzen stoßen, Defizite aufweisen oder Konflikte verursachen. Natürlich können KI-Systeme innerhalb bzw. trotz dieser Grenzen genutzt werden, Prozesse können so auch komplett technisiert werden. Jedoch ist dies mit spezifischen Voraussetzungen (vorgelagert), veränderten Nutzungs- (parallel) und Anschlussprozessen (nachgelagert) verbunden – und damit mit anderen Ressourcen, Bedingungen und Folgen. Diese Änderungen werden durch die Spezifik der Selektivität technisch-formaler Repräsentation (Objektivierung) bestimmt. Sie erleichtert technische Lösungen und Anschlussfähigkeit, geht aber mit möglichen Passungsproblemen einher. Die Kenntnis dieser Selektivität sensibilisiert damit aber auch für Mensch-Technik-Differenzen und ermöglicht so eventuell neue Perspektiven für die Befähigung des Menschen durch KI bzw. für eine komplementäre Technikgestaltung (vgl. Huchler 2022).

### **3.4 Selektivität durch Verfügbarkeit, Auswahl und Qualität der Daten (Datengrundlage)**

Bias-Problematiken in Bezug auf Daten fangen schon bei der Verfügbarkeit von Daten an, also bei der Frage, worüber Daten vorliegen und worüber nicht. Daten werden in spezifischen Kontexten selektiv erzeugt und verbunden mit Interessen verfügbar gemacht. Daher gilt es, den Entstehungshintergrund von Daten zu

betrachten, um deren Qualität bzw. Aussagekraft beurteilen zu können und um einschätzen zu können, wie stark sie bereits dadurch selektieren, dass sie nur einen Teil der Wirklichkeit repräsentieren, und dies auf eine spezielle Weise. Ein weiterer entscheidender selektiver Schritt ist die Auswahl der Daten, die für ein KI-System als Grundlage genutzt werden. Die Kriterien für die Auswahl sollten reflektiert werden. Das Gleiche gilt für die Qualität der Daten und deren Passung zum Gegenstand und den anvisierten Funktionen des KI-Systems (vgl. Pfeiffer 2021, S. 284). Oftmals wird zudem darauf verwiesen, wie aufwendig die Herstellung einer sauberen und passenden Datengrundlage ist und dass in diesem Zusammenhang oftmals auf global verteilte unsichtbare Arbeit bzw. „ghost work“ (Gray und Suri 2019) zurückgegriffen wird. In diesem Fall wird im Herstellungsprozess soziale Ungleichheit ausgenutzt und verfestigt.

### **3.5 Selektivität sozial verzerrter Daten (Datenstruktur/-inhalt)**

Eine breit diskutierte Quelle für selektive Wirkungen von KI ist die Fortschreibung von Verzerrungen, die bereits in den Daten angelegt sind (Mittelstadt et al. 2016; Friedman und Nissenbaum 1996). Die Datenbasis von KI-Systemen ist nicht nur immer begrenzt und wird gezielt ausgewählt, sie enthält (insbesondere in sozio-technischen Anwendungskontexten) auch oft sozial generierte und damit mehrfach verzerrte Daten (ebd.). Soziale Daten beruhen bereits bei ihrer Entstehung auf immer unvollständigen und teilweise (kulturell, interessen-geleitet, wissensbasiert etc.) verzerrten Operationalisierungen und bilden so immer nur einen sozial konstruierten Ausschnitt von Wirklichkeit ab. Zugleich ist auch ihr Gegenstand (also die Gesellschaft) durchdrungen von Verzerrungen, wie sie zum Beispiel entlang von Diskriminierung (u. a. in Bezug auf Geschlecht, Abstammung, Zugehörigkeit) diskutiert werden. Diese Diskussionen sind jedoch thematisch sehr eng auf bestimmte, gut bekannte Verzerrungen fokussiert. In den Blick genommen werden sollten auch die Fortschreibungen weniger offensichtlicher Ungleichheiten, Machtverhältnisse, Erwartungen und Deutungen sowie der Artefakte (inklusive der vermeintlich objektiven und damit Legitimität generierenden Daten), die aus Datensätzen reproduziert, verstärkt oder erzeugt werden; zum Beispiel über Wertschöpfung, Umwelteinflüsse, Kriminalität usw. Verschärft wird die Problematik dadurch, dass KI-Systeme nicht nur solche artifiziellen Zusammenhänge fortschreiben (anonymisiert und als Verfahren legitimiert), sondern die statistischen Verfahren auch dazu tendieren, gefundene Korrelationen hervorzuheben und zu verstärken bzw. die Vielfalt an den Rändern der Verteilung

abzuschneiden, worauf selten mit sachgemäßen Gewichtungen gegengesteuert wird (vgl. Pfeiffer 2021, S. 284). Damit wird Masse betont und Besonderheit ausgeblendet, was den standardisierenden und normierenden Charakter von KI ausmacht. KI schreibt also nicht nur Selektivitäten fort, sondern verstärkt sie auch auf eigene Weise.

### **3.6 Selektivität durch Trainingsprozesse (Training)**

Die Wirkungen von KI werden auch durch das Training selektiv beeinflusst, zum Beispiel durch die Beurteilung von Resultaten (supervised learning) oder auch durch entsprechende Gewichtungen (reinforced learning). Auch so kann Bias in Form von Unvollständigkeiten, impliziten Annahmen, Erwartungen, Interessen und Deutungen in KI-Systeme eingeschrieben werden (Diakopoulos 2015). Werden KI-Systeme anhand von Simulationen trainiert, können sich diese Problematiken zusätzlich automatisch verstärken, da Verzerrungen (Bias, Fehlannahmen oder einfach nur Unvollständigkeiten und Abstrahierungen) der Simulationen reproduziert und überbetont werden. Die aufwendigen Trainingsprozesse werden in der Regel an billige Arbeitskräfte externalisiert (Gray und Suri 2019), die mehr oder weniger gut eingewiesen werden, oft unter sehr schlechten Bedingungen arbeiten und keinen direkten praktischen oder fachlichen Bezug zum späteren Anwendungskontext aufweisen – das gilt teilweise selbst beim Trainieren von Bildererkennungstechnik in der Medizin. Die durch Trainingsvorgaben und den Trainingsprozess implizit sich einspielenden Strukturierungen gilt es genauer in den Blick zu nehmen.

### **3.7 Selektivität durch Wahrscheinlichkeitsorientierung (Stochastik)**

Auch die mathematischen/statistischen Methoden subsymbolischer/konnektionistischer KI bergen eigene Selektivitäten bzw. Verzerrungen, die selten diskutiert werden. Im Gegensatz zur symbolischen KI basiert das Innenleben der subsymbolischen KI nicht auf Modellannahmen und Expertenwissen, sondern im Kern auf Korrelationen bzw. „Funktionsapproximation“ (Brödner 2022). Mit diesem Ansatz übernimmt die KI alle Probleme und Grenzen der Sozialstatistik, d. h. einer Statistik, die versucht, komplexe sozio-technische Zusammenhänge (und nicht nur komplizierte abstrakte Prozesse) zu erfassen. Dazu gehören

zahlreiche methodische Einzelprobleme, etwa Scheinkorrelationen im Datenrauschen, statistische Verzerrungen (z. B. Tendenz zur Mitte, Nichtberücksichtigung von Ausreißern etc.), Passung der Skalenniveaus, Selbstverstärkungseffekte und Probleme im Umgang mit notwendigerweise unvollständigen Daten (vgl. Pfeiffer 2021, S. 284). Zu diesem Zweck wurden in der Statistik verschiedene Qualitätskriterien und Methoden der Qualitätsprüfung entwickelt, die sich jedoch nicht einfach auf das Vorgehen subsymbolischer KI übertragen lassen. Darüber hinaus gibt es – zumindest in der Forschung – Leitsätze zur Qualitätssicherung, wie zum Beispiel: erst (theoriegeleitet) Hypothesen bilden und dann nach signifikanten Korrelationen suchen – und nicht umgekehrt. Dies schließt sowohl eine permanente einfache Variation bestehender Annahmen als auch die massenhafte stichprobenartige Überprüfung möglicher Korrelationen (als automatisierte Hypothesenbildung und -überprüfung) aus, da beides theoriefrei erfolgt. Lernende KI-Systeme hingegen basieren auf einer „theoriefreien“ Massensuche nach Korrelationen, auf deren Grundlage Kategorien gebildet werden (Brödner 2022). Auf diese Weise kommen sie oft zu sehr stabilen Ergebnissen. Dies ist jedoch mit Qualitätsrisiken verbunden. Zum Beispiel kann die massenhafte Verknüpfung von selektiven Korrelationen am Ende einzelne Fehlkorrelationen zwar zufällig ausgleichen, aber auch aufschaukeln und so falsche Ergebnisse erzeugen. Auch kann der richtig gerechnete Durchschnitt an der Sache vorbeigehen; zum Beispiel wenn aus sehr vielen Literaturquellen eine wahrscheinlichste Durchschnittsquelle kreiert wird, die es nicht gibt (was irreführenderweise oft als „Halluzinieren“ bezeichnet wird). Hinsichtlich der strukturierenden Wirkungen, die von diesen systemimmanenten Selektivitäten der subsymbolischen KI-Methoden ausgehen, besteht noch Forschungsbedarf.

### **3.8 Selektivität technischer Adaptivität (Starrheit)**

Technische Systeme strukturieren ihre Nutzungspraxis in gewisser Weise vor; nicht nur dadurch, dass sie einen gewissen Anwendungshorizont vorgeben, sondern durch die in der Regel geringe Veränderbarkeit ihrer Funktionalitäten. In der Regel passen die Nutzer\*innen ihre Anwendungspraxis an die technischen Systeme an oder umgehen deren Restriktionen. Eher selten ist der umgekehrte Prozess, dass sich also die technischen Systeme in der bzw. durch die Nutzungspraxis anpassen und weiterentwickeln. Selbst interaktive KI-Systeme, die flexibel auf das Gegenüber reagieren und sich durch maschinelles Lernen anpassen, erweitern auf diese Art in der Regel nicht ihre grundlegenden Funktionalitäten und damit die Anwendungsmöglichkeiten jenseits vorbestimmter Optionen.

Insofern lassen sich zwei Formen der Anpassungsfähigkeit von KI-Systemen unterscheiden: a) die „assimilierende Adaptivität“ (Huchler 2019), die sich entlang der gesetzten Eigenlogik bzw. der Funktionen der KI-Systeme gestaltet, d. h. das System wird so gestaltet, dass es auch bei unbekanntem Input reagieren und seine Funktionalitäten aufrechterhalten kann; und b) die „komplementäre Adaptivität“ (ebd.), was bedeuten würde, dass KI-Systeme ihre Funktionalitäten in der Interaktion verändern und weiterentwickeln und (zumindest zum Teil) durch das Gegenüber gestaltet werden (können). Damit geht komplementäre Adaptivität weit über die klassische Werkzeugmetapher hinaus, da die Nutzenden im Nutzungsprozess zu Werkzeugmacher\*innen würden. Eine komplementäre Adaptivität ist extrem aufwendig, fragil und birgt Risiken; sie ist aber zugleich mit großen Innovations- und Befähigungspotenzialen verbunden. Die Möglichkeit einer solchen Selbstveränderung (bzw. Umprogrammierung) ist in KI-Systemen zumindest potenziell angelegt.

### **3.9 Selektivität bei der Implementierung (Nutzungskontext)**

KI-Systeme werden in einem konkreten Nutzungskontext realisiert bzw. wirksam; sei es, dass sie in eine konkrete Anwendung integriert sind, mit einer Dienstleistung verbunden werden oder in einen Arbeits-/Wertschöpfungsprozess integriert sind. Die Art der Implementierung in einen Nutzungskontext strukturiert auch die Nutzungsmöglichkeiten. Welche Anschlussmöglichkeiten werden nahegelegt bzw. eröffnet? Welche Nutzungsformen werden (z. B. durch Funktionalität) belohnt, welche ausgeschlossen? Oftmals entscheidet die Implementierung in einen Nutzungs- und in der Regel auch Verwertungszusammenhang über die gewollten und ungewollten Wirkungen von KI-Systemen; wie zum Beispiel, ob sie Denk- und Handlungsräume eröffnen oder einengen, qualifizieren oder dequalifizieren, nützlich und unterstützend sind oder anstrengend und belastend. Die sozialen Selektivitäten von KI wirken also nicht nur über die Entwicklung der Systeme, die Daten und das Training, sondern auch über die Art und Weise ihres Einsatzes.



### **3.10 Selektivität bei der sozialen Einbettung (Nutzungspraxis)**

Nicht zuletzt hängt die soziale Wirkung von KI von der individuellen Nutzungspraxis ab, die teilweise stark abweichen kann von den Intentionen und Erwartungen beim Design und bei der Implementierung der Systeme in einen Anwendungskontext. Es können unabsichtliche und unerkannte Fehler gemacht werden. KI-Systeme können aber auch absichtlich anders als vorgesehen genutzt werden – in einem als schädlich angesehenen Sinne oder auch als positiver innovativer Schritt. Die Vermittlung zwischen den über das Design und die Implementierung in die KI-Systeme (in Abstimmung mit ihren jeweiligen technischen Eigenschaften) eingeschriebenen Funktionalitäten und Erwartungen einerseits und der (individuellen und kollektiven) Nutzungspraxis andererseits kann unterschiedliche Formen annehmen; wie zum Beispiel a) konfliktlose Anpassung an die vorgesehenen Nutzungsprozesse, b) Konflikt und in der Folge reduzierte Akzeptanz oder Ablehnung, c) Selbstermächtigung durch Technikaneignung, das heißt zum Beispiel eine innovative Nutzung, um Handlungsräume und Wissen zu erweitern – gegebenenfalls auch über die vorgesehenen Funktionalitäten hinaus.

### **3.11 Selektivitäten in gesellschaftlichen Anpassungsprozessen (Formierung)**

Wie bei anderen Techniken auch passen sich Gesellschaft und Individuen an die Anforderungen von Technik an, indem für die Entfaltung ihres Nutzens die notwendigen bzw. vorgesehenen Bedingungen geschaffen werden. Die „richtige“ Nutzung von technischen Systemen wird im Alltag in der Regel – bzw. wenn alles gut geht – durch Funktionalität und Einfachheit belohnt. Nicht nur die Menschen formen die Technik, sondern auch die Technik die Menschen. KI-Systeme wirken dabei jedoch nicht nur auf der Ebene der Handlungsprozesse, sondern vor allem latent auf kognitiver und sozialer Ebene, etwa in Bezug auf die Aufrechterhaltung und Weiterentwicklung von Wissen, Kompetenzen und Erfahrung oder die Ausdrucksformen von Sozialität, Kreativität, Ästhetik und Emotionen, zum Beispiel bei Kommunikations- und Emotionserkennungssystemen. KI-Selektivitäten werden auch insofern latent relevant, als sie sich im KI-Einsatz niederschwellig manifestieren und Nutzer\*innen wie Nutzungskontext verändern. KI-Systeme werden einerseits nach gesellschaftlichen und individuellen Bedürfnissen gestaltet, andererseits aber passen sich umgekehrt auch die Systemumgebung und die individuelle Nutzungspraxis an die Bedürfnisse der

technischen Systeme an, damit diese ihre Wirkung entfalten können. So ließe sich das Autonome Fahren umso leichter realisieren, je mehr die Komplexität des Verkehrssystems und -verhaltens reduziert würde. KI kann umso passendere Bilder, Texte und Musikstücke erstellen, je konformistischer entsprechende Erwartungen an Ästhetik, Inhalt, Ausdruck und Stil sind. Diese Standardisierungsaspekte von KI-Selektivitäten sind im Hinblick auf die „transformativen Effekte“ (Mittelstadt et al. 2016) von KI zu berücksichtigen. Ein typischer Effekt des Strebens nach technologischer Beherrschung von Komplexität und Unsicherheit ist die Standardisierung des entsprechenden sozialen Umfelds bzw. der entsprechenden Praxis. Doch die Herstellung von Anschlussfähigkeit durch Anpassung und Beherrschung der sozialen Umwelt/Praxis im Sinne der Anforderungen formal-technischer Systeme kann auf Kosten von Vielfalt, Qualität und individueller Freiheit gehen (Huchler 2019).

Auch wenn in KI-Anwendungen in der Regel Ideen über Nutzer\*innen-Führung (und teilweise auch -erziehung und -manipulation) eingeschrieben sind, beruhen Prozesse der Gestaltung und Standardisierung sozialer Praxis durch KI selten auf expliziten Entscheidungen. Vielmehr schleichen sie sich eher unbemerkt in den Nutzungsprozess ein. Wenn beispielsweise emotionssensitive KI auf Basis von Sprach- und Gesichtserkennung für Einstellungstests eingesetzt wird, wenn sie Lernergebnisse im Unterricht misst oder die Gesichter der Kunden liest, um individuelle Wünsche beim Einsteigen in ein fahrerloses Taxi zu erkennen, verändert dies die Wahrnehmung von und den Umgang mit Emotionen durch Individuen und Gesellschaften. KI-Systeme könnten darauf konditionieren, leicht erkennbare oder positiv sanktionierte Sprache und Gesichtsausdrücke zu produzieren, was die Vielfalt verringert, von authentischer Emotionalität entfremdet und zu einem instrumentellen Selbstbezug tendiert. Berater\*innen könnten Menschen auf wichtige Einstellungsverfahren vorbereiten, indem sie sie für die entsprechenden KI-Systeme trainieren; Schüler\*innen würden KI-Systemen Aufmerksamkeit vortäuschen; standardisierte Gesten würden an der KI-Schnittstelle zur Gewohnheit. Die vielfältigen Formen, in denen Menschen Emotionen wie Freude, Trauer, Glück, Wut usw. ausdrücken können, stehen im Gegensatz zu den begrenzten Möglichkeiten sowohl der Perzeption (Repräsentanz in Daten, Qualität und Möglichkeiten des Trainings etc.) wie auch der Anschlussfähigkeit. KI basiert auf einer (sehr hohen, aber) begrenzten Anzahl von Einordnungen und lässt keine „Grauzonen“ zu (Brödner 2020 und 2022). Gleichzeitig belohnen interaktive KI-Systeme (soziale) Kompatibilität mit Funktionalität. Durch die Orientierung an Vergangenheitsdaten und an Häufigkeitsverteilungen tendiert KI zudem dazu, vorgefundene Tendenzen zu verstärken – sei es in Richtung Monopolisierung oder in Richtung Polarisierung. Bei interaktiven KI-Systemen spielt also nicht nur die

Frage eine Rolle, wie die Systeme an das Verhalten und die Umgebung der Menschen angepasst werden können, sondern es schwingt zugleich immer auch der umgekehrte Prozess mit: Wie machen sich Individuen und Gesellschaft – oftmals latent und schleichend, aber auch durch explizite Entscheidungen – anschlussfähig an KI? Dies birgt typische Gefahren der Selbst- oder Fremddisziplinierung, der Einengung von Denk- und Handlungsspielräumen sowie der Verfestigung von Pfadabhängigkeiten und Machtasymmetrien – sei es in Form antizipierender Konformität vor der Nutzung oder unmerklich im Prozess der Nutzung. Diese Form der KI-Selektivitäten hat latente, aber hoch relevante soziale Auswirkungen.

### **3.12 Zusammenfassung: Technische und soziale Selektivitäten**

Die elf chronologisch beschriebenen Selektivitäten von KI lassen sich clustern in a) soziale Selektivitäten bei der sozio-technischen Gestaltung und Nutzung von KI, b) technikimmanente Selektivitäten bei der Beherrschung sozialer Komplexität zum Beispiel durch digitale Technologien und Automatisierung, c) KI-spezifische Selektivitäten, die dem Verfahren konnektionistischer subsymbolischer KI inhärent sind, d) latente Selektivitäten durch individuelle und gesellschaftliche Formierungsprozesse bzw. durch die Anpassung der sozialen Umwelt an die Bedingungen oder Anforderungen von KI (Tab. 1).

Damit wirken KI-spezifische manifeste Formen technik- und sozialinduzierter Selektivitäten zusammen mit eher latenten Selektivitäten, die sich auf technische Verfahren oder den Umgang mit Technik allgemein beziehen. Entsprechend lässt sich dies in vier Feldern systematisieren (Tab. 2):

---

## **4 Folgerungen für das Mensch-KI-Verhältnis**

Die hier systematisierten Selektivitäten zielen auf die latenten sozialen Wirkungen von KI und hieraus entstehende Konflikte und unintendierte Nebenfolgen – angelegt in der Entwicklung und der Funktionsweise von KI sowie in ihrer (auch längerfristigen) Nutzung. Entsprechend verweisen sie auf verschiedene Formen der Komplexitätsreduktion an unterschiedlichen Stellen im KI-Entwicklungs- und Anwendungsprozess sowie auf Gestaltungsfelder und potenzielle Bruchstellen und Konflikte. Selektivitäten dienen der Herstellung von Handlungsfähigkeit durch Komplexitätsreduktion. Sie eröffnen spezifische Handlungsräume und schränken dabei andere ein. Ein Bewusstwerden der Selektivitäten von KI und ein

**Tab. 1** Cluster der Selektivitäten von KI

KI-Selektivität		
KI-Gestaltung und Nutzung	Instrumentelle Gestaltung (1)	Die Funktionalität und der Entwicklungsprozess von KI werden interessen- und erwartungsbasiert gestaltet, z. B. durch die in sie eingeschriebenen Zielsetzungen
	Annahmen und Modelle (2)	Auch Anwendungen subsymbolischer KI werden erst auf Basis von ergänzenden (notwendigerweise selektiven) Annahmen, Modellen, Theorien und Operationalisierungen sozial wirksam bzw. erhalten so einen funktionalen Kontext
	Datengrundlage (4)	Selektivitäten gehen auch mit der Qualität der Datengrundlage, einer limitierten Existenz von und eingeschränktem Zugang zu Daten, der Auswahl sowie der Aufbereitung der Datengrundlage einher. Auch der Vergangenheitsbezug von Daten schränkt Interpretationen ein
	Datenstruktur (5)	Der Daten-Bias bzw. soziale und kulturelle Verzerrungen in den verwendeten Daten sind eine Form von Selektivität

(Fortsetzung)

**Tab. 1** (Fortsetzung)

KI-Selektivität		
	Training (6)	Mit dem Training von KI sind zahlreiche vorstrukturierende Selektivitäten verbunden, z. B. durch die Organisation des Trainings (Annahmen, Zielbeschreibung und Vorlagen, Trainingsdatensatz, Qualitätskontrolle etc.) und durch den reproduzierten sozial-kulturellen Bias
	Nutzungskontext (8)	Ein KI-Output wird in einem gezielt gestalteten Anwendungskontext wirksam gemacht mit selektiven Anschlussmöglichkeiten; von der Gestaltung von Nutzer-Schnittstellen bis zur Implementierung
	Nutzungspraxis (9)	KI-Systeme treffen zugleich auf eine konkrete Nutzungspraxis, die zwischen den Polen Anpassung und Ablehnung liegen, aber auch aktive Aneignungsprozesse einschließen kann
	Starrheit (10)	Selbst adaptive Anwendungen von KI eröffnen i. d. R. nur vorgesehene Räume der Anpassung, die nicht in grundlegende Funktionsweisen hineinreichen und Technikaneignung begrenzen

(Fortsetzung)

**Tab. 1** (Fortsetzung)

KI-Selektivität		
technik-immanent	Objektivierung (3) (von Wissen)	Die auf Wahrscheinlichkeiten und Statistik beruhende spezielle Funktionsweise subsymbolischer KI (z. B. Funktionsapproximation) geht mit eigenen (verzerrenden) Selektionsprozessen einher
(Symbolische) KI immanent	Stochastik (7) (Funktionsweise)	Wie andere (Digitalisierungs- und Automatisierungs-) Technologien ist subsymbolische KI auf eine Explizierung von Wissen angewiesen (Dateninput, Verarbeitungsprozesse, Output), um Wirklichkeit informationstechnisch zu repräsentieren und rechenbar zu machen. Damit einher geht eine notwendige Komplexitätsreduktion bzw. Selektivität entlang der Grenzen der Formalisierung

(Fortsetzung)

**Tab. 1** (Fortsetzung)

KI-Selektivität		
Latente soziale Formierungen rund um Technik	Formierung Individuum-KI (11a)	Menschen richten ihr Handeln (und damit auch in gewissem Maße ihre Kommunikation und ihr Denken) an den antizipierten (limitierten) Optionen und Funktionsweisen von technischen Systemen wie KI aus. Diese latenten Anpassungsprozesse werden u. a. durch Funktionalität belohnt
	Formierung Gesellschaft-KI (11b)	Gesellschaftliche Strukturen und Prozesse richten sich auch an den Bedarfen technischer Systeme wie KI aus, um so bestimmte Funktionalitäten zu ermöglichen. Es bestehen strukturierende Wechselwirkungen zwischen KI-System und Systemumwelt

**Tab. 2** Selektivitäten bei KI und allgemeiner Technik

Selektivitäten	speziell KI	Technik allgemein
Technikinduziert	Subsymbolische KI als Verfahren	Technik (v. a. Digitalisierung/Automatisierung)
Sozial induziert	KI-Gestaltung und Nutzung	Soziale Formierungen rund um Technik

reflexiver Umgang mit ihnen sensibilisiert für die Gestaltung und Gestaltbarkeit des Mensch-Technik-Verhältnisses und wirkt damit den gesellschaftlichen Folgeproblemen entgegen, die entstehen, wenn KI „als per se nicht gestaltungsfähige Technik gesehen wird und daher alle für ihren konstruktiven Einsatz notwendigen Gestaltungs- und Entscheidungsfragen erst gar nicht formuliert werden“ (Pfeiffer 2021, S. 284). Die Selektivitäten verweisen also auf die Möglichkeit und Notwendigkeit von Gestaltung. Sie eröffnen damit Räume und Zugriffspunkte für die Bearbeitung weiterer wichtiger Herausforderungen im Umgang

mit KI als soziales Phänomen. Diese sollen nun abschließend entlang von zwei Abschnitten kurz umrissen werden. Erstens wird für einen nüchternen Blick auf das Mensch-KI-Verhältnis jenseits von Anthropomorphisierungen plädiert. Und zweitens wird dem häufig verkürzten Blick auf KI-Automatisierung als Ersatz von menschlichen Tätigkeiten und Kompetenzen ein alternatives Leitbild für ein künftiges Mensch-KI-Verhältnis gegenübergestellt, das auf Komplementarität und Koevolution beruht. Dabei werden gesellschaftliche Gestaltungsfelder mit Blick auf soziale Implikationen von KI kurz umrissen.

Zum Abschluss dieses Abschnitts können fünf zusammenführende Thesen für einen nüchternen Blick auf das Mensch-KI-Verhältnis aufgestellt werden:

- Daten sind nur ein Ausschnitt sozialer Praxis.
- Wahrscheinlichkeiten sind eine Reduktion von Komplexität.
- Simulation ist nur eine Annäherung an den Gegenstand.
- Variation ist nur eine Vorstufe von Innovation.
- Erkennen von Mustern ist nur ein kleiner Teilaspekt von Sinn/Sinnverstehen.

#### **4.1 Ein nüchterner Blick auf das Mensch-KI-Verhältnis – jenseits von Anthropomorphisierungen**

Die begrifflichen Unklarheiten und Anthropomorphisierungen, also die Vermenschlichung von KI, waren, wie eingangs erwähnt, schon bei der Lancierung des Begriffs „Künstliche Intelligenz“ angelegt. Die technische Nachahmung oder der technisch-funktionale Ersatz von menschlichen Tätigkeiten und Kompetenzen durch KI werden nicht entsprechend begrifflich gerahmt, sondern Nachahmung und Ersatz werden als Entsprechung inszeniert. Hier können unzählige Beispiele genannt werden. So wird von emotionalen Systemen anstelle von emotionserkennenden Systemen, von Handeln anstelle von Ausführen, von Wahrnehmen anstelle von Sensorik, von Lernen anstatt z. B. Programmieren, von Training anstelle von möglicherweise Kalibrieren und auch von Halluzinieren anstelle von ungeeigneten Korrelationen gesprochen. Dies erschwert einen nüchternen Blick auf KI vor allem im öffentlichen Diskurs.

Auf dem Weg zu einem fundierten Blick auf das Mensch-KI-Verhältnis jenseits von potenziell fehlleitenden Anthropomorphisierungen können folgende Fragen erste Richtungshinweise geben: Inwieweit ist Wissen ohne Verständnis möglich? Eignet sich KI menschliche Kompetenzen an? Inwieweit kann



KI als Akteur verstanden werden? In welcher Form können KI-Systeme Träger von Vertrauen und Verantwortung sein? Diese grundlegenden Fragen können hier natürlich nicht ausreichend diskutiert werden, doch können hier einige grundsätzliche Überlegungen dazu angestellt werden.

Inwieweit ist Wissen ohne Verständnis möglich? Subsymbolische KI zeichnet sich gerade dadurch aus, dass sie ohne Expert\*innen- und Kontextwissen über den jeweiligen Bearbeitungsgegenstand auskommt und dabei oftmals recht gute Ergebnisse erzeugt. Auch wenn die technischen Grundlagen subsymbolischer KI-Verfahren skeptisch eingeschätzt werden können (z. B. Brödner 2022), eröffnet subsymbolische KI eine neue Phase des technikzentrierten Umgangs mit Komplexität. Zwar kann weiterhin konstatiert werden: „Aus Information Wissen zu machen und Wissen mit Praxis zu verbinden, das bleibt eine vom Subjekt nicht ablösbare intellektuelle Leistung“ (Schmiede 2006, S. 473). Jedoch übernimmt KI nun Teilschritte in diesem Prozess. Subsymbolische KI reproduziert vorhandenen Sinn, teils auch in neuen Variationen, ohne Verständnis für den Inhalt und simuliert bzw. inszeniert so Wissen und Handeln allein auf Basis von Wahrscheinlichkeiten und Korrelationen. Entsprechend bilden KI-Verfahren auch keine theoriegeleiteten Hypothesen, sondern variieren und wiederholen Prozesse. Im Hinblick auf die Verfasstheit zukünftiger Mensch-KI-Verhältnisse stellt sich die Frage, wie diese weiterhin explizierende und abstrahierende, aber nun von Kontextwissen unabhängig erscheinende Form des Umgangs mit Wissen mit anderen Wissensformen wie zum Beispiel tacit knowing (Polanyi 1985) und Erfahrungswissen (Böhle 2009) zusammengebracht werden kann. Es existieren umfangreiche Forschungen zum Zusammenhang von Sinnverstehen, Körperlichkeit und Wissen, die hier aufgegriffen werden müssten. Aber auch jenseits von Körperlichkeit als Grundlage sozialen Handelns (z. B. Böhle und Wehrich 2010) und von Identität, Interessen, Gefühlen, Emotionen und hierauf basierenden Handlungsformen beruhen informationstechnische Verfahren wie das oben beschriebene der subsymbolischen KI auf der Notwendigkeit der Objektivierung von Wissen in Form von Daten. Damit geht eine Abstrahierung und Reduktion von Komplexität einher, die auch Wissensformen betrifft, die sich nicht vollständig explizieren lassen, wie das implizite Wissen (Polanyi 1985).

Hier gibt es Anknüpfungspunkte für die Diskussion um KI. Zum einen betrifft dies die Idee, dass mit zunehmender Vermessung und Erfassung von Welt bzw. Durchdringung und Verfügbarmachung des Sozialen mit Daten ein wachsendes ‚Weltwissen‘ entstehe, auf das KI-Verfahren immer umfänglicher zurückgreifen können. Hierbei muss beachtet werden, dass eine solche Formalisierung immer ausschnitthaft bleibt, selbst neues Nicht-Wissen (Schmiede 2006) erzeugt, das

Wissen zudem veraltet und sich der Gegenstand permanent wandelt. Die Selektivitäten subsymbolischer KI zeigen auf, dass KI-Wissen selbst selektiv ist und bestimmte Zugänge zu Wissen nicht eröffnet. Zum anderen liegt eine begriffliche Unklarheit in der Bezeichnung von latenten bzw. bislang noch nicht aufgedeckten Zusammenhängen in Datensätzen als „implizites Wissen“. Im Sinne von Polanyi (1985) und der breiten Forschung zu implizitem Wissen wird damit jedoch nicht das *noch* nicht explizierte Wissen gefasst, sondern das nicht-explizierbare Wissen, das zum Beispiel als Erfahrungswissen in der sozialen Praxis angeeignet werden kann (Böhle 2009). Auch auf die Idee, dass sich ein solches reichhaltiges Wissen mittels Trainingsprozessen durch erfahrene Personen digital objektivieren ließe, muss einschränkend reagiert werden. Denn ganzheitliches Erfahrungswissen kann auch auf diese Art nur (zeitlich und inhaltlich) fragmentiert in Form von expliziten Einzelinformationen (ohne umfassenden Kontextbezug und ohne Sinn für Dynamiken bzw. inhaltliche Übertragbarkeit) an das KI-System übertragen und nicht als umfassende Kompetenz, wie zum Beispiel im Sinne des erfahrungsgeliteten Handelns (Böhle 2009), das zwar am Gegenstand entwickelt wird, jedoch auch auf neue unbekannte Kontexte adaptiv anwendbar ist.

Eignet sich KI menschliche Kompetenzen an? Wenn KI menschliche Tätigkeiten ausführt, handelt es sich dabei in der Regel um ein funktionales Äquivalent oder höchstens (selten) um eine Simulation (Huchler 2019). KI-Systeme haben keinen Zugang zu verkörpertem Wissen oder Identität, keine Interessen, keine Sinnlichkeit und kein Sinnverstehen und damit kein Verständnis von Bedeutung, keine Empathie und keine Kreativität. Jedoch können viele Situationen, in denen diese menschlichen Kompetenzen gefragt sind, mit mehr oder weniger großem technischem Aufwand auch durch KI-Verfahren angegangen werden, wie etwa das auch auf Intuition basierende Go-Spielen. Damit hat sich KI jedoch nicht die entsprechenden Kompetenzen angeeignet (z. B. Intuition), sondern es wurde mit der Technik ein anderer Lösungsweg für dasselbe Ziel gefunden. Entsprechende funktionale Äquivalente können auch in Form von Simulationen menschlichen Handelns auftreten bis hin zur Täuschung. Und selbst eine perfekte Täuschung bzw. technische Nachahmung aller relevanten Interaktionen würde eine solche bleiben, da sie auf andere (oftmals sehr aufwendige) Rahmenbedingungen zu ihrer Herstellung und Aufrechterhaltung angewiesen ist und damit permanent entzaubert werden kann. Umgekehrt können Menschen natürlich Objekte ‚subjektivieren‘ bzw. wie Subjekte behandeln und zugleich die Grundfähigkeit zur Reflexion und Distanzierung behalten, um Objekte als solche zu erkennen und entsprechend zu handeln. Hierfür braucht es nicht unbedingt Ähnlichkeit oder gar die Vortäuschung einer menschlichen Interaktion, obwohl dies den Aufbau von

Objekt-Beziehungen erleichtern kann. Aneignungs- und Ermächtigungsprozesse spielen hier mindestens ebenso eine Rolle.

Inwieweit kann KI als Akteur verstanden werden? Analytisch kann es durchaus Sinn machen, die soziale Praxis von KI als Resultat von Interaktionen und Relationen von menschlichen und technischen Akteuren zu verstehen ([Heinlein]), gerade mit Blick auf die durch subsymbolische KI erweiterten Kontingenzen (ebd.) im Mensch-Technik-Verhältnis. Um das soziale Phänomen subsymbolische KI zu fassen, reicht die Metapher des passiven Werkzeugs nicht mehr aus (Rammert 1999). Dies hat damit zu tun, dass KI auch in komplizierteren Settings und bei unvollständigen Informationen nicht vorausgeplante bzw. unerwartete anschlussfähige Aktionen hervorbringt. Der Akteurscharakter macht sich zudem daran fest, dass lernende KI durch die Art, wie sie zu ihrem Output gelangt, wie eine Blackbox erscheint. Die Kompliziertheit bzw. teilweise extrem große Menge an generierten Verknüpfungen im ‚neuronalen Netzwerk‘ macht es extrem aufwendig bis unmöglich, diese im Detail nachzuvollziehen. So entsteht eine Unsicherheitszone, die nicht nur Fragen der Verantwortung, Zurechenbarkeit und Regulierbarkeit aufwirft, sondern einer Anthropomorphisierung und Mystifizierung der Technologie Vorschub leisten kann, indem ihr pauschal eigenständiges Handeln zugeschrieben wird. Entsprechend wichtig ist es, die Kontingenzen von KI als soziale Praxis rund um Mensch-KI-Relationen zu verstehen ([Heinlein]). Hilfreich sind hier Brückenkonzepte wie das der verteilten Handlungsträgerschaft (Rammert und Schulz-Schaeffer 2002), die verschiedene Relationen in Mensch-Technik-Interaktionen differenzieren und dabei die Möglichkeiten der Programmierung, der Einschreibung von Zielen und der Delegation von Entscheidungen an Technik berücksichtigen (Schulz-Schaeffer 2008). So lässt sich KI auch als Abkehr von einer linearen Wenn-dann-Programmierung hin zu einer zielorientierten Um-zu-Programmierung verstehen (Huchler 2023). Die oben systematisierten Selektivitäten zeigen Handlungs- und Konfliktfelder mit unterschiedlich gewichteten Mensch-KI-Relationen entlang des sozio-technischen Entwicklungs-, Verarbeitungs- und Nutzungsprozesses von KI auf.

In welcher Form können KI-Systeme Träger von Vertrauen und Verantwortung sein? Ähnlich wie KI-Systeme nicht denken oder handeln, aber als Akteure in einer sozialen Praxis verstanden werden können ([Heinlein]), können KI-Systeme Träger von Vertrauen und Verantwortung sein und Vertrauen in der Interaktion generieren, ohne selbst die hierfür notwendigen Eigenschaften zu besitzen. KI-Systeme bieten eine breite Projektionsfläche für Systemvertrauen (Luhmann 1979), sei es als Vertrauen (oder Misstrauen) in Technik bzw. KI allgemein oder in relevante Institutionen (Anbieter, Hersteller, Gesetzgeber bzw. Qualität,

Normierung, Zulassungsverfahren etc.). Dieses Vertrauen kann als sozialer Prozess (Mayer et al. 1995) verstanden werden, es kann sich im Prozess beweisen und schrittweise aufbauen (Böhle et al. 2014). So kann auch eine objektspezifische Vertrauensbeziehung entstehen, die sich jedoch in grundlegenden Aspekten von Personen- wie Systemvertrauen (Luhmann 1979) unterscheidet. Während der Aspekt der Fähigkeit bzw. des Könnens („ability“) als Teil der im Prozess generierten bzw. wahrgenommenen Vertrauenswürdigkeit („perceived trustworthiness“) (Mayer et al. 1995, S. 715) dem KI-System als Interaktionspartner zugeschrieben werden kann, liegt bei den weiteren Aspekten von Vertrauenswürdigkeit wie „benevolence“ und „integrity“ (ebd.) bereits der Schritt zum Systemvertrauen und damit eine institutionelle Abstrahierung vom einzelnen Objekt nahe. Wesentlich bedeutender ist jedoch, dass interpersonelles Vertrauen (Mayer et al. 1995) und auch Vertrauen in Organisationen als wechselseitiger Herstellungsprozess in sozialen Interaktionen konzipiert wird (Böhle et al. 2014) und ein Risiko bzw. eine Verwundbarkeit („vulnerability“, Mayer et al. 1995, S. 712) voraussetzt. Verwundbarkeit ist auch eine Voraussetzung für das Tragen von Verantwortung oder die Möglichkeit zu sanktionieren. Beides trifft nur auf die Repräsentanten des KI-Artefakts zu und nicht auf das konkrete Objekt selbst.<sup>3</sup> Indem die aufgezeigten Selektivitäten auf die Gestaltung von KI hinweisen, wirken sie einer undifferenzierten und anthropomorphisierenden Konzeption von Vertrauen in KI entgegen.

## 4.2 Ein alternatives Leitbild für das künftige Mensch-KI-Verhältnis: Komplementarität und Koevolution

Die öffentlichen Diskussionen um KI sind geprägt durch die Konstatierung eines Konkurrenzverhältnisses zwischen menschlichen und technischen Fähigkeiten bzw. durch eine Ausrichtung auf das Ersetzen menschlicher Tätigkeiten. Natürlich werden auch Gegenbilder gezeichnet, vor allem im Kontext von Arbeit, zum Beispiel die Forderung ‚human in the loop‘ oder Konzepte der Befähigung bzw. Augmentierung der Menschen durch „human-centered AI“ (Shneiderman 2022). Jedoch wird dies oftmals als normative Forderung formuliert und nicht

---

<sup>3</sup> So hat eine (potenzielle oder ausgeführte) ‚Bestrafung‘ eines einzelnen KI-Systems zum Beispiel durch Abschaltung oder Zerstörung keine abschreckenden Wirkungen für andere KI-Systeme, sondern wird nur über den menschlichen Interaktionspartner in nächste Interaktionen weitergetragen. Anders sieht es mit der ‚Bestrafung‘ einer Technologie aus, indem zum Beispiel eine entsprechende Zulassung entzogen wird; ähnlich einer Sippenhaft.

als ökonomisch und gesamtgesellschaftlich funktionales Leitbild für ein allgemeines Mensch-KI-Verhältnis. Ein mögliches weittragendes Gegenbild wäre das der komplementären Koevolution zwischen Mensch und KI. Das heißt, KI-Entwicklung und KI-Nutzung müssten auf eine parallele, sich wechselseitig bedingende und verstärkende Weiterentwicklung von Mensch und KI ausgerichtet sein; zum Beispiel in der Form, dass mit zunehmendem Einsatz von KI und Delegation an KI die Handlungsmöglichkeiten und Gestaltungskompetenzen der Menschen entsprechend steigen. Eine solche Koevolution benötigt ein weitreichendes Verständnis für Mensch-Technik- bzw. Mensch-KI-Differenzen, um hieraus permanent neue Komplementaritätsverhältnisse zu generieren (vgl. Huchler 2022). Ein solches potenzial- und entwicklungsorientiertes Leitbild könnte unter anderem sozialen Zusammenhalt und Vertrauen fördern, wäre individuell förderlich und könnte auch die Passgenauigkeit bzw. Gebrauchstauglichkeit und ökonomische und soziale Nachhaltigkeit von KI-Lösungen unterstützen. Dazu sind jedoch einige Voraussetzungen zu schaffen. Komplementarität und Koevolution müssten im Design von interaktiven KI-Systemen angelegt sein. Dies betrifft vor allem die Entwicklungs- und Innovationsförderlichkeit der Interaktion mit KI; zum Beispiel im Sinne einer erfahrungs- und lernförderlichen Interaktionsgestaltung, die Aneignungs- und Gestaltungsmöglichkeiten eröffnet. Weiterentwicklung kann natürlich nicht ausschließlich der Interaktion überantwortet werden, sondern eine solche Koevolution verweist auch auf die Begleitung durch Qualifizierung. Dabei geht es im Sinne einer breiten nachhaltigen sozialen Praxis mit KI weniger um Programmierfähigkeiten als um ein Verständnis für die grundlegenden Funktionsweisen von KI (z. B. Statistik), um die Stärken und Schwächen, Anwendungsmöglichkeiten und Resultate von KI einschätzen und zum eigenen Wissen und Handeln in Bezug setzen zu können. Dazu gehört, „mit Ergebnissen um[zu]gehen, die als eine fixe und objektive Wahrheit erscheinen mögen, in Wirklichkeit aber nur eine Prognose mit einer gewissen Aussagewahrscheinlichkeit sind (also im Einzelfall immer auch falsch-positiv oder falsch-negativ sein können)“ (Pfeiffer 2021, S. 284). Insgesamt benötigen KI-Kompetenzen eine individuelle und gesellschaftliche Reflexionsfähigkeit in Bezug auf KI als soziales Phänomen. Ein Verständnis für die hier aufgezeigten Selektivitäten wie auch für die Kontingenzen von KI ([Heinlein]) kann dabei ein Baustein sein und ein Bewusstsein für die latenten sozialen Folgen von KI fördern.

In der Anwendungspraxis öffnet KI neue Wissens- und Handlungsspielräume und schließt zugleich andere aus. Diese selektierenden und strukturierenden Wirkungen gilt es im Blick zu behalten, vor allem wenn die Gefahr besteht, dass sich auf diese Art rekursive Wissenszirkel bzw. Selbstbestätigungs- und

Verfestigungsspiralen ausbilden, die mit Wissens- und Kompetenzverlust verbunden sind; zum Beispiel wenn textgenerierende KI zunehmend auf KI-generierte Texte zurückgreift oder wenn KI-Systeme auf Basis von KI-unterstützten Simulationen trainiert werden. Im Sinne einer Steigerung der gesellschaftlichen Reflexionsfähigkeit müssten Prozesse unterstützt werden, die einer Verengung, Monopolisierung und Konformierung von Wissens- und Handlungsräumen systematisch entgegenwirken. Dazu gehört es, kritische Diskussionen, wie Hinweise auf Ungleichheiten, Interessenkonflikte, mögliche Folgen etc., als Innovations-treiber zu fördern und nicht als Neid- und Angstdiskurse von Bedenkenträgern abzutun und damit die Debatte um KI zu polarisieren. Vielmehr gilt es, den oftmals den Regeln der Aufmerksamkeitsökonomie folgenden, durch Polarisierung, Überspitzung, Heilsversprechen, Dystopien und latente Drohungen sowie Überlegenheitsinszenierungen geprägten öffentlichen Diskurs zu KI zu erden. Dazu gehört letztlich auch eine Abkehr von der vereinfachenden Ersetzungsrhetorik hin zu einem dynamischen Verständnis des Mensch-KI-Verhältnisses – zum Beispiel von Automatisierung als permanentem sozio-technischem Wandlungsprozess, der sich auch nach dem Prinzip der komplementären Koevolution ausgestalten ließe. Mittel- und langfristige Überlegungen zum Wandel des Mensch-KI-Verhältnisses brauchen empirische Erkenntnisse und Konzepte zu Formen und sozialen Mechanismen der (permanenten) Neuschneidung der ‚Arbeitsteilung‘ zwischen Mensch und KI und den damit verbundenen Tätigkeiten und Kompetenzen. Dies betrifft das gesamte soziotechnische Setting rund um KI, also die Technikgestaltung, Interaktion und Kommunikation, Organisation und Institutionen; wie zum Beispiel Bildung und die Wertigkeit von (kognitiven und manuellen) Tätigkeiten, Wissen und Kompetenzen.

Insgesamt besteht noch großer Forschungsbedarf hinsichtlich des sozialen Phänomens KI und eines sich wandelnden Mensch-KI-Verhältnisses; zum Beispiel hinsichtlich der alten, aber nicht ausreichend beantworteten Frage, was KI bzw. Technik und was der Mensch gut kann und was nicht (Fitts 1962; Dekker und Woods 2002). Es braucht noch mehr Erkenntnisse darüber, wie KI soziotechnische Konstellationen verändert und wie sich in der konkreten sozialen Praxis von KI Wissen und Handeln, Erfahrung und Kompetenzen, Einstellungen und Kommunikation, Wissens- und Gestaltungsräume neu konstellieren. Dies schließt Konzepte zur Neustrukturierung einer Arbeitsteilung zwischen Mensch und KI sowie Erkenntnisse zu Automatisierungs- und Wandlungsdynamiken ebenso ein wie empirische und konzeptionelle Erkenntnisse zu sozialen Einbettungsprozessen und ihren latenten sozialen Folgen. Insbesondere gilt es, die permanente Rekonstitution des Sozialen rund um technisierte Prozesse auf eine

Weise herauszuarbeiten, dass sie sich auch in den Diskursen um KI wiederfinden lässt. Insgesamt braucht es noch mehr Forschungen zu den direkten und latenten, kurz- und langfristigen sozialen Implikationen von unterschiedlichen KI-Einsatzformen für die Menschen und die Gesellschaft, um KI als soziales Phänomen zu verstehen und den Lösungshorizont für eine gesellschaftlich nachhaltige Gestaltung zu verbreitern.

---

## Literatur

- Amoore, L., und V. Piotukh (Hrsg.) 2016. *Algorithmic Life. Calculative Devices in the Age of Big Data*. London.
- Bainbridge, L. 1983. Ironies of automation, *Automatica*, 19 (6): 775–779.
- Berger, U. und C. Offe 1980. Das Rationalisierungsdilemma der Angestelltenarbeit. Arbeitssoziologische Überlegungen zur Erklärung des Status von kaufmännischen Angestellten aus der Eigenschaft ihrer Arbeit als ‚Dienstleistungsarbeit‘. *Geschichte und Gesellschaft*, Sonderheft 7, Vandenhoeck & Ruprecht: 39–58.
- Böhle, F. 2009. Weder rationale Reflexion noch präreflexive Praktik. Erfahrungsgeleitet-subjektivierendes Handeln. In *Handeln unter Unsicherheit*, hrsg. F. Böhle, & M. Wehrich, 203–230, Wiesbaden.
- Böhle, F. und M. Wehrich (Hrsg.) 2010. *Die Körperlichkeit sozialen Handelns*. Transcript.
- Böhle, F., A. Bolte, N. Huchler, J. Neumer, S. Porschen-Hueck, S. Sauer 2014. *Vertrauen und Vertrauenswürdigkeit. Arbeitsgestaltung und Arbeitspolitik jenseits formeller Regulierung*, Springer VS.
- Brödner, P. 2019. Grenzen und Widersprüche der Entwicklung und Anwendung „Autonomer Systeme“. In *Autonome Systeme und Arbeit. Perspektiven, Herausforderungen und Grenzen der Künstlichen Intelligenz in der Arbeitswelt*, hrsg. H. Hirsch-Kreinsen, & A. Karačić, 69–97, Bielefeld.
- Brödner, P. 2020. Paradoxien der Ko-Aktion von Experten und adaptiven Systemen. In *Zukunft der Arbeit – Soziotechnische Gestaltung der Arbeitswelt im Zeichen von »Digitalisierung« und »Künstlicher Intelligenz«*, hrsg. P. Brödner & K. Fuchs-Kittowski, Abhandlungen der Leibniz-Sozietät der Wissenschaften 67, 143–159, trafo Wissenschaftsverlag.
- Brödner, P. 2022. Die Illusionsfabrik der „KI“-Narrative. *FifF-Kommunikation*, 2022 (2): 32–36.
- Crawford, K. 2022. *Atlas of AI – Power, Politics and the Planetary Costs of Artificial Intelligence*. New Haven.
- Danaher, J. 2016. The Threat of Algocracy: Reality, Resistance and Accommodation. *Philosophy & Technology*, 29 (3): 245–268.
- Dekker S.W.A., und D.D. Woods 2002. MABA-MABA or Abracadabra? Progress on Human-Automation Coordination. *Cognition, Technology & Work*. (4): 240–244.
- Diakopoulos, N. 2015. Algorithmic accountability: Journalistic investigation of computational power structures. *Digital Journalism*, 3 (3): 398–415.

- Europäisches Parlament 2020. *Was ist künstliche Intelligenz und wie wird sie genutzt?* <https://www.europarl.europa.eu/news/de/headlines/society/20200827STO85804/was-ist-kunstliche-intelligenz-und-wie-wird-sie-genutzt> (letzter Abruf 01.08.2023).
- Fitts, P.M. 1962. Functions of Man in Complex Systems. *Aerospace Engineering*, 21 (1): 34–39.
- Fodor, J.A. und Z.W. Pylyshyn 1988. Connectionism and Cognitive Architecture: A Critical Analysis. *Cognition*, 28 (1–2): 3–71.
- Friedman, B. und H. Nissenbaum 1996. Bias in computer systems. *ACM Transactions on Information Systems*. 14 (3): 330–347.
- Gray, M.L. und S. Suri 2019. *Ghost Work. How to Stop Silicon Valley from Building a New Global Underclass*. Boston.
- Hirsch-Kreinsen, H. 2023. *Das Versprechen der Künstlichen Intelligenz. Gesellschaftliche Dynamik einer Schlüsseltechnologie*. Frankfurt a. M./New York.
- Huchler, N. 2019. Assimilierende vs. komplementäre Adaptivität. Grenzen teil-autonomer Systeme, Autonome Systeme und Arbeit. Perspektiven. In *Herausforderungen und Grenzen der Künstlichen Intelligenz in der Arbeitswelt*, hrsg. H. Hirsch-Kreinsen, & A. Karačić, 139–180, Bielefeld.
- Huchler, N. 2022. Komplementäre Arbeitsgestaltung. Grundrisse eines Konzepts zur Humanisierung der Arbeit mit KI. *Zeitschrift für Arbeitswissenschaft*. 76 (6), 158–175.
- Huchler, N. 2023. Arbeit und Künstliche Intelligenz. In: *Lexikon der Arbeits- und Industriegesellschaft*, hrsg. R. Bohn, H. Hirsch-Kreinsen, S. Pfeiffer, & M. Will-Zocholl, 3., aktualis. Aufl., 45–49, Baden-Baden.
- Kämpf, T. und B. Langes 2021. Künstliche Intelligenz in der digitalen Arbeitswelt. Erste Befunde einer empirischen Bestandsaufnahme. *Denk-doch-Mal.de*: 03–21.
- Luhmann, N. 1979. *Trust and Power: Two Works*. Wiley.
- Lutz, B. 1987. *Technik und sozialer Wandel*. Campus.
- Mayer, R.C., J.H. Davis, und F.D. Schoorman 1995. An integrative model of organizational trust. *Academic Management Review*, 20: 709–734.
- Mittelstadt, B.D., P. Allo, M. Taddeo, S. Wachter, und L. Floridi 2016. The ethics of algorithms: Mapping the debate. *Big Data & Society*, 3 (2).
- McCarthy, J., M. L. Minsky, N. Rochester, und C. E. Shannon, C. E. 1955. *A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence*. <https://www-formal.stanford.edu/jmc/history/dartmouth/dartmouth.html> (letzter Abruf 01.02.2023).
- Neuweg, G. H. 2015. *Das Schweigen der Könnner*. Gesammelte Schriften zu implizitem Wissen. Münster/New York.
- Pfeiffer, S. 2020. Kontext und KI: Zum Potenzial der Beschäftigten für Künstliche Intelligenz und Machine-Learning. *HMD Praxis der Wirtschaftsinformatik*, 57: 465–479.
- Pfeiffer, S. 2021. *Digitalisierung als Distributivkraft. Über das Neue am digitalen Kapitalismus*. Bielefeld.
- Polanyi, M. 1985. *Implizites Wissen*. Suhrkamp.
- Rammert, W. 1999. Relations That Constitute Technology and Media That Make a Difference: Toward a Social Pragmatic Theory of Technicization. *Society for Philosophy and Technology Quarterly*, 4 (3): 165–177.
- Rammert, W. 2003. Technik in Aktion: verteiltes Handeln in soziotechnischen Konstellationen. *TUTS – Working Papers*, 2–2003. Technische Universität Berlin, Fak. VI Planen,



- Bauen, Umwelt, Institut für Soziologie, Fachgebiet Techniksoziologie. <https://nbn-resolving.org/urn:nbn:de:0168-ssoar-11573> (letzter Zugriff: 01.08.2023).
- Rammert, W., und I. Schulz-Schaeffer 2002. *Können Maschinen handeln? Soziologische Beiträge zum Verhältnis von Mensch und Technik*. Frankfurt a. M/New York.
- Rohde, M., P. Brödner, G. Stevens, M. Betz, und V. Wulf 2017. Grounded design – a praxeological IS research perspective. *Journal of Information Technology*, 32 (2): 163–179.
- Russell, S. J. 2010. *Artificial intelligence a modern approach*. Pearson Education, Inc.
- Salles, A., K. Evers, und M. Farisco 2020. Anthropomorphism, AI. *AJOB Neuroscience*, 11 (2): 88–95.
- Schmiede, R. 2006. Wissen und Arbeit im „Informational Capitalism“. In: *Informatisierung der Arbeit – Gesellschaft im Umbruch*, hrsg. A. Baukrowitz, T. Berker, A. Boes, S. Pfeiffer, R. Schmiede, M. Will-Zocholl, 457–490, Berlin.
- Schulz-Schaeffer, I. 2007. Technik als sozialer Akteur und als soziale Institution: Sozialität von Technik statt Postsozialität. *TUTS – Working Papers*, 3–2007. Technische Universität Berlin, Fak. VI Planen, Bauen, Umwelt, Institut für Soziologie, Fachgebiet Techniksoziologie. <https://nbn-resolving.org/urn:nbn:de:0168-ssoar-12139> (letzter Zugriff: 01.08.2023).
- Schulz-Schaeffer, I. 2008. Deutung und Delegation: Handlungsträgerschaft von Technik als doppeltes Zuschreibungsphänomen. In: *Die Natur der Gesellschaft: Verhandlungen des 33. Kongresses der Deutschen Gesellschaft für Soziologie in Kassel 2006*, hrsg. K.-S. Rehberg, 3135–3144, Campus.
- Seising, R. 2021. *Es denkt nicht! Die vergessenen Geschichten der KI*, 1. Aufl., Büchergilde Gutenberg.
- Shneiderman, B. 2022. *Human-centered AI*. Oxford University Press.
- Smolensky, P. 1988. On the Proper Treatment of Connectionism. *Behavioral and Brain Sciences*, 11: 1–74.
- Spiestersbach, K. 2023. Update: Geheimnis gelüftet! Soviele Parameter hat GPT-4. *SEARCH ONE*. (<https://www.search-one.de/gpt-4-anzahl-parameter/>)
- Watson, D. 2019. The Rhetoric and Reality of Anthropomorphism in Artificial Intelligence. *Minds & Machines*, 29: 417–440.
- Zuboff, S. 2015. Big Other: Surveillance Capitalism and the Prospects of an Information Civilization. *Journal of Information Technology*, 30 (1): 75–89.



# Von Interaktion zur Transformaktion: Die Folgen von Künstlicher Intelligenz für Theorien sozialen Handelns

Valentin Rauer

## Zusammenfassung

Aus sozialwissenschaftlicher Perspektive ist Künstliche Intelligenz (KI) eine Technologie, die zum einen eine Infrastruktur bereitstellt und zum anderen selbst aktiv Handlungsträgerschaften übernimmt. Als Infrastruktur vernetzt KI Daten aus Datenbanken, Inputs von Sensoren, Rückmeldungen von Aktoren und Chatbots, die in Interaktion mit menschlichen Akteure stehen etc. In diesen Infrastrukturen werden sinnorientierte Handlungsprogramme konstituiert und soziale Beziehungen generiert. KI wird so zu einem Handlungsträger neben menschlichen Handlungsträgern und ko-konstituiert damit sozialen Sinn, kollektive Typologien und gesellschaftliche Kategorisierungen. Insgesamt bildet KI damit einen Bestandteil der sozialer Strukturbildung und der sinnhaft sozialen Interaktivitäten. Diesem zweiten Aspekt, den Folgen von KI für unsere Vorstellungen sozialer sinnhafter Interaktivität widmet sich dieser Aufsatz. Soziales Handeln wird im Zuge der aktuellen KI in Interaktionssituationen zunehmend durch nicht-menschliche Handlungsträgerschaften und Kommunikationsressourcen ko-konstituiert und damit grundlegend transformiert. Basierend auf Grundannahmen der Akteur-Netzwerk-Theorie wird vorgeschlagen, diesen Wandel als Übergang von *Interaktion* zu *Transformaktion* zu beschreiben.

---

V. Rauer (✉)

Türkisch-deutsche Universität Istanbul, Istanbul, Türkei

E-Mail: [rauer@tau.edu.tr](mailto:rauer@tau.edu.tr)

© Der/die Autor(en), exklusiv lizenziert an Springer Fachmedien Wiesbaden GmbH, ein Teil von Springer Nature 2024

M. Heinlein und N. Huchler (Hrsg.), *Künstliche Intelligenz, Mensch und Gesellschaft*, [https://doi.org/10.1007/978-3-658-43521-9\\_18](https://doi.org/10.1007/978-3-658-43521-9_18)

---

**Schlüsselwörter**

Soziale Interaktion • Handlungstheorie • Akteure • Aktanten •  
Akteur-Netzwerk-Theorie • Transformer • Deep Learning • Digitalisierung •  
Handeln

---

## 1 **Problemstellung: Besteht ein Mensch auch in Zukunft noch den Turing Test?**

Die Debatte um KI hat sich in den letzten zwei Jahren gewandelt und intensiviert. Momentan kreist sie um bedeutungsproduzierende Plattformen wie ChatGPT<sup>1</sup> oder Googles Bard.<sup>2</sup> Dabei handelt es sich um netzbasierte Infrastrukturen, auf denen Nutzende Textsegmente oder Fragen eintippen können, die Plattform führt dann den Text weiter bzw. macht einen Textvorschlag oder aber sie beantwortet die gestellten Fragen. Solche Plattformen werden auch Chatbots<sup>3</sup> genannt. Dieser Begriff ist eine Kombination aus der Kurzform von ‚Bot‘ (Robot) und ‚Chat‘. Chats sind Internetmedien, auf denen Nutzende interaktiv kommunizieren und Informationen austauschen. Der Begriff ‚Chatbot‘ kombiniert diese beiden Ansätze, d. h. ein menschlicher kommunizierender Nutzender wird durch ein kommunizierendes KI-Programm ersetzt. Chatbots basieren auf Sprachmodellen, die auf großen Datenmengen trainiert wurden. Insbesondere Googles Bard nutzt für dieses Training sämtliche von Nutzenden kommunizierten Inhalte und Texte, wenn eine Einwilligung erfolgte.<sup>4</sup>

Debattiert werden Gefahren von diesen KI basierten Chatsbots insbesondere mit Blick auf die zu erwartende Unterlegenheit menschlicher Kompetenzen im Bereich der sinnproduzierenden und abstrahierenden Verarbeitung großer Datensätze und Informationsressourcen. Indirekt scheint die Frage zu sein, ob in zukünftigen Turing-Testverfahren nicht mehr getestet werden muss, ob es einem Roboter gelingt, den Menschen perfekt zu imitieren, sondern ob es einem Menschen noch gelingt, den Roboter bzw. den Chatbot erfolgreich zu imitieren.

Diese aktuell diskutierten KI-Plattformen oder Chatbots sind in der Lage auf Nachfrage hin spezifische Texte zu einem gewünschten Thema zu verfassen, Examina zu bestehen und Wissen zu kategorisieren und zusammenzufassen. Ihre KI konstituiert offenbar damit eigene Sinn- und Bedeutungssysteme. Basierend

---

<sup>1</sup> <https://openai.com/blog/chatgpt>.

<sup>2</sup> <https://bard.google.com/>

<sup>3</sup> <https://www.sueddeutsche.de/wirtschaft/google-bard-chatgpt-bing-chatbot-deutschland-1.6025188> (letzter Download: 23.07.2023).

<sup>4</sup> Ebd.

auf sogenannten Trainings an großen Datenmengen von Prädiktionsalgorithmen initiieren sie soziale Klassifikations- und Interpretationszusammenhänge in Textform bis hin zu Lösungsvorschlägen für die Erstellung neuer Algorithmen im Codingprozess selbst.

Klassische Trennungen zwischen menschlicher und nicht-menschlicher Auto-renschaft werden obsolet. Allein durch den Einsatz von KI-basierten Textplatt- formen lassen sich menschliche und nicht-menschliche Sinnproduktionen immer weniger auf einzelne Individuen zurückführen. Diese soziale, sinngenerierende KI, wie beispielsweise die aktuellen ‚Deep Learning‘ Plattformen auch interpretiert werden können, haben damit Folgen für das, was wir bis dahin unter rein menschlichem sozialen Handeln verstanden haben.

Um ein Beispiel zu geben: Die folgende Textpassage handelt von einem Überblick zu den Folgen von KI für soziales Handeln. Die Textpassage wurde von einer KI verfasst. Auf der Plattform <https://bard.google.com/> wurde der folgende Satz von mir als Verfasser und menschlichen Handlungsträger selbst getippt: „The consequences of KI for our concepts of human action“. Nach einer Eingabe der Return-Taste erschien auf der Plattform Bard der folgende, hier nicht modifizierte Text:

“The development of artificial intelligence (AI) has the potential to challenge our understanding of human action in a number of ways. For example, as AI systems become more sophisticated, they may be able to perform actions that were previously thought to be uniquely human, such as creativity, problem-solving, and even empathy. This could lead us to question the nature of what it means to be human, and to rethink our concepts of agency and free will.

In addition, AI systems could potentially have a profound impact on our social and economic structures. For example, AI-powered automation could lead to widespread job displacement, as machines become capable of performing many of the tasks that are currently done by humans. This could lead to social unrest and economic instability.

On the other hand, AI could also be used to improve our lives in many ways. For example, AI-powered healthcare systems could provide more personalized and effective care, and AI-powered transportation systems could make our commutes faster and more efficient. Ultimately, the consequences of AI for our concepts of human action will depend on how we choose to develop and use this technology.

Here are some specific examples of how AI could challenge our concepts of human action:

- **Creativity:** AI systems are already being used to generate creative content, such as music, art, and literature. In the future, it is possible that AI systems could become so creative that they are indistinguishable from human artists. This could lead us to question the value of human creativity, and to wonder if machines can truly be original.
- **Problem-solving:** AI systems are already being used to solve complex problems in a variety of fields, such as medicine, finance, and engineering. In the future, it is possible that AI systems could become so good at problem-solving that they are able to outperform humans in even the most challenging tasks. This could lead us to question our own intelligence and ability to think critically.
- **Empathy:** AI systems are still in their early stages of development, but some researchers believe that it is possible to create AI systems that are capable of empathy. This would mean that AI systems could understand and respond to the emotions of others, just like humans. This could have a profound impact on our social interactions, as we would need to learn to interact with machines in a way that is sensitive to their emotional needs.

Overall, the development of AI has the potential to challenge our understanding of human action in a number of ways. It is important to be aware of these potential challenges so that we can make informed decisions about how to develop and use AI technology.”<sup>5</sup>

Falls diejenigen, die den obigen KI-Text gelesen haben, ein menschlicher Handlungsträger sind, bietet sich eine kurze Reflexion der textuellen Lesepraxis an. Exemplarisch werde ich in Ich-Form eine solche Lesereflexion präsentieren:

*Erstens*, habe ich den Text grundsätzlich nur als Beispiel für einen KI generierten Text lesen können? Oder sprang das eine oder andere Mal ein inhaltlicher Interpretationsgedanke während des Lesens auf?

*Zweitens*, habe ich nicht doch vor allem auf Fehler der KI gelauert, um meine Vorannahme zu einer KI zu bestätigen, die davon ausgeht, dass sie bei der subjektiven Sinnproduktion letztlich doch stets der menschlichen Autorenschaft unterlegen ist? Hat mich bei der Suche ein Gefühl der Selbstbestätigung heimgesucht, ah, hier haben wir den Fehler, die KI irrt doch und wir als menschliche Handlungsträger behalten unsere Superiorität bei der Produktion subjektiven Sinns bei?

---

<sup>5</sup> Text von der KI-Plattform: <https://bard.google.com/>. Account persönlich von Valentin Rauer. Basisanfrage: Valentin Rauer, Auswahl der Antwort: Valentin Rauer. Datum der Texterstellung: 10.07.2023.

*Drittens*, bin ich nicht doch ab und zu unmerklich während des Lesens von der Ebene des Lesens eines ‚Beispiels für KI‘ zur Ebene eines ‚Textes über KI‘ verwechselt? Schwankte ich zwischen der Dekodierung eines KI-Textbeispiels zu einer inhaltlichen Aussage über die Folgen von KI für soziales Handeln generell? Habe ich diesen Rahmenwechsel, oder mit Erving Goffman (1974) gesprochen, dieses *Keying* zwischen unterschiedlichen Rahmungsebenen vollständig kontrollieren können oder ereigneten sich nicht hier und da kurze Rahmenwechsel zwischen Beispieltext und Aussagentext?

*Viertens*, fand sich wiederholt nicht eine kurze Interpretationsphase, in der der Text so von mir kodiert wurde, als ob er von rein menschlichen Handlungsträgern verfasst worden sei? Warum und wie war dieses Wissen über die schreibenden Handlungsträgerinnen relevant während meines Lesevorgangs? Beispielsweise interessiert es mich nicht doch auch inhaltlich, ob das, was die KI zu ‚Empathie‘ schreibt, sinnhaft argumentiert und theoretisch nachvollziehbar ist? Was bedeutet dieses ‚rein inhaltlich‘ beim Dekodieren dieses von KI generierten Sinnproduktes im Vergleich zu einem ‚rein inhaltlichen‘ Lesen eines Textes, der rein von menschlichen Handlungsträgern verfasst wurde? Nehme ich unbewusst den von menschlichen Handlungsträgern verfassten Text womöglich ernster und bewerte seine Aussagekraft höher als die der KI – und warum?

*Fünftens*: Habe ich mich nicht vielleicht kurzzeitig dabei ertappt, den einen oder anderen Gedanken der KI selbst weiter zu verfolgen oder gar umformuliert und modifiziert zu übernehmen? Hat mir der Text nicht auch ein wenig entmutigt, mich selbst an die mühsame Arbeit des Zusammenfassens zu setzen, wenn es die KI-Plattform nicht viel schneller erledigt?

*Sechstens*: Wie strukturiere ich meinen Text über die handlungstheoretischen Folgen von KI, wenn diese KI-Plattform das Thema in die drei Themenblöcke 1. Kreativität, 2. Problemlösung und 3. Empathie strukturiert? ‚Machen‘ diese Strukturierungen denn nicht auch tatsächlich ‚Sinn‘, schließlich wertet die KI quantitativ ungleich mehr Sekundärliteratur zu dem Thema aus, als ich überhaupt in der Lage bin, in der kurzen Zeit bis zu meiner Textabgabe zu lesen. Ist der Text-Transformer von Googles Bard aufgrund des radikalen Wandels der KI-Entwicklung nicht tatsächlich besser geeignet, diesen großen neuen Textkorpus zu überblicken? Vielleicht ist diese Struktur dieser drei thematischen Grundfragen adäquater als mein eigener aktueller Wissensstand der Debatte? Ist es dann nicht unwissenschaftlich, auf den Gebrauch dieser KI beim Verfassen dieses Textes zu verzichten? Warum verzichtet heute noch jemand auf die KI? Ist das nicht einfach nur traditional legitimiertes Handeln qua Gewohnheit im Sinne Max Webers?

*Siebtens:* Ich könnte diese Strukturierung des Themas der KI als meine eigene übernehmen und anschließend mit meinen eigenen Inhalten und Beispielen erläutern. Aber wer schreibt dann diesen Text in diesem Sammelband? Ist es Valentin Rauer oder ist es Valentin Rauer, der sich von <https://bard.google.com/> inspirieren lässt? Aber die gute Eingangsfrage mit den zehn Wörtern „The consequences of KI for our concepts of human action“ habe doch ich selbst verfasst. Also ohne diesen geschickt gewählten allgemeinen, von mir selbst verfassten Satzanfang hätte die KI nicht diesen Textabschnitt derartig erfasst. Das würde für meine eigene, die KI-Textualisierung relevant übersteigende Schöpfungstiefe eines solchen Texts sprechen – oder neigt die Handlungsträgerschaft der Schöpfungstiefe nicht doch auf die Seite der KI?

*Achtens:* Meine Frage auf der KI-Plattform <https://bard.google.com/> ist mit meinen E-Mail-Plattformen bei Google verlinkt, sowohl meine private als auch meine berufliche E-Mail-Adresse ist bei Alphabet/Google. Hat die KI meine dort verlinkten Texte und versendeten.pdf-Dateien eventuell ausgewertet, um seine Vorhersage, was ich nach meinem 10-Worte-Satz lesen möchte besser präzisieren zu können? Habe ich nicht so etwas Ähnliches wie Datenschutzrichtlinien gelesen, als ich mich für Bard registriert habe? Wäre dann der obige Bard-Text nicht doch auch ein Teil meiner eigenen wissenschaftlichen Textproduktion und deren Zusammenfassung durch die KI? Also wäre der Text eher meiner eigenen Schöpfungstiefe und Handlungsträgerschaft als der der KI geschuldet. Die KI hätte lediglich qua meines Google Accounts einen besseren, präziseren Archivzugang zu meinen eigenen Texten, aber die Texte bleiben ja meine Texte. Fasst also die KI meine Texte nur in Relation zu N-anderen Texten zusammen? Das bedeutet, was ich bereits in anderen Kontexten und anderen Sprachen verfasst habe wird in Relation zum Forschungsstand gestellt. Aber nichts anderes habe ich als Autor doch auch immer schon gemacht? Wer ist dann der Autor dieses Textes – die Plattform, die meine eigenen Textproduktionen besser zusammenfasst als ich selbst? Oder die meine Textproduktionen der Vergangenheit mit den Textproduktionen anderer rein menschlicher Handlungsträgerschaften abgleicht? Schließlich: Was bedeutet es in diesem Kontext der Textproduktionen überhaupt noch von ‚rein menschlicher‘ Handlungsträgerschaft zu sprechen? Ist das nicht obsolet? Wo ist letztlich der Unterschied, ob ich mich für meinen Text von einem anderen rein unter menschlicher Handlungsträgerschaft verfassten Text inspirieren lasse oder von einem statistisch zusammengefassten KI-Text?

*Neuntens:* Auch zuvor habe ich mich bei der Texterstellung auf die KI der Google-Suchfunktion verlassen, um Themen zu bearbeiten, Literatur zu finden etc. Doch die jeweiligen Gewichtungen und Strukturen der Texte sowie deren Wortlaut hatte ich bisher selbst geschrieben. Ist diese Eigenleistung noch

wünschenswert oder nicht bereits ebenfalls Ausdruck eines inzwischen obsolet gewordenen, humanzentrierten Traditionalismus?

*Zehntens:* Kann ich im Jahre 2023 ohne den Einsatz von KI über KI überhaupt noch schreiben? Verletze ich dann nicht die inzwischen geltenden wissenschaftlichen Standards und methodologischen Anforderungen adäquater Debattenzusammenfassung?

Die hier als Liste und in autologischer Frageform formulierten Reflexionen kreisen unsystematisch um die Frage nach den handlungstheoretischen Folgen von KI. Sie ähnelt dem in der Wissenschaft bekannt gewordenen sogenannten ‚Turing Test‘ und unterscheidet sich von diesem, wie gleich zu zeigen sein wird, in einem zentralen Aspekt. Eine Darstellung dieser Forschungslage zum Turing-Test muss aus Platzgründen leider unterbleiben (vgl. Pinar Saygin et al. 2000).

Alain Turing hatte in seinem ‚Imitationsspiel‘ genannten, später als Turing-Test rezipierten Gedankenexperiment die Bedingung einer erfolgreichen KI formuliert (ebd.). Es sei die Aufgabe einer erfolgreichen KI, dass der Roboter in der Interaktion einen menschlichen Handlungsträger derartig perfekt imitiere, dass dieser nicht mehr von menschlichen Interaktionspartner\*innen unterschieden werden könne (Turing 1950). Im Turing-Test soll der Roboter also den Menschen imitieren. Der Mensch bleibt gleich, der Roboter wird so lange optimiert, bis er in der Lage ist, wie ein Mensch zu interagieren. Der Test war asymmetrisch angelegt: der menschliche Handlungsträger wurde als stabil gedacht, der robotische Handlungsträger musste optimiert werden.

Das obige Beispiel aus einer realen Interaktionssituation mit einer KI zeigt einen entscheidenden Unterschied zum Turing-Test: die symmetrischen Veränderungspotentiale. Nicht nur der Roboter verändert sich, auch der Mensch. Der Mensch übernimmt Deutungsangebote von der interagierenden KI und verändert sein Handeln. Ein menschlicher Handlungsträger verfasst seine Texte angesichts der Interaktion mit einer Texte verfassenden KI anders als zuvor.

Der Turing-Test war asymmetrisch anthropozentrisch ausgelegt. Nahezu narzisstisch gingen die Theorien davon aus, dass die menschlichen Interaktionspartner mit sich selbst in der Interaktionssituation mit einer KI identisch bleiben könnten. Die obige Reflexion zu einer Texte verfassenden KI zeigt, wie fehlleitend eine solche Annahme ist. Interaktionsplattformen wie Googles Bard oder ChatGPT sind keine Imitationsmaschinen mit anthropomorphen Körpergestalten oder festen materiellen Grenzen. Vielmehr ereignet sich die Interaktion von Menschen mit der KI auf Akteurs-Daten-Netzwerken im Sinne sozialer Konnektivitäten (van Dijck 2013). Dies sind keine Maschinen, die imitieren, sondern Maschinen die menschliche, sinngenerierende Handlungen der Vergangenheit sensorisch registrieren, verarbeiten, gewichtet und abstrahierend reproduzieren



und verbinden. Und die über Algorithmen verfügen, die an großen Datenmengen trainiert wurden. Eine menschliche Autorin interagiert mit Googles Bard nicht, als ob sie sich an einen Roboter wende, sondern sie stellt eine Suchanfrage an einen interaktiven Handlungsträger, d. h. an eine wissensgenierende Maschine, die mit statistisch mathematischen Kompetenzen vergangene sinnhafte Handlungsprodukte menschlicher und nicht-menschlicher Akteure aggregiert und probabilistisch auswertet. Die Kompetenz dieser KI ist das, was die menschliche Interaktionspartnerin berücksichtigt. Ob die robotische Maschine einem Menschen gleicht oder nicht, ist irrelevant. Die Handlungspragmatik entscheidet über die Folgen von KI für menschliches Handeln, nicht die Ontologie einer vermeintlich anthropomorphen Maschine.

Der Turing-Test ging noch von einer binären sozialen Interaktionssituation aus. Im Turing-Test interagiert ein Mensch mit einem Roboter. Beide Handlungsträgerschaften – Mensch und Roboter – stehen sich als jeweils ‚rein‘ unterscheidbare Akteursträgerschaften gegenüber. Die menschlichen Handlungsträger sind ‚reine‘ Menschen, die robotischen Handlungsträger sind ‚reine‘ Maschinen. Doch wie das obige Beispiel zeigt, entspricht eine so konzipierte, binär unterscheidbare Interaktionssituation von zwei binären, angeblich ontologisch vollständig unterscheidbaren Akteursträgerschaften eines Turing-Tests nicht der sozialen Realität der digitalen Gesellschaft. Wie im obigen Beispiel einer Interaktion mit Bard, interagieren die Lesenden eines robotisch erstellten Textes mit Sinnprodukten, die von divers verteilten Handlungsträgerschaften stammen (Hutchins 1995; Rauer 2017b). Die ersten initiiierenden Wörter stammen von einem menschlichen Handlungsträger. Der auf die zehn Wörter antwortende Text von Bard orientiert sich aber nicht nur an den zehn Wörtern, sondern aggregiert vergangene Texte und deren Strukturierungen und Kategorisierungen. Die Strukturierungen in drei Themenschwerpunkte beispielsweise, also 1. Kreativität, 2. Problemlösungskompetenz, 3. Empathie, hat Bard selbstständig vorgenommen. Gleichzeitig ist diese ‚Selbstständigkeit‘ auch nur eine Scheinautonomie der Maschine. Die Aggregationsarbeit ist die einer Maschine, aber die einzelnen Inhalte sind auch von nicht-maschinellen Handlungsträgern geprägt. Und zwar die wissenschaftlichen Artikel, an denen der Bard-Algorithmus trainiert wurde und die Algorithmen, die von menschlichen Handlungsträgern kodiert wurden. Die Trainingsdaten, d. h. die Texte über Robotik und soziales Handeln, stammen auch von menschlichen Handlungsträgern. Es stehen sich nicht Menschen und Roboter gegenüber, wie der Turing-Test suggeriert, vielmehr sind die Grenzen zwischen beiden Handlungsträgern durchlässig.

Denn aktuell lassen sich auch immer mehr Texte nur noch zu Teilen von rein menschlichen Handlungsträgern herleiten. Der Bard-Algorithmus ermittelt aus

seinen Trainingsprozessen stochastisch unter diesen zumeist aktuell noch wohl in der Mehrheit von menschlichen Handlungsträgern stammenden Texten die wahrscheinlichste thematische Strukturierung in drei Bereiche. Vermutlich hat der Bard-Algorithmus festgestellt, dass eine Strukturierung in drei thematische Bereiche, die bei einem solchen Thema wahrscheinlichste Typisierung bildet. Doch unter den Trainingsdaten können sich auch bereits KI-produzierte Texte befinden. Die Grenze zwischen Mensch und Maschine lässt sich schlicht nicht mehr ziehen.

Insgesamt handelt es sich bei der oben gezeigten Leseerfahrung um eine kollektive, sowohl menschliche als auch nicht-menschliche Handlungsträgerschaften aggregierende Textantwort auf einen Satz. Die statistische Auswertung der aktuellen Texte, die zum Thema KI und soziales Handeln publiziert wurden, ist kein ‚Akteur‘ im Sinne eines verkörperten ‚Roboters‘ einer Turing-Maschine. Vielmehr handelt es sich um ein weit verzweigtes Akteur-Daten-Netzwerk, das algorithmischen Trainings und Auswertungen unterzogen wird. Weder bestimmen die Kodierenden der Algorithmen den Inhalt dieses oben zitierten Textes von Bard noch die Nutzenden von Googles Bard-Plattform oder gar die Trainingsdaten allein die inhaltliche Struktur des Textes. Stattdessen handelt es sich um das dynamische Zusammenspiel all dieser Elemente, das den Text als KI-Output konstituiert.

Schließlich ergibt sich die Frage nach dem Determinismus und der Kontingenz einer solchen KI. Roboter wurden gemeinhin in der Forschung oftmals als letztlich doch determinierte Maschine beschrieben, die nur scheinbar wie ein autonom handelnder menschlicher Handlungsträger agiert. Kontingenz, d. h. das Handlungsvermögen, sich möglicherweise aber nicht notwendig für oder gegen etwas zu entscheiden, wurde aufseiten der Menschen verortet. Doch wie Michael Heine in diesem Band argumentiert, ist KI ebenfalls hochgradig in kontingente Handlungspraktiken eingebunden (Heine 2023). Auch bei KI, so wird inzwischen deutlich, ist Kontingenz ähnlich wie bei rein menschlichen Akteuren grundsätzlich in den Prozess sinnhafter und kulturell relevanter Bedeutungsproduktionen eingelassen (Nieborg und Poell 2018). Die symmetrische Verteilung von Kontingenz zwischen menschlichen Handlungsträgern und KI ist eine weitere Folge von KI für Theorien des sozialen Handelns. Der Maschinenanteil der Interaktion wurde klassischerweise mit Notwendigkeit assoziiert und nicht mit Kontingenz konzipiert. Übergänge bilden die techniksoziologischen Ansätze zu KI von Werner Rammert und Ingo Schulz-Schaeffer (2006). Eine symmetrische Verteilung von Kontingenz als Voraussetzung beider Seiten der Interaktion ist jedoch erst noch in den Anfängen, wie Heine (2023) zeigt.

KI ist in diesem handlungstheoretischen Verständnis keine robotische Maschine, die notwendigen Mechanismen folgt, sondern eine Wahrscheinlichkeit und Kontingenz prozessierendes, dynamisches Sensor-Daten-Software-Aktuator-Netzwerk. Dynamisiert wird es durch eine Kontingenz inkludierende statistische Speicherung, Datenauswertung, algorithmischen, ‚non-supervised‘ Trainings großer Datenmengen und mathematischer probabilistischer Berechnungen bei der Interaktion mit menschlichen Handlungsträgern auf Plattformebene.

Wichtig ist dabei, dass nicht nur die Produkte der KI symmetrisch und dynamisiert zu verstehen sind. Die Grenzen zwischen jeweiligen beteiligten Handlungsträgerschaften selbst sind durchlässig. Nutzende sind zugleich Datenliefernde, an denen KI weiter trainiert wird. Interaktionssituationen sind gleichzeitig Schnittstellen etc. Die Grenzen der Handlungsträgerschaften sind fluide und multivokal. Der Text lässt den Roboter nicht als Mensch erscheinen, sondern menschliche und nicht-menschliche Handlungsträgerschaften werden in dem Text ununterscheidbar ineinander verwoben. Im Turing-Test war diese Ununterscheidbarkeit noch asymmetrisch begrenzbar angelegt. Der Turing-Test sollte dann als erfolgreich gelten, wenn der interagierende Roboter sich nicht mehr von einem Menschen unterscheiden lasse, d. h. vollständig menschenähnlich wird. Aktuelle KI-Systeme sind demgegenüber symmetrisch. Von Menschen verfasste Texte lassen sich nicht mehr eindeutig in jeder Hinsicht von maschinellen Texten unterscheiden, weil sich die Grenzen zwischen menschlich und nicht-menschlich verfassten Texten in vielen Einzelschritten der kognitiven Verknüpfungen neu verteilen. Ein mit der Suchfunktion von Google sich inspirierender Text hat bereits eine Ko-Autorenschaft mit einer KI.

Anders formuliert: Bei der aktuellen plattformgestützten Textautorenschaft stammt eine der in den Text eingeschriebenen Sinnfolgen möglicherweise von Menschen, andere wurden von maschinellen Handlungsträgern geprägt, anschließend von Menschen leicht modifiziert, je nach Interessenlage eines menschlichen Handlungsträgers nochmals transformiert und erneut einem maschinellen Handlungsträger zur Interaktion und Verarbeitung vorgelegt. Dieser komplexe Prozess, der sich über große Datenmengen, das Synthetisieren und singuläre Output-Produkte erstreckt, kann in strengem Sinne nicht mehr als ‚soziale Interaktionssituation‘ bezeichnet werden. Es handelt es sich auch nicht bloß um ein Netzwerk. Anders als in einem Netzwerk, werden konkrete binäre Frage-Antwort-Konstellationen realisiert, wie die obige Textanfrage, die einer Interaktionssituation ähnelt.

Oder nochmals anders formuliert: Googles Bard wird sich bei dem Prozessieren seines Interaktionsprozesses und bei seinen Trainings auf menschlich

verfasste Texte beziehen, aber eben nicht ausschließlich auf diese Texte. Zusätzlich wird er sich auch auf maschinell erstellte Textprodukte oder unterschiedlich graduierte Mischformen aus beidem beziehen. Um ein Beispiel zu geben: Angenommen das obige Beispiel von Bards Textausschnitt wird gedruckt und als digitale Textform lesbar – etwa, wenn dieser Text in einem Sammelband als open access publiziert wird. In der Folge wird auch der von Bard erstellte Textausschnitt wiederum von Bard zur statistischen Prädiktion herangezogen, so wird auch dieses Textzitat wiederum Grundlage weiterer Textproduktionen.<sup>6</sup> Die Tatsache, dass der Bard-Text in einem wissenschaftlich legitimierten Sammelband erscheint, wird die Gütebewertung der KI zur Berücksichtigung des Textes in einer Vorschlagssituation mit Nutzenden erhöhen. Ob Bard den Text zuvor selbst erstellt hat oder nicht, ist sekundär. Primär ist der ‚Publikationserfolg‘ des Textausschnitts, um ein höheres Ranking bei der Vorschlagsauswahl zu erreichen. Diese Höherbewertung wäre auch in diesem Fall dadurch bedingt, dass dieses Textzitat von menschlichen Handlungsträgern ausgewählt und mit Blick auf die geäußerten Geltungsansprüche überprüft worden ist. Es wird von dem menschlichen Autor dieses Textes zitiert und von den Herausgebenden des Sammelbandes zur Publikation akzeptiert. Damit unterläuft der maschinell erstellte Text einen weiteren Zwischenschritt maschinell-menschlicher Ko-Autorenschaft.

Das binäre Modell der Turing-Tests, bei dem sich die Maschine einseitig an den Menschen anpassen sollte und in einer singulären Interaktionssituation mit festen Grenzen der Handlungsträgerschaften getestet wird, ist handlungstheoretisch endgültig obsolet, wenn wir die handlungstheoretischen Folgen von KI analysieren. Stattdessen ist ein Verlust einer eindeutig menschlich reservierten Sphäre sinnhaft kommunikativer Handlung zu beobachten. Nicht nur die Roboter ändern sich, sondern auch die Menschen. Was sich zeigt, ist eine dynamische, symmetrisch verteilte kontingenzermöglichende Interaktionssituation, die durch zunehmende Verteilung und Distribution menschlicher und nicht-menschlicher Handlungsträgerschaften in-, unter und übereinander charakterisiert ist.

---

<sup>6</sup> Es finden sich im Internet inzwischen Berichte, dass Google sich Klagen gegenüber sieht, weil das Unternehmen mutmaßlich seinen Algorithmus an illegalen Textquellen trainiert habe, die nicht open access waren (<https://wsvn.com/news/us-world/google-hit-with-lawsuit-alleging-it-stole-data-from-millions-of-users-to-train-its-ai-tools/> Zugriff am 14.07.2023). Ein solches Training hätte Google vergüten müssen. Dass der Algorithmus von Bard auch Textkonvolute von nicht als open access klassifizierten Büchern einbezieht, über international verfügbare illegal öffentlich gemachte Plattformen, wäre eine weitere Erklärung, warum aktuell Bard derartig vollständige wissenschaftlich strukturierte Texte liefert. Doch dies muss hier aufgrund der aktuell noch fehlenden Gerichtsverfahren und eines laufenden Untersuchungsprozesses als reine Verdachtsvermutung offengelassen werden.

Die soweit dargelegten verteilten Handlungsträgerschaften (Rauer 2017a) und Handlungselemente lassen eine Neujustierung von bestimmten Aspekten klassischer soziologischer Handlungstheorien notwendig werden. Eine Darstellung dieser klassischen Handlungstheorien kann aus Platzgründen nicht erfolgen und wird hier vorausgesetzt. Soziales Handeln wird seit Max Weber allgemein als ein interaktives Handeln definiert (Weber 1984). Die Handlungsgründe und Motive orientieren sich an einem subjektiven Handlungssinn. D. h. in einer Interaktion orientiert sich Akteur A an den von ihm interpretierten vermeintlichen Handlungssinn von Akteur B und legt diese Vermutung seinem Handeln als zugrunde. Während ökonomisches Handeln sich an einer vermeintlichen Nutzenmaximierung von Akteur A gegenüber Akteur B und vice versa orientiert, orientiert sich soziales Handeln an sinnhaften Bedeutungen, situativen Rahmungen, Interpretationen und Werten. Soziales Handeln ist nach Max Weber also sinnhaftes Handeln, es ist kein bloßes Verhalten und Reagieren auf Reize oder Anreize auf materielle Vorkommnisse. Sozial handeln nur menschliche Akteure, alles andere ist bloßes Verhalten. Diese Voraussetzung ist zentral für die anschließende Argumentation, wie KI in Interaktionssituationen diesen subjektiven Sinn formt und transformiert (vgl. dazu allgemein Sebald 2014). In der klassischen handlungstheoretischen Sichtweise waren Interaktionssituationen zudem auf das räumlich-zeitliche Hier und Jetzt begrenzt. In einer sozialen Situation wurden Geltungsansprüche kommunikativ verbal oder gestisch-indexikal wechselseitig bestätigt oder verworfen (Goffman 1967). Die Folgen von KI sind, dass sich diese Grundannahmen so nicht mehr halten lassen. Zum einen müssen materiale Handlungsträger in die sinnhafte Interaktionssituation mit einbezogen werden, zum anderen sind mikrosoziale Interaktionssituationen (Rammert und Schubert 2006) nicht mehr auf das Hier und Jetzt abgrenzbar.

Diese Folgen lassen sich anhand der obigen Interaktionssituation mit Bard erläutern. Auf einen Satz als Suchanfrage, der aus zehn von mir als menschlicher Handlungsträger verfassten Worten bestand, reagiert die KI-Plattform Bard mit einem Text aus etwa einhundert Wörtern. Diese Reaktion als eine Art ‚Textantwort‘ auf meine ‚Textfrage‘ wurde aus milliardenfach vorhandenen Datensätzen, d. h. Textbeispielen, probabilistisch generiert. Die Situation ist interaktiv, aber die vergangenen Textproduktionen sind direkt anwesend in der singulären Interaktionssituation und der Anfrage zur Schreibsituation dieses Textes auf der Plattform.

Im vor-digitalen Zeitalter konstituierte sich diese Anwesenheit der Vergangenheit in Re-Präsentationen, d. h. die situative Anwesenheit wurde durch das soziale Gedächtnis ermöglicht, welches hochselektiv umfassenden Reduktions-

und Vergessensprozessen unterlag (Dimbath und Heinlein 2014). Das vergangene Wissen war in Texten in Form von Büchern und sozialer Gedächtnisbildung geronnen und kodiert und konnte bei Bedarf in der Gegenwart selektiert und verwendet werden. Im Zuge der Nutzung von KI-Plattformen wie ChatGPT oder Google Bard ist diese Anwesenheit maschinisiert und stochastisch präzisiert worden. Soziale Interaktion im Bereich gesellschaftlichen Wissens und sozialer Gedächtnisse ist nicht mehr auf menschliche Handlungsträger und soziale Gedächtnisträger begrenzt. Die Archive selbst sind dynamisiert, sie interagieren in der jeweiligen Interaktionssituation unmittelbar mit anfragenden menschlichen Handlungsträgerinnen.

Welche theoretischen Herausforderungen ergeben sich also konkret für die Interaktionssituation bei der Integration von KI-basierten Handlungsträgerschaften in soziale Sinnbildungsprozesse? Zunächst ist dies eine Frage nach den KI-bedingten Veränderungen von subjektiver Sinnbildung und handlungskonstituierenden Bedeutungssystemen. Während Roboter die menschliche ‚körperliche‘ Arbeit transformieren, transformiert die aktuelle sinngenerierende KI die menschliche ‚geistige‘ Arbeit. So wie Roboter zunächst bei hochgradig repetitiven Tätigkeiten menschlicher körperlicher Arbeit ersetzen, so haben auch KI-Systeme Menschen zunächst bei geistig hoch repetitiven intellektuellen Tätigkeiten assistiert, beispielsweise bei der Kodierung von großen Datenmengen und der statistischen Auswertung. Doch inzwischen sind diese Auswertungen interaktiv mit den Nutzenden dynamisch verbunden: Die Nutzenden tragen aufgrund ihrer Nutzungsmuster zur Optimierung der Ergebnisse bei. Damit ist die Interaktionssituation zugleich sowohl eine wechselseitige Begegnung als auch eine Produktionssituation und Evaluation von KI. Interaktion, Produktion und Konsumption werden damit immer weniger unterscheidbar. Gleiches gilt für die Differenz von Autorenschaft und Leserschaft, die ebenfalls zunehmend ununterscheidbar wird. Selbst basale Interaktionseinheiten wie Ego und Alter scheinen sich zunehmend zu überlagern.

Ein solcher Entdifferenzierungsprozess stünde in starkem Kontrast zu den normativen Ordnungen moderner Gesellschaften. Denn diese beruhen auf einem handlungstheoretischen Modell, das von bestimmten Voraussetzungen ausgeht, wie beispielsweise die Fähigkeit zu autonomem Handeln und Entscheiden menschlicher Akteure und deren unzweifelhafter Identifizierbarkeit als verantwortliches Individuum (Teubner 2007, Beckers und Teubner 2023). Wenn sich Handlungsträgerschaften über menschliche und maschinelle Träger verteilen und als Translationssystem sozialen Sinn in spezifischen singulären Interaktionssituationen generieren, dann steht dieses klassische Modell samt seiner normativ

kulturellen Folgen vor einer großen Herausforderung. Falls zunehmend KI-Systeme in diesen Prozessen konstituierende Effekte entfalten, dann bedroht dies die Grundlagen normativer Ordnungen moderner Gesellschaften (Rauer 2017a).

Allgemein bedeutet dies für die hier verhandelte Problemstellung jedoch nur, dass KI soziologisch gesprochen nicht als abgeschlossene Akteureinheit ‚Maschine‘ mit Menschen in sozialen Situationen interagiert, sondern es zu Grenzverlusten innerhalb des Interaktionszusammenhangs kommt. KI meint hier also mehr als das, was bislang als menschlich-maschinelle Interaktion im Zuge der Robotik thematisiert wurde (Rammert 1995, 2007, 2012). Die aktuelle sinn-generierende KI zielt auf Fragen systemischer Datengenerierung, algorithmischer Datenklassifizierungen und davon abgeleiteten statistisch ermittelten Vorhersagen ab. Anders als kausal deterministische Maschinen, die Menschen bedienen mussten, verarbeitet und exekutiert die aktuelle sinn-generierende KI soziale Bedeutungen als Wahrscheinlichkeiten. Der erste Schritt besteht also erst einmal darin, den maschinellen Wandel von klassisch determinierten Maschinen zu wahrscheinlichkeitsorientierten maschinellen Systemen als sinn-generierende Interaktionspartner der Menschen anzuerkennen. Wie diese Handlungsträgerschaft soziologisch zu fassen ist, ist eine zweite Frage. Doch der Mensch ist nicht mehr das Maß, an der diese Handlungsträgerschaft gemessen werden muss. Eine reine Imitation wie im klassischen Turing-Test verkehrt die Kompetenzverteilungen. In Zukunft stellt der Turing-Test vielleicht eher die Frage, wie lange noch ein Mensch in der Lage war, die sinn-generierende Leistung einer KI so zu imitieren, dass sie wie eine sinn-generierende KI erschien.

---

## **2 KI als bedeutungskonstituierende Handlungsträgerschaft**

Welche Folgen hat die sinn-generierende KI für die soziologischen, handlungstheoretischen Grundlagen? Zunächst lässt sich allgemein feststellen, dass im Zuge der sinn-generierenden KI Akteure ihren sozialen Handlungssinn nicht nur in zwischenmenschlichen Interaktionen mit anderen konkreten Menschen, sondern zunehmend in Interaktionen mit KI-prozessierenden Plattformen bzw. Chatbots generieren. Plattformen wie Googles ‚Bard‘ oder ChatGPT erfassen die Produkte menschlicher Einzelhandlungen, archivieren diese in großen Datensätzen und interpretieren darin Regelmäßigkeiten, um auf deren Grundlage zukünftig Handlungsabläufe besser vorhersagen zu können. ChatGPT und Bard bieten auf diese Weise den Menschen Textabschnitte an, die diese dann zu einem gewissen

Wahrscheinlichkeitsgrad verwenden oder verwerfen. Diese Daten werden wiederum archiviert, interpretiert etc. KI-Systeme sind somit nicht statisch, sondern dynamisch und interaktiv. Nicht vermeintliche Roboter im Sinne einer Entität sind gemeint, sondern sinngenerierende digitale Plattformen, die aus den Texten und Handlungen von Menschen ‚lernen‘ wie es leider etwas verwirrend heißt (Rogalla 2023).

Bisher galt als leitende Vorstellung, dass KI eine von menschlichen Interaktionen weitgehend getrennt agierende (semi-)autonome operierende Maschine sei. Von der Maschine getrennt sind die menschlichen Praktiken zu verorten. Von den Menschen getrennt seien die algorithmisch prozessierenden, auf virtueller Realität ‚intraaktiv‘ operierende Roboter im Sinne einer „distributed agency“ (Rammert 2012, S. 89) zu verorten (vgl. auch Hutchins 1995).

Das Konzept des ‚Roboters‘, eine sich auf das russische Wort für Arbeiter beziehende Metapher, war die leitende Begrifflichkeit für diese duale Sichtweise. Zwar wurden diesen Robotern zunehmend autonome Handlungsträgerschaften in menschlich-maschinellen Interaktionen zugestanden (Rammert und Schulz-Schaeffer 2002), doch die Interaktion wurde stets asymmetrisch gedacht. Die Menschen produzieren die Roboter und bestimmen die algorithmischen Abläufe. Die Praxis des non-supervised learning aktueller KI-Systeme wurde in diesen frühen soziologischen Ansätzen noch nicht berücksichtigt. Stattdessen blieb den Menschen im situativen Loop nach wie vor so gut wie jede relevante Agency vorbehalten (Rauer 2017b).

Der aktuelle Erfolg von KI-Systemen wie die auf Deep Learning Algorithmen beruhenden Plattformen ist nicht mehr mit diesen dualen handlungstheoretischen Interaktionsmodellen zu beschreiben. Vielmehr zeigt sich eine permanent wandelnde menschlich-maschinelle aggregierende Dynamik, die im Zuge der sukzessiven, menschlich-maschinellen Interaktionsschritte die Auswertungskriterien der interagierenden Handlungsträgerschaften transformiert. Und diese Transformation ist nicht mehr asymmetrisch auf den Menschen als Maßstab der zu erreichenden Kompetenz ausgerichtet.

Beispielsweise interagieren die aktuellen KI-Systeme wie Bard oder ChatGPT nicht nur mit menschlichen Praktiken, sondern sie archivieren die jeweiligen Interaktionsprozesse, aggregieren die Daten mit anderen Interaktionsprozessen und passen die jeweiligen Prädiktionsalgorithmen dynamisch an. Handlungstheoretisch gesprochen, stehen sich also nicht mehr nur zwei stabile Akteure – Mensch und Maschine – gegenüber, sondern es handelt sich um dynamisch datengenerierende, archivierende und auswertende menschliche und nicht-menschliche Handlungsträgerschaften. In einem permanenten dynamischen Kreislauf von



Erfassung, Auswertung und Wiederangebot sind Kontingenzelemente eingelassen (Heinlein 2023). Die Letztentscheidung über die bessere Handlungserwartung liegt nicht mehr bei den rein menschlich prozessierenden Handlungsträgerschaften, sondern erfolgt selbst in transformativen Interaktionsprozessen zwischen KI-System und Menschen und jeweils kontingenten Bewertungsprozessen. Kontingenz, d. h. die Möglichkeit etwas zu tun, ohne die Notwendigkeit dies tun zu müssen, ist kein Privileg menschlicher Akteure mehr.

Diese permanent dynamisch und kontingent sich transformierenden Handlungsträgerschaften sind nicht adäquat als Subjekte im klassischen handlungstheoretischen Sinne beschreibbar. Passende sprachliche Konzepte fehlen jedoch bisher, für das, was als aktive und dynamisierte Verdichtungszone daten- und sinngenerierenden Akteur-Netzwerken ähnelt. Aktiv vernetzt können darin menschliche Handlungsträger sein, die Handlungsverdichtungszone können aber auch durch Bots und andere algorithmische Interaktionsprogramme konstituiert, geprägt und transformiert werden. Plattformen wie Google oder ChatGPT sind in diesem Bild als kollektiv vernetzte, dynamische Aktivitätszentren oder als ‚Interaktionswolken‘ im Sinne einer ‚interaction-cloud‘ zu verstehen. ‚Clouds‘ sind netzbasierte Datenspeicher, die inzwischen soziologisch „als eine grundlegende Infrastruktur für vernetztes Leben und Arbeiten“ (Eggert und Kerpen 2015, S. 1770) beschrieben wurden. Ähnliches gilt für die lernbasierten dynamisierten Interaktionsplattformen. Plattformen ersetzen dabei offenbar zunehmend die theoretische Position, die vormals klassisch sozialen Organisationen, Institutionen oder kollektiven Akteuren in der Soziologie zugemessen wurde (Dolata 2013; Davis 2016; Gillespie 2016; Kenney und Zysman 2016). Die Forschung zu diesen neuen kollektiven digitalen Akteursträgerschaften nimmt aktuell immer umfassendere Züge an und kann hier nicht erschöpfend dargestellt werden (Barwise und Watkins 2018; Nieborg und Poell 2018; Dolata und Schrape 2018; Dolata 2019; Dolata und Schrape 2022).

Beschränkt auf die Mikroanalyse sozialen Handelns, bedeutet diese neue Form dynamisch interaktiver Infrastrukturen, dass auch der klassische Begriff ‚sozialer Interaktion‘ zu kurz greift, da er stets nur Individuen in Gestalt menschlicher Handlungsträgerschaften berücksichtigt. So wie in makrobasierten Studien das Konzept der Organisation durch Plattformkonzepte ergänzt werden muss, so muss in mikrobasierten Studien der Interaktionsbegriff durch KI-basierte, interaktive Handlungsträgerschaften ergänzt werden. Das Problem ist dabei, dass nicht mehr präzise erkennbar ist, wer die autorisierenden, produzierenden und nutzenden Handlungsträgerschaften in kommunikativen Interaktionssituationen sind.

Im Zuge des 20. Jahrhunderts begannen, wie bereits gesagt, komplexe Maschinen, sogenannte Roboter, menschliche Arbeiter zu ersetzen. Zu nennen sind

multiple Einsatzbereiche in der Pflegearbeit, Polizeiarbeit, beim Katastrophenschutz und im Militär (Knoll und Christaller 2003). Auch in diesen Bereichen setzen sich zunehmend semi-robotische Verfahren durch, mit denen menschlich-maschinelle Interaktionsmöglichkeiten zur Anpassung an situative Anforderungen etabliert werden (Müller et al. 2017). Auch in der klassischen Robotik wird die Interaktion inzwischen als semi-robotisch durch unklare Grenzziehungen zwischen menschlichen und KI-basierten Handlungsträgerschaften charakterisiert. Aus einer soziologisch-handlungstheoretischen Perspektive stellt sich damit ein Problem. Denn die Grenze zwischen menschlichem und maschinellern Handeln war stets für die soziologische Theoriebildung konstitutiv. Seit Max Webers einflussreichen Definitionen sozialen Handelns als sinnhaft subjektivem Handeln, das dem bloßen mechanisch reaktiven Verhalten gegenüberstehe, wurde eine Abgrenzung gegen alles Nicht-Menschliche etabliert. Diese Grenze ist nun theoretisch selbst das Problem, nicht deren technische Überschreitung (Latour 2007).

Zunächst fällt bei der Analyse dieses Problems auf, dass sich Menschen bei der Lösung von intelligiblen Problemen schon immer Werkzeugen bedienen. Streng genommen müsste jeder auf einem Blatt mit Stift ausgeführte Rechenalgorithmus als ‚künstlich‘ und damit als nicht menschlich-sinnhaft gelten. Warum erst die elektronischen Hilfsmittel zu einer scharfen Grenze zwischen künstlicher und natürlicher Intelligenz führten, die menschliches Handeln streng von nicht-menschlichen Praktiken unterscheidet, ist auch ohne KI-Systeme bereits nicht überzeugend (Latour 2007). Eine solche linear gezogene Grenze verführt dazu, menschliches Handeln und menschliche Praktiken stets auf deren intentionale und motivationale Anteile zurückzuführen und tendiert dann dazu, den materiellen Bereich menschlicher Handlungsträgerschaften auszublenden. Das führt, so die These aus dem Umkreis der Akteur-Netzwerk-Theorien, zu einer reduktiven scholastischen Sozialwissenschaft (Latour 1991). Stattdessen gelte es einen symmetrischen Blick auf alle Handlungsabschnitte zu werfen, der explizit die technisch erfolgten Praktiken gleichwertig miteinbezieht (Latour 2001, 2007; Laux 2011).

Vereinfacht ausgedrückt: Wenn menschliche Akteure durch technologische Handlungsanteile befähigt werden, etwas zu tun oder zu unterlassen, dann sind diese technischen Anteile an dem Praxisverlauf mit-ursächlich beteiligt. Nicht nur die motivationale und intentionale Lage entscheidet über Handlung und Unterlassung, sondern auch die technologische Befähigung (Rauer 2017b). Warum dieser zweite Aspekt so lange aus der soziologischen Theoriebildung exkludiert wurde, hat Gründe, deren Nennung hier zu weit führen würde. Für die handlungstheoretische Frage nach den Folgen von KI ist zunächst festzuhalten, dass

die Unterscheidung von künstlicher versus natürlicher Intelligenz fälschlich dazu verleitet, die maschinisierten Anteile sozialer Interaktionsprozesse zu übersehen. Die Unterscheidung von künstlicher versus natürlicher Intelligenz wirkt scheinbar apriorisch als gesetzt und verleitet zu einer binären theoretischen Konzeptualisierung von menschlicher Intelligenz als natürliche und robotischer Intelligenz als etwas künstlich Geschaffenes. Der von Menschen im Kopf rechnerisch ermittelte Problemlösungsweg einer mathematischen Aufgabe ist soziologisch nicht natürlicher als der maschinell ermittelte Lösungsweg. Die Rechenalgorithmen sind identisch, sie sind weder Natur noch Kultur. Die Grenze zwischen künstlicher und natürlicher Intelligenz ist aus sozialtheoretischer Sicht nicht schlüssig rechtfertigbar (vgl. dazu ausführlicher Rauer 2014).

Werner Rammert (Rammert 1995) hat in seinem Klassiker zur Soziologie der Künstlichen Intelligenz wiederholt auf die Agency und menschlich-maschinelle Intraaktivitäten dieser Systeme hingewiesen (Rammert und Schulz-Schaeffer 2002). Verlässt man die Dichotomie von Natürlichkeit und Künstlichkeit und blickt stattdessen auf die Grenzziehungen von Handlungsträgerschaften, so wird indes deutlich, dass künstliche Intelligenz eher einen systemisch zonalen Bereich maschineller Wissenswahrnehmung, Speicherung, Analyse und Prädiktion bezeichnet. Die Grenzen dieses System sind weder natürlich noch künstlich, sondern dynamisch und kontingent. Denn in jedem der einzelnen Schritte werden menschliche und nicht-menschliche, d. h. in der alten Begrifflichkeit ‚natürliche‘ und ‚künstliche‘ Wissensformen miteinander ununterscheidbar verschränkt. Dabei kann nicht mehr sinnhaft zwischen Menschen auf der einen Seite und Robotern auf der anderen Seite unterschieden werden. Vielmehr geraten zonal verdichtete Aktivitätszentren in den Blick, die maschinelle und menschliche Handlungspartikel interaktiv miteinander in Relation setzen, auswerten und transformieren. Sie wären auch als Interaktionssysteme des Messens, Aggregierens, Sortierens, Kategorisierens und Präzidierens sowie als Sensorik und Aktorik zu bezeichnen. Angesichts der jüngsten Entwicklung der KI bleibt daher nach wie vor die Frage unbeantwortet, wie sich soziologische Handlungstheorien so modifizieren lassen, dass die prozessualen, transformierenden und dynamischen Anteile von KI sichtbarer und empirisch besser in den analytischen Blick rücken.

In diesem Zusammenhang muss noch eine Besonderheit der Sozialwissenschaften benannt werden. In den Kultur- und Sozialwissenschaften wird menschliches soziales Handeln in zwei verschiedene Bereiche gegliedert, für die je unterschiedliche Bedingungen ihrer Maschinisierbarkeit gelten. Zudem stellen diese beiden Bereiche unterschiedliche Anforderungen an die Sinn- und Verantwortungsformen des menschlichen Handelns (Beck 2016): erstere werden soziale Interaktionstheorien genannt, letztere Verhaltenstheorien. Die sozialen

Interaktionstheorien beruhen auf der Annahme, dass Motive und Intentionen sich innerhalb von Interaktionsprozessen qua Emergenz entwickeln und verändern (können). Dabei wird von einer intersubjektiven Sinnerzeugung im Zuge des Interaktionsprozesses ausgegangen (Luckmann 1992; Sebald 2014). Die Verhaltenstheorien gehen demgegenüber stärker von einer stabilen und vorab bereits gefestigten Intentionsordnung aus. In der realen Handlungssituation werden diese Selektionskriterien nicht Emergenz bilden, sondern abgerufen oder leicht adaptiert werden. So wurde bislang, wie bereits erwähnt, die Mensch-Maschine-Interaktion als instrumentelles Handeln oder im Sinne Max Webers (1984) als ‚bloßes Verhalten‘ bezeichnet. Mensch-Maschine-Interaktionen wurden somit stets zur zweiten Gruppe, d. h. zu den vorab determinierten Verhaltensprogrammen gerechnet. Die auf subjektiven wechselseitigen Sinnprozessen beruhenden Theorien schienen bisher den rein menschlichen Interaktionspaaren vorbehalten. Auch diese scharfe Grenzziehung subjektiv sinnhaften Handeln und bloßem Verhalten muss bei der Analyse der handlungstheoretischen Folgen von KI verworfen werden – gerade auch, weil die jüngsten Versuche zur Soziologie von KI und algorithmischen Handlungen immer noch die Dichotomie zwischen sinnhaftem Handeln der Menschen und programmiertem Verhalten der Maschinen zumindest latent ihren Analysen zugrunde legen (Mau 2018; Nassehi 2019).

Ein weiteres Problem stellt der Interaktionsbegriff selbst dar. So wurde vonseiten der Physikerin und Philosophin Karen Barad argumentiert, dass angesichts der Relativitätstheorie und der Beobachterabhängigkeit von Theorien nicht mehr von Interaktionen, sondern auch in den Sozialwissenschaften von *Intraaktionen* ausgegangen werden sollte (Barad 2012). Letztlich ist dies ein starker Appell an den bereits in der Soziologie existierenden Ansatz von einer Intersubjektivität erinnernden Theorieschule. Intersubjektivität geht von im Prozess emergent sich entwickelnden Subjektformationen aus, wie sie seit Max Webers frühen Bestimmungen des subjektiv gemeinten Sinnes verstanden werden. Barad weitet diesen intersubjektiv sich entwickelnden Sinn auch auf die Ebene des technischen und instrumentellen ‚bloßen Verhaltens‘ aus. Dieser entgrenzende Schritt ist folgerichtig, insbesondere wenn die dynamische KI in Interaktionssituationen in den theoretischen Blick genommen werden soll. Barads Konzept der ‚Intraaktion‘ wurde von ihr eingeführt, um die Wechselwirkung von Beobachtung und Beobachteten in Experimenten theoretisch zu berücksichtigen. Das Konzept versucht die physikalische Beobachtung von der Beeinflussung von Beobachtetem und Beobachtem handlungs- und subjekttheoretisch zu integrieren. Intraaktion beeinflusst die Beobachtung und die Intentionsstruktur der Interagierenden, die Interagierenden können demnach nicht mehr als eine unabhängige Variable des sozialen Handelns in Interaktionen gelten. In analoger Weise finden sich bei

dynamischen KI-Plattformen permanente Beeinflussungen zwischen den sensorisch erfassten Texten und den ‚aktorisch‘ produzierten Texten. Handelnde und Handlung, d. h. Aktanten und Aktoren, bilden und verändern sich im Handlungsprozess sozialer Situationen stets rückbezüglich je nach interaktiven Wegen und Prozessen. Für die Debatte um KI ist dieser Ansatz bedeutsam, da damit die technisch realisierte Entgrenzung sozialer Interaktionssituationen auf so genannten ‚Plattformen‘ realistischer beschrieben werden kann.

---

### **3 Umriss einer Soziologie der ‚Transformaktion‘ – theoretische Anknüpfungspunkte**

Die Interaktion von Menschen mit maschinellen Kommunikationsprogrammen und sinngenerierender KI ist keine Interaktion im klassisch soziologischen Verständnis. Im Folgenden wird der Vorschlag gemacht, diese als Transformaktion zu bezeichnen. Dieses Kunstwort soll dazu dienen, die zunehmend verschränkte Handlungslogik von digitalen und menschlichen Handlungsträgerschaften nicht mehr getrennt voneinander theoretisch zu analysieren. Der Begriff selbst findet sich bei David Gutman und bezeichnet dort ein transformatives Handlungspotential außerhalb digitaler Kontexte (Gutman 2009). Das hier nun vorgestellte Konzept der Transformaktion schließt in einer digitalen Verortung an Ansätzen der Akteur-Netzwerk-Theorie (ANT) (Belliger und Krieger 2006) und der Science and Technology Studies (STS) an (Law und Hassard 1999; Bauer et al. 2017).

In diesen sich teilweise überlappenden Ansätzen wird intensiv mit wechselseitigem Blick über die Sozialität von technischen Handlungsprogrammen geforscht. So gilt in der ANT soziales Handeln als ein soziotechnisches Arrangement, in dem sowohl menschliche Akteure als auch nicht-menschliche Aktanten Handlungsträgerschaft übernehmen. Eines der zentralen Beispiele ist der sogenannte ‚Berliner Schlüssel‘ (Latour 1996). Dies ist ein Schließmechanismus, der die Bewohner eines Mehrparteienhauses dazu zwingt, stets die Eingangstür zu verschließen. Die Bewohnerinnen und Bewohner sind nicht mehr auf ihr Gedächtnis angewiesen, sie müssen auch keinen Gebotsschildern folgen, sondern sie werden qua Technologie dazu gezwungen, das Schloss zu verschließen, wenn sie ihren Schlüssel an sich nehmen wollen. Der Berliner Schlüssel reduziert Kontingenz. Latour beschreibt diese menschlich-materiellen Formen von sozialen Handlungsketten als ‚programme of action‘. Diese ähneln algorithmischen Vorschriften und Determinationen (Rauer 2014). Gleichzeitig zeigt sich jedoch auch, dass in die

Handlungsprogramme widerständige Formen eingelassen sind, die die Determination unterlaufen und menschliche Gegenprogramme des Handelns ermöglichen (Latour et al. 1992). Hinsichtlich dieser Handlungsprogramme schreibt der Techniksoziologe Ingo Schulz-Schaeffer: „‘Programme of action’ is a general term introduced by Latour to denote goal-directed behavior of human actors as also of technological artefacts. In this terminology a certain human strategy of goal attainment and an algorithm determining the behavior of a technological artefact are both programmes of action.“ (Schulz-Schaeffer 2006, S. 131) Doch wie sich bei Latours originalem Text zeigt, finden sich sogar in Handlungsprogrammen mit starken deterministisch-technischen Anteilen häufig kontingente Spielräume der Möglichkeiten, auch anders zu handeln. Kontingenz und Determination sind zwei Pole eines Kontinuums, keine Entweder-oder-Bedingung sozialen Handelns.

Diese erste Fassung einer soziologischen handlungsorientierten Akteur-Netzwerk-Theorie hat zahlreiche Weiterentwicklungen erfahren. So hat Latour selbst angesichts der konstituierenden Funktion materialer Handlungsträgerschaften den Begriff der „Interobjektivität“ (Latour 2001, S. 237) geprägt. Sein Argument besagt, dass in jeder sozialen Situation stets materiale Bedingungen eine ko-konstitutive Rolle in der Interaktionsdynamik spielen, die in der Soziologie, die sich stets nur für die Motive der menschlichen Handlungsträger interessiert, fälschlich ausgeblendet werden (Rauer 2017b). Diese handlungstheoretische Wende in der Soziologie hin zur Einbeziehung von materialen Handlungsträgern bildet eine gute Voraussetzung für eine Analyse der Folgen von KI für Handlungstheorien (Rauer 2017a). Denn was bei analogen Schlüsselssystemen vielleicht noch etwas konstruiert als ein determinierender ‚Handlungsträger‘ erscheint, ist bei einer textverfassenden KI mehr als evident auch offen für Kontingenz (Rauer 2012). Die nicht-menschlichen Handlungsträger interagieren mit menschlichen Handlungsträgern und prägen deren Wissen und Entscheidungsbedingungen.

Entscheidend bei dieser soziologischen Konzeptualisierung ist, dass hier KI nicht als intelligibles künstliches Bewusstsein eines Roboters, sondern als Handlungsprogramm im Sinne der ANT beschrieben wird. Nicht der Ontologie der Intelligenz oder KI als Bewusstsein, sondern KI als Programm einer Pragmatik des sozialen Handelns gilt der analytische Blick. Der Begriff der ‚Interobjektivität‘ dient bei Latour als Markierung dieser Blickverschiebung. Allerdings muss hier einschränkend auch erwähnt werden, dass der Terminus Interobjektivität wie in einem Determinismus erscheint – im Sinne von Objektivität und Kausalität. Damit würde Kontingenz im Interaktionsprozess ausgeschlossen werden und der transformaktive Aspekt von dynamisch interaktiver KI gerade wiederum aus dem Blickfeld. Algorithmen können deterministisch programmiert sein, aber in den

neuen Handlungsprogrammen, die ‚auf deep learning‘ Programmen beruhen, sind Kontingenzeffekte ebenso eingelassen (Heinlein 2023).

Mit anderen Schwerpunkten definiert der Techniksoziologe Werner Rammert interagierende Roboter mit Bezug auf Bettina Heintz‘ Geschichte des Computers (Heintz 1993) als „geregelte Problemverarbeitungsmaschinen“ (Rammert, 1995, S. 155). Eine Problemverarbeitungsmaschine ist zunächst an den Problemlösungen orientiert und nicht an der Wesenheit der ausführenden Akteure. Rammert nimmt eine dediziert pragmatisch-theoretische Perspektive auf KI ein. Die Problemlösung von Menschen im Zuge von KI sollte für Rammert als „sitierte Kooperation“ (Rammert 2007, S. 111) bezeichnet werden, nicht als Determination. So heißt es bei Rammert zu Robotern: „Serviceroboter, welche bei einem koordinierten Einsatz im Kanalsystem oder im Katastrophenfall tauchen können, und fußballspielende Roboter, welche, statt allein bis vors Tor zu stürmen die Kunst des Doppelpasses beherrschen, basieren zwar weiterhin auf Algorithmen, aber ihre Verhaltensabläufe lassen sich nicht als determinierte Operationen, sondern angemessener als situierte Kooperationen [...] beschreiben.“ (ebd.)

Die situierte Kooperation markiert eine erste Hinwendung zu einer KI adressierenden Handlungstheorie, die ein gewisses Maß an Kontingenz in Interaktionssituationen mitdenkt und erlaubt, ohne es explizit allerdings so auch zu nennen. Wichtig ist hier die pragmatische Perspektive auf einer im ‚Einsatz‘ sich befindenden KI-basierten Handlungsträgerschaft. In der jeweiligen Interaktion wird mit der KI in Ko-Handlungsträgerschaft ein Problem gelöst, sei es im Katastrophenfall oder im Spiel. Situative Kooperation ist eine jeweils neu bestimmte Kooperation, die zwar in den robotischen Systemen angelegt ist, aber nicht in ihrer tatsächlichen Verwendung vollständig vorab determiniert werden kann. Solche situativen kooperativen Algorithmen finden sich auch in analogen sozialen Situationen (Rauer 2017b). Auf diese Weise ergibt sich zumindest bereits indirekt eine gewisse Kontingenzsensitivität von KI. Hier zeigt sich bei Rammert bereits die Problematik, die später dann bei den sogenannten Lernalgorithmen virulent wird, ohne dass sie auf einer theoretisch abstrakten Ebene ausformuliert wurde. Zunehmend verschleift sich die Differenz zwischen sensitiv-aktorischen robotischen Systemen und menschlichen Akteuren im Bereich situativ kooperativer Problemlösungspraktiken (ebd.). Der Begriff der Interobjektivität suggerierte noch zu viel an Determinismus. Das Konzept der situativen Kooperation weist bereits die Richtung, die eine auf KI antwortende Handlungstheorie einschlagen sollte.

Des Weiteren liegen eine Reihe an Versuchen vor, die Frage nach der Bedeutungskonstruktion von Algorithmen mit Bezug auf ihre ‚Macht‘ (Beer 2009) ‚Kultur‘ (Roberge und Seyfert 2017; Levermann 2018) oder ‚Definitionsmacht‘

(Gillespie 2017) zu analysieren, die hier nur kurz erwähnt werden sollen. Diese Ansätze weisen ebenfalls Grundbedingungen einer Konzeptualisierung auf, die die transformative handlungstheoretische Kontingenz von KI integriert.

Eine Typologie von KI-gestützten Definitions- und Bewertungspraktiken besonders exemplarisch mit Blick auf Internetkommunikation verweist zumindest indirekt auf nicht deterministische Handlungsanteile. Typologien lassen in einem geringen Ausmaß Abweichungen und Kontingenzeffekte dynamisch transaktiver Interaktion zu. Diese gelten dann als Devianz des intendierten algorithmischen Effektes, ohne das typologische Modell selbst infrage zu stellen. Im Internet sind Algorithmen, so fasst es beispielsweise Dominique Cardon zusammen, „[...] die Grundlage aller Werkzeuge, die die Aufmerksamkeit der Internetnutzer kanalisieren sollen“ (Cardon 2017, S. 131). Dabei lassen sich laut Cardon vier „Typen der Informationsberechnung“ (ebd., S. 134) mit dem Ziel der Aufmerksamkeitskanalisierung unterscheiden: Erstens, „Popularität“, zweitens „Autorität“, drittens „Reputation“ und viertens „Prognose“ (ebd.). Der erste Typus befasst sich mit der errechneten Ermittlung der Popularität eines Inhalts ähnlich wie die bereits beschriebenen Trending-Algorithmen. Der zweite Typus beschreibt die Erstellung von Rankings, d. h. autoritären, nicht von den Nutzenden überprüfbar bewertungen. Der dritte Typus erzeugt eine Reputation eines Inhalts qua Messung der ‚likes‘. Und der vierte Typus prognostiziert wahrscheinliche Entwicklungen angesichts des vergangenen Verhaltens der Nutzenden.

Doch diese Algorithmen determinieren nicht die Aufmerksamkeit, sondern lassen Unaufmerksamkeit und Abweichungen zu. Die ersten drei Typen haben die entscheidende Schwäche, nur das jeweilig subjektive Bewerten der Nutzenden zu erfassen, nicht deren tatsächliches Handeln im Netz oder deren wahrscheinliches zukünftiges Handeln. Sie ‚messen‘, was Nutzende meinen, nicht, was sie tatsächlich tun. Daher sind die Messungen erheblichen Kontingenzspielräumen ausgesetzt. Etwas zu liken, heißt nicht, auch danach zu entscheiden oder das Produkt in der analogen Welt dann tatsächlich sich anzuzeigen. Erst die Prognose versucht dieses Problem zu lösen. Die aktuellen Big-Data Plattformen wie ChatGPT oder Googles Bard sind Weiterentwicklungen dieser Praxis: Die KI-basierte, von Cardon so genannte „Prognose“ (ebd., S. 134–135). Um die Manipulation der Aufmerksamkeitsdaten zu minimieren, die in scheinbaren ‚Likes‘ oder künstlich geschaffenen ‚Twitter-Follower‘ (Twitter Bots etc.) suggeriert werden, wird hier versucht, das jeweils aktuelle soziale Handeln mit den Spuren des vergangenen sozialen Handelns im Netz abzugleichen, um eine Voraussage des wahrscheinlichsten Handlungsverlaufes zu erstellen. Aufgrund von Spuren wird ermittelt, wie die Nutzenden zuvor im Netz und anderen digitalen Medien interagiert hatten. Aus dieser ermittelten Vergangenheit lassen sich



wahrscheinliche von unwahrscheinlichen Interaktionsmuster unterscheiden und Kontingenz minimieren. Es sind diese die soziale Situationen zeitlich expandierenden Algorithmen, die vorgeben, „dass man Vertrauen nur in das echte Verhalten von Individuen haben könne und nicht in ihre Selbstverlautbarungen auf sozialen Medienplattformen“ (ebd., S. 141). Die Prognosealgorithmen berechnen also nicht nur, was Nutzenden subjektiv gefällt und welchen Inhalten sie folgen wollen, sondern sie beziehen auch das in die Berechnung mit ein, was diese Nutzenden ansonsten in der Vergangenheit im Netz interaktiv tatsächlich getan und nicht getan haben. Die Prognosealgorithmen lernen etwas über die Vergangenheit der Nutzenden und verwenden dieses Wissen zur Plausibilitätskalkulation von deren aktuellem und zukünftigem sozialen Handeln.

Neuere theoretische Ansätze beschäftigen sich inzwischen auch mit dem reinen Akteursstatus von Algorithmen im Sinne einer Sozialtheorie. So versucht Armin Nassehi KI unter dem Titel der „Lernenden Technik“ (Nassehi 2019, S. 228–262) in die Sozialtheorie zu integrieren. Zwar gibt er mit Latour zu, dass die KI-basierten Algorithmen wie „Akteure“ (ebd., S. 248) handeln würden, dennoch unterschieden sie sich von menschlichen Akteuren bezüglich der Fähigkeit zum Erleben und Handeln. Sie handeln, aber sie ‚erleben‘ nicht wie menschliche Akteure. Daher ergibt sich bei Nassehi ein klares Primat der Unterscheidung von Technik und Mensch im Sinne einer Aufrechterhaltung dieser binären theoretischen Figur. Die transaktiven Handlungsketten, in denen jeweils die Interaktion zu einer Transaktion und, theoretisch einheitlich gesprochen, zu einer Transformation führt, gerät auch mit diesem neuen jüngsten Ansatz der Sozialtheorie allerdings leider wieder aus dem Blick und ist nicht anschlussfähig. Es zeigt sich, dass auch begrifflich ein Bedarf an einer Markierung der handlungstheoretischen Folgen von KI besteht.

---

## 4 Soziale Transformation – eine Konzeptualisierung

In diesem Unterkapitel soll der konzeptuelle Vorschlag, statt von Interaktion im Zusammenhang mit KI von Transformation zu sprechen, ausführlicher inhaltlich anhand eines Fallbeispiels erläutert werden. Wie gezeigt wurde, hatte Latour angesichts der Integration von nicht-menschlichen Handlungsträgerschaften in die soziologische Analyse vorgeschlagen, statt von Intersubjektivität von ‚Interobjektivität‘ zu sprechen. Barad schlug den Begriff ‚Intraaktivität‘ vor und Rammert das Konzept einer intraaktiven, situativen Kooperation. Alle drei Vorschläge weisen in die Richtung, die auch das Konzept der ‚Transformation‘ einschlägt. Anders als der Interobjektivitätsbegriff wird allerdings weniger das

Objekt selbst in den Vordergrund gerückt, sondern die permanente Wandlung und Veränderung im Zuge der Interaktionsprozesse. Die Vorsilbe des ‚Trans‘ in dem Wort Transformaktion verweist auf die Dynamik auf der Ebene der Handlungsträger als auch auf der Ebene der Handlungsobjekte. Weniger die Subjekte und die Objekte als statische Elemente sind bei KI anzutreffen, sondern Dynamiken, Reflexivitäten und Optimierungen. Auch Barads Begriff der Intraaktivität passt nur bedingt auf diese Handlungssituation. Intraaktivität nimmt in ihrem Verständnis stärker die Inversion von Beobachtung und Beobachteten in den Blick als die transformierende Dynamik. Auch dies findet sich im Kontext von KI-basierten Interaktionssituationen, ist jedoch nicht zentral. Das Umschlagen und die wechselseitige Abhängigkeit von Beobachtung und Beobachteten erfassen nicht erschöpfend die transformierenden Handlungspotentiale der aktuellen KI-Plattformen, sondern werden hier eher als Ununterscheidbarkeit von Handlungsträgerschaften systematisiert. Schließlich bildet die situative Kooperation von Rammert wichtige Einblicke in KI-basierte Pragmatik und Kontingenzsensitivität (Heinlein 2023). Doch situative Kooperation ist nur ein Ziel unter vielen anderen einer KI-basierter Interaktion. Der Begriff bezieht die Bedeutung von mit Big Data trainierten Lernalgorithmen, die situative Kooperation übersteigen, nicht mit ein. Zudem fehlt dabei die Berücksichtigung der die Möglichkeit permanenter Rückkopplung von KI-basierter Interaktionen.

Was ist mit dem Konzept der *Transformaktion* also konkret gemeint? Zunächst soll dies wieder anhand eines Fallbeispiels erläutert werden. Dieses Beispiel stammt von dem sogenannten ‚Huggingface Transformer‘. Dies ist eine Plattform, die das ‚natural language processing‘ Lernprogramm Huggingface zur Nutzung anbietet. Diese Plattform existierte zeitlich bereits vor ChatGPT und Googles Bard, arbeitete jedoch bereits mit den ähnlichen KI-basierten Sprachmodellen.

Der folgende zitierte Textabschnitt bezeugt eine Interaktion von mir als Nutzer und dem Huggingface Transformer. Dies wurde dazu entwickelt, um Amateuren eine Interaktion mit den aktuellen NLP Apps zu ermöglichen, bevor ChatGPT und Googles Bard diese Anforderungen bedienen konnten. Der Vorteil dieser Plattform ist, dass die kontingenten Interaktionsschritte im Sinnproduktionsprozess noch sichtbarer sind als bei Googles Bard. Die nicht-fett gedruckten Worte stammen von mir, die fettgedruckten Teilsätze stammen von einem dreifachen Vorschlagsmenü von Huggingface:

“This paper aims to see AI through **its natural language features**. Strong AI is not the topic **of this paper**. Strong AI has a tradition in **research, where experiments are designed to test on real life**. Theoretical approaches to strong AI **are based on the assumption that humans will make decisions** based on **knowledge and emotion**.

Classical debates stem from **using knowledge- and emotion-based models** on the lack of human conciseness. Weak AI instead **uses the state of the art neural models**. Machine learning **approaches focus on the ability of a machine to learn from the data and then apply its knowledge to improve the machine model to create a better algorithm**. In these data processes however, humans are **still responsible for selecting the most probable target for an algorithm**. Humans are still in the loop to **get the best result by creating a better algorithm**. Two and three is **the essence of weak AI.**<sup>7</sup>

Als erstes stellt sich hier die einfache pragmatische Frage der Texturheberschaft in Gestalt der Fußnote 5. Was soll ich hier angeben? Wer ist der Autor dieses Textes? Ein pragmatischer Vorschlag könnte lauten: „Zwei Quellen: Nicht-Fett markierte Textteile: Valentin Rauer. Fett markierte Textteile: Huggingface transformer. Auswahl der Huggingface Satzteile aus jeweils drei Optionen: Valentin Rauer. Entscheidung über eine direkte/indirekte Anschlusssatzbildung direkt an Huggingface Satzteile: Valentin Rauer. <https://transformer.huggingface.co/doc/arkiv-nlp> (Interaktion am 15.11.2021, 11:00–11:15 Uhr).“

Beim Schreiben mit dem KI-Transformer haben sich folgende Interaktionsmuster eingestellt: Den Sinn meines Satzes habe ich in der ersten Satzhälfte selbst formuliert (This paper aims to see AI through ...). Anschließend habe ich mir mit der Tabulatortaste einen Autocomplete Satzteil anbieten lassen. Bei dieser Funktion ist, anders als bei Googles Bard, noch eine stärkere Interaktivität und Optionalität eingestellt. Selbstverständlich könnte diese kleinteiligere Optionalität auch mit Bard durch Modifikationen in der Programmführung realisiert werden, darum geht es hier jedoch nicht. Wichtig ist stattdessen, die jeweilige interaktiv geformte Entscheidungsmöglichkeit bei dem Prozessieren des Texts durch die KI und die menschlichen Handlungsträger in Kooperation sichtbar zu machen als bei dem Eingangsbeispiel von Bard. Auch ist die Kontingenz der jeweiligen Textentscheidung in einer solchen optionsbasierten Einstellung transparenter.

Die Entscheidung für eine der drei Varianten ermöglicht Auswahl und realisiert damit eine menschlich-maschinelle Kontingenzsituation. Der Algorithmus determiniert nicht den Text, sondern interagiert mit den anfragenden menschlichen Handlungsträgern. Doch die Kontingenz ist auch auf drei Varianten reduziert. Von Determination im Zusammenhang von KI-basierten Interaktion zu sprechen, geht an den realen Bedingungen solcher Interaktionspraktiken vorbei.

---

<sup>7</sup> Quelle: Nicht-Fett markierte Textteile: Valentin Rauer. Fett markierte Textteile: Huggingface transformer. Auswahl der Huggingface Satzteile aus jeweils drei Optionen: Valentin Rauer. Entscheidung über eine direkte/indirekte Anschlusssatzbildung direkt an Huggingface Satzteile: Valentin Rauer. <https://transformer.huggingface.co/doc/arkiv-nlp> (Interaktion am 15.11.2021, 11:00–11:15 Uhr).

Algorithms are defined in computation as

“: ... A

: follows.

: the processing of rules by an”

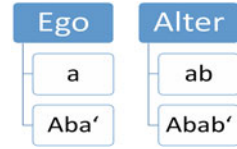
**Abb. 1** Optionsmöglichkeiten. (Quelle: Quelle: Erste Zeile: Valentin Rauer. Zweite bis vierte Zeile: Huggingface transformer. <https://transformer.huggingface.co/doc/arxiv-nlp> (Interaktion am 15.11.2021, 11:00 – 11:15 Uhr).)

Stattdessen findet sich eine offene, aber kontingenzzreduzierte Semi-Determination in Gestalt dreier Optionsmöglichkeiten (Abb. 1):

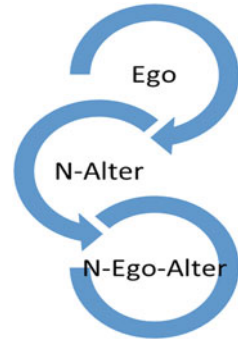
Die Auswahl aus drei Angeboten habe also ich als menschlicher Handlungsträger getroffen. Der Transformer ermittelt diese Satzteile nicht über Sinnverstehen meiner Intention (Schreiben über ‚weak AI‘), sondern über die ermittelten Wahrscheinlichkeiten ähnlicher Satzanfänge in einem Milliarden Dokumente umfassenden Daten- bzw. Textkorpus. Der Sinn der jeweiligen Autor\*innen dieses Korpus fließt stochastisch in die drei Vorschläge mit ein. Ich als Autor interagiere damit mit aggregierten und probabilistisch gewichteten Satzfolgen anderer menschlicher und nicht-menschlicher Akteure. Denn wenn ich eine Satzfolge akzeptiere, dann wird diese Akzeptanz zur Grundlage weiterer Vorschläge etc.

Den Unterschied zwischen einer sozialen Interaktionssituation und einer menschlich-nicht-menschlichen Transformaktion sind in Abb. 2 und 3. schematisch dargestellt. Die Interaktionssituation in Abb. 2 zeigt die Interaktionspartner Ego und Alter. Zwischen diesen beiden menschlichen Akteuren ist eine klar definierte Grenze. Wenn Ego die Aussage a Alter mitteilt, dann reagiert Alter mit einer Antwort ab. Auf diese Antwort reagiert wiederum Ego mit der Aussage Aba‘ und Alter mit Abab‘ usf. Entscheidend ist, dass Ego und Alter als Adressaten und Subjekte der kommunikativen Handlungen gelten und sich im Zuge der Interaktionssituation nicht verändern. Die Handlungsträger unterliegen in ihrer Identität und Grenze keiner Transformation und keiner Kontingenz. Kontingenz ist lediglich das, was sie jeweils als Aussage interpretieren. Wenn Ego a als Aussage formuliert, so ist in der Interpretation dieses a durch Alter ein Kontingenzzfreiraum eingelassen. Alter kann sich nicht absolut gewiss sein, dass Ego tatsächlich dieses a gemeint hat. Alter antwortet daher mit ab, d. h. sein b reagiert auf ein vermutetes a von Alter. Die Antwort von Ego auf Alter ist nun mit einer

**Abb. 2** Soziale Interaktionssituation: Ego und Alter. (Quelle: Eigene Darstellung)



**Abb. 3** Menschlich-nicht-menschliche Transformaktion. (Quelle: Eigene Darstellung)



zweifachen Kontingenz konfrontiert. Die Interpretation von a durch Alter ist für Ego kontingent als auch das b, das auf a antwortet usw. Diese kontingenzbedingte Interaktionsproblematik wird auch in Nachfolge von Talcott Parsons und Niklas Luhmann als Problem der doppelten Kontingenz beschrieben (Kärtner 2015).

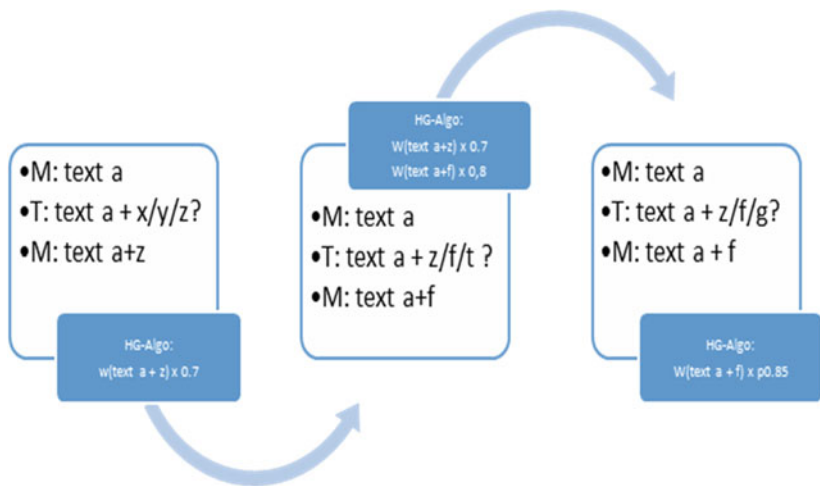
Klassisch soziales Handeln orientiert sich also an einem probabilistisch aggregierten und gewichteten kollektivierten sozialen Sinn. In dieser ersten Fassung wäre die Folge von KI also eine Kollektivierung sozialer Sinnprodukte und deren ‚Interagierbarkeit‘.

Doch das trifft nur einen kleinen Bereich solcher sinngenerierender KI. So muss zweitens die Interaktionssituation dieses Textes selbst als Lernprozess berücksichtigt werden. Dieser kleine Text ist Teil des sogenannten ‚humans in the loop‘ dieses KI-Systems. Jeder der Transformer-Satzteile wird von dem fragstellenden Nutzenden aus drei Alternativen selbst ausgewählt. Das System merkt sich die Auswahl und korrigiert in seinen neuronalen Netzwerkknoten eine Gewichtungsfunktion leicht. Diese Transformaktion ist in Abb. 3 dargestellt.

Abb. 4 zeigt eine Interaktionsfolge von einem menschlichen Handlungsträger, den Userinnen M, und der KI-Plattform, in diesem Fall dem Huggingface Transformer T. Dabei könnte es sich auch um ChatGPT oder Googles Bard handeln. In diesem Fall schlägt M eine Wortfolge a vor. Der Transformer reagiert mit drei verschiedenen Wortfolgevorschlägen: a + x/y/z. Die Wahl einer

dieser drei Vorschläge durch M ist teilweise kontingent und orientiert sich teilweise an den sozialen Sinnstrukturen oder den sozialen Rahmungen, die M mit in die Interaktionssituation bringt. Um den Text zu vervollständigen, muss M eine der drei Varianten wählen. Der Transformer ‚lernt‘ aufgrund dieser Wahl, welche der drei Varianten die erfolgreichste war und ändert im Zuge dieser Interaktion seine Prädiktions- und Gewichtungsmuster. In dem Schema wird dies metaphorisch mit der Symbolfolge  $p(\text{text } a + z) \times 0.7$  symbolisiert. Damit soll ausgedrückt werden, dass die Interaktionsfolge unmittelbar bewertet und für die nächste Interaktionssituation dem Algorithmus bereitgestellt werden kann.

Der mittlere Block von Abb. 4 zeigt nun eine solche Folgeinteraktion. Eine menschlicher Handlungsträger M bietet wieder den Text a an. Diesmal bietet der Transformer T als Antwort die z-Variante bereits an erster Stelle der Liste an und führt mit f und t z-ähnlichere Wortvarianten in die Vorschlagsliste ein. M hat nun andere Wahloptionen. Gleichzeitig ist diese Wahl von der KI vorge schlagen und ausgewählt sowie an der kontingenten Wahl des Vorgängers von M aus dem ersten linken Block orientiert. Die zur Wahl stehenden Textabschnitte stellt die KI, doch die jeweiligen Varianten sind durch die kontingent sich entscheidenden ehemaligen menschlichen Handlungsträger ko-beeinflusst. Weder die KI noch die menschlichen Handlungsträger interagieren als unabhängige Einheiten in einer autonomen sozialen Situation. Die Situation ist gebunden an



**Abb. 4** Transformaktion (M = Mensch, T = Transformer). (Quelle: Eigene Darstellung)

vorhergehende Situationen. Der Transformer T als Interaktionspartner wandelt sich mit jedem Interaktionsschritt dynamisch. Die KI determiniert nicht ein für alle Mal die textlichen Angebote, sondern entwickelt aufgrund ihrer permanenten Interaktionserfahrung an der Plattform neue Textvarianten.

Die menschliche Interaktionspartnerin M kann diese Auswahl, die sie einmal bekommen hat, nicht reproduzieren. Das System ist im permanenten Wandel und stark dynamisiert. Daher muss auch nicht nur der Tag, sondern auch die Minute der Schreibinteraktion als Quelle angegeben werden. M interagiert bei einem Reproduktionsversuch mit ihren eigenen sozialen Handlungen selbst, zumindest in ihrem stochastisch gewichteten Anteil. Die Gesamtkette ist keine Interaktion zwischen zwei unabhängigen Akteuren, sondern eine *Transformaktion*, in der ein Transformer sich selbst und die jeweiligen Interaktanten transformiert.

Klassische Interaktionssituationen bestehen aus mindestens zwei menschlichen Akteuren und einer zeitlich und räumlich begrenzten Situation. Transformaktions-situationen bestehen aus mindestens zwei menschlichen und nicht-menschlichen Akteuren und räumlich und zeitlich entgrenzten Situationen.

---

## 5 **Ausblick: Asymmetrie als potentielles Risiko von Transformaktivitäten**

In diesem Ausblick sollen kurz potentielle Überlegungen zu Risiken der sinn-generierenden KI-Transformaktionen skizziert und angedacht werden. Eine Hauptquelle für potentielle Risiken ist in der opaken räumlich-zeitlich entgrenzten Transformaktionssituation begründet. Die interagierenden menschlichen Akteure sind nicht in der Lage, selbst die Grenzen der sozialen Situation wahrzunehmen oder mitzubestimmen, wenn sie mit sinn-generierenden KI-Plattformen interagieren. Das, was in analogen Interaktionssituationen den Interagierenden zur Aushandlung über die Definition einer sozialen Situation zumindest epistemisch allen Beteiligten offensteht (Rauer 2017b), ist bei Transformaktionen infrastrukturell bedingt den Plattformen vorbehalten. Die einzige kontingente Mitbestimmung besteht in der Exit-Option aus der Transformationssituation ganz. Zu fragen wäre in Zukunft, welche Mit-Bestimmungen denkbar seien und wie die Grenzbestimmung sozialer Transformaktionssituationen symmetrischer verteilt werden könnte.

Zu denken wäre klassisch zunächst an Setting-Einstellungen. Doch diese Bestimmungen haben nicht den gleichen Gewissheitsgrad wie indexikale

räumlich-zeitliche Situationswahrnehmungen. Die Settings sind stets opaker als die räumlich-zeitliche Anwesenheit interagierender Akteure. Bei KI-Transformaktionen schlägt nicht eine vorab definierte und determinierte KI etwas vor, was dann reproduziert wird. Vielmehr werden die Suchanfragen und Vorschläge in einer permanenten prozessualen ‚Interaktivität‘ zwischen menschlichen Nutzerinnen und KI-Gewichtungen generiert. Das Risiko dabei ist, dass die Grenzen dieser Interaktivität immer intransparenter werden. Beispielsweise stellt sich die Frage, ob oder inwieweit die Nutzungsgeschichte des Plattform-Nutzenden einbezogen wurde, wenn er oder sie einen Text über die KI erstellen lässt. Diese Fragen sind stets prekär, sie können, müssen aber nicht transparent sein. Damit geht ein Verlust an der Definitionsmacht sozialer Situationen von Sinnproduktionen einher. Die Definitionsmacht sozialer Situationen wird von den individuell beteiligten Transformationspartner\*innen auf die KI-Plattformen übertragen. Die situative Grenzbestimmungen und damit die Bedingung der Möglichkeit zur kooperativen sinnorientierten Definition einer sozialen Situation werden asymmetrisch aufseiten der KI-Plattformen transferiert.

Das Problem ist, anders formuliert, dass ein Teil der Bedingungen, über den Sinn der Transformaktionssituation zu entscheiden, asymmetrisch verteilt ist. Nicht alle Handlungsträgerschaften und Transformaktionssituationen lassen sich von menschlichen Akteuren als solche auch lokalisieren und identifizieren. Denn selbst die eigenen Textprodukte können in die Sinnproduktion einer Transformaktion integriert sein, müssen es aber nicht. Es ist der KI vorbehalten darüber zu entscheiden, welche Datensätze im Falle einer transaktiven Sinnproduktion einbezogen werden und welche nicht. Damit bestimmt die KI einseitig über die Grenzen einer sinnorientierten Transformaktionssituation. Die zur Orientierung herangezogenen relevanten Deutungsgrundlagen sind nicht symmetrisch für alle Beteiligten der Situation verfügbar. An den situativ zur Definition notwendigen Grenzziehungen der Transformationsaktion sind die Nutzenden nicht selbst beteiligt. Damit ist das Handlungsvermögen in Transformaktionssituationen asymmetrisch verteilt, zulasten der Nutzenden.

Potentiell wären begleitende KI-gestützte Beobachtungskonzepte zu entwerfen, die nicht a priori verbieten und vorschreiben, wie es in Zeiten der analogen Technologien noch möglich war, sondern die selbst reflexiv auf die interoperativ gelernten, digitalen KI-Systeme *transformativ* reagieren. Es wird wohl in Zukunft darum gehen, KI-basierte Strategien zu entwickeln, die zu starke Asymmetrien in Transformaktionssituationen ausgleichen könnten. Zu denken wäre an eine KI, die den zum Training eingesetzten Datensatz und dessen Reichweite unabhängig von den Plattformen prüft und kontrolliert. Möglicherweise wären auch staatliche



KI-Plattformen denkbar, in denen transparenter über die Datenqualität und Reichweite informiert werden würde als es bei den aktuellen privaten Plattformen der Fall ist. Die Herausforderung ist dabei, die internen Grenzbedingungen von Transformaktionen aufzuspüren, öffentlich zu benennen und mit einer transparenten KI die Asymmetrien abzubauen.

---

## Literatur

- Barad, Karen. 2012. *Agentieller Realismus*. Berlin: Suhrkamp.
- Barwise, P., und Watkins, L. 2018. The evolution of digital dominance: How and why we got to GAFa. In *Digital dominance. The power of Google, Amazon, Facebook, and Apple*, hrsg. M. Moore & D. Tambini, 21–49. Oxford: Oxford University Press.
- Bauer, Susanne, Thorsten Heinemann und Thomas Lemke (Hg.). 2017. *Science and Technology Studies. Klassische Positionen und aktuelle Perspektiven*. Berlin: Suhrkamp.
- Beck, Susanne. 2016. The problem of ascribing legal responsibility in the case of robotics. In *AI & Society*, 31 (4), 473–481.
- Beckers, Anna, und Gunther Teubner. 2023. Die digitale Verantwortungslücke: Vorschläge zur Haftung für algorithmisches Fehlverhalten. In *Künstliche Intelligenz, Mensch und Gesellschaft*, hrsg. Michael Heinlein und Nobert Huchler. Wiesbaden: Springer VS.
- Beer, David. 2009. Power through the algorithm? Participatory web cultures and the technological unconscious. In *New Media & Society*, 11 (6), S. 985–1002.
- Belliger, Andréa und David J. Krieger (Hg.). 2006. *ANThology. Ein einführendes Handbuch zur Akteur-Netzwerk-Theorie*. Bielefeld: transcript.
- Cardon, Dominique. 2017. Den Algorithmus dekonstruieren. Vier Typen digitaler Informationsverarbeitung. In *Algorithuskulturen. Über die rechnerische Konstruktion der Wirklichkeit*, hrsg. Robert Seyfert und Jonathan Roberge, 131–150. Bielefeld: transcript.
- Dimbath, Oliver und Heinlein, Michael. 2014. *Die Sozialität des Erinnerns. Beiträge zur Arbeit an einer Theorie des sozialen Gedächtnisses*. Wiesbaden: Springer.
- Davis, G. F. 2016. What might replace the modern corporation? Uberization and the web page enterprise. *Seattle University Law Review*, 39: 501–515.
- Dolata, U. 2013. *The transformative capacity of new technologies. A theory of sociotechnical change*. London: Routledge.
- Dolata, U., und Schrape, J.-F. 2018. *Collectivity and power on the Internet. A sociological perspective*. Cham: Springer.
- Dolata, U. 2019. Plattform-Regulierung. Koordination von Märkten und Kuratierung von Sozialität im Internet. *Berlin: J Soziol* 29: 179–206
- Dolata, U. und Schrape, J.F. 2022. Internet, Big Data und digitale Plattformen: Politische Ökonomie – Kommunikation – Regulierung. *Köln Z Soziol* 74 (Suppl 1), 1–9.
- Eggert, Michael und Kerpen, Daniel. 2015. Trust by design? Vertrauen als zentrale Ressource für die Cloud. In *Routinen der Krise – Krise der Routinen*, hrsg. Stephan Lessenich, 1170.1783. Verhandlungen des 37. Kongresses der Deutschen Gesellschaft für Soziologie in Trier 2014.

- Gillespie, T. 2016. Regulation of and by Platforms. In *The SAGE Handbook of social media*, hrsg. J. Burgess, T. Poell & A. Marwick, 254–278. Los Angeles: Sage.
- Gillespie, Tarleton. 2017. #trendingistrending. In *Algorithmuskulturen. Über die rechnerische Konstruktion der Wirklichkeit*, hrsg. Robert Seyfert und Jonathan Roberge. 75–106. Bielefeld: transcript.
- Goffman, Erving. 1974. *Frame analysis: An essay on the organization of experience*. New York: Harper & Row.
- Gutman, David. 2009. *From Transformation to Transformaction. Methods and Practices*. London: Routledge.
- Heinlein, Michael. 2023. Künstliche Intelligenz als kontingenzerzeugende Technologie: Eine praxistheoretische Perspektive. In *Künstliche Intelligenz, Mensch und Gesellschaft*, hrsg. Michael Heinlein und Nobert Huchler. Wiesbaden: Springer VS.
- Heintz, Bettina. 1993. *Die Herrschaft der Regel. Zur Grundlagengeschichte des Computers*. Frankfurt, New York: Campus.
- Hutchins, Edwin. 1995. *Cognition in the Wild*. Cambridge, Mass. U. a: MIT Press.
- Kärtner, Jurit. 2015. Das Problem der doppelten Kontingenz als Ausgangsproblem des Sozialen und der soziologischen Theorie. Vorschlag zu einer Systematisierung der soziologischen Systemtheorie Niklas Luhmanns. In *Zeitschrift für theoretische Soziologie* 4/1: 60–88.
- Kenney, M., und Zysman, J. 2016. The rise of the platform economy. *Issues in Science and Technology*, 32(3): 61–69.
- Knoll, Alois und Christaller, Thomas. 2003. *Robotik: Autonome Agenten. Künstliche Intelligenz. Sensorik. Embodiment. Maschinelles Lernen. Serviceroboter. Roboter in der Medizin. Navigationssysteme. Neuronale Netze. RoboCup. Architekturen*. Fischer: Frankfurt am Main.
- Latour, Bruno. 1991. *We Have Never Been Modern*. Cambridge: Harvard University Press.
- Latour, Bruno, Philippe Mauguin und Geneviève Teil. 1992. A Note on Socio-Technical Graphs. In: *Social Studies of Science*, 22 (1): 33–57.
- Latour, Bruno. 1996. *Der Berliner Schlüssel. Erkundungen eines Liebhabers der Wissenschaften*. Berlin: Akademie-Verlag.
- Latour, Bruno. 2001. Eine Soziologie ohne Objekt? Anmerkungen zur Interobjektivität. In *Berliner Journal für Soziologie*, 11 (2): 237–252.
- Latour, Bruno. 2007. *Eine neue Soziologie für eine neue Gesellschaft. Einführung in die Akteur-Netzwerk-Theorie*. Frankfurt am Main: Suhrkamp.
- Laux, Henning. 2011. Latours Akteure. Ein Beitrag zur Neuvermessung der Handlungstheorie. In *Akteur, Individuum, Subjekt. Fragen zu ‚Personalität‘ und ‚Sozialität‘*, hrsg. Nico Lüdtke et al. 275–300. Wiesbaden: VS, Verlag für Sozialwissenschaften.
- Law, John und John Hassard (Hg.). 1999. *Actor Network Theory and after*. Malden, CA: Blackwell Publishers.
- Levermann, Thomas. 2018. Wie Algorithmen eine Kultur der Digitalität konstituieren: Über die kulturelle Wirkmacht automatisierter Handlungsanweisungen in der Infosphäre. In *Journal für kooperative Kommunikation*, 7 (2): 31–42.
- Luckmann, Thomas. 1992. *Theorie des sozialen Handelns*. Berlin: Walter de Gruyter.
- Mau, Steffen. 2018. *Das metrische Wir. Über die Quantifizierung des Sozialen*. Berlin, Suhrkamp.

- Müller, Rainer, Vette, Matthias und Geenen, Aaron. 2017. Skill-based Dynamic Task Allocation in Human-Robot-Cooperation with the Example of Welding Application, *Procedia Manufacturing*, Volume 11, Pp 13–21.
- Nassehi, Armin. 2019. *Muster: Theorie der digitalen Gesellschaft*. München: C.H. Beck.
- Nieborg, D. B., und Poell, T. 2018. The platformization of cultural production: Theorizing the contingent cultural commodity. *New Media & Society*, 20 (11): 4275–4292.
- Pinar Saygin, A., Cicekli, I. und Akman, V. 2000. Turing Test: 50 Years Later. *Minds and Machines* 10: 463–518.
- Rammert, Werner. 1995. *Soziologie und Künstliche Intelligenz: Produkte und Probleme einer Hochtechnologie*. Frankfurt am Main: Campus.
- Rammert, Werner und Ingo Schulz-Schaeffer (Hg.). 2002. Können Maschinen handeln? Soziologische Beiträge zum Verhältnis von Mensch und Technik. Tagung ‚Handlungsträgerschaft von Technik? Oder: Wieviel Eigenmächtigkeit und Interaktionsfähigkeit messen wir technischen Artefakten zu?‘. Frankfurt am Main: Campus.
- Rammert, Werner und Cornelius Schubert (Hg.). 2006. *Technografie. Zur Mikrosoziologie der Technik*. Frankfurt am Main: Campus.
- Rammert, Werner. 2007. *Technik Handeln Wissen. Zu einer pragmatistischen Technik- und Sozialtheorie*. Wiesbaden: VS Verlag für Sozialwissenschaften.
- Rammert, Werner. 2012. Distributed Agency and Advanced Technology. Or: how to Analyse Constellations of Collective Inter-Active Agency. In *Agency without Actors? New Approaches to Collective Action*, hrsg. Jan-Hendrik Passoth, Birgit Peucker und Michael Schillmeier, 89–112. London: Routledge.
- Rauer, Valentin. 2012. Interobjektivität. Sicherheitskultur aus Sicht der Akteur-Netzwerk-Theorie. In *Sicherheitskultur. Soziale und politische Praktiken der Gefahrenabwehr*, hrsg. Christopher Daase, Philipp Offermann und Valentin Rauer, 69–93. Frankfurt am Main: Campus.
- Rauer, Valentin. 2017a. Distribuierte Handlungsträgerschaft. Verantwortungsdiffusion als Problem der Digitalisierung sozialen Handelns. In *Politische Vierteljahresschrift*, Sonderheft 52, 436–453.
- Rauer, Valentin. 2017b. Drohnen: zur Materialisierung von Algorithmen. In *Algorithmenkulturen. Über die rechnerische Konstruktion der Wirklichkeit*, hrsg. Jonathan Roberge und Robert Seyfert, 189–214. Bielefeld: transcript.
- Rauer, Valentin. 2014. *Aufteilung des Sozialen*. Eine Theorie algorithmischen Handelns. Monographie im Begutachtungsprozess.
- Roberge, Jonathan und Robert Seyfert. 2017. Was sind Algorithmenkulturen? In *Algorithmenkulturen. Über die rechnerische Konstruktion der Wirklichkeit*, hrsg. Robert Seyfert und Jonathan Roberge, 7–40. Bielefeld: transcript.
- Rogalla, Irmhild. 2023. Maschinen lernen nicht! – „Machine learning“-Algorithmen entzaubert. In *Künstliche Intelligenz, Mensch und Gesellschaft*, hrsg. Michael Heinlein und Nobert Huchler. Wiesbaden: Springer VS.
- Schulz-Schaeffer, Ingo. 2006. Who Is the Actor and Whose Goals Will Be Pursued? Rethinking Some Concepts of Actor Network Theory. In *Prenatal Testing: Individual Decision or Distributed Action?*, hrsg. Bernhard Wieser, Sandra Karner und Wilhelm Berger, 131–158. München/Wien: Profil.
- Sebald, Gerd. 2014. *Generalisierung und Sinn. Überlegung zur Formung sozialer Gedächtnisse des Sozialen*. Konstanz: UVK.

- 
- Teubner, Gunther. 2007. Elektronische Agenten und große Menschenaffen: Zur Ausweitung des Akteursstatus in Recht und Politik. In *Interdisziplinäre Wege in der juristischen Grundlagenforschung*, hrsg. Paolo Becchi, 1–29. Zürich, Basel, Genf: Schulthess.
- Turing, Alan .1950. *Computing Machinery and Intelligence*, *Mind*, LIX (236), 433–460.
- Van Dijck, J. 2013. *The culture of connectivity. A critical history of social media*. Oxford: Oxford University Press.
- Weber, Max. 1984. *Soziologische Grundbegriffe*. Tübingen: Mohr.



# Künstliche Intelligenz: Eine Methode für alles? Sozialwissenschaftliche Methodologie der KI-Forschung, ihre Herausforderungen und Möglichkeiten

Peter Kahlert, Maryam Tatari, Suzette Kahlert, Silvan Pollozek, Johan Buchholz, Benedict Lang und Jan-Hendrik Passoth

## Zusammenfassung

Künstliche Intelligenz (KI) ist als Mittel oder Herausforderung der Forschung methodisch relevant. Unser Beitrag untersucht die Bandbreite methodischer KI Forschung über und mit KI aus sozialwissenschaftlicher Perspektive. Aus unseren digitalen und Grounded Theory Methoden und Kartierungen, ergibt

P. Kahlert (✉) · S. Kahlert · S. Pollozek · J. Buchholz · B. Lang · J.-H. Passoth  
Europa Universität Viadrina, Frankfurt (Oder), Deutschland  
E-Mail: [kahlert@europa-uni.de](mailto:kahlert@europa-uni.de)

S. Kahlert  
E-Mail: [skahlert@europa-uni.de](mailto:skahlert@europa-uni.de)

S. Pollozek  
E-Mail: [pollozek@europa-uni.de](mailto:pollozek@europa-uni.de)

J. Buchholz  
E-Mail: [johan.buchholz@tum.de](mailto:johan.buchholz@tum.de)

B. Lang  
E-Mail: [lang@europa-uni.de](mailto:lang@europa-uni.de)

J.-H. Passoth  
E-Mail: [passoth@europa-uni.de](mailto:passoth@europa-uni.de)

M. Tatari  
Technische Universität München, München, Deutschland  
E-Mail: [tatari@europa-uni.de](mailto:tatari@europa-uni.de)

sich Einsicht in die disziplinäre und wissenschaftspolitische Bandbreite der KI-Forschung, sowie ihr verbindendes Potenzial zwischen Disziplinen, Traditionen, und Themen. Ausgehend dieser diskursiv-reflexiven und kollektiv-bildenden Eigenschaft, argumentieren wir für eine Differenzierung von KI als Gegenstand und Mittel. Es ist genau zu unterscheiden, um welche Art, Form und Konzept von KI es sich handelt: Computational brute force, Expert system, Neural net, oder Dispositiv und Imaginäres – und inwiefern dabei auch konzeptionell idente Technologien vergleichbar sind. In Konsequenz gehen wir intensiv auf unsere eigene Methode, insbesondere die verwendeten Mapping-Algorithmen ein. Forschung und Lehre müssen sich für diese Unterschiede sensibilisieren, um KI-Technologien sinnvoll nutzen und untersuchen zu können.

---

**Schlüsselwörter**

Methodologie • Grounded Theory • Mapping KI • Sozialwissenschaft

---

## **1 Intro: Ist Künstliche Intelligenz ein Forschungsgegenstand?**

Wie werden die sozialwissenschaftliche Forschung und ihre Methoden durch die fortschreitende Entwicklung und Implementierung von Systemen und Anwendungen der künstlichen Intelligenz (KI) herausgefordert? Was können die Methoden der KI-Forschung beitragen und was ist oder könnte ihr Anteil an dem Problem sein? Diese Fragen bildeten den Ausgangspunkt für unsere Forschung. Die folgenden Ausführungen zu unseren Bemühungen und Erkenntnissen müssen jedoch ihren Ursprung in eben dieser Forschungsfrage negieren.

Wie ein Toolkit installieren diese Forschungsfragen ihre ontologischen und metaphysischen Annahmen, um mit jeder Instanziierung der KI-Forschungsmethodik geladen zu werden. Sie reproduzieren eher gesellschaftliche Vorstellung von KI und entsprechende Erwartung im sozialen Handeln als einen realen Forschungsgegenstand. Wie bei solchen Vorstellungen üblich, wird durch Kommunikation und Denken ein imaginäres Objekt der Vollständigkeit konstruiert und suggeriert, das den Mangel an gewünschter Einheit kompensiert. Im Falle der KI handelt es sich eher um eine politische und wirtschaftliche Institution zur Sammlung, Vermittlung und Fokussierung von Interessen, als um ein Forschungsobjekt. Als Forschungsgegenstand verliert sich die KI in unzähligen konkreten Anwendungsfällen, Problemen oder Diskursen. In diesem Sinne hat das Center

on Privacy & Technology an der Georgetown Law erklärt, dass Begriffe wie “artificial intelligence,” ‘AI’ and ‘machine learning,’ [as they] placeholder everywhere for the scrupulous descriptions that would make the technologies they refer to transparent for the average person” (Tucker 2022).

Es gibt viele methodische Ideen, Konzepte und Verfahren innerhalb und außerhalb der Sozialwissenschaften, die sich mit dem Verständnis oder der Erklärung von KI befassen. Dazu gehören Ansätze, Taktiken und Methoden aus verschiedenen Disziplinen wie Datenwissenschaft, Mensch-Computer-Interaktion, Psychologie und sogar Betriebswirtschaft, sowie auch Aktivismus und Journalismus, um nur einige zu nennen. In ähnlichem Maße begegnet man der KI in vielfältigen Formen, Bildern, Vorstellungen und ‚Gespenstern‘: Künstliche Intelligenz als ein bestimmter Stil der Bayesschen Statistik, der rechnerischen Ausbeutung in Bootstrapping-Verfahren, als Turing-Test-Benchmark, als Einsatz von Schichten heuristischer Netzwerke, selbstorganisierender Karten, natürlicher Sprachverarbeitung, algorithmischer Entscheidungssysteme oder als eine weitere diskursive Darstellung. Insbesondere Sozialwissenschaftler\*innen wissen, dass es viele Möglichkeiten gibt, ein solches terminologisches Durcheinander von Darstellungen, Ansprüchen, Interessen und Annahmen zu strukturieren und zu konzeptualisieren. Daher ist jeder Versuch, dies zu tun, eine eigene epistemologische Leistung und eine eigene ontologische Positionierung. Obwohl diese Unordnung keineswegs nur für KI gilt, sondern auch für Vorstellungen, Institutionen, Normen und andere leere, schwebende (vgl. Laclau 2005) und möglicherweise offene Signifikanten (Kornak 2015), ist KI nicht einfach eine weitere soziale Signifikantenkette oder Spur (vgl. Derrida 1988, S. 90), die sich in einer grammatologischen Dekonstruktion (vgl. Derrida 1983) und ontosemantischen Deklination erschöpft. Sie ist auch ein Mittel der Abgrenzung im Sinne einer Reinigung (*purification*) (Latour 1993). Das gleiche System, z. B. der Empfehlungsdienst einer Inhaltsplattform, kann ein Mittel und ein Problem der Medienkuration, der Diskriminierung, der Infrastruktur, der Datenontologie, der Zeitpolitik, des Gatekeeping oder des Ersatzes (von menschlichen Akteuren) sein. Dementsprechend können sich die Herausforderungen und Anforderungen drastisch ändern. Die Suche nach den verborgenen Arbeiten, Perspektiven, Interessen und Normen, die durch die Praxis und die Ordnung des Wissens unsichtbar gemacht werden, ist jedoch eine besondere Stärke der Geistes- und Sozialwissenschaften. Die Sozialwissenschaft hat eine lange und starke Tradition im Umgang mit und sogar in der Nutzung von Komplexität. Unabhängig davon, ob man die Sozialwissenschaft als ‚Spezialwissenschaft‘ betrachtet oder ihr eine eigene Ontologie zugesteht (Durkheim 1976), sind Sozialwissenschaftler

per se darin geschult, mit der Undurchsichtigkeit von Objekten, Akteuren, Situationen, Semantiken, Strukturen und Schichten der gesellschaftlichen Ontologie selbst umzugehen.

Die Schärfung des Blicks und der Sensibilität für KI darf nicht zu einer Verwirrung über die ‚tatsächliche‘ Beteiligung von KI an Prozessen führen; da KI auch als (Marketing-)Label genutzt wird, ist die Frage nach der jeweiligen Rolle von KI ihr als Gegenstand inhärent. Erst recht müssen Forscher über die spezifische Beteiligung von KI nachdenken, die sie vermuten, beobachten und interpretieren, und die ihrer Methodik und Untersuchung zugrunde liegt. Darüber hinaus darf sich dies nicht auf positive Fälle von KI-Forschung oder das Hinterfragen der Beteiligung von KI beschränken: Da sich Fälle von KI in allen Bereichen und Sphären des Sozialen ausbreiten, kann die daraus resultierende Normalisierung KI auch für den Forscher unsichtbar machen. Derartige negative Fehlerkapazitäten sind jedoch schwer zu bewältigen und können möglicherweise nie ganz überwunden werden.

Wie wir in der Schlussfolgerung des Kapitels argumentieren werden, haben wir uns zwar der Frage gewidmet, wie Sozialwissenschaften KI-Forschung für oder mit ihren Methoden unterstützen kann. Sie müssen nichtsdestoweniger an die Interaktion zwischen einer solchen Methodik und dem Forschungsgegenstand erinnern, erst recht, insofern KI ein expliziter Begriff und ein diskursives Vehikel der Politik, der Wirtschaft, der (Pop)Kultur und der Forschung ist (z. B. wie Sie sie in diesem Buch finden); mit sowohl präzisen, technisch differenzierten Definitionen als auch unscharfen und leeren. Wissenschaftler, unabhängig von ihrer jeweiligen Disziplin, können den Diskurs, der hier beschrieben wird, nicht verlassen – was die Disziplinen hier unterscheidet ist nicht die Beteiligung, sondern die disziplinäre Bedeutung der Sache.

Im Folgenden werden wir die methodischen Herausforderungen der KI-Forschung untersuchen, indem wir über unsere eigene Methodik und Vorgehensweise berichten, wobei wir uns sowohl auf die Erhebung und Kodierung von Literatur als auch auf die digitalen Kartierungs-Techniken konzentrieren, die wir zur Verbesserung unserer heuristischen Sicht auf das vorliegende Material einsetzen. Die dabei aufgetretenen Hindernisse sollen dann anhand ausgewählter Beispiele aufgezeigt werden. Wir werden argumentieren, dass der Pragmatismus, wie er den Sozialwissenschaften – sowohl als methodische Attitüde als auch im Sinne eines philosophischen Unterbaus – inhärent ist, kritische Lektionen in Bezug auf Methode und Perspektive für die Bemühungen der KI-Forschung enthält.



## 2 Methodologie und Methoden: KI-Forschung erforschen

Sich in dieser Komplexität zu bewegen, lässt sich als soziologische Leidenschaft bezeichnet. Über methodisches Vorgehen in der wissenschaftlich geforderten Klarheit zu berichten, gehört nichtsdestoweniger zu ihren Pflichten. Was aber haben wir getan, um mit der oben nur grob skizzierten Komplexität umzugehen? Man kann die Geschichte unserer eigenen Methodik erzählen, indem man das folgende Problem beschreibt: Die Bewertung des derzeitigen Stands und der Kapazitäten sozialwissenschaftlicher Methoden und Verfahren für die KI-Forschung scheitert in beiden klassischen Szenarien:

- a. Ein quantitativer, digital automatisierter Ansatz (z. B. Scraping von Datenbanken zur Beschreibung und Indexierung der ermittelten Literatur) kann nicht zur ursprünglichen Forschungsfrage beitragen. Er würde die positivistische Naivität imitieren, die den automatisierten Verfahren der KI innewohnt. Der passive Konsum objektiver Datensätze und ihrer jeweiligen Metadaten ist kein echter methodischer Beitrag, sondern eine Reproduktion und Bestätigung der digitalen Methode selbst.
- b. Eine qualitative, sorgfältige Interpretation des Standes der Technik und des Methodendiskurses würde das aktuelle Terrain der KI-Forschung grundsätzlich nicht abdecken. Den klassischen Methoden der Literaturrecherche, des Interviews, der Beobachtung fehlt die Effizienz und Interoperabilität digitaler Methoden auf Feldebene.

Daher haben wir einen Mixed-Methods-Ansatz verfolgt, der unser persönliches Fachwissen und unsere qualitative Sensibilität mit digitalen und quantitativen Analyse-, Interpretations- und Kategorisierungsmethoden verbindet. Die erfasste Literatur umfasst einen Publikationszeitraum von 1983 bis 2021 (was auch den statistischen Modus des Datenkorpus darstellt). Das meiste, über 80, stellt Literatur im engeren Sinne dar, es sind aber im Übrigen auch einige Ressourcen wie Code-Repositoryn oder Plattformen darunter. Insgesamt wurden 85 Gegenstände gelistet, die empirische (also Anwendung von) oder methodologische Publikationen darstellen, oder sich als solche verwenden lassen. Da das Material vom unserem Team gesichtet und organisch erweitert wurde, haben wir zwar ausschließlich mit der Akquise methodologischer Veröffentlichungen im oben beschriebenen Sinne begonnen, aber dieses bald um weitere Genre erweitert. So fallen schlussendlich in den Untersuchungsbereich auch Propädeutik und

Überblicksartikel. Obschon unsere Forschungsgruppe von sich aus einen sozialwissenschaftlichen Bias in die Sammlung einbringt, haben wir die Literatur nicht auf rein sozialwissenschaftliche Arbeiten beschränkt, sondern die Entscheidung für oder gegen Aufnahme in unseren Korpus anhand der sozialwissenschaftlichen Beschaffenheit von Forschungsfrage oder -gegenstand bestimmt.

Letztendlich haben wir etwa 85 Artikel aus der (Social Media-)Plattform- und Empfehlungsforschung (Wall Street Journal 2021; Bhandari und Bimo 2020; van Es 2017; Möller et al. 2018; und Francisco und Ruhela 2021), Industrie 4.0 (Passoth und Kahlert 2018a; und Passoth und Kahlert 2018b), Netzwerkforschung (Centola 2018; Hine 2005; Callon et al. 1983; Özyer et al. 2013; und Gießmann 2009), digitale und experimentelle Methoden (Murthy et al. 2013; Marres und Weltevrede 2013; Niederer 2016; Scott 2017; und Weltevrede 2016), Responsible Engineering und Ethik (Ananny 2016; Buolamwini und Gebru 2018; und Linardatos et al. 2021), Big Data (Mützel 2015; Kitchin 2014; und Boellstorff und Maurer 2015), Internetforschung (Rogers, Digital Methods 2009; Rieder 2012; und Rogers und Marres 2000), Statistik und Datenwissenschaft (Bandy und Diakopoulos 2020; und Little und Rubin 2013) sowie Smart Cities (Yigitcanlar et al. 2020) gesammelt, um nur die wichtigsten Stränge zu nennen.

In den daraus resultierenden methodischen Kombinationen und vor allem in deren Austausch liegt ein großes Potenzial (vgl. Marres 2012). Unter anderem ist ein solcher Ansatz durch die ‚Grounded Theory‘ repräsentiert. Daher sind wir dem von Juliet Corbin und Anselm Strauss (1990) ausgearbeiteten Kodierparadigma gefolgt, welches ein geeignetes Instrument war, um die Sammlung und Kodierung der Literatur auf das gesamte Team zu verteilen. Das Kodierparadigma in der qualitativen Sozialforschung drückt die der Grounded Theory (vgl. Strauss 1998) zugrunde liegende Annahme aus, dass empirisches Material nach Codes, Konzepten, und Kategorien geordnet werden kann, die den im Phänomen – bestehend aus Bedingungen, Konsequenzen, Strategien und Kontexten – gegebenen Zusammenhängen zwischen Objekten, Tropen, Interaktionen, Handlungen, Positionen und Rollen durch eingehende Auseinandersetzung mit dem empirischen Beobachtungs- und Studienmaterial in einer Weise organisiert werden kann, dass sie eine beschreibende Erzählung mit wissenschaftlichem Mehrwert bildet. „Das Kodierparadigma soll dazu verhelfen, systematisch über die Daten (Texte, Textstellen) nachzudenken und komplexe Beziehungen herauszuarbeiten“ (Boehm 1994, S. 131). Für unsere Untersuchung gehen wir dafür aus, dass auch ein Literaturüberblick auf diese Art gestaltet werden kann, insbesondere wenn nicht nach einem bestimmten methodischen Sachstand gefragt wird, sondern eine reflexive, kritische Untersuchung der breiteren Forschungsprogrammatisierung angestrebt wird. Zu diesem Zweck eignet sich eine diskursensible

qualitative Analyse besser, als die massenstatistische Konstruktion eines Kanons. Die Masse und das Durcheinander offener Codes, die auf diese Weise entstanden sind, konnten in Kodierungsworkshops reduziert und verdichtet werden, in denen wir Kategorien und Dimensionen für unsere Codes entwickelt haben. Wie Adele Clarke (2003) argumentiert, impliziert eine solche Kodierung einen erkenntnistheoretischen Konstruktivismus, der den sozialen Konfliktfeldern und der zeitgenössischen Heterogenität nicht besonders viel Aufmerksamkeit widmet, insofern der klassische Kodieransatz der Grounded Theory vor allem auf die Produktion *eines* Narrativs abstellt. Ihre Kartierungsverfahren fokussieren entsprechend Widersprüchlichkeiten und konfligierende Beobachtungsperspektiven. Dieser Einwand geht über die Auseinandersetzung der traditionellen Grounded Theory mit unterschiedlichen Kontrasten hinaus, wie sie z. B. durch die Methode des Vergleichs maximaler und minimaler Unterschiede, oder anhand der ‚theoretischen Sättigung‘ – also wenn weitere Beobachtungen keine analytischen Irritationen mehr bieten – angestrebt werden. Ähnlich wie Clarkes methodischer Vorschlag haben wir ein digitales Kartierungs-Verfahren angewandt, um mehr Perspektiven und Wahrnehmungen in den Daten zu entdecken und zu diskutieren, unterstützt durch Computer. Auf diese Weise haben wir gehofft, unsere Daten und ihre kodierte Struktur nicht nur in einer subjektiven Erzählung, sondern sie in einer räumlichen Ordnung zu Vektoren verschiedener Anwendungs- und Problemfelder anzureichern. Im Sinne von Marres‘ (2015) Gedanken zum digitalen ‚Controversy Mapping‘ haben wir versucht zu verstehen, welche methodischen Verstrickungen und Konstellationen sich ergeben können.

In Bezug auf das Verfahren kann unsere Forschung in drei Phasen unterteilt werden: a) die erste Phase des Sammelns erster Lektüre und des Verfolgens von Referenzen und thematischen Spuren, b) eine zweite, iterative Phase des Codierens, der gemeinsamen Diskussion und Bewertung unserer Codes und Daten, einschließlich weiterer Ergänzungen des Literaturkorpus, c) eine letzte Phase, in der wir den Literaturkorpus gelegentlich ergänzen und reduzieren, Codes und Dimensionen ordnen und unsere methodischen Entscheidungen im Hinblick auf die Forschungsergebnisse artikulieren. Wir haben mit dem Ausgangszustand am Frontend der Datenerhebung mit unserem bereits gegebenen Verständnis von künstlicher Intelligenz begonnen, nämlich in sozialwissenschaftlicher Manier überall dorthin zu schauen, wo Menschen behaupten, künstliche Intelligenz einzusetzen, und auch dorthin, wo Maschinen aus Daten gewonnene Modelle nutzen, um Entscheidungen als Handlungsanleitung oder Zuordnung zur Verfügung zu stellen (darin etwa Empfehlungssysteme, Digitale Zwillinge, Unterstützung Bildgebender Verfahren, etc.), und haben den Umfang der Literaturrecherche entsprechend erweitert (vgl. Bechmann/Bowker 2009 oder Searle

1980). Wir haben Themen und spezialisierte Teildisziplinen recherchiert, die explizit oder implizit künstliche Intelligenz berührten. Darüber hinaus haben wir selbstständig in einschlägigen Repositorien wie Google Scholar, Web of Science, unseren Universitätsbibliotheken und dergleichen nach Schlüsselbegriffen gesucht. Unser Team hat ebenso die Ressourcen und das selektive Bewusstsein der Mitglieder für das Thema aus ihren eigenen Forschungsschwerpunkten sowie aus abgeschlossenen (z. B. das IMPROVE Projekt, in welchem wir die Entwicklung digitaler Assistenzsysteme für die industrielle Folienfertigung sozialwissenschaftlich betreut haben) und noch laufenden Forschungsprojekten (z.B. HIRING, in welchem wir die Entwicklung einer Lernplattform für den Einsatz von künstlicher Intelligenz in der Personalwirtschaft unterstützt haben) zusammengetragen. Dabei hat sich schnell herausgestellt, dass es sich um eher lose verbundene, heterogene Gegenstandsbereiche handelte. Gleiches gilt für den umgekehrten Fall fehlender (expliziter oder fixer) KI-Bezüge und Inhalte, obwohl diese von den Autor\*innen in Titel und/oder Abstract angekündigt oder anderweitig angegeben bzw. nahegelegt wurden. Alle selektierten Literaturen, Projekte und Ressourcen wurden zusätzlich auf enthaltene methodologische Verweisungen und referenzierte Autor\*innen selbst untersucht, sofern sie sich für unseren eigenen, allgemeineren Bereich der methodischen Forschung als vielversprechend und anwendbar qualifizierten. Zu diesem Zweck wurden Abstracts und selektive Abschnitte aus den referenzierten Papieren daraufhin untersucht, ob sie zu unserem Materialkorpus passten. Unpassende Elemente wurden nur dann aussortiert, wenn ihre fehlende Relevanz augenfällig war, wie etwa Studien ohne Einsatz oder Untersuchung von digitaler Technik. Desweiteren haben wir Netzwerkanalyse- und Mapping-Literatur in unsere Forschung einbezogen, da wir uns Synergieeffekte versprochen haben (z. B. das Auffinden von methodischen Ressourcen für unsere eigenen Strukturierungsversuche des resultierenden Literaturkorpus) und neugierig auf das Zufälligkeitspotenzial dieser Strategie waren; schließlich gibt es bei der KI neuronale Netze, selbstorganisierende Karten (SOM) und das gemeinsame Problem, große (in inhaltlichem und relationalem Umfang), unstrukturierte (z. B. nominale) oder wenig strukturierte (z. B. semantische Skalen) Daten zu behandeln. Das so zusammengetragene Material war durch persönliche und referenzielle Bindung zusammenhängend, aber auch durch Fremdverweise und die Heterogenität unseres Teams diffus. Mit dieser Kontextannahme im Hinterkopf haben wir den Gesamteindruck der Codes und der dazugehörigen Kodiererfahrungen diskutiert, um allgemeinere Einsichten zu konstruieren, welche – ironischer Weise – vor allem in der diagnostizierten Partikularität von Gegenständen, Perspektiven und Methoden bestehen.

Zunächst haben wir das gesammelte Material offen kodiert. Wir haben den semantischen Bereich unserer Lesarten zu einer flachen Ontologie abstrahiert. In anderen Worten haben wir versucht, auf konzeptionelle Hierarchien zu verzichten oder diese zu kompensieren, sofern sie in Methodologien implementiert waren. Es wurde also keine techno- oder anthropozentrische Perspektive bevorzugt, oder wurden individualistische über oder unter kollektivistische, ökologische oder kontextualistische Ansätze geordnet. Durch die so bewahrte struktur-konzeptionelle Offenheit konnte der aggregierte Korpus auf unterschiedliche Weise arrangiert und erkundet werden – in einer Vielfalt, wie sie der Vielfalt des Gegenstands gerecht wird. Dennoch muss man sich bewusst bleiben, dass all diese Kodierungsaktivitäten selbst eine spezifische soziale Praxis sind, mehr noch, eine Ad-hoc Praxis (vgl. Garfinkel 1967, S. 20 ff.).

Während der Kodierungsdiskussionen haben wir auch versucht, unsere Kodierungsontologie erneut zu dehierarchisieren, um unsere Entscheidungen zu hinterfragen, mit Kodierungsinkonsistenzen umzugehen und Möglichkeiten für eine weitere Literatursammlung zu erkunden, wie z. B. das Hinzufügen weiterer Online-Ressourcen, die mehr gebrauchsfertige Tools (für mehr und weniger computererfahrene Forscher), Anleitungen und Beispiele für digitale Methoden bieten, die für KI-Forschungsszenarien geeignet sind. Diese Art von Daten besteht aus Blogbeiträgen, Git-Repositorien und – was noch wichtiger ist – aus Berichten und Projekten von Journalisten und Aktivisten. Diese ‚iterativen Schleifen‘ des Schließens und Öffnens unserer Kodieretiketten und -strukturen sind so bekannt wie üblich in den qualitativen Sozialwissenschaften und der Grounded Theory im Besonderen. Sie sind jedoch nicht nur in der qualitativen Forschung anzutreffen: Wie bei der explorativen quantitativen Methodik (z. B. bei der Arbeit mit großen oder sogar Big Data Datensätzen) besteht ein solcher Forschungsprozess aus einer Abfolge von Versuchsschritten und Mikroexperimenten innerhalb der Daten, die selbst in der Statistik nicht durch ‚rohe Gewalt‘ ersetzt werden können (zumindest solange es um ordinale oder nominale Daten geht). Daraus ergibt sich eine Klasse von Forschungsszenarien, die abduktive (vgl. Peirce 1931 oder Reichertz 2010) Aktivitäten des Spielens, der Entwicklung von Forschungsstrategien, und des Umdenkens erfordern, um die untersuchten Phänomene nachvollziehbar zu verstehen. Während unsere individuellen Aktivitäten zur Literatursammlung kontinuierlich und verstreut stattgefunden haben, waren unsere gemeinsamen Diskussionen zur Kodierung, Kartierung und narrativen Integration das Ergebnis umfangreicher En-Bloc-Workshops, die zwischen drei und fünf Stunden dauerten.

Die Ergebnisse dieses ersten Kodierungsverfahrens haben wir in zwei umfangreichen Workshops gemeinsam diskutiert. Das Hauptziel dieser Workshops war die Vereinheitlichung der Variationen, die trotz unserer Online-Zusammenarbeit

während der individuellen Kodierungsphase übriggeblieben waren. Die Vereinheitlichung der einzelnen Codes betraf auch solche, die einfach zu spezifisch, verstreut und selten waren, um sich als eindeutige Marker zu erweisen. Codes, die zu unzusammenhängend oder willkürlich waren, wurden gestrichen. Auf der eher konstruktiven Seite der Workshops haben wir Dimensionen und Achsen für unsere Labels bzw. Bezeichnungen entwickelt, um den Inhalt unserer Datentabellen zu organisieren. In der hier besprochenen Phase haben wir versucht, über diese vorläufigen Ergebnisse nachzudenken und Verzerrungen auszugleichen. Wir haben den Daten, die größtenteils aus der wissenschaftlichen Literatur stammten, weitere Tools, Dienste und Projekte hinzugefügt und die vorhandene Forschung über YouTube und seine Voreingenommenheit um mindestens eine weitere Plattform ergänzt. Dabei haben wir uns stärker auf eine dieser Plattformen konzentriert, nämlich TikTok, den Social-Media-Kurzvideodienst mit der derzeit bekanntesten und meistdiskutierten KI-Technologie. In einem zweiten Workshop haben wir über und mit möglichen vorläufigen Zuordnungen des Materials unter Verwendung der erstellten Codes und anderer Kriterien (wie Autorschaft, Verlag oder Zeit) diskutiert und experimentiert. In einem abschließenden Workshop haben wir unsere Labels in bestimmten Dimensionen reduziert und organisiert. Die daraus resultierenden Dimensionen und Kategorien können Abb. 1 entnommen werden.

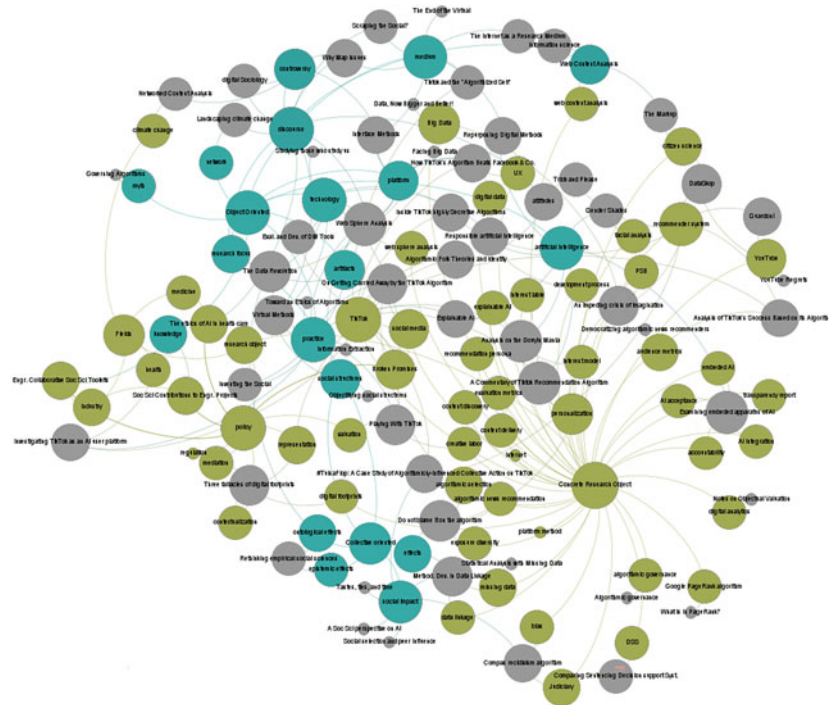
Die obige Abbildung zeigt unsere Code-Struktur, sowie ihre Achsen und Dimensionen. Alle Codes wurden einer gemeinsamen erkenntnistheoretischen Achse zugeordnet, mit Ausnahme einiger Überbleibsel, die als Überlaufkategorie für neue, nicht ausgerichtete Codes dienten. Diese sollten bei späteren Workshops eingeführt oder aufgelöst werden. Darüber hinaus wurden die elementaren Codes in ordinale, semantische oder nominale Gruppen eingeteilt. Wie oben gezeigt, gibt es kleine, sogar unmittelbare Code-Kategorien, wie z. B. den disziplinären Code ‚Sozialwissenschaft‘, und übermäßig große Code-Gruppen wie unsere Reihe von Forschungsobjekten, die wir als ‚konkrete Forschungsobjekte‘ zusammengefasst haben.

Abb. 2 zeigt eine mögliche Kartierung. Wir haben sie ausgewählt, um die besondere und differenzierte Vielfalt der verschiedenen Forschungsschwerpunkte und -objekte zu verdeutlichen. Sie alle erfordern eine bestimmte Methode – ob man sich für die Politik und die tatsächliche Kuratierung von Medien auf Plattformen oder für KI-gesteuerte Nutzerbewegungen oder -verhalten interessiert oder ob man das für eine Plattform wie YouTube, Facebook, TikTok oder Amazon untersucht. Die Fokussierung auf Arbeit, Ungleichheit und Inklusion kann die Sichtweise auf ein Forschungsobjekt stark verändern, nicht nur bei der Untersuchung von KI. Die hier mögliche Auflösung und Detaillierung der

▶ Type of methods:	{...}	▼ Research focus:	
▼ Research discipline:		▶ Object oriented:	[...]
▶ socialscience:	[...]	▶ Collective oriented:	[...]
▶ Between humanities and STEM:	[...]	▶ Discourse:	[...]
▶ Hard(er) sciences:	[...]	▼ Research object:	
▼ Data collection:		▶ Concrete research object:	[...]
▶ digital (data collection):	[...]	▶ Policy:	[...]
▶ Analogue:	[...]	▶ Fields:	[...]
▼ Data analysing:		▼ Research approach on AI:	
▶ automatically:	[...]	▶ Design/Construction:	[...]
▶ Visually:	[...]	▶ Critique:	[...]
▶ Manually:	[...]	▶ Genealogy:	[...]
▼ Research focus:		▼ Approach on AI:	
▶ Object oriented:	[...]	▶ affirmative:	[...]
▶ Collective oriented:	[...]	▶ reflexive:	[...]
▶ Discourse:	[...]	▶ extrinsic:	[...]
▼ Research object:		▼ Content-type:	
▶ Concrete research object:	[...]	▶ empirical study:	[...]
▶ Policy:	[...]	▶ Methodology:	[...]
▶ Fields:	[...]	▶ Overview:	[...]
▼ Research approach on AI:		▶ Perspective:	[...]
▶ Design/Construction:	[...]	▼ Item-type:	
▶ Critique:	[...]	▶ publication:	[...]
▶ Genealogy:	[...]	▶ Project:	[...]
▼ Approach on AI:		▶ Ressource:	[...]
▶ affirmative:	[...]	▶ Other:	[...]
▶ reflexive:	[...]	▶ Theory/method label:	{...}
▶ extrinsic:	[...]	▶ Method approach:	{...}

**Abb. 1** Dimensionen und Kategorien der codierten Literatur, wie sie in unserer JSON-Datei enthalten sind. (Quelle: Eigene Darstellung)

Karte erlaubt leider nur eine konzeptionelle Veranschaulichung unseres Ergebnisses. Zwar lassen sich einfach an dem Straw-farbenen Cluster rechts unten der Vielzahl an Überbegriffen enthaltener Forschungsgegenstände die Vielfalt eben dieser erkennen. Die untere Hälfte enthält vor allem sozialtheoretische, reflexive und politische Essays, die obere hingegen konkrete Studien mit einer Tendenz zu Forschungen über social media Plattformen. Am Rand finden sich neben einzelnen Ausreißern (etwa eine TikTok Studie, die sich in unserem Datensatz durch die Verbindung von ‚social media‘ und ‚health‘ auffällt, vor allem Ressourcen



**Abb. 2** Die beiden Kategorien ‚Forschungsobjekt‘ in Cyan und ‚Forschungsschwerpunkt‘ in Straw werden zusammen abgebildet. (Quelle: Eigene Darstellung)

und Vorlagen, die für die Erforschung digitaler Umwelten von künstlicher Intelligenz genutzt werden können (z. B. zum Sammeln von Recommender-Daten). Ihr weiteres Potenzial offenbart eine solche Kartierung und der ihr zugrunde liegende Datensatz jedoch zweifelsohne vor allem in ihrer Sortier- und Filterbarkeit, wie sie in anderen und insbesondere digitalen Medien besser dargestellt werden können. So lassen sich gezielt Themen, Perspektiven, Methoden und Schnittpunkte identifizieren und abbilden – und selbstverständlich auch erweitern.



### **3 Die Werkzeuge, aus denen Werkzeugkästen (gemacht) sind**

Wie haben wir nun die Karte aus Abb. 2 (und die vielen anderen, die wir aus unseren Literaturdaten erstellen können) erstellt und verwendet? Es lohnt sich immer, Wörter in einem Wörterbuch oder einer Enzyklopädie nachzuschlagen, in diesem Fall in Wikipedia. Wir haben dort den Begriff ‚Werkzeugkasten‘ (Toolbox) nachgeschlagen und seine einschlägigen aber auch latenten, spezifischen Eigenschaften gefunden. Der Werkzeugkasten umfasst Systeme, die Werkzeuge lagern und organisieren, ihn an verschiedene Orte transportierbar machen sowie den Inhalt schützen (Wikipedia 2021). Ein Werkzeugkasten ist jedoch nie vollständig. Er ist stets unter Bearbeitung, sozusagen ein ‚work-in-progress‘, immer auf der Suche nach neuen Werkzeugen und Geräten. Ein Werkzeugkasten für die akademische Forschung weist dieselben allgemeinen Merkmale auf: Er organisiert akademische Ressourcen, macht sie für verschiedene Projekte nutzbar und ist in Arbeit. Allerdings spielt die Komplexität des Themas neben dem Zweck, den es erfüllt, eine wesentliche Rolle bei der Gestaltung. Wie könnte also ein Werkzeugkasten für die KI-Forschung aussehen?

Welche Merkmale und Funktionalitäten weist er auf? Was wird einbezogen oder ausgeschlossen? Und was kann er dem Nutzer – dem Sozialwissenschaftler – bieten? Neben der Beantwortung dieser Fragen werden wir in diesem Abschnitt vor allem untersuchen, wie die Prozesse ablaufen, die zur Entstehung eines solchen Werkzeugkastens führen. Was sind die Ressourcen, die wir für den Aufbau heranziehen? Wie sind sie konzeptionell miteinander verbunden? Welche Art von Konzepten und Fällen stellen sie dar und welche Erhebungs- und Analysestrategien kommen zum Einsatz?

Wie in unserem Methodik-Kapitel beschrieben, besteht unser primärer Ansatz zur Erstellung des Werkzeugkastens in der Anwendung einer qualitativen/quantitativen Forschungsmethodik. Daher beginnen wir mit unserer Datenerhebung und den Strategien, die hinter der Zusammenführung von Literatur aus verschiedenen Bereichen und der Berücksichtigung der für die KI verwendeten Methoden stehen. Die von uns gesammelte Literatur stammt aus den Bereichen Soziologie, Psychologie, Wirtschaft, Wissenschaft und Technikforschung sowie Informatik. Folglich werden wir die Erhebung und Klassifizierung von ‚Code Knoten‘ – anstelle von Artikel-Schlüsselwörtern – als analytische Strategie für unsere Literatur diskutieren. Schließlich exportieren wir all diese Daten in die Gephi-Software und visualisieren sie mithilfe bestimmter Layouts, die zeigen können, wie bestimmte Konzepte und Methoden innerhalb des Literaturkorpus an Bedeutung gewonnen haben oder gewachsen sind. Hier streben wir einen Zustand

an, den Latour (1985) als ‚optische Konsistenz‘ (optical consistency) beschreibt, bei dem wir vertikale und horizontale Verflechtungen, unterschiedliche Praktiken und unterschiedliche Realitäten erkennen können.

Eine wesentliche Funktion dieses Werkzeugkastens ist es, eine Brücke oder eine Form der Kontinuität zu erzeugen zwischen der klassischen sozialwissenschaftlichen qualitativen Forschungsmethodik und den ‚innovativen‘ digitalen Werkzeugen, die wir zur Visualisierung verwenden. In diesem Abschnitt wird erklärt, wie man diese Brücke in 5 Schritten schlägt: 1 – Erstellung des Baums und Definition der Beziehungen, 2 – Import des Baums und Aufbau von Knoten- und Kanten-Tabellen, 3 – Auswahl des Layouts, 4 – Ästhetik und 5 – Analyse.

### 3.1 Tabellen und Layouts

Zunächst müssen wir die Code-Knoten-Tabelle umwandeln, die wir im Rahmen unserer Dokumentenanalyse erstellt haben. Jede Kategorie beginnt hier einen Baum, wird zur Wurzel dieses Baums und nimmt verschiedene Code-Knoten-Gruppen als seine Kinder auf. Die verschiedenen Klassen der unterschiedlichen Kategorien befinden sich alle auf derselben Ebene, ebenso wie die Codes, die sie enthalten. Jede Besonderheit dieser Qualitätsgruppen wird zu einem Kinder gleichen Ranges. Durch diese Struktur wird der Baum horizontal ausgedehnt, was zu der Flachheit beiträgt, die von uns angestrebt wurde. Wie weiter oben erläutert, wollten wir Potenziale zu Analyse und Applikation bewahren. Komplexere Kartierungen mit drei (und mehr) Dimensionen, oder vorauslaufenden Unterscheidungen und Ordnungen (z. B. in einen quantitativen und qualitativen) Korpus enthalten zwar mehr analytischen Gehalt, allerdings nur für fortschreitend spezifische Fälle analytischer Anwendung. Die Auswertung würde sich beispielsweise dann auf ein bestimmtes differenziales Verständnis von quantitativer, qualitativer und mixed (oder residualer) Methodik kondensieren, oder müsste sich auf einen spezifischen Konzeptfall von KI und KI-Forschung fokussieren. Das wollten wir angesichts des bedeutend allgemeineren Forschungsauftrags vermeiden. Zwar enthalten unsere Kartierungen durchaus mehr Dimensionen (bzw. Informationen), die nicht räumlich sind, wie etwa die Farbe, welche die ‚Art des Knotens‘ repräsentiert, allerdings handelt es sich hierbei um nicht weiter hinterfragte Designationen von Gegenständen (z. B. als Konzept oder Fall, als Teil des Korpus oder des Codes), dessen Annahme selbstverständlich sein sollte und dessen skeptizistische Demontage gewiss intellektuell nicht uninteressant, wohl aber wissenschaftlich wenig fruchtbar ist.

Bevor wir weitere Schritte unternehmen, müssen wir die Teile beschreiben, aus denen dieser Werkzeugkasten besteht. Unsere Datensätze werden in Excel und einem.csv-Format gesammelt. Bevor wir sie in eine Illustrationssoftware wie Gephi exportieren, müssen wir zwei grundlegende Tabellen erstellen: die ‚Knotentabelle‘ und die ‚Kantentabelle‘. Zu Beginn bezeichnen wir jede Verbindung als ‚Kante‘ und jede Entität am Anfang oder Ende einer Kante als ‚Knoten‘. Da eine Kante eine Beziehung zwischen zwei Knoten darstellt, kann sie ein Gewicht haben, das die Dichte, also Häufigkeit, angibt. Außerdem sind die Kanten in diesem Werkzeugkasten ‚ausgerichtet‘, da sie zeigen, welcher Knoten ein Quellknoten und welcher ein Zielknoten ist. Das Design der Datenstruktur impliziert, dass jeder Quellknoten ein Elternteil wird und die Kante von mindestens einem Zielknoten erhält. Der Zielknoten kann entweder eine Klasse, ein Codeknoten oder ein Artikel sein. Die Knotentabelle besteht aus den Namen der Knoten als ‚Label‘, und die Kantentabelle enthält die Beziehung zwischen jeweils zwei Knoten, indem zwei separate Spalten als (die) Quelle (Knoten) und (das) Ziel (Knoten) gebildet werden. Das Gewicht jeder Kante kann ebenfalls als Eigenschaft zu dieser Datei hinzugefügt werden, will man etwa bestimmte Arten der Zusammenhänge hervorheben. In unserem Vorgehen war eine solche, künstliche Emphase nicht erforderlich.

Um sie schließlich zu visualisieren, werden diese Dateien in Gephi importiert, eine Software, „simulat[ing] a physical system to spatialize a network“ (Jacomy, et al. 2014). Die Software nimmt physikalische Systeme als Grundlage für die räumliche Darstellung der Eingabedaten, indem sie sie in Algorithmen definiert. Diese Algorithmen sind dann für die Daten als Layouts nachvollziehbar. Jedes Layout entspricht dem ursprünglichen physikalischen System und macht sich dessen Haupteigenschaften zu Eigen. Zu diesen gehören primäre physikalische Attribute wie Gravitationskraft, Temperatur und Entfernung, die dem Benutzer Gestaltungsmöglichkeiten bieten.

Das Layout, das wir für diesen Werkzeugkasten verwenden, heißt ‚Multigravitational Force Atlas‘. Es ist ein Abkömmling des kräftegesteuerten ForceAtlas-Layouts. Das bedeutet, dass die räumliche Darstellung des Netzwerks auf der Grundlage von Knotenabstoßung und Kantenzug erfolgt. Es ist inspiriert von der Tatsache, „[that] in physical systems, forces depend on the distance between the interacting entities“ (ibid). Daher bilden in dieser Familie von Layouts verwandte Knoten eine Gemeinschaft oder einen Ort in der Nähe des anderen, während diejenigen, die keine Beziehungen haben, sich gegenseitig wegstoßen.

Darüber hinaus bietet das Layout Eigenschaften wie Ling-Long-Force (Hervorhebung der Modularität und Strukturierung engerer Cluster), Hub-Dissuasion

(Hervorhebung der Knoten mit höheren In-Graden als Hub-Zentren und Verdrängung an die Peripherie) und die Zuweisung von Schwerkraft an Zentren. Diese drei Eigenschaften zusammen konfigurieren die Struktur, weil sie es ermöglichen, zwischen verschiedenen Kategorien als Knotenpunkte zu navigieren. Wir können sehen, ob ein Code-Knoten als Knotenpunkt in der Modularität der Kategorien und Klassen auf unterer Ebene fungiert. Da kein Knoten als einziges Zentrum zugewiesen wird, wird die gegenseitige Abhängigkeit zwischen den Knoten und die Notwendigkeit des Navigierens und Springens zwischen den Knotenpunkten für eine ganzheitlichere Sicht hervorgehoben.

### 3.2 Digitale Lösungen wie Gephi

Bevor wir mit der Kartenanalyse fortfahren, sollten wir uns fragen, welche anderen grafischen Merkmale wir bei der Analyse der Karte berücksichtigen sollten. Wir sprechen hier von zwei geometrischen Merkmalen – Nähe und Verteilung – und einem Größenmerkmal – Knotengröße –, das ebenfalls von den geometrischen Merkmalen abhängt. Zusammengenommen bilden sie unser Verständnis der Visualisierung. Es ist wichtig zu erkennen, dass diese grafischen Merkmale die Essenz unseres Werkzeugkastens sind. In der Praxis sind sie in den Daten und der Software eingebettet und werden in der endgültigen Visualisierung dargestellt. Darüber hinaus erleichtern sie die kleinen Änderungen, die man vornimmt, um die Daten besser lesbar zu machen.

Zunächst einmal verbindet die Umgebung die Datenordnung und die dahinterstehende Kategorisierung zu Clustern. Während wir das Merkmal durch Einfärbung beobachtbar machen können, übersetzt es die Kantentabelle in grafische Cluster, die aus Knoten und Kanten bestehen. Nähe entsteht, wenn Knoten sich verbinden, wenn sie als Quell- und Zielknoten in Beziehung zueinanderstehen, und die Kante kann auch als ausgerichtet oder nicht ausgerichtet definiert werden. Diese Definition fördert jedoch nur Nachbarn mit dem Grad Null – keine anderen Knoten liegen dazwischen. Auch zwischen Knoten mit höheren Nachbarschaftsgraden gibt es eine Nähe. Während in vielen Netzwerkvisualisierungen ein großer Teil der Knoten mit nur wenigen anderen verbunden ist und daraus eine gewisse Übersichtlichkeit erwächst, wird in stärker strukturierten Netzwerken, insbesondere bei systematischen Literaturübersichten, die Verbindung zwischen Codes und Kategorien wichtiger für ihre Betrachtung; so machen Nachbarschaftsgrade hier spezifische Machtdynamiken und Strukturen geltend.

Während ein Cluster allein durch die Betrachtung der Codes und Klassen verdichtet wirken kann, ist die Funktion ‚Nähe/Umgebung‘ nicht repräsentativ für

die örtliche Annäherung des Clusters. Stattdessen berücksichtigt ein Feature wie die Verteilung mehr Elemente, um die Knoten auf der Leinwand festzulegen. Wenn wir über die Verteilung in einem Multi-Gravitations-Layout diskutieren, sollten wir bedenken, dass wir im Gegensatz zu einem Single-Gravitations-Zentrum-Layout das Zentrum nicht als ordnende Kraft ansehen. Daher werden die Kategorisierungscluster durch ihr kumulatives Gewicht in einer ausgewogenen und harmonischen Weise verteilt. Wenn wir dem Netzwerk kategorisierte Knoten hinzufügen, würde eine reaktive Verteilungskraft das Gleichgewicht stören und neue Verflechtungen schaffen. Das bedeutet, dass Knoten desselben Kategorisierungsclusters nicht zusammenbleiben würden, und die Vermischung würde uns eine ganzheitliche Sicht darauf geben, wie Kategorien und kategorisierte Knoten miteinander in Beziehung stehen. Der Grund für diese Veränderung ist das Ungleichgewicht der Verbindungen für kategorisierte Knoten (hier die uns zur Verfügung stehende Literatur). In diesem Moment kann ein Artikel zu einer bestimmten Kategorisierung neigen. Wenn jedoch die Neigung auf verschiedene Gruppen aufgeteilt werden muss, verwendet die Verteilung Elemente der Kanten- und Knotengewichte, um die Position eines bestimmten Knotens zu bestimmen. So wie die Nachbarschaft die Vorteile der Kanten bei der Bildung von Clustern und der Machtdynamik nutzt, mobilisiert die Verteilung sie auch, um Knoten verschiedener Cluster nebeneinander zu platzieren. Der Akt der Verteilung scheint eine Verhandlung zwischen verschiedenen Kräften zu sein. Daher hören die Knoten in einigen Einzel- und Multigravitations-Layouts wie Force Atlas kaum auf, sich zu bewegen. Die Fortsetzung zeigt die ständigen Austauschbeziehungen zwischen den Knoten und wie eine Entscheidung für einen Knoten zu einer neuen Welle von Veränderungen innerhalb des Netzes führen würde.

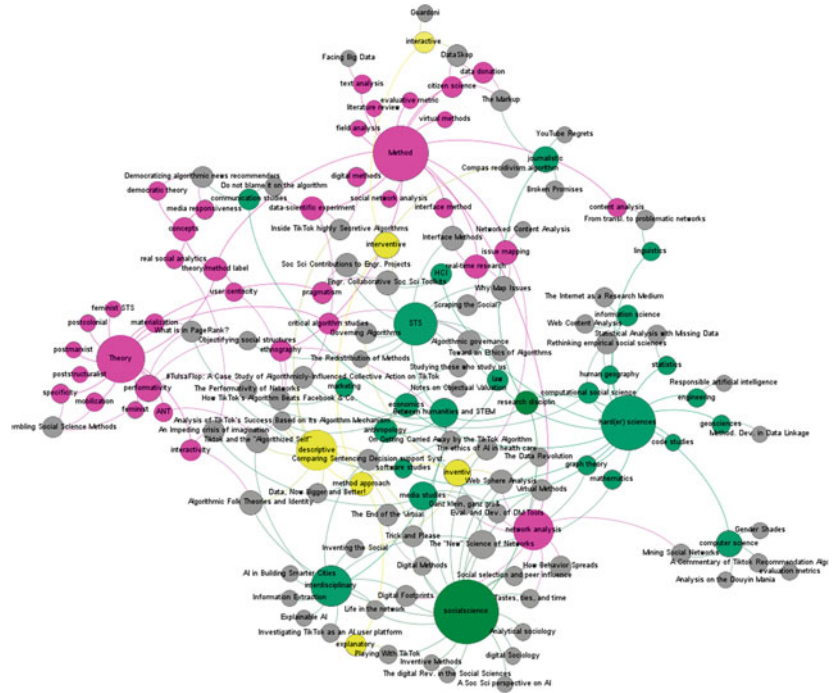
Die letzte zu diskutierende Funktion betrifft die Größe der Knoten. Obwohl sie in der Anfangsphase der Visualisierung alle die gleiche Größe haben, kann eine Größenänderung entsprechend einer spezifischen Größenfunktion der Software erfolgen. Als Teil der ‚Appearance‘-Funktion von Gephi können die Größe der Knoten und die Gewichte je nach der gewählten Option, den Bereichsgrenzen und der jeweiligen Funktion für das Größenwachstum – linear, exponentiell, polynomial oder trigonometrisch – geändert werden. Für unseren speziellen Zweck ändern wir nur die Knotengröße und legen für jede Kante eine solide Gewichtung fest. Außerdem nehmen wir als Koeffizienten für die Größe nicht nur die In-Grade oder Out-Grade, sondern wählen den ‚Degree‘, der den Durchschnitt der In-Grade und Out-Grade eines Knotens darstellt. Außerdem wählen wir auf der Grundlage der Struktur der Kategorien einen Größenbereich zwischen 16 und 64

und eine Exponentialfunktion, um schrittweise Größen zu erhalten. Die Größenfunktion als rein grafische Funktion visualisiert daher die verschiedenen Knoten besser.

### 3.3 Visualisierungen

Das Hauptziel hinter der Visualisierung ist es, ein verständliches Maß dafür zu schaffen, wie unsere gesammelte Literatur miteinander verbunden ist, und verschiedene Diskurse und Themen zusammenzubringen. Während eine ganzheitliche Karte wie in Abb. 4 eine Zusammenführung aller Kategorien und Literatur darstellt, beginnen wir unsere Kartierung und Visualisierung mit weniger komplexen Illustrationen der Beziehungen und Allianzen zwischen verschiedenen Kategorien. Auf diese Weise spielt unsere Literatur eine wesentliche Rolle, um uns bei der Navigation durch diese Verbindungen behilflich zu sein. Je mehr Literatur wir zwischen zwei Kategorien oder Codeknoten erkennen können, desto mehr Ähnlichkeiten haben sie. Der Nachteil dieses Zuordnungsansatzes ist jedoch, dass bestimmte Artikel möglicherweise keine Verbindung zu bestimmten Kategorien aufweisen. Daher haben sie einige Zuordnungspraktiken ausgelassen. Es kann auch der Fall sein, dass sie nur minimale Verbindungen aufweisen, sodass sie nur am Rande der Karte zu finden sind.

Dennoch sind die Information und das Verständnis für uns wertvoll, welche Art von Verbindungen wir finden können, indem wir verschiedene Schichten der ganzheitlichen Karte erstellen. Dabei orientieren wir uns an Literatur und Codes, die mehr oder weniger als ‚innere Knoten‘ positioniert sind. Zum Beispiel in Abb. 3, wo wir die verwandte Literatur mit den drei Kategorien ‚Theorie/Methodenbezeichnung‘, ‚Forschungsdisziplin‘ und ‚Methodenansatz‘ kartiert haben. Unser Ziel ist es, die Literatur daraufhin zu untersuchen, welche Methoden verwendet werden, um spezifische Ansätze, Innovationen und gesellschaftliche Folgen der Künstlichen Intelligenz zu erforschen. Diese Artikel sind meist mit intervenierenden, erfinderischen und deskriptiven Ansätzen verbunden – Knoten in Türkis – und befassen sich mit den Wurzeln und Hintergründen der Forschung zwischen Geisteswissenschaften und MINT (STEM), d. h. STS, Anthropologie und Kommunikationswissenschaften. Einerseits können wir auf der ‚Theorie‘-Seite sehen, dass kritische Studien zu Algorithmen, Pragmatismus, ANT und Performativität zu den bemerkenswerten Theorien gehören, die diese Karte beherrschen. Andererseits können wir bei einem Blick auf den Knotenpunkt ‚Methode‘ feststellen, dass Ethnografie, Issue Mapping, Inhaltsanalyse und Netzwerkanalyse am häufigsten als Forschungsmethoden verwendet werden.



**Abb. 3** Drei Kategorien – ‚Forschungsdisziplin‘ in Grün, ‚Theorie/Methodenbezeichnung‘ in Neon-pink und ‚Methodenansatz‘ in Gelb – werden mit Gephi abgebildet. (Quelle: Eigene Darstellung)

Eine interessante Beobachtung ist jedoch die Nähe von Methoden und Disziplinen. So ist beispielsweise die als Wissenschafts- und Technikforschung (STS) kodierte Literatur konzeptionell mit kritischen Studien zu Algorithmen, Performativität und Pragmatismus verbunden. Während jedoch Issue Mapping und Echtzeitforschung in mehreren Artikeln als methodische Ansätze verwendet werden, sind diese Artikel nicht mit einem theoretischen Hintergrund verbunden. Diese Beobachtung lässt uns ein Versagen in der allgegenwärtigen Verbindung zwischen verschiedenen Knotenpunkten erkennen.

Unser nächster Versuch, den Werkzeugkasten zu bereichern, besteht darin, zwei weitere Kategorien zusammenzufassen und die Bildung von Beziehungen zwischen ihren Codes zu beobachten. Diese beiden Kategorien sind ‚Forschungsobjekt‘ und ‚Forschungsschwerpunkt‘. Wir haben die Visualisierung bereits in

Abb. 2 zeigt. Einerseits gibt es in der Kategorie ‚Forschungsobjekt‘ Codes, die sich auf konkrete Forschungsobjekte beziehen, wie TikTok, YouTube, Medizin, Industrie 4.0. Andererseits können wir Forschungsperspektiven und -foki wie objektorientiert, kollektivorientiert, policy, design, intervention oder Diskurs sehen, die die Verteilung der Literatur verstärken. Ebenso haben wir den Ansätzen, wo ersichtlich und möglich, theoretische Affiliationen gegeben, wie etwa ‚postcolonial‘, ‚Feminist‘, oder pragmatist. Selbstverständlich sind auch methodische Kategorien, wie quantitative, qualitative, oder digitale Methoden enthalten.

Im Gegensatz zu der eher räumlichen Normalverteilung der Knoten in Abb. 3 beherrscht spezifische Literatur die Visualisierungseinstellungen, indem sie signifikante Verbindungen zu verschiedenen Knoten herstellt. Insbesondere erhalten sie ihren Einfluss, indem sie Verbindungen zu verschiedenen Kategorien herstellen. Zu diesen Artikeln gehören zum Beispiel „The ethics of AI in Healthcare“ (Morley et al. 2020) oder „How TikTok algorithm beats Facebook and Co“ (Rach und Peter 2021). Ihre Verbindung zu Forschungsobjekten wie ‚TikTok‘ und ‚digitale Daten‘ löst sie aus dem großen Kreis konkreter Forschungsobjekte heraus und bringt sie in Kontakt mit sozialen Strukturen, Plattformen und Technologien. Diese Trennung von einem der Zentren führt zur Bildung einer neuen Nachbarschaft. In Abb. 2 schlägt die Literatur in der Mitte jedoch eine Brücke zwischen den beiden Hauptkategorien und macht auf weniger beachtete Orientierungen wie Ethik, d. h. Gender Shades (Buolamwini und Gebru 2018) und erklärbare KI (Linardatos et al. 2021) aufmerksam.

Möglicherweise besteht kein direkter Zusammenhang zwischen den verschiedenen Codes der einzelnen Kategorien. Der letzte Schritt besteht jedoch darin, die Karte als Ganzes zu betrachten und zu erkennen, wie die algorithmische Kraft hinter den Elementen von Nähe, Verteilung und Größe die Karte formt. In der Folge konstruieren diese Elemente eine Örtlichkeit in der Karte, in der unverbundene Knoten zusammengefügt werden. Dies ist ein Ansatz, so störend er auch sein mag, der in der Literatur weiter untersucht werden sollte. Haben wir eine Methode, einen Fokus oder ein Objekt übersehen? Sollten wir einen anderen Blickwinkel wählen? Oder wie können wir verschiedene Literatur zusammenbringen, hervorheben und ihre Gesamtheit verstehen?

Während die obigen Karten uns die Möglichkeit geben, zu sehen, wie sich spezifische Literatur in kleineren Kategorien durchsetzt, ist es wichtig, sie als Ganzes zu sehen und sie zu den Wurzeln zurückzuverfolgen. Die ganzheitliche Karte zeigt uns, welche verschiedenen Methoden, Ansätze und Theorien an der Untersuchung eines bestimmten KI-bezogenen Gegenstands beteiligt waren, wie



sie Verbindungen und Allianzen aufbauen und wie man sich zwischen diesen Verbindungen bewegt.

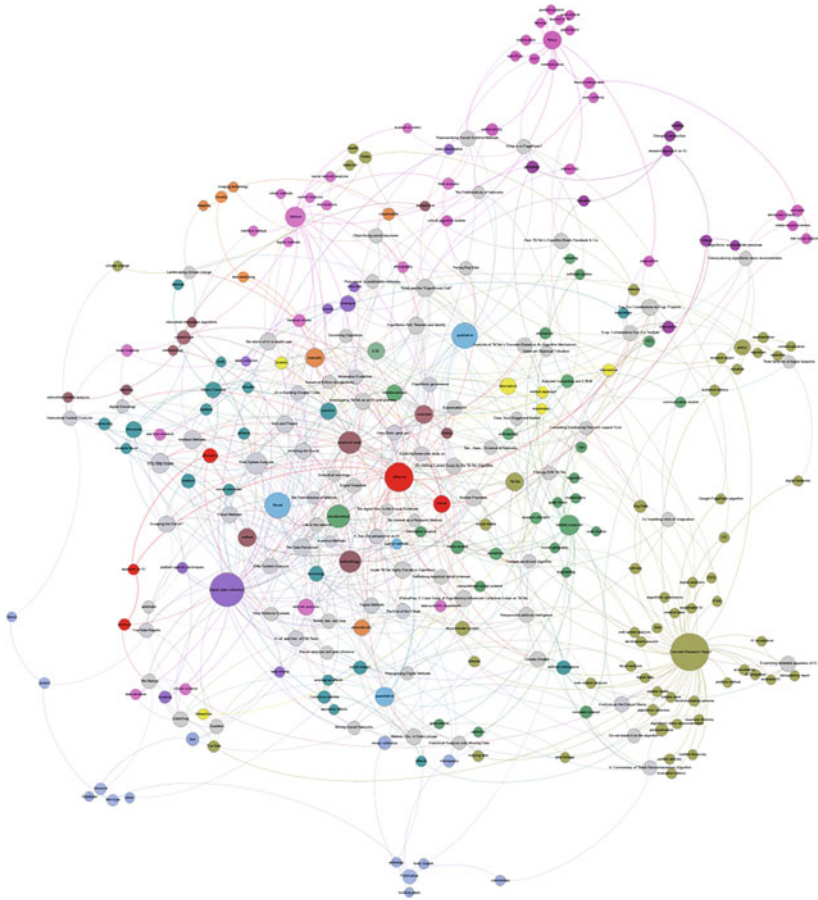
Ein wesentliches Merkmal eines Werkzeugkastens ist schließlich, dass er eine Konfiguration von Werkzeugen anbietet und sicherstellt, dass der Benutzer das Material leicht sammeln, pflegen, finden und verwenden kann. Unser Hauptbeitrag besteht darin, verwandte Literatur aus zahlreichen akademischen Datenbanken zu sammeln, ihr allgemeines Thema und ihren Ansatz zu bewerten, Datensätze zu kategorisieren und aufzubauen und verschiedene Versionen der Karte zu erstellen. Alle diese Schritte, zusammen mit spezifischen Elementen und Merkmalen der Karte, helfen uns, zu navigieren und zu zeigen, wie verschiedene Kombinationen von Kategorien Allianzen hervorheben werden, die wir vorhergesehen haben oder nicht (Abb. 4).

---

## 4 Bemerkungen, Einsichten und Beispiele

Im Folgenden stellen wir Beispiele und konzeptionelle Einsichten vor, die wir aus unserer Auseinandersetzung mit methodologischen Herausforderungen in der KI-Forschung gewonnen haben. Wir finden auch sozialtheoretische Konzepte und Methoden für deren Einsatz in Forschungsszenarien. Dass sozialwissenschaftliche Methoden für Forschungs- und Studienbemühungen in Bezug auf KI wertvoll sind, war für uns weder einer Aporie noch eine Frage ohne ‚lebendigen Zweifel‘, also ohne lebensweltlichen, praktischen Bezug. Dass KI selbst, so die stehende Annahme, keinen „living doubt“ (vgl. Peirce 1975 S. 70) empfinden kann, bedeutet nicht zwangsläufig, dass sie nicht Dinge sehen könnte, die menschlichen Beobachtern aufgrund eben dieses mangelnden Interesses verborgen bleiben. Dieser Funke des wissenschaftlichen Interesses ist aber notwendig für iterative, open end Prozesse und um kreatives Experimentieren zu ermöglichen. Das Ergebnis erschöpft sich weder in starrer Permutation noch ist es im Voraus beschreibbar: Die sozialwissenschaftliche Fähigkeit, diese Offenheit nicht nur auszuhalten, sondern an ihr zu wachsen, wurde durch die oben beschriebenen methodischen Herausforderungen nahegelegt und soll durch die folgenden Beiträge beantwortet werden, die wir als sozialwissenschaftlichen Pragmatismus verstehen.

Obwohl wir es für wichtig halten, die Erwartungen an ein methodisches Konzept oder eine Reihe von Methoden zur Lösung oder Verbesserung der KI-Forschung zu überwinden und zu ignorieren, kann die Sozialwissenschaft in der Tat einen Beitrag zur Bewältigung der oben beschriebenen Probleme bei der Erforschung von KI leisten. Es gibt eine Art Pragmatismus, der den Sozialwissenschaften inhärent ist, ein Pragmatismus, der in der autoreflexiven



**Abb. 4** Eine ganzheitliche Ansicht der Karte. Zusätzlich zu den anderen Kategorien und zuvor verwendeten Farben ist ‚Datenerhebung‘ in Violett, ‚Gegenstandstyp‘ in Aero, ‚Art der Methoden‘ in Maximal Blau, ‚Datenanalyse‘ in Orange, ‚Inhaltstyp‘ in Lila Rot, ‚KI-Ansatz‘ in Rot und ‚KI-Forschungsansatz‘ in Imperial. (Quelle: Eigene Darstellung)

Reichweite der Disziplin und ihrer Verankerung in sozialen Bezügen und kultureller Bedeutung begründet ist. Dieser sozialwissenschaftliche Pragmatismus enthält Merkmale, die helfen, die oben beschriebenen methodischen Probleme zu bewältigen. Wir nennen sie Merkmale der ‚flachen Beschreibung‘, der ‚Relationalität‘,

der ‚historischen Sensibilität‘ und der ‚multiparadigmatischen methodologischen Spezifität‘.

Unsere eigene Methode, die den semantischen Overflow der KI kompensieren und die breite Besonderheit ihrer Bereiche und Felder hervorheben sollte, ist selbst von der ‚multiparadigmatischen methodologischen Spezifität‘ der Sozialwissenschaften geprägt; nicht nur durch ihren inhärenten Eklektizismus, sondern durch ihr Verständnis mehrerer Disziplinen, Perspektiven und Methoden. Die große Bandbreite unterschiedlicher Forschungstraditionen und paradigmatischer Schulen innerhalb der Sozialwissenschaften ist kein bloßes Manko der soft sciences, Erkenntnisse in einen kohärenten Kanon zu integrieren, sondern eine Ausdifferenzierung von Wissenschaft und Forschung auf bestimmte, spezifische Gegenstände und Interessen. Ein gutes Beispiel für diese Besonderheit ist die Untersuchung ähnlicher Plattformen wie YouTube und TikTok. Obwohl sie sich recht ähnlich sind, da es sich bei beiden um Social-Media-Plattformen für nutzergenerierte Inhalte handelt, die sich auf ein Empfehlungssystem konzentrieren, unterscheiden sie sich dennoch stark hinsichtlich Nutzerpraktiken und, was noch wichtiger ist, in Bezug auf die nativen und messbaren Daten, die sie beinhalten. Sie benötigen daher unterschiedliche digitale Werkzeuge (vgl. <https://tracking.exposed/>) und unterschiedliche Ansätze (z. B. Netzwerkanalyse, Diskursanalyse, multivariate Datenanalyse). In soziotechnischen Gefügen wie solchen actor networks (Akteursnetzwerken) aus menschlichen Individuen, Maschinen, Software und Daten kann man beobachten, dass Daten nicht durch ihre Rollen und Funktionen innerhalb eines Geräts verstanden werden können, sondern als ein Produkt sozialer Interaktion, Konflikte und Koordination verstanden werden müssen, die oft stark durch den sie umgebenden organisatorischen Kontext geprägt sind (vgl. Pöchhacker im Erscheinen).

Die Merkmale der ‚flachen Beschreibung‘ (und Ontologie), der ‚Relationalität‘ und der ‚historischen Sensibilität‘ gehören gewissermaßen als Kernkompetenzen des sozialwissenschaftlichen Pragmatikers zusammen. Es geht um die Idee nicht neutraler, sondern flacher Beschreibungen, um multi-sided Ansätze, die den Verlust von one-sided Forschungsfeldern verkraften können (vgl. Marcus 1995). Ebenso vertreten dies Ansätze einer flachen Ontologie, die eher mit ontologischen Ansprüchen, Traditionen, Erwartungen und Gewohnheiten spielen (Mol 2010) und andere Ansätze, die mittels deskriptiver Reproduktion mit ihrer vorinterpretierten Welt (vgl. Schütz 1993) zu brechen suchen (vgl. z. B. Lynch 2013). Dies setzt natürlich eine Sensibilität für Kontingenz voraus und ermöglicht so eine kritische Perspektive der Geschichtlichkeit, die Pfadabhängigkeiten und selbstverständliche Vorstellungen und Wahrnehmungen auf den Prüfstand stellt und die gegebenen Erfahrungen ihren Voraussetzungen gegenüberstellt.

Ein Beispiel für ein relationales Konzept, das sich mit Mehrdeutigkeit und Objektivität in der Interaktion beschäftigt, sind „boundary objects“ (vgl. Star und Griesemer 1989). Dabei handelt es sich um Objekte, die zur disziplinären oder feldspezifischen Grenzziehung genutzt werden, aber zugleich, als Objekte, stabil genug sind, um von allen Seiten der Grenzen als solches referenziert zu werden. So kann beispielsweise Verschlüsselung von Computerexpert\*innen als eine konkrete Schwelle zur Überwindung von Sicherheitsmechanismen verstanden werden, während Alltagsnutzer\*innen darin eher ein mit Vertrauensvorschuss ausgestattetes Versprechen und eine Erwartungshaltung verbinden. Für Technologieinvestor\*innen handelt es sich um den Verwertungsaspekt eines Produkts.

Selbstverständlich ist die Diskursanalyse und -theorie ein weiteres Beispiel für einen methodologischen Zweig, der sich dem Umgang mit terminologischen Interessenkonflikten, Wissen und Ideologie widmet. Solche pragmatistischen Merkmale sind jedoch schwach, wenn es um das Verständnis von Kausalität geht, das vor allem bei der Untersuchung von KI gefragt ist – was ist der Ursprung einer Voreingenommenheit? Obwohl die von uns vorgeschlagene Methodik Fragen der Kausalität ausschließt, können sie sich mit Dringlichkeit stellen, zum Beispiel im Fall von Entscheidungsunterstützungssystemen, die Subjekte aufgrund der gegenwärtigen Tendenz zur Überwachung kriminalisieren oder sie sogar auf der Grundlage von Modellen beurteilen, die anhand von Daten aus totalen Institutionen wie Strafvollzugsanstalten trainiert wurden (vgl. Angwin et al. 2016). In diesen Fällen sind die Opfer schwer auszumachen, und selbst die Entscheidung über unabhängige und abhängige Variablen, Proxies und Mediatoren kann zu dauerhaften, spannungsgeladenen Debatten werden (vgl. Kahlert [né Müller] und Pöchhacker 2019). Die gleiche historische Sensibilität und der inhärente Relationalismus machen es den Sozialwissenschaften schwer, positivistisch zu argumentieren. Selbst wenn man es versucht, kann die Selbstreferenzialität der Sozialwissenschaft ihren konstruktiven Schlussfolgerungen nicht entkommen: Wenn die Dinge als rationaler Akt zu erklären sind, ist es auch die Erklärung selbst. Die Sozialwissenschaft kann also ihren Konstruktivismus leugnen oder vergessen, aber sie kann den Konsequenzen des berühmten Thomas-Theorems (Merton 1995) nicht entkommen. Dementsprechend betont Wolbring (2020), der auch einen prägnanten Überblick über die Auswirkungen von und auf die digitale Soziologie verfasst hat, die fortwährende Notwendigkeit einer sozialwissenschaftlichen Theorie und eines Bewusstseins, anstatt Big Data naiv in ‚Wissen‘ zu verwandeln (vgl. Anderson 2008).

Betrachten wir ein weiteres Konzept, das flache Ontologien verwendet und die Relationalität hervorhebt, wie sie im Pragmatismus und in der Praxeologie

üblich ist: stillschweigendes Wissen (tacit knowledge). KI ist ein gewöhnliches Element von Hilfsmitteln wie Entscheidungsunterstützungssystemen, die einen hohen Grad an Vollständigkeit der Wissensmodelle erfordern, die sie beinhalten. Hierarchische Ontologien, wie sie in Konzeptionen von ‚mental maps‘ oder ‚think aloud‘-Methoden integriert sind, lassen das meiste, wenn nicht das gesamte inkorporierte Wissen vermissen, was im schlimmsten Fall zu Geräten führen kann, die menschliche Fähigkeiten eher unterdrücken als fördern (vgl. Passoth und Kahlert 2018b). Obwohl dies auch für Assistenztechnologien gilt, die keine KI-Funktionen besitzen, stellt stillschweigendes, inkorporiertes Wissen eine besondere Herausforderung dar, die von digitalen KI-Systemen berücksichtigt werden muss. Dies betrifft sowohl Fragen der Schnittstellengestaltung als auch der Modellierungspraktiken und -verfahren und beweist einmal mehr, wie wichtig das Verständnis digitaler Technologien ist, wenn es darum geht, sie zu untersuchen.

Abschließend sei noch ein Hinweis auf typische Fälle des ‚Übersehens‘ von KI gegeben, obwohl nur eine gewisse Art von Paranoia sicherstellen kann, dass dieser Fehler vermieden wird: das Beispiel der datengetriebenen Netzwerkforschung in den sozialen Medien und die Annahme der Objektivität (oder interobjektiven Unabhängigkeit), die der Begriff ‚digitaler Fußabdruck‘ impliziert (vgl. Golder und Macy 2014, Lewis et al. 2008). Aber auch hier ist KI nicht unbedingt das Problem, sondern eines von vielen (vgl. Lewis 2015, Kinder-Kurlanda 2020), wenn es darum geht, die Fälschung von Daten zu ignorieren (vgl. Gitelman 2013).

---

## 5      **Outro: KI-Forschung zwischen Mustern und Einzigartigkeiten**

Zusammenfassend lässt sich sagen, dass unsere Methode weder zu einer Sättigung in Bezug auf die Codes geführt hat noch diese jemals erreichen konnte. Im Sinne der Grounded Theory finden wir jedoch eine stabile, in ihren Code-Dimensionen bzw. Kategorien gesättigte Struktur, die auch weitere Dateninputs aufnehmen und verarbeiten kann, die bei solchen Forschungen zu sozialwissenschaftlichen KI-Forschungsmethoden anfallen können. Wir sind uns bewusst, dass es viele vernachlässigte Forschungsbereiche gibt, die während unserer Datenerhebung nicht vorkamen und daher im sich entwickelnden Forschungsprozess praktisch ausgeschlossen wurden. Wir halten die Achsen, Kategorien und Einbettungsdimensionen für nützlich, um unsere Forschungsbemühungen zu erweitern und die Ergebnisse in aufschlussreichen Karten und Werkzeugkästen zu sortieren, da sie sich auf Bereiche, Objekte und Werkzeuge der Forschung, Didaktik,

Erweiterung und Disziplin beziehen. Während die Code-Genese also fortgesetzt werden kann, erlaubt unsere Kodierung bereits, die vorliegenden Daten hinsichtlich des methodischen Problems und der methodologischen Frage zu organisieren oder daraus ein brauchbares Werkzeug zu extrahieren, dessen Anwendbarkeit, wie wir es für KI allgemein postulieren, selbstverständlich für jeden Anwendungsfall einzeln zu prüfen und zu bewerten ist. Es ist eben nur ein weiterer Ausgangspunkt, da es sich um unsere spezifische Reflexion und Betrachtung von KI-Instanzierungen handelt.

Wie wir gesehen haben, gibt es keine einfache Antwort darauf, wie Phänomene der künstlichen Intelligenz erforscht werden können. Schließlich ist künstliche Intelligenz eher ein politischer Begriff, der Praxis-, Diskurs-, Interessen- und Konfliktkollektive auf den Plan rufen und deuten kann. Daher ist es nicht verwunderlich, dass die Erforschung von KI alle Arten von Disziplinen und Professionen, Forschungsobjekten und Gebieten umfasst. In diesem Artikel geben wir einen Überblick und eine Einführung in einige der Fälle, Ansätze und Perspektiven, die wir für diese Expertise gesammelt haben.

Sie bestehen aus viel mehr als den erwähnten Instanzen der Diskursanalyse, Beobachtungen, digitalen Methoden, Datenverarbeitung und statistischen Erhebungs- und Analysestrategien. Wir haben in groben Zügen die Probleme beschrieben, auf die wir gestoßen sind, als wir die riesigen, sich gegenseitig durchdringenden und relationalen Komplexe der künstlichen Intelligenz kartiert und interpretiert haben – ganz zu schweigen von dem üppigen Reichtum an analytischen und theoretischen Konzepten wie Infrastruktur, Interaktion, stillschweigendes Wissen usw., die kein monolithisches, übergreifendes Verhalten in Sachen KI-Forschung erlauben.

Die Herausforderung eines so weiten und empirisch überfrachteten Feldes erlaubt keine orthodoxen Verfahren wie die maschinelle Massenerhebung von methodologischer Literatur oder eine spezifizierte qualitative Bewertung des Standes der Technik. Zu naiv, zu unverständlich wäre Ersteres – zu klein, zu spezifisch Letzteres, und beides letztlich zu fahrlässig.

Unser Exkurs im Bereich der KI-Forschungsmethodik ist also weder als ein Bericht über den Stand der Technik zu verstehen noch als das Hinzufügen eines weiteren monolithischen Anspruchs der Methodologie, noch als die Entdeckung oder Betonung eines weiteren Paradigmenwechsels, der sich durch KI andeutet. Unser Beitrag zielt vielmehr darauf ab, diese wissenschaftlichen Vorstellungen und Erwartungen zu überwinden und sie durch eine multiparadigmatische Spezifik der KI-Forschung zu ersetzen, die sensibel ist für disziplinäre Perspektiven, konzeptionelle Vielfalt, Partikularität der Gegenstände und die (Un)-Sichtbarkeit der Fantasien der KI. Die Künstliche Intelligenz ist auf dem Vormarsch und fasst

eine Vielzahl von Ereignissen, Phänomenen, Artefakten und Praktiken zusammen. Eine akademische Didaktik und Anforderungen an die wissenschaftliche und ingenieurwissenschaftliche Forschung, die eine solche Wachsamkeit einschließt, ist die unabdingbare Voraussetzung für die Bewältigung dieser sozio-technischen Entwicklungen.

---

## Literatur

- Ananny, Mike. 2016. Toward an Ethics of Algorithms: Convening, Observation, Probability, and Timeliness. *Science, Technology, & Human Values* 41: 93–117.
- Anderson, Chris. 2008. The end of theory: The data deluge makes the scientific method obsolete. *Wired*, June 23. <https://www.wired.com/2008/06/pb-theory/>. Accessed 03 Juni 2022.
- Angwin, Julia et al. 2016. Machine Bias. *ProPublica*, May 23. <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>. Accessed 03 May 2022.
- Bandy, Jack, und Nicholas Diakopoulos. 2020. #TulsaFloP: A Case Study of Algorithmically-Influenced Collective Action on TikTok. [arXiv:2012.07716 \[cs\]](https://arxiv.org/abs/2012.07716).
- Bhandari, Aparajita, und Sara Bimo. 2020. Tiktok and the „Algorithmized Self“: a New Model of Online Interaction. *AoIR Selected Papers of Internet Research*.
- Boehm, Andreas. 1994. Grounded Theory – wie aus Texten Modelle und Theorien gemacht werden. In *Texte verstehen: Konzepte, Methoden, Werkzeuge*, hrsg. A. Mengel & T. Muhr, 121–140. München/Tübingen: UVK.
- Boellstorff, Tom, und Bill Maurer. 2015. *Data, Now Bigger and Better!*. Chicago: Prickly Paradigm Press.
- Bechmann, Anja, und Geoffrey C. Bowker. 2019. Unsupervised by any other name: Hidden layers of knowledge production in artificial intelligence on social media. *Big Data & Society* 6: 1–11.
- Buolamwini, Joy, und Timnit Gebru. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, ACM 77–91.
- Callon, Michel et al. 1983. From Translations to Problematic Networks: An Introduction to Co-Word Analysis. *Social Science Information* 22: 191–235.
- Centola, Damon. 2018. *How Behavior Spreads: The Science of Complex Contagions*. New Jersey: Princeton University Press.
- Clarke, Adele E. 2003. Situational Analyses: Grounded Theory Mapping After the Postmodern Turn. *Symbolic Interaction* 26: 553–576.
- Corbin, Juliet M., und Anselm Strauss. 1990. Grounded Theory Research: Procedures, Canons and Evaluative Criteria. *Zeitschrift für Soziologie* 19: 418–427.
- Derrida, Jacques. 1983. *Grammatologie*. Berlin: Suhrkamp.
- Derrida, Jacques. 1988. Ouisa und gramme. In *Randgänge der Philosophie*, hrsg. ebd., 57–92. Wien: Passagen.
- Durkheim, Emile. 1976. *Soziologie und Philosophie*. Berlin: Suhrkamp.

- Francisco, Mark Eugene Z., und Sonakshi Ruhela. 2021. Investigating TikTok as an AI User Platform. *2nd International Conference on Computation, Automation and Knowledge Management (ICCAKM)*, 293–298.
- Garfinkel, Harold. 1967. *Studies in Ethn methodology*. New Jersey: Prentice-Hall.
- Gießmann, Sebastian. 2009. Ganz klein, ganz groß. Jacob Levy Moreno und die Geschichte des Netzwerkdiagramms. In *Medienumbrüche; Medien in Raum und Zeit*: 267–291. Bielefeld: transcript.
- Gitelman, Lisa. 2013. *“Raw Data” Is an Oxymoron*. Cambridge: MIT Press.
- Golder, Scott A., und Michael W. Macy. 2014. Digital Footprints: Opportunities and Challenges for Online Research. *Annual Review of Sociology* 40:129–152.
- Hine, Christine. 2005. *Virtual Methods: Issues in Social Research on the Internet*. Oxford: Berg Publishers.
- Jacomy, Mathieu et al. 2014. ForceAtlas2, a Continuous Graph Layout Algorithm for Handy Network Visualization Designed for the Gephi Software. *PLoS ONE*.
- Kahlert [né Müller], Peter, und Nikolaus Pöchhacker. 2019. Algorithmic Risk Assessment als Medium des Rechts. *Medientechnische Entwicklungen und institutionelle Verschiebungen aus Sicht einer Techniksoziologie des Rechts. Österreichische Zeitschrift für Soziologie* 44: 157–179.
- Kinder-Kurlanda, Katharina. 2020. Big Social Media Data als epistemologische Herausforderung für die Soziologie. *Soziale Welt Sonderband Soziologie des Digitalen – Digitale Soziologie?*: 109–133.
- Kitchin, Robo. 2014. *The Data Revolution*. New York: Sage.
- Kornak, Jacek. 2015. Judith Butler’s Queer Conceptual Politics. *Redescriptions* 18: 52–73.
- Laclau, Ernesto. 2005. *Introduction to Marcel Mauss*. Oxfordshire: Routledge.
- Latour, Bruno. 1985. Visualization and Cognition: Drawing things together. In *Knowledge and Society Studies in the Sociology of Culture Past and Present* 6: 1–40.
- Latour, Bruno. 1993. *We have Never Been Modern*. Cambridge: Harvard University Press.
- Lewis, Kevin. 2015. Three Fallacies of Digital Footprints. *Big Data and Society* 2: 1–4.
- Lewis, Kevin et al. 2008. Tastes, Ties, and Time: A New Social Network Dataset Using Facebook.Com. *Social Networks* 30: 330–342.
- Linardatos, Pantelis, Vasilis Papastefanopoulos, und Sotiris Kotsiantis. 2021. Explainable AI: A Review of Machine Learning Interpretability Methods. *Entropy* 23: 18.
- Little, Roderick J. A., und Donald B. Rubin. 2013. *Statistical Analysis with Missing Data*. 2nd. New Jersey: Wiley.
- Lynch, Michael. 2013. Ontography: Investigating the production of things deflating ontology. *Social Studies of Science* 43:444–462.
- Marcus, George E. 1995. Ethnography in/of the World System: The Emergence of Multi-Sited Ethnography. *Annual Review of Anthropology* 24: 95–117.
- Marres, Noortje. 2015. Why Map Issues? On Controversy Analysis as a Digital Method. *Science, Technology, & Human Values* 40: 655–686.
- Marres, Noortje, und Esther Weltevrede. 2013. Scraping the Social? Issues in live social research. *Journal of Cultural Economy* 6: 313–335.
- Marres, Noortje. 2012. On Some Uses and Abuses of Topology in the Social Analysis of Technology (Or the Problem with Smart Meters). *Theory, Culture & Society* 29: 288–310.
- Merton, Robert K. 1995. The Thomas Theorem and the Matthew Effect. *Social Forces* 74: 379–424.



- Möller, Judith et al. 2018. Do Not Blame It on the Algorithm: An Empirical Assessment of Multiple Recommender Systems and Their Impact on Content Diversity. *Information, Communication & Society* 21: 959–977.
- Mol, Annemarie. 2010. Actor-Network Theory: Sensitive Terms and Enduring Tensions. *Kölner Zeitschrift für Soziologie und Sozialpsychologie* 50: 253–269.
- Morley, Jessica et al. 2020. The Ethics of AI in Health Care: A Mapping Review. *Social Science & Medicine* 260: 1–14.
- Murthy, Dhiraj et al. 2013. Evaluation and Development of Data Mining Tools for Social Network Analysis. In *Mining Social Networks and Security Informatics*, hrsg. Tansel Özyer, Zeki Erdem, Jon Rokne und Suheil Khoury, 183–202. Berlin/Heidelberg: Springer.
- Mützel, Sophie. 2015. Facing Big Data: Making sociology relevant. *Big Data & Society* 1: 1–4.
- Niederer, Sabine M. C. 2016. *Networked Content Analysis: The Case of Climate Change*. PhD Thesis, Amsterdam, The Netherlands: Amsterdam School for Cultural Analysis.
- Özyer, Tansel et al. 2013. *Mining Social Networks and Security Informatics*. Dordrecht: Springer Netherlands.
- Passoth, Jan-Hendrik, und Peter Kahlert. 2018. Engineering Collaborative Social Science Toolkits. STS Methods and Concepts as Devices for Interdisciplinary Diplomacy. In *Developing Support Systems*, hrsg. A. Karafillidis und R. Weidner, 137–145. Berlin/Heidelberg: Springer.
- Passoth, Jan-Hendrik, und Peter Kahlert. 2018. Social Science Contributions to Engineering Projects: Looking Beyond Explicit Knowledge Through the Lenses of Social Theory. In *IMPROVE – Innovative Modelling Approaches for Production Systems to Raise Validatable Efficiency. Intelligent Methods for the Factory of the Future*, hrsg. O Niggemann und P Schüller, 19–36. Berlin/Heidelberg: Springer.
- Peirce, Charles S. 1931. Pragmatism and Abduction. In *Charles Sanders Peirce. The Collected Papers Volume 5*.
- Peirce, Charles S. 1975 [1877]. Die Festlegung einer Überzeugung [The Fication of Belief]. In *Philosophie des Pragmatismus*, 61–98. Leipzig: Reclam.
- Pöchlhacker, Nikolaus. *Democratic Algorithms. Ethnography of a Public Recommender System*. Meson Press, im Erscheinen.
- Rach, Markus, und Marc K. Peter. 2021. How TikTok’s Algorithm Beats Facebook & Co. for Attention Under the Theory of Escapism: A Network Sample Analysis of Austrian, German and Swiss Users. In *Advances in Digital Marketing and eCommerce*, hrsg. F. J Martínez-López und D López López, 137–143. Berlin: Springer International Publishing.
- Reichert, Jo. 2010. Abduction: The Logic of Discovery of Grounded Theory. *Forum Qualitative Sozialforschung* 11.
- Rieder, Bernhard. 2012. What Is in PageRank? A Historical and Conceptual Investigation of a Recursive Status Index. *Computational Culture* 2.
- Rogers, Richard. 2009. *Digital Methods*. Cambridge: MIT Press.
- Rogers, Richard, und Noortje Marres. 2000. Landscaping climate change: a mapping technique for understanding science and technology debates on the World Wide Web. *Public Understanding of Science* 9: 141–163.
- Schütz, Alfred. 1993. *Der sinnhafte Aufbau der sozialen Welt. Eine Einleitung in die verstehende Soziologie*. Berlin: Suhrkamp.
- Scott, John. 2017. *Social Network Analysis*. New York City: SAGE Publications Ltd.

- Searle, John R. 1980. Minds, brains, and programs. *Behavioral and Brain Sciences* 3: 417–424.
- Star, Susan Leigh, und James R. Griesemer. 1989. Institutional Ecology, ‘Translations’ and Boundary Objects: Amateurs and Professionals in Berkeley’s Museum of Vertebrate Zoology, 1907–39. *Social Studies of Science* 19(3). 418–427.
- Strauss, Anselm L. 1998. *Grundlagen qualitativer Sozialforschung*. Stuttgart: UTB.
- Tucker, E. 2022. Artifice and Intelligence. *medium.com*, März 08. <https://medium.com/center-on-privacy-technology/artifice-and-intelligence1-f00da128d3cd>. Accessed 03 June 2022.
- van Es, Karin. 2017. An Impending Crisis of Imagination: Data-Driven Personalization in Public Service Broadcasters. Working Paper.
- Wall Street Journal. 2021. Investigation: How TikTok’s Algorithm Figures Out Your Deepest Desires. <https://www.wsj.com/video/series/inside-tiktoks-highly-secretive-algorithm/investigation-how-tiktok-algorithm-figures-out-your-deepest-desires/6C0C2040-FF25-4827-8528-2BD6612E3796>.
- Weltevrede, Esther. 2016. *Repurposing digital methods*. PhD Thesis, Amsterdam, The Netherlands: Amsterdam School for Cultural Analysis.
- Wikipedia. 2021. *Toolbox – Wikipedia*. <https://en.wikipedia.org/wiki/Toolbox>. Accessed 02 September 2022.
- Wolbring, Tobias. 2020. The Digital Revolution in the Social Sciences. *Soziale Welt. Sonderband: Soziologie des Digitalen – Digitale Soziologie?*: 60–72.
- Yigitcanlar, Tan et al. 2020. Contributions and Risks of Artificial Intelligence (AI) in Building Smarter Cities: Insights from a Systematic Review of the Literature. *Energies* 13: 1–36.