Sabine Ammon
Birgit Beck
Christoph Benzmüller
Aljoscha Burchardt
Marie Lena Heidingsfelder
Simone Kaiser
Bertram Lomfeld
Rainer Mühlhoff
Peter Remmers
Martina Schraudner

3|2020

#VerantwortungKI – Künstliche Intelligenz und gesellschaftliche Folgen

KI als Laboratorium? Ethik als Aufgabe!

Eine Schriftenreihe der interdisziplinären Arbeitsgruppe
Verantwortung: Maschinelles Lernen und Künstliche Intelligenz

berlin-brandenburgische
AKADEMIE DER WISSENSCHAFTEN

Berlin-Brandenburgische Akademie der Wissenschaften (BBAW)

KI ALS LABORATORIUM? ETHIK ALS AUFGABE!

KI ALS LABORATORIUM ?
ETHIK ALS AUFGABE !

Sabine Ammon
Birgit Beck
Christoph Benzmüller
Aljoscha Burchardt
Marie Lena Heidingsfelder
Simone Kaiser
Bertram Lomfeld
Rainer Mühlhoff
Peter Remmers
Martina Schraudner

#VERANTWORTUNG KI – KÜNSTLICHE INTELLIGENZ
UND GESELLSCHAFTLICHE FOLGEN 3|2020

Herausgeberin: Interdisziplinäre Arbeitsgruppe *Verantwortung: Maschinelles Lernen und Künstliche Intelligenz* der Berlin-Brandenburgischen Akademie der Wissenschaften.

Redaktion: Isabella Hermann und Ute Tintemann
Grafik: Thorsten Probst/angenehme gestaltung
Druck: bud Brandenburgische Universitätsdruckerei und Verlagsgesellschaft Potsdam mbh

Jägerstraße 22–23, 10117 Berlin, www.bbaw.de

ISBN: 978-3-939818-93-9

INHALTSVERZEICHNIS

AUTORINNEN UND AUTOREN

Sabine Ammon: Professorin für Wissensdynamik und Nachhaltigkeit in den Technikwissenschaften und Leiterin des Berlin Ethics Lab, Institut für Werkzeugmaschinen und Fabrikbetrieb sowie Institut für Philosophie, Literatur-, Wissenschafts- und Technikgeschichte an der Technischen Universität Berlin.*

Birgit Beck: Juniorprofessorin und Leiterin des Fachgebiets Ethik und Technikphilosophie an der Technischen Universität Berlin.

Christoph Benzmüller: Professor am Dahlem Center for Machine Learning and Robotics der Freien Universität Berlin.

Aljoscha Burchardt: Research Fellow und stellvertretender Standortsprecher des Deutschen Forschungszentrums für Künstliche Intelligenz (DFKI) in Berlin.

Marie Lena Heidingsfelder: Teamleiterin, Fraunhofer Center for Responsible Research and Innovation (CeRRI) in Berlin.

Simone Kaiser: Stellvertretende Leiterin des Fraunhofer Center for Responsible Research and Innovation (CeRRI) in Berlin.

Bertram Lomfeld: Juniorprofessor für Privatrecht und Grundlagen an der Freien Universität Berlin.

Rainer Mühlhoff: Postdoktorand am Excellence Cluster „Science of Intelligence" an der Technischen Universität Berlin.

Peter Remmers: Wissenschaftlicher Mitarbeiter am Institut für Philosophie, Literatur-, Wissenschafts- und Technikgeschichte an der Technischen Universität Berlin.

Martina Schraudner: Professorin für Gender und Diversity in der Technik und Produktentwicklung an der Technischen Universität Berlin; Leiterin des Fraunhofer Centers for Responsible Research and Innovation (CeRRI) in Berlin.

* Mitglied der interdisziplinären Arbeitsgruppe *Verantwortung: Maschinelles Lernen und Künstliche Intelligenz* der Berlin-Brandenburgischen Akademie der Wissenschaften.

VORWORT

Künstliche Intelligenz (KI) verspricht viele Vorteile, die unser Leben erleichtern können: Die Technologie hätte ansonsten kaum eine so große Aufmerksamkeit erfahren, wie es in den letzten Jahren der Fall war. Doch konkrete KI-Anwendungen wie beispielsweise das Scoring von Personen, um die Kreditwürdigkeit oder Reputation zu bestimmen, laufen auch Gefahr, dass Menschen ungerechtfertigt diskriminiert werden oder zu Schaden kommen. Gewichtige Probleme hierbei sind, dass die Verfahrensweisen von algorithmischen Entscheidungssystemen nicht oder nur schwer nachzuvollziehen sind. Zudem ist nicht klar geregelt, wer die ethische und rechtliche Verantwortung bei der Verwendung solcher Systeme trägt. Und wer kann und muss den Einsatz automatisierter Entscheidungssysteme überprüfen und regulieren? Diesen Fragen gehen die Mitglieder der interdisziplinären Arbeitsgruppe (IAG) „Verantwortung: Maschinelles Lernen und Künstliche Intelligenz" der Berlin-Brandenburgischen Akademie der Wissenschaften nach.

Und diese Fragen beziehen sich selbstredend nicht nur auf das jeweilige Individuum, sondern auf die Gesellschaft als Ganzes. Könnten nicht systematische Ungerechtigkeiten, die sich in die vermeintlich objektiven Maschinen eingeschlichen haben, ganze gesellschaftliche Gruppen benachteiligen? Durchbrüche im Bereich von KI werden oft als disruptiv und transformativ gefeiert, als Erneuerungen, die die Märkte auf den Kopf stellen und die Wirtschaft grundlegend verändern. Doch dass die Anwendung solcher Systeme immer in ein soziotechnisches System eingebettet ist, wird leider viel zu häufig vernachlässigt. Insbesondere mit den gesellschaftlichen Herausforderungen von KI-Anwendungen befasst sich das Berlin Ethics Lab (BEL) an der Technischen Universität Berlin, das unser IAG-Mitglied Sabine Ammon leitet. Das BEL und unsere IAG eint dabei in herausragender Weise die Überzeugung, dass die ethische Auseinandersetzung mit KI-Systemen in einer interdisziplinären Umgebung stattfinden muss.

Das BEL versteht sich dabei als Labor, Zusammenschluss und Plattform für Expertinnen und Experten unterschiedlichster Disziplinen von den Geistes- bis zu den Computerwissenschaften. Es beschäftigt sich mit den sozialen Anforderungen von KI-Anwendungen und es hat das Ziel, vor allem praktikable Lösungsansätze vom Beginn der Entwicklung an in die Technik zu implementieren. In unserer aktuellen dritten, in Kooperation mit den BEL entstandenen Ausgabe der Publikationsreihe „#VerantwortungKI – Künstliche Intelligenz und gesellschaftliche Folgen" mit dem Titel „KI als Laboratorium? Ethik als Aufgabe!" setzen sich die Autorinnen und Autoren mit verschiedenen Fragestellungen rund um verantwortungsvolle KI und guter Mensch-Maschine-Interaktion auseinander.

Den Anfang macht Sabine Ammon mit ihrem Beitrag „Ethical Vision Design im Berlin Ethics Lab: Technologievisionen in der Entwicklung verantwortlicher KI und verantwortlicher Mensch-Maschine-Interaktion". Sie zeigt, wie das BEL dazu beitragen will, dass KI-Systeme so gestaltet werden, dass sie den gesellschaftlichen Zusammenhalt befördern. In dem darauffolgenden Text „Ethische Aspekte der Mensch-Maschine-Interaktion" geht Peter Remmers spezifisch auf den Ansatz des BEL hinsichtlich der Mensch-Roboter-Interaktion ein. Er betont, dass ethische Standards ganz konkret in die Gestaltung von Technologie übersetzt werden müssen, um keine Lippenbekenntnisse zu bleiben. Birgit Beck und Aljoscha Burchardt entschlüsseln im Anschluss in ihrem Text „Alle reden von ethischer KI – aber was meinen sie damit?" mögliche Missverständnisse durch unterschiedliche Begriffsverwendungen, wenn von „ethischer KI" die Rede ist. Christoph Benzmüller und Bertram Lomfeld schlagen in ihrem Beitrag „Träumen vernünftige Maschinen von Gründen? Eine reale Utopie" als eine Lösung für fehlende Transparenz von KI-Systemen vor, dass Systeme sozial akzeptable Gründe für ihre Entscheidungen kommunizieren sollten. In seinem Text zum Thema „Prädiktive Privatheit: Warum wir alle ‚etwas zu verbergen haben'" zeigt Rainer Mühlhoff, dass im Zeitalter von Big Data das individualistische Datenschutz-Prinzip an seine Grenzen stößt und erweitert werden sollte. Zum Schluss plädieren Marie Lena Heidingsfelder, Simone Kaiser und Martina Schraudner in ihrem Beitrag „If you want to go far, go together: Gesellschafts-Foresight und Zukunftsbilder als Schlüssel für verantwortliche KI-Gestaltung" für einen stärkeren Einbezug der Zivilgesellschaft im Hinblick auf verantwortliche KI-Gestaltung durch Foresight-Methoden.

Abschließend möchten wir nicht nur den Mitgliedern unserer interdisziplinären Arbeitsgruppe für ihr Engagement danken, sondern vor allem auch den Autorinnen und Autoren rund um das Berlin Ethics Lab für ihre überaus inspirierenden Beiträge

in dieser Ausgabe. Mit dem Heft möchten wir an der Diskussion um gemeinsame Werte und Normen mitwirken, die sicherstellen, dass wir jetzt und in der Zukunft die Vorteile von Künstlicher Intelligenz nutzen können, ohne dass Menschen benachteiligt werden oder zu Schaden kommen.

Christoph Markschies
Sprecher der interdisziplinären Arbeitsgruppe „Verantwortung: Maschinelles Lernen und Künstliche Intelligenz" und Präsident der BBAW

Isabella Hermann
Wissenschaftliche Koordinatorin der interdisziplinären Arbeitsgruppe „Verantwortung: Maschinelles Lernen und Künstliche Intelligenz" der BBAW

Sabine Ammon

ETHICAL VISION DESIGN IM BERLIN ETHICS LAB: TECHNOLOGIEVISIONEN BEI DER ENTWICKLUNG VERANTWORTLICHER KI UND VERANTWORTLICHER MENSCH-MASCHINE-INTERAKTION

Berliner Zeitung, 12. Juli 2035. Rassistischer Rettungsroboter RESCUE-3 im Einsatz der Berliner Feuerwehr.

Eine vom Bundesministerium für Gleichstellung und Diversität beauftragte wissenschaftliche Studie bescheinigt es nun offiziell, was Insider schon lange vermutet haben. Der intelligente Rettungsroboter RESCUE-3 zeigt diskriminierendes Verhalten bei seinen Einsätzen. Eine Auswertung von Großbränden der letzten fünf Jahre bestätigt, dass der selbstlernende Roboter zunehmend ethnische Minderheiten in seiner Erkennungssoftware ignoriert. Damit verschlechterten sich die Chancen auf Rettung für Bevölkerungsgruppen mit dunkler Hautfarbe drastisch.

Dieser schockierende Befund ist ein schwerer Schlag für den Berliner Exzellenzverbund. Das einstige Vorzeigeprojekt ging aus einer Public-Private-Partnership hervor und stand wie keine andere Technologieentwicklung der letzten zwei Jahrzehnte für die internationale Konkurrenzfähigkeit des deutschen Wissenschafts- und Wirtschaftsstandorts. Als Feuertaufe und Durchbruch galt der Großbrand der Berliner Charité im Dezember 2030, bei dem RESCUE-3 einen Großteil der Patient*innen in Sicherheit bringen konnte.

Der Exportschlager RESCUE-3 verzichtet im Betrieb bewusst auf datenschutzrechtlich problematische Cloudlösungen und setzt auf ein kontinuierliches, lokales Training seiner Künstlichen Intelligenz. Doch genau das wurde für den einst gefeierten Algorithmus zur Identifizierung von Menschen unter Brandbedingungen zum Problem. Der verzerrte Datensatz der Berliner Realsituation führte dazu, dass der Roboter hellhäutige Menschen immer besser erkennen konnte.

Große Durchbrüche in der Künstlichen Intelligenz versprechen tiefgreifende Innovationen, die in den nächsten Jahren immer weiter in Anwendungsfelder vordringen. Kaum ein Industriezweig, der nicht vom Erneuerungsschub des Maschinellen Lernens und der Datenbankanalyse profitieren und Milliardengewinne erzielen soll, wenn die Prognosen führender Marktforschungsunternehmen zutreffen. Schon jetzt ist deutlich, dass diese Technologien tief in das Sozialgefüge eingreifen werden. Durch ihren Einfluss auf unsere sozialen Beziehungen und die gesellschaftliche Dynamik gelten Technologien im Bereich der Künstlichen Intelligenz und der intelligenten Mensch-Maschine-Interaktion als sozial disruptiv. Sie haben das Potential, eine erhebliche normative Wirkung zu entfalten. Ob die derzeit technologiegetriebene Entwicklung Produkte hervorbringen wird, die sich zum Wohle der Gesellschaft auswirken, ist nicht selbstverständlich. Wie kann sichergestellt werden, dass die Ausgestaltung der intelligenten Systeme im Einklang mit unseren gesellschaftlichen Werten stattfindet?

Wer versucht, diese Frage zu beantworten, befindet sich in einem Dilemma.[1] Wir können heute wenig darüber sagen, wie domänenspezifische Technologien in zehn, 20 oder 30 Jahren aussehen werden und welche gesellschaftlichen Auswirkungen zu erwarten sind. Doch heute, in der Frühphase der Entwicklungen, gibt es die größten Gestaltungsspielräume. Gerade jetzt könnten wir besonders gut eingreifen, um unerwünschte Folgen abzufangen. Sind Entwicklungspfade erst einmal gesetzt, Investitionen getätigt und neue Technologien in der Anwendung, gibt es aufgrund unumkehrbarer Grundsatzentscheidungen oder hoher Kosten kaum noch Steuerungsmöglichkeiten. Diese Lehre aus 50 Jahren Technikfolgenabschätzung macht es umso dringlicher, Verfahren zu entwickeln, die eine ethische Mitgestaltung bereits in der Grundlagenforschung und in der frühen Technologieentwicklung erlauben.

Wie können wir auf diese methodische Herausforderung reagieren? Zwar fehlen in frühen Entwicklungsstufen neuer Technologien konkrete Produkte und Anwendungen, deren Auswirkungen wir abschätzen könnten. Was aber diese Phasen prägt, sind Ideen, Konzepte und Vorstellungen. Ihre Wirkung auf die Ausprägung der Technologien können nicht hoch genug bewertet werden. Von der Raumfahrt bis zur Nutzung der Atomkraft kann die Technikgeschichte auf viele Beispiele verweisen, in denen Visionen zu Treibern der Technikentwicklung wurden.

1 Die Grundform dieses Dilemmas wird auf David Collingridge zurückgeführt (David Collingridge (1980): The Social Control of Technology. London: Pinter).

Sie aufzudecken, kritisch zu diskutieren und zu bewerten, ist der Schlüssel für eine aktive ethische Gestaltung und die Eröffnung neuer Entwicklungsoptionen. Kehren wir zum Eingangsszenario zurück. Hier führte das kontinuierliche Lernen im Einsatz zu einer Verhaltensänderung des Roboters, die gegen ein zentrales Prinzip unseres Zusammenlebens verstößt: der Gleichbehandlung von Menschen unabhängig von ihrer ethnischen Herkunft. Wie hätte verhindert werden können, dass der Rettungsroboter ein derart diskriminierendes Verhalten entwickelt? Sobald wir diese Frage stellen, können wir gezielt nach neuen Entwicklungsoptionen fragen. Muss der Lernprozess überwacht und angeleitet werden, damit die Gleichbehandlung dauerhaft gewährleistet ist? Können regelmäßige Tests sicherstellen, dass das Roboterverhalten unseren gesellschaftlichen Werten entspricht? Sollten ethische Normen direkt in das System implementiert werden? Einen derartigen Frageprozess in Gang zu setzen, lässt die Analyse der Vision zu einem wirkungsvollen Werkzeug werden. In der Suche nach Antworten können nun geeignete Entwicklungspfade erarbeitet werden, um eine spätere, ungewollte Diskriminierung zu verhindern.

Die Implementierung von ethischen Normen, die gezielte Überwachung und Anleitung des Lernprozesses oder regelmäßiges Testen sind Maßnahmen, die auf der Ebene der Algorithmen angesiedelt sind. Doch die Analyse der Visionen erschließt noch weitere Betrachtungsebenen. Denn Visionen erlauben, den Blick auf neue Technologien zu weiten, wodurch tieferliegende Problemstellungen in den Fokus rücken. Das Szenario des Rettungsroboters setzt eine Annahme voraus, die alles andere als selbstverständlich ist: Ein Rettungsroboter ersetzt das Personal der Feuerwehr und bringt die Bewohnerinnen und Bewohner aus dem Gebäude. Doch sollte das Design der intelligenten Maschine überhaupt so ausgerichtet sein, dass es menschliche Akteure ersetzt? Oder wäre die Orientierung der Maschinenentwicklung an einem unterstützenden Werkzeug, bei dem die Hoheit über die Rettungsstrategie und die Auswahl der zu Rettenden beim Menschen verbleibt, das eigentlich Wünschenswerte? Wie sollte die Arbeitsteilung zwischen Menschen und Maschine aussehen? Fragen dieser Art sind auf der Ebene der Beziehung zwischen Menschen und intelligenten Maschinen angesiedelt, die in ihrer Interaktion komplexe soziotechnische Milieus ausbilden.

Noch grundsätzlichere Fragen stellt die Visionenanalyse auf der Ebene der gesellschaftlichen Problemstellung. Sie versucht, das Ausgangsproblem in die Betrachtung zurückzuholen und kritisch zu hinterfragen. Was geht dem Ziel, einen Rettungsroboter zu entwickeln, voraus? Speist es sich aus dem Wunsch zu

verhindern, dass Menschen in Gebäuden verbrennen? Der Rettungsroboter des Szenarios tritt auf den Plan, wenn der verheerende Großbrand bereits um sich greift. Eine tiefergehende Rahmung des Entwurfsproblems erlaubt es, neue Stellschrauben zu erkennen. Wie müsste das Gebäude beschaffen sein, damit sich ein Brand nicht weiter ausbreiten kann oder erst gar nicht entsteht? Bedarf es dazu smarter Gebäude, neuer intelligenter Materialien oder einfach nur intelligenter Entwurfswerkzeuge, die menschliche Planungsfehler verhindern?

Entscheidend ist es, die Analyse der Visionen direkt in den Entwicklungsprozess von Technologien einzubetten. Ethische Überlegungen können so direkt bei der Präzisierung der Anforderungen und Spezifikationen einfließen. Dazu braucht es einen engen interdisziplinären Zusammenschluss zwischen technikwissenschaftlichem Domänenwissen und ethischem Reflexionswissen. Um langfristige Perspektiven und eine Vielfalt der Sichtweisen zu integrieren, braucht es zugleich einen transdisziplinären Entwicklungsansatz, der gesellschaftliche Akteure durch Formen des Co-Designs partizipativ mitgestalten lässt. Die Folgen von Technologien können im Jahr 2030 andere sein als im Jahr 2040 oder 2060, ebenso wie die Ausprägungen der Technologie sich in unterschiedlichen kulturellen Kontexten deutlich unterscheiden können. Insbesondere dort, wo KI-Technologien den Charakter von Infrastrukturen annehmen, werden sie eine lange Wirkungszeit entfalten. Das Ziel muss es sein, sozial robuste und resiliente Technologien zu entwickeln, die zugleich so flexibel sind, dass sie gesellschaftlichen Wandel ermöglichen und angemessen auf sich wandelnde Gesellschaften reagieren. Den sozialen Zusammenhalt und das Gemeinwohl zu stärken, spielt hierbei eine zentrale Rolle. Schlecht gestaltete Formen der Künstlichen Intelligenz und der Mensch-Maschine-Interaktion drohen, den gesellschaftlichen Zusammenhalt auszuhöhlen. Einrichtungen wie das Berlin Ethics Lab (BEL) können helfen, dieses disruptive Potential zu erkennen und ungewollte Technikfolgen zu verhindern. Je früher mögliche Probleme in den Blick geraten, desto größer ist der Spielraum für Veränderungen, um die Entwicklung in eine gesellschaftlich erwünschte Richtung zu lenken. Insbesondere das projektbegleitende Vorgehen des *Ethical Vision Design* erlaubt die frühe Orientierung des Entwicklungsprozesses an gesellschaftlichen Werten und dem Gemeinwohl. Dabei bauen in iterativen Schritten die Analyse von Visionen, ethische Interventionen, die Ausarbeitung neuer Entwicklungsoptionen sowie ethisches Testen aufeinander auf. Integriert in ein umfassenderes Portfolio an Maßnahmen, kann *Ethical Vision Design* ein wichtiger Baustein für die aktive Gestaltung von Zukunft im Einklang mit unseren gesellschaftlichen Werten werden.

Postskriptum: Zum Großbrand der Berliner Charité kam es 2030 nicht. Durch massive Investitionen in intelligente Entwurfswerkzeuge fanden in den 2020er Jahren umfangreiche Baumaßnahmen zur Verhinderung von Feuerausbrüchen im Universitätsklinikum statt. Unterstützt durch Smart Building Technologien auf der Basis intelligenter Materialien werden seitdem Brandgefahren frühzeitig erkannt und unmittelbar eingedämmt. Durch die frühe Einbettung ethischer Überlegungen in den Entwicklungsprozess des Rettungsroboters RESCUE-unlimited gelang es, veränderte Designpfade zu implementieren. Gut zu wissen, dass der Berliner Feuerwehr ein intelligenter und nicht-diskriminierender Rettungsroboter im Notfall zur Verfügung steht, aber nicht gebraucht wird.

Peter Remmers

ETHISCHE ASPEKTE DER MENSCH-MASCHINE-INTERAKTION

MENSCH-ROBOTER-INTERAKTION

Im weiten Feld der Digitalisierung sind es die unmittelbaren Interaktionen mit digitalen technischen Artefakten, die besonders neuartig und für jeden Einzelnen direkt erfahrbar sind. Seien es die Tipp- und Wischbewegungen der Smartphones, die sich in vergleichsweise kurzer Zeit zu geläufigen Gesten der Steuerung von Bildschirminhalten entwickelt haben oder die Gepflogenheiten der Online-Videokonferenzen, die sich in der aktuellen Pandemie-Situation aus dem Bedürfnis nach sozialen Kontakten bei physischer Distanz etablieren: Die einfach verfügbare Vielfalt und schnelle Anpassbarkeit digitaler Medien und virtueller Inhalte ermöglichen komplexe Interaktionen mit und durch Technologien.

Ethische und philosophische Aspekte der Mensch-Maschine-Interaktion bilden einen Schwerpunkt im Berlin Ethics Lab (BEL). Besonders eine gewissermaßen „hybride" Form der Mensch-Maschine-Interaktion erfährt in der Öffentlichkeit und im BEL besondere Aufmerksamkeit: Die Mensch-Roboter-Interaktion (MRI). „Hybrid" ist diese Interaktion in zweierlei Hinsicht: Erstens sind Roboter software-gesteuerte „Bewegungsautomaten",[1] die in der einen oder anderen Form physisch präsent sind. Hier verbindet sich die tendenziell als immateriell und entsprechend plastisch vorgestellte Welt des Digitalen mit der greifbaren Wirklichkeit mechanischer Materialität. Die Interaktion mit Robotern ist daher nicht unbedingt auf die Informationsvermittlung per Bildschirm oder auf andere Ein- und Ausgabegeräte angewiesen. Ein interagierender Roboter ist vielmehr ein verkörpertes Programm, dessen interne Datenverarbeitung sich unmittelbar in Bewegungen und Manipulation äußert – vergleichbar mit Handlungen und Verhaltensweisen.

1 Vgl. die Definition des (Industrie-)Roboters gemäß der VDI-Richtlinie 2860: „Industrieroboter sind universell einsetzbare Bewegungsautomaten mit mehreren Achsen, deren Bewegungen hinsichtlich Bewegungsfolge und Wegen bzw. Winkeln frei (d. h. ohne mechanischen bzw. menschlichen Eingriff) programmierbar und gegebenenfalls sensorgeführt sind. Sie sind mit Greifern, Werkzeugen oder anderen Fertigungsmitteln ausrüstbar und können Handhabungs- und/oder Fertigungsaufgaben ausführen."

Zweitens orientiert sich die Entwicklung von Mensch-Roboter-Interaktionen tendenziell an geläufigen und eingespielten Interaktionen zwischen Menschen (oder auch zwischen Mensch und Tier). Hier überlagert sich die Steuerung von Maschinen mit den Handlungen, Verhaltensweisen und Kommunikationsmitteln, die wir von zwischenmenschlichen Interaktionen kennen. Das hat im Idealfall den Vorteil, dass sich entsprechende technische Lösungen nicht mehr nur an Expert*innen oder geschulte Anwender*innen richten. Man muss sich (im Idealfall) nicht erst mit einer Bedienungsanleitung auseinandersetzen, sondern kann an geläufige Handlungsmuster und vertraute Kommunikationsweisen anknüpfen. Diese Vorstellung geht auf den erfolgreichen Ansatz einer ‚intuitiven' Benutzerschnittstelle zurück, die Technologien wie beispielsweise Smartphones und Navigationsassistenten so erfolgreich gemacht hat.[2]

Zur Bezeichnung dieser besonderen Form der Interaktion mit Robotern hat sich inzwischen der Begriff der *Kollaboration* etabliert, um sie von anderen Mensch-Maschine-Interaktionen wie z. B. der Programmierung oder der Fernsteuerung zu unterscheiden. Die technischen Herausforderungen für die Robotik sind enorm und werden in interdisziplinären Projekten zur „human-robot-interaction" (HRI) erforscht, zumeist in spezialisierten HRI-Labs. Neben der technischen Realisierung sind aktuell noch schwierige sicherheitsrelevante und nicht zuletzt rechtliche Anforderungen zu erfüllen. An diesen offenen Punkten wird allerdings intensiv gearbeitet, und es besteht kein Grund zu der Annahme, dass es sich um unüberwindbare Schwierigkeiten handelt. Welche Rolle spielt nun aber die Ethik für die Mensch-Roboter-Interaktion?

WAS BEDEUTET HIER „ETHIK"?

Um die Herangehensweise des BEL an die Ethik der MRI zu erläutern, muss ich etwas weiter ausholen, denn hier hängt viel vom Ethik-Verständnis ab. Natürlich soll jede Technologie irgendwie „ethisch" sein, alles andere führt mindestens zu einem schlechten Ruf. Aber viele Selbstverständlichkeiten in der Rede über „Ethik"

2 „Intuitiv" bedeutet in diesem Kontext allerdings nicht „ohne Vorkenntnisse bedienbar". Vielmehr werden weit verbreitete implizite Fähigkeiten und Kompetenzen durch das Design der Technologie abgerufen. Selbstverständlich hat dieser Ansatz auch Grenzen, z. B. für Menschen, die nicht mit dem Bedienparadigma der Technik seit Fernseher, Videorekorder und HiFi-Anlage vertraut sind.

bleiben oft unausgesprochen; nicht zuletzt deswegen schleichen sich immer wieder Übervereinfachungen, Missverständnisse und auch Etikettenschwindel ein.[3]

Ein Problem besteht beispielsweise darin, dass in ethischen Überlegungen häufig die „harten" Auswirkungen einer Technologie in den Vordergrund gerückt werden. Das sind in erster Linie drohende Gefahren für Leib und Leben. Beispielsweise wurden in den letzten Jahren ausführlich bestimmte moralische Dilemmata diskutiert, die sich beim Betrieb automatisierter Fahrzeuge im Straßenverkehr ergeben könnten. In den entsprechenden Gedankenexperimenten geht es üblicherweise um Personenschäden.[4] Auf den ethischen Wert der körperlichen Unversehrtheit können wir uns problemlos einigen (bis hin zur Überführung in geltendes Recht); wir zweifeln auch keinen Moment daran, dass ein kollaborativer Roboter nicht unmittelbar Personen verletzen darf. Aber auch wenn diese Diskussion einen ethischen Kerngedanken plakativ hervorhebt, so lenkt sie zugleich von vielen ungleich schwierigeren und dringenden ethischen Fragen ab. Denn im Vergleich zu den harten Fakten der existenziellen Extreme behandelt Ethik in weiten Teilen ein „weiches" Themenfeld, das – trotz der inzwischen vorangeschrittenen philosophischen und institutionellen Professionalisierung – erheblichen Spielraum für verschiedene Perspektiven und Interpretationen bietet.[5] Was ist beispielsweise mit Werten, deren konkrete Realisierungen in der MRI nicht so klar auf der Hand liegen? Welche Handlungs-, Verhaltens- und Denkmuster werden durch Interaktionen mit Robotern unterstützt, gehemmt oder gar unmöglich gemacht? Wie verändert sich das alltägliche Leben, die Wahrnehmung der Welt und das Miteinander, wenn Kollaborationen mit Robotern zur Normalität werden?

Offene Fragen wie diese lassen sich mit einer bekannten Formulierung zusammenfassen: Es geht darum zu bestimmen, wie wir mit Robotern arbeiten und leben wollen. Die Antworten darauf sind notwendigerweise unsicher, tastend, vorläufig – aber nur so können gesellschaftliche Bedürfnisse nach Orientierung, Selbstbestimmung und (Mit-)Gestaltung der Technikentwicklung aufrichtig angesprochen werden. Zugleich wird sich ethische Arbeit in diesem Sinne nicht ganz ohne bestimmte politische Positionierungen durchführen lassen, denn bereits die

3 Vgl. hierzu auch den Beitrag von Birgit Beck und Aljoscha Burchardt „Alle reden von ethischer KI – aber was meinen sie?" in diesem Band.

4 Awad, E et al. (2018): The Moral Machine Experiment. In: Nature 563, Nr. 7729 (November 2018), S. 59–64, DOI: 10.1038/s41586-018-0637-6.

5 Van der Burg, S; Swierstra, T (2013): Ethics on the Laboratory Floor. Basingstoke: Palgrave Macmillan.

Idee, in den eigengesetzlichen Verlauf der technologischen Entwicklung einzugreifen, muss auch politisch gewollt und gefördert werden.

Bevor wir also Fragen danach stellen können, welche konkreten Auswirkungen Mensch-Roboter-Kollaborationen haben werden, brauchen wir zuerst eine klare Vorstellung davon, wo wir genau suchen müssen. Denn für eine verantwortungsvolle Gestaltung der technologischen Zukunft müssen Interaktionen bewertet werden, obwohl wir bisher kaum Erfahrungen damit machen können. Im BEL werden daher Methoden und Formate entwickelt und erprobt, um Phänomene zu beschreiben, mit denen wir uns eigentlich noch gar nicht auskennen können. Das funktioniert nur in enger Zusammenarbeit mit den verantwortlichen Techniker*innen, aber auch mit Forscher*innen aus verschiedenen wissenschaftlichen Disziplinen und nicht zuletzt mit den Nutzer*innen selbst.

Im Übrigen stehen diese vergleichsweise „weichen" Aspekte der ethischen Forschung häufig in einem engen Zusammenhang mit den „harten" Risiken, was am Beispiel der Sicherheitsmaßnahmen in der Robotik veranschaulicht werden kann: Hier stellt sich grundsätzlich die Frage, wie weit Maßnahmen zur funktionalen Sicherheit eines interagierenden Roboters eigentlich gehen sollten. Ab welchem Punkt schränken Sicherheitsmaßnahmen den Menschen in der Interaktion ein, etwa indem sie die interagierenden Personen bevormunden und keinen Spielraum für einen selbstbestimmten verantwortungsvollen Umgang lassen? Auch für einen scheinbar sehr klaren Wert wie physische Sicherheit ergeben sich folglich Fragen, die ethisch abgewogen und entschieden werden sollten.[6]

Neben der einseitigen Betrachtung der existenziellen Dimension der Ethik gibt es ein weiteres Missverständnis, wenn ethische Implikationen von Technik als Akzeptanzprobleme definiert werden. In meiner bisherigen Zusammenarbeit mit Ingenieur*innen war ich gelegentlich mit der Auffassung konfrontiert, dass die entwickelten Roboter grundsätzlich eine gute Sache seien, dass es allerdings noch an der Akzeptanz der Nutzer*innen mangelt. Der Ethik wird dann die Aufgabe zugeordnet, Berührungsängste zu zerstreuen und zur Förderung der Technikakzeptanz beizutragen, etwa in Form von Gestaltungsempfehlungen, die eher die äußere Erscheinung und weniger den Kern der technologischen Entwicklung betreffen sollen.

6 Remmers, P (2020): Ethische Perspektiven der Mensch-Roboter-Kollaboration. In: Buxbaum, H-J (Hrsg.): Mensch-Roboter-Kollaboration. Wiesbaden: Springer Gabler, 2020, S. 55–68. DOI: 10.1007/978-3-658-28307-0_4.

Natürlich spielt die Akzeptanz der Nutzer*innen eine wichtige Rolle in ethischen Interventionen, insbesondere weil jene nicht immer frei zwischen Nutzung und Nicht-Nutzung entscheiden können. Aber eine Engführung von Ethik auf Akzeptanzfragen knüpft an eine problematische Tendenz an, sofern dadurch die Technik selbst von der ethischen Intervention ausgenommen wird. Dahinter kann sich die Annahme verbergen, dass Technik eigentlich im Kern ethisch neutral ist und ihre ethische Relevanz erst im Kontext der Nutzung erhält. Diese Auffassung technischer Artefakte ist allerdings in der Technikethik inzwischen stark differenziert worden, wodurch der Ansatz einer ethischen Technikgestaltung letztendlich seine Begründung erfährt.[7] Es geht nicht (nur) um die Akzeptanz oder um die guten oder bösen Absichten des Nutzers oder der Nutzerin, sondern um die Möglichkeiten, Grenzen und Suggestionen, die mit der Technologie selbst gesetzt werden.

Echte ethische Eingriffe in die Technik sind schließlich noch aus einem weiteren Grund wichtig. In aktuellen Diskussionen über Technologien der Künstlichen Intelligenz und der Robotik wird permanent eine ethische Haltung angemahnt. Doch die öffentlichkeitswirksame Hervorhebung der Ethik in neuen Technologien läuft Gefahr, als Feigenblatt des Innovationsdrucks missbraucht zu werden. Wenn der aktuelle Ethik-Hype folgenlos bleibt – und das bedeutet: wenn gut gemeinte Bekenntnisse nicht in konkrete Gestaltungen von Technologien übersetzt werden – , dann können beispielsweise die vielerorts formulierten Ethik-Kodizes zur Künstlichen Intelligenz ihre Wirksamkeit in einer leerlaufenden Debatte verlieren.[8] Aus diesem Grund sind unabhängige, wissenschaftlich solide und kritische Interventionen für jede aufrichtige Forderung nach verantwortungsvoller Technikgestaltung notwendig.[9]

MENSCH-MASCHINE-INTERAKTION IM BERLIN ETHICS LAB

Aktuell sind bereits einige Projekte zur Mensch-Roboter-Interaktion im BEL vernetzt. Beispielsweise geht es im vom Bundesministerium für Bildung und Forschung (BMBF) geförderten Berliner Verbundprojekt „Roboterunterstützung bei Routineaufgaben zur Stärkung des Miteinanders in Pflegeeinrichtungen"

7 Kroes, P; Verbeek, P-P (2014) (Hrsg.): The Moral Status of Technical Artefacts. Dordrecht: Springer.

8 Jobin, A; Ienca, M; Vayena, E (2019): The Global Landscape of AI Ethics Guidelines. In: Nature Machine Intelligence 1, Nr. 9 (September 2019), S. 389–99. DOI: 10.1038/s42256-019-0088-2.

9 Bietti, E (2020): From ethics washing to ethics bashing: a view on tech ethics from within moral philosophy. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, FAT* '20 (New York, NY, USA: Association for Computing Machinery, 2020), 210–219. DOI: 10.1145/3351095.3372860.2020.

(RoMi) um die Entwicklung eines Robotersystems für den Einsatz im Pflegeheim. Im Unterschied zu anderen Projekten, die eine assistierende Interaktion von Robotern mit pflegebedürftigen Personen anstreben, geht es in RoMi vor allem darum, den Roboter als Assistenz für die Pflegekräfte zu gestalten. Denn ein ethisches Ziel des Projekts besteht in der Förderung der pflegerischen Interaktion zwischen Pflegekräften und Pflegebedürftigen, wobei der Roboter nur als technisches Hilfsmittel, nicht aber als „Begleiter" oder „Ansprechpartner" stilisiert werden soll.[10]

Allerdings dreht das BEL die üblichen institutionellen Verhältnisse um. Das BEL ist kein MRI-Lab, in dem verschiedene nicht-technische Forschungsdisziplinen an der robotischen Grundlagenforschung mitarbeiten; es ist vielmehr ein Labor für ethische, technikphilosophische und sozialwissenschaftliche Grundlagenforschung, an der Ingenieur*innen, Computerwissenschaftler*innen und andere nicht-geistes- und sozialwissenschaftliche Forscher*innen mitarbeiten. Aus den Erkenntnissen und Erfahrungen der technologischen Forschung und Entwicklung werden im Labor Methoden und Formate erarbeitet, die wiederum in der Zusammenarbeit mit Techniker*innen erprobt und verfeinert werden. Dabei orientiert sich die Arbeit im Lab u.a. am Konzept des Real-Labors und am Paradigma der *Responsible Research and Innovation* (RRI), also an dem Ansatz, technologische Entwicklungsprozesse möglichst früh und kontinuierlich durch ethische, philosophische und sozialwissenschaftliche Arbeit anzureichern sowie nicht-technische Stakeholder einzubeziehen. Das Ziel besteht nicht nur in der wissenschaftlichen Erforschung der begleiteten Prozesse, sondern vor allem auch in der Entwicklung von konstruktiven Werkzeugen zur Intervention. Denn selbst wenn viele Parameter bereits durch technische, soziale und politische Dynamiken gesetzt sind, gibt es noch eine Menge Anknüpfungspunkte für kritische Reflexionen im Sinne der Ethik. Die Hoffnung ist, dadurch eine effektive ethische Entwicklung neuer Technologien anzustoßen – eine Idee, die sich bisher noch kaum bewähren konnte, weil sie erst seit vergleichsweise kurzer Zeit in einem umfassenden Rahmen durchgeführt und erprobt wird.

Schließlich beschränkt sich der Ansatz des BEL nicht auf das Forschungsfeld der Mensch-Roboter-Interaktion. Die Idee einer Interaktion nach dem Vorbild der Kollaboration eröffnet in vielerlei Hinsicht weitreichende Potentiale für praktische Mittel zur Reflexion neuer Technologien, beispielsweise im Bereich der Künstlichen

10 Das Projekt folgt damit dem technikethischen Ansatz des Care Centered Value Sensitive Designs, vgl. Van Wynsberghe, A (2015): Healthcare robots: ethics design and implementation, Burlington, VT: Ashgate.

Intelligenz. Hier soll es nicht um die Entwicklung von Benutzerschnittstellen gehen, sondern um die konkrete Nutzung von interaktiven Design-Ansätzen für ethische Reflexionen und sinnstiftende Perspektiven. Ein Beispiel: KI-Anwendungen und auch der Einsatz von Robotern in der Industrie sind für die meisten Menschen unsichtbar und manchmal auch kaum begreiflich. Ein realistisches Bild der Technik und eine informierte Beurteilung der ethischen Aspekte ist deswegen schwierig. Ein Ansatz zur Überbrückung dieser „Erfahrungslücke" besteht darin, mittels virtueller Interaktionen eine Verbindung zwischen der Technologie und einer konkreten Erfahrungsperspektive zu schaffen. So kann die Technologie nicht nur sichtbar und verständlich, sondern auch für eine Bewertung und eine bessere Anwendungsperspektive verfügbar gemacht werden.

Eine weitere Möglichkeit für einen derartigen Eingriff besteht darin, die hinter der Technologie stehenden Visionen herauszuarbeiten, zu explizieren und gemeinsam mit den prägenden Akteuren der Entwicklung zu thematisieren. Um diese Visionen aber nicht nur als abstrakte Vorstellungen zu diskutieren, können wir in Experimenten oder virtuellen Umgebungen konkrete Interaktionen entwickeln, um mit den entworfenen Zukünften in Interaktion zu treten. Hier können Methoden und Inhalte der experimentellen Philosophie, der philosophischen Phänomenologie, aber auch aus Virtual- bzw. Augmented-Reality-Technologien oder aus der Performance Art zum Einsatz kommen. Visionen und Zukünfte sollen im wörtlichen Sinne (virtuell) gebaut und ausprobiert werden; auf diese Weise können sie adäquat veranschaulicht, (be-)greifbar und konkret abschätzbar gemacht werden. Ziel ist die Vermittlung von Zusammenhängen zwischen Ideen und Wirklichkeiten, zwischen abstrakten Werten und konkreten Praktiken und zwischen großen Visionen und den kleinen Schritten, die dahin führen können.

Birgit Beck und Aljoscha Burchardt

ALLE REDEN VON ETHISCHER KI – ABER WAS MEINEN SIE DAMIT?

Wirft man einen Blick auf öffentliche – und teils auch wissenschaftliche – Debatten über Digitalisierung und speziell Künstliche Intelligenz (KI), könnte man folgenden Eindruck erhalten: Wo früher ausschließlich Menschen anhand ethischer Standards moralisch gut handeln mussten, kommt heute gefühlt noch ein technischer Akteur mit ins Spiel. Muss man nun Maschinen so „programmieren", dass sie (für uns) moralisch handeln? Dass und aus welchen Gründen diese Frage nicht einfach wörtlich zu nehmen ist, möchten wir im Folgenden in einem knappen Überblick aufzeigen. Mögliche Missverständnisse der Rede von „ethischer KI" ergeben sich, so unsere These, durch eine unreflektierte Begriffsverwendung sowie unplausible Erwartungen an die Funktionalität technischer Artefakte bzw. Prozesse wie KI-basierte Systeme. Nach einer begrifflichen Differenzierung zwischen Ethik und Moral werden wir danach fragen, was die Rede von „ethischer KI" überhaupt bedeuten kann, und drei mögliche Lesarten unterscheiden, die wir anhand praktischer Beispiele explizieren. Das Ergebnis unserer Überlegungen wird sein: Anstatt pauschal von „ethischer KI" zu sprechen, sollte im Einzelfall präzisiert werden, was jeweils im vorliegenden Kontext mit diesem Schlagwort gemeint ist, um terminologischen Missverständnissen und vor allem der Gefahr einer (weiteren) Erosion individueller und gesellschaftlicher Verantwortung im Zuge des technischen Wandels vorzubeugen.

ETHIK UND MORAL

Im Alltag werden die Begriffe „Ethik" und „Moral" bzw. die Adjektive „ethisch" und „moralisch (gut)" häufig synonym gebraucht. Man spricht zum Beispiel von „unethischem Verhalten", wenn man eigentlich unmoralische, also moralisch schlechte bzw. verbotene Handlungen meint. Auch in wissenschaftlichen Kontexten liest man etwa, eine bestimmte Praxis müsse „ethisch" gestaltet sein, und in der

Regel stört sich niemand an diesem Begriffsgebrauch.[1] In der Wissenschaft ist diese Gleichsetzung mutmaßlich einer wörtlichen Übertragung aus dem Englischen geschuldet, wo die Adjektive *moral* und *ethical* anders als im Deutschen synonym zu gebrauchen sind.[2] Aus philosophischer Perspektive gilt es aber, zwischen Moral und Ethik zu unterscheiden:[3] Moral umfasst in einem weit verbreiteten modernen Verständnis die faktisch vorherrschenden normativen (nicht unbedingt kodifizierten) Regeln, Sitten und Gebräuche in einer gegebenen Gemeinschaft. Moral grenzt sich dabei einerseits von positiv rechtlichen Vorgaben (strafbewehrten Gesetzen) und andererseits von reinen Konventionen etwa der Höflichkeit oder der Etikette ab. Ethik hingegen wird als die (normative)[4] Wissenschaft von der Moral angesehen, als Reflexionswissenschaft, die sich in die normative allgemeine Ethik und die normative angewandte Ethik und ihre verschiedenen Bereichsethiken aufgliedert, zum Beispiel die Bioethik, Medizinethik, Umweltethik, Technikethik oder speziell Maschinen- bzw. Roboterethik.

Der Vollständigkeit halber wollen wir noch ein anderes gebräuchliches Verständnis erwähnen, demzufolge sich der Begriff Moral auf all diejenigen normativen Regeln, Rechte und Pflichten bezieht, die für alle Moraladressat*innen gleichermaßen (kategorische, überzeitliche und universelle) Gültigkeit beanspruchen. Demgegenüber bezieht sich hier der Begriff Ethik in einem an antike Konzeptionen angelehnten Verständnis lediglich auf partikulär (individuell oder

1 So wird zum Beispiel in den 2019 veröffentlichten Ethik-Leitlinien für eine vertrauenswürdige KI der Hochrangigen Expertengruppe für künstliche Intelligenz der EU-Kommission gefordert, KI müsse „**ethisch** sein und somit die Einhaltung ethischer Grundsätze und Werte garantieren" (Hochrangige Expertengruppe für Künstliche Intelligenz (2019): Ethik-Leitlinien für eine vertrauenswürdige KI, Brüssel, S. 2; Hervorhebung im Original). Diese Ausdrucksweise ist mehrfach missverständlich. Erstens *ist* KI nicht *ethisch* (kein konzeptionelles und normatives wissenschaftliches Unterfangen), sondern höchstens *Gegenstand* ethischer Reflexion, Rechtfertigung oder Kritik. Zweitens kann die Vorgabe unterschiedlich gedeutet werden; wahrscheinlich ist damit gemeint, der Einsatz von KI müsse *ethisch gerechtfertigt* oder zumindest *rechtfertigbar* sein, bzw. die (jeweilige) KI müsse (ausschließlich) zu *moralisch guten Zwecken* eingesetzt werden. Drittens kann keine KI die *Einhaltung* ethischer Grundsätze und Werte *garantieren*, außer, sie manipulierte Menschen so, dass sie nicht mehr in der Lage wären, von diesen Grundsätzen und Werten abweichend zu entscheiden und zu handeln (vgl. das Gedankenexperiment der „Gottesmaschine" von Savulescu, J; Persson, I (2012): Moral Enhancement, Freedom and the God Machine, Monist. 2012 Jul; 95(3), S. 399–421). Darüber hinaus muss natürlich spezifiziert werden, um welche Grundsätze und Werte es sich hierbei handelt und wie diese begründet sind.

2 Vgl. Hübner, D (2018): Einführung in die Philosophische Ethik. 2. Auflage, Göttingen: Vandenhoeck & Ruprecht, S. 20.

3 Vgl. Ach, J S; Siep, L (2011): Ethik – Zur Einführung. In: Ach, J S/Bayertz, K/Siep, L (Hrsg.): Grundkurs Ethik. Band I: Grundlagen. 2. Auflage, Paderborn: mentis, S. 9–30.

4 Von der normativen Ethik wird die deskriptive Ethik abgegrenzt, die sich mit der *empirischen* Erhebung und Analyse faktisch vorherrschender moralischer Sitten und Gebräuche beschäftigt. Dies ist keine genuin philosophische Aufgabe, sondern eher in der Moralsoziologie, Moralpsychologie oder auch in historischen Wissenschaften angesiedelt.

für eine bestimmte Gemeinschaft) gültige Ratschläge für ein gutes und gelingendes Leben. Obwohl also auch innerhalb der Philosophie nicht immer eine einheitliche Begriffsverwendung vorliegt, herrscht jedenfalls Konsens darüber, dass Moral und Ethik auf verschiedenen systematischen Ebenen angesiedelt sind und entsprechend die Adjektive „moralisch" und „ethisch" sich auf verschiedene Sachverhalte und Akteure beziehen.

WAS KANN „ETHISCHE KI" BEDEUTEN?

Im Lichte dieser konzeptionellen Differenzierung wird deutlich, dass der Ausdruck „ethische KI" mindestens mehrdeutig (wenn nicht schlicht ein Kategorienfehler) ist. Ob es darüber hinaus „moralische Maschinen" in einem substanziellen Sinn von „moralisch" gibt, ob Maschinen also einerseits moralische Subjekte (*moral agents*) und andererseits Objekte moralischer Berücksichtigungswürdigkeit (*moral patients*) sein können, ist ebenfalls umstritten.[5] „Künstliche Intelligenz" ist zunächst einmal ein Sammelbegriff für mindestens ein Gebiet der Informatik sowie eine Menge von Technologien und darauf mehr oder weniger basierende Anwendungen. Wenn wir das Gebiet der Informatik einmal ausnehmen, geht es bei KI in jedem Fall um die *empirische* Entwicklung, Produktion, Implementation und Anwendung technischer Artefakte bzw. Prozesse oder Systeme. Ethik dagegen ist, wie beschrieben, ein *theoretisches*, *reflexives, konzeptionelles* und *normatives* Unterfangen, das sich mit der hermeneutischen Rekonstruktion, Analyse und Bewertung moralischer Normen bzw. Vorstellungen des guten Lebens beschäftigt. Auf den ersten Blick haben also KI und Ethik nicht viel gemeinsam. KI *ist* insofern nicht ethisch, als KI selbst keine Weise ist, theoretische bzw. normative Wissenschaft zu betreiben. Man kann aber dennoch versuchen, der Rede von „ethischer KI" einen Sinn zu verleihen. Dazu bieten sich verschiedene Lesarten an:

1) Der Ausdruck „ethische KI" könnte in dem Sinne gebraucht werden, dass mittels der Entwicklung und Implementation KI-basierter Systeme moralisch gute Zwecke realisiert werden (sollen).

2) Der Ausdruck „ethische KI" könnte in dem Sinne gebraucht werden, dass in den Entwicklungs- und Implementationsprozess derartiger Systeme oder in deren Nutzung ethische Überlegungen einfließen (sollen).

5 Vgl. Misselhorn, C (2018): Grundfragen der Maschinenethik. 2. Auflage, Ditzingen: Reclam.

3) Der Ausdruck „ethische KI" könnte in dem Sinne gebraucht werden, dass KI-basierte Systeme zu dem Zweck entwickelt und implementiert werden (sollen), ethische Deliberation, Begründung und die moralische Entscheidungsfindung zu unterstützen bzw. zu erleichtern.

1. ETHISCHE KI IM SINNE DES ZIELS DER UMSETZUNG MORALISCH GUTER ZWECKE

Die Rede von „ethischer KI" könnte in der ersten Lesart so verstanden werden, dass man sich damit auf die Realisierung bestimmter, ethisch gerechtfertigter und moralisch guter Zwecke bezieht. Solche Zwecke sind zum Beispiel Gerechtigkeit, Gesundheit oder Sicherheit. KI könnte dafür eingesetzt werden, innovative Mittel zur quantitativ (mehr verschiedene Maßnahmen) und qualitativ (mehr Effizienz) besseren Umsetzung derartiger Zwecke zu entwickeln und zu implementieren. In der Tat gibt es bereits Bestrebungen und vielfältige Beispiele dafür, solche Zwecke mittels KI-basierter Systeme umzusetzen, wobei der Erreichungsgrad des Ziels abhängig vom heutigen Stand der Technik, den verfügbaren Daten und nicht zuletzt den Möglichkeiten der einzelnen Projekte und Beteiligten stark variiert.

So wurde beispielsweise vorgeschlagen, die Auswahl von Bewerber*innen für Arbeitsplätze an Algorithmen zu delegieren, da man der Ansicht war, dass diese unvoreingenommener als menschliche Personalbeauftragte entscheiden würden und somit mehr Gerechtigkeit gewährleistet werden könne. Die Annahme, dass Algorithmen automatisch weniger *Bias* aufweisen würden als menschliche Entscheidungsträger*innen, hat sich jedoch als trügerisch herausgestellt („Algorithms inevitably make biased decisions").[6]

Von der Anwendung KI-basierter Systeme und dem Einsatz von Big Data im Gesundheitssektor verspricht man sich unter anderem Fortschritte auf dem Feld personalisierter diagnostischer und therapeutischer Angebote, neue und effizientere Präventionsmöglichkeiten sowie die Verbesserung biomedizinischer Forschung und klinischer Studien.[7] Auch in KI-basierte Geräte bzw. Programme wie Fitness-

6 Vgl. Mittelstadt, B D; Allo, P; Taddeo, M; Wachter, S; Floridi, L (2016): The ethics of algorithms: Mapping the debate. Big Data & Society, July-December 2016, S. 1–21 (Zitat siehe S. 7). DOI: 10.1177/2053951716679679.

7 Vgl. Deutscher Ethikrat (2017): Big Data und Gesundheit – Datensouveränität als informationelle Freiheitsgestaltung. Stellungnahme, abrufbar unter https://www.ethikrat.org/fileadmin/Publikationen/Stellungnahmen/deutsch/stellungnahme-big-data-und-gesundheit.pdf [25.10.2020].

tracker und Health Apps und in den Einsatz von Robotik in der Pflege[8] werden große Hoffnungen gesetzt. Die individuelle wie öffentliche Gesundheit soll durch den Einsatz KI-basierter Systeme einfacher und effektiver gefördert werden können.

Der Einsatz von KI-basierten Systemen wird auch als probates Mittel angesehen, die Sicherheit im öffentlichen wie privaten Raum zu erhöhen. So sollen Gesichtserkennungskameras an öffentlichen Plätzen zur Verbrechensprävention beitragen und autonome Fahrzeuge die Unfallgefahr im Straßenverkehr reduzieren.[9] Smarte Umgebungen (sog. *ambient intelligence* oder *smart homes*) sollen dafür sorgen, dass zum Beispiel ältere, pflegebedürftige Personen länger selbständig in ihren eigenen vier Wänden wohnen können, da durch technische Überwachung Unfällen im Haushalt vorgebeugt und im Notfall schneller Hilfe geholt werden können sollen.

2. ETHISCHE KI IM SINNE DER INTEGRATION ETHISCHER ÜBERLEGUNGEN IN DEN PRODUKTIONS- UND NUTZUNGSPROZESS

Zweitens kann die Rede von „ethischer KI" auch dahingehend verstanden werden, dass bereits im Prozess des Designs und der Entwicklung von KI-Systemen (und in Folge auch in deren Nutzung) ethische Reflexion und Beurteilung einfließen sollen. Dieses Ziel kann nur durch interdisziplinäre Anstrengungen verfolgt werden. Es trägt der technikphilosophischen Einsicht Rechnung, dass Technik – und damit auch KI – niemals nur ein *neutrales* Werkzeug ist, sondern im Prozess der Entstehung bereits diverse, in der Regel implizite (soziale, kulturelle, aber ggf. auch idiosynkratische) Werte und Normen einfließen,[10] die im Idealfall transparent gemacht und einer ethischen Evaluation unterzogen werden sollten. Dies wird zum Beispiel in Ansätzen des sog. *Value Sensitive Design* angestrebt.[11] Dabei sollte darauf geachtet werden, welche Werte und Normen in die Entwicklung einfließen, ob diese ethisch gut begründet sind, ob bei der Implementation der fraglichen KI-Technik diese Werte respektiert und geschützt werden, und wer ggf. ethische und rechtliche Verantwortung für Fehlfunktionen und potenziell schädliche Auswirkungen übernimmt.

8 Vgl. Bendel, O (Hrsg.) (2018): Pflegeroboter. Wiesbaden: Springer Gabler.

9 Vgl. Maurer, M; Gerdes, J C; Lenz, B; Winner, H (Hrsg.) (2015): Autonomes Fahren. Technische, rechtliche und gesellschaftliche Aspekte. Berlin/Heidelberg: Springer Vieweg.

10 Vgl. Verbeek, P-P (2011): Moralizing Technology. Understanding and Designing the Morality of Things. Chicago/London: The University of Chicago Press.

11 Vgl. Friedman, B; Hendry, D G (2019): Value Sensitive Design: Shaping Technology with Moral Imagination. Cambridge, MA: MIT Press.

Eine Besonderheit der KI-Entwicklung liegt hierbei in der relativen Unabhängigkeit verschiedener Bestandteile wie zum Beispiel dem zugrundeliegenden Lernalgorithmus, dem durch Daten trainierten Modell und der eigentlichen Anwendung. Diese können durch verschiedene Akteure zu ganz verschiedenen Zeiten beauftragt und erzeugt worden sein, sodass der Entstehungsprozess und die Aufladung mit Werten beliebig schwer nachvollziehbar sein können.[12] Bereits hier verschwimmen die Grenzen von „Implementation" und „Nutzung", welche bei im Gebrauch weiterlernenden Systemen endgültig an Bedeutung einbüßen.

3. ETHISCHE KI IM SINNE DES ZIELS DER UNTERSTÜTZUNG ETHISCHER DELIBERATION UND (INDIVIDUELLER) ENTSCHEIDUNGSFINDUNG

Obwohl wir oben festgestellt haben, dass KI insofern nicht ethisch ist, als damit keine Ethik betrieben wird, gibt es auch Beispiele für diese letzte Lesart des Ausdrucks „ethische KI". Eines davon ist der von Alberto Giubilini und Julian Savulescu eingebrachte Vorschlag eines *Artificial Moral Advisor (AMA)*, also eines künstlichen moralischen Ratgebers. Der AMA ist den Autoren zufolge „a form of moral artificial intelligence that could be used to improve human moral decision-making".[13] Er zeichnet sich durch Unparteilichkeit sowie durch Interesse- und Affektlosigkeit aus und soll bei der Entscheidungsfindung in einer konkreten Situation die jeweiligen moralischen Überzeugungen, ethischen Prinzipien und Werte der Nutzer*innen einbeziehen. Der AMA ist eine „ethische KI", da er (selbstlernend, auf der Basis einer nutzer*innenspezifischen Programmierung bzw. eines entsprechenden Trainings) eine ethische Abwägung durchführt und dadurch die Nutzer*innen bei der Entscheidung für eine moralisch richtige Handlung in einer gegebenen Situation unterstützt.

Warum sollten wir einen solchen künstlichen moralischen Ratgeber brauchen? Die Autoren weisen darauf hin, dass wir häufig aufgrund mangelnder zeitlicher und kognitiver Ressourcen sowie impliziter affektiver Beeinflussung Entscheidungen treffen, die unseren *eigenen* moralischen Standards und rationalen ethischen Begründungen zuwiderlaufen.[14] Der AMA soll uns dabei unterstützen, derartige Voreingenommenheiten und Defizite zu umgehen und dadurch bessere moralische

12 Vgl. https://algorithmenethik.de/2018/02/05/wo-maschinen-irren-koennen-fehlerquellen-und-verantwortlichkeiten-in-prozessen-algorithmischer-entscheidungsfindung/ [30.10.2020].

13 Giubilini, A; Savulescu, J (2018): The Artificial Moral Advisor. The "Ideal Observer" Meets Artificial Intelligence. Philosophy & Technology 31, S. 169–188, hier: S. 169.

14 Ebd., S. 170.

Entscheidungen zu treffen, indem er uns das mühsame Geschäft der ethischen Überlegung, Begründung und Abwägung abnimmt. Er ist also insofern eine „ethische KI", als er tatsächlich *Ethik betreibt*. Er wirkt als moralischer Berater oder sogar „persuader"[15] und überzeugt uns von einer Handlungsoption, die wir infolgedessen umsetzen – während wir uns darauf verlassen, dass diese Handlung moralisch richtig und ethisch gut begründet ist. Die Notwendigkeit einer solchen technischen Unterstützung wirkt indes nur unter der Annahme plausibel, dass wir selbst *unzulängliche* moralische Akteur*innen sind, weil wir die Konsequenzen unserer Handlungen nicht umfassend vorhersehen und gegeneinander aufrechnen können und unfähig sind, von unseren affektiven Einstellungen in unparteilicher Weise Abstand zu nehmen. Unter anderen ethischen und anthropologischen Vorzeichen, beispielsweise im Rahmen einer Tugendethik, in der es nicht darum geht, Handlungsfolgen zu beurteilen, sondern das moralische „Know-how" von Personen,[16] ist der Vorstellung einer solchen „ethischen KI" wenig abzugewinnen.

FAZIT

Die Rede von „ethischer KI" ist, wie wir veranschaulicht haben, mindestens mehrdeutig, weshalb die jeweilige Bedeutung dieses Ausdrucks in einem bestimmten Kontext nicht auf den ersten Blick ersichtlich ist. Daher sollte man in öffentlichen und wissenschaftlichen Diskussionen immer genau hinsehen und ggf. nachfragen, was denn im vorliegenden Fall genau mit „ethischer KI" gemeint ist. Dass damit nicht gemeint sein kann, dass ethische und rechtliche Verantwortung einfach an die Technik delegiert werden kann, sollte sich eigentlich von selbst verstehen. Maschinen betreiben (außer in Gedankenexperimenten) keine Ethik und sie sind auch nicht dazu geeignet, uns moralisch relevante Überlegungen, Entscheidungen oder gar Handlungen einfach und bequem abzunehmen. Bis auf Weiteres müssen wir selbst dafür Sorge tragen, moralisch angemessen zu handeln und eine ethisch vertretbare Technikentwicklung zu betreiben, am besten auf der Grundlage reflektierter ethischer Deliberation und interdisziplinärer Zusammenarbeit. Dazu zwingen uns die Maschinen – durch ihre Anwesenheit.

15 Ebd., S. 182.

16 Vgl. Borchers, D (2011): Moralische Exzellenz – Einführung in die Tugendethik. In: Ach, J S; Bayertz, K; Siep, L (Hrsg.): Grundkurs Ethik. Band I: Grundlagen. 2. Auflage, Paderborn: mentis, S. 33–48.

Christoph Benzmüller und Bertram Lomfeld

TRÄUMEN VERNÜNFTIGE MASCHINEN VON GRÜNDEN? EINE REALE UTOPIE[1]

Aktuelle Forderungen nach „moralischen Maschinen"[2] stellen kein Science-Fiction-Szenario dar, sondern sind vielmehr ein aktuelles praktisches Anliegen. Durch das Eindringen autonomer Systeme in nahezu alle Lebensbereiche, einschließlich hochkomplexer und ethisch-kritischer Anwendungen, wie selbstfahrende Autos, medizinische Roboter oder auch Kredit-Scoring-Tools, werden KI-Systeme unausweichlich mit moralischen und rechtlichen Fragen konfrontiert und müssen über diese entscheiden. Ein Kernproblem für die Beurteilung ethisch-rechtlicher Verantwortlichkeit oder gar die ethisch-rechtliche Steuerung autonomer Systeme sind dabei die verdeckten Entscheidungsprozesse moderner (subsymbolischer) KI-Technologien (sog. Black-Box); diese stellen ein Hindernis sowohl in punkto Transparenz als auch für direkte Interventionen dar. Ein einfaches Einfordern von Transparenz lässt allerdings technologische Realitäten außer Acht oder behindert sogar dringend notwendige Weiterentwicklungen.[3]

Die Vielzahl von KI-Anwendungen u. a. in Industrie, Mobilität und Verkehr, im Gesundheitswesen, Finanzbereich und Militär oder auch in der öffentlichen Verwaltung führt zu einer historischen Übergangsphase mit einer beispiellosen Innovationsdynamik und mit zum Teil schwer prognostizierbaren Auswirkungen. Politik, Kontrollorgane und die Gesellschaft als Ganzes stehen deshalb vor der Herausforderung, mit den teilweise disruptiven Entwicklungen Schritt zu halten und den eingeläuteten Wandel vorausschauend zu gestalten. Es gilt ökonomische und gesellschaftliche Chancen von moderner KI-Technologie optimal zu nutzen, dabei aber gleichzeitig Gefahren und Risiken zu minimieren. Die Vision eines Gleichgewichts zwischen Chancen und Risiken findet sich in vielen aktuellen Strategiepapieren zum Thema KI auf nationaler[4] und europäischer Ebene.[5]

1 Eine englische Variante des Textes erschien als Benzmüller, C; Lomfeld, B (2020): Reasonable Machines: A Research Manifesto. In: Schmidt,U; Wolter, D; Klügl, F (Hrsg): KI 2020: Advances in Artificial Intelligence – 43rd German Conference on Artificial Intelligence, Bamberg, Germany, September 21–25, 2020. DOI: 10.1007/978-3030-58285-2_20.

2 Wallach, W; Allen, C (2008): Moral machines: Teaching robots right from wrong. Oxford University Press.

3 Wahlster, W (2020): Deep Learning alleine reicht nicht. Gastbeitrag in der Frankfurter Allgemeine Zeitung vom 10.09.2020.

4 Deutsche Bundesregierung (2018): Strategie Künstliche Intelligenz der Bundesregierung, Berlin, online unter: https://www.bmbf.de/files/Nationale_KI-Strategie.pdf [20.10.2020].

5 High-Level Expert Group on Artificial Intelligence (2019): Ethics Guidelines for trustworthy AI, Brüssel, 08.04.2019.

Das aktuelle Weißbuch der Europäischen Kommission zu KI schlägt in diesem Zusammenhang die Schaffung eines „Ökosystems der Exzellenz" in Kombination mit einem „Ökosystem des Vertrauens" vor.[6]

VERTRAUEN DURCH KOMMUNIKATION

Um das Vertrauen in moderne KI-Systeme zu fördern, schlagen wir die Erforschung von hybriden Methoden vor, die intelligente Systeme dazu befähigen, „echte Gründe" für ihre Handlungen und Entscheidungen zu generieren und zu kommunizieren. Während „klassische" Ansätze interpretierbarer KI nach transparenten Erklärungen für (subsymbolische) Maschinenprozesse suchen, möchten wir „Vertrauenswürdigkeit durch rationale Kommunikation" erreichen, d. h. Maschinen sollen in realer sozialer Interaktion Gründe für ihr Handeln austauschen. Dieser interdisziplinäre Forschungsvorschlag führt innovative Ansätze aus symbolischer KI, Maschinellem Lernen, Mensch-Maschine-Interaktion sowie Recht und Philosophie zusammen.

Seit seinen Anfängen unterscheidet das Gebiet der KI bei der Modellierung und Erklärung von intelligentem Verhalten zwischen dem konnektionistischen (subsymbolischen) und dem symbolischen Paradigma.[7] Subsymbolische Ansätze modellieren intellektuelle Fähigkeiten typischerweise mit Hilfe von künstlichen neuronalen Netzen, d. h. durch Netzwerke von Berechnungseinheiten ohne jegliche semantische Bedeutung. Dagegen geht der symbolische Ansatz davon aus, dass Intelligenz aus der Manipulation von abstrakten kompositionalen und bedeutungstragenden Darstellungen resultiert. Zu den Techniken, die im symbolischen Bereich verwendet werden, gehören regelbasierte Systeme und formale Logik. Beide Paradigmen haben bekannte Stärken und Schwächen, und die Debatte, ob menschliche Intelligenz durch das konnektionistische oder symbolische Paradigma plausibler modelliert und erklärt werden kann, hat eine lange Tradition.

6 Europäische Kommission (2020): White Paper on Artificial Intelligence: A European approach to excellence and trust, Brüssel, 19.02.2020.

7 Die Begriffe „intelligentes Verhalten" und „Künstliche Intelligenz" werden in der Philosophie, Psychologie, Kognitionswissenschaft und Informatik weiterhin kontrovers diskutiert; vgl. z. B. Bringsjord, S; Govindarajulu, N S (2020): Artificial Intelligence. In: Zalta, E N (Hrsg.): The Stanford Encyclopedia of Philosophy, Metaphysics Research Lab, Stanford University, Summer 2020. Wir gehen hier davon aus, dass Rationalität und die Fähigkeit zur Selbstreflexion (bis zu einem gewissen Grad) wichtige Elemente intelligenten Verhaltens sind.

In unserem Forschungsvorhaben, bei dem es eher um einen praktischen Durchbruch als um eine biologisch plausible Modellierung geht, möchten wir Möglichkeiten der Hybridisierung symbolischer und subsymbolischer Techniken untersuchen, um die Vorteile beider Paradigmen zu nutzen.[8] Wir wollen zeigen, dass hybride Ansätze die Operationalisierung europäischer und nationaler Forderungen in Richtung erklärbare und vertrauenswürdige KI besser ermöglichen als andere Lösungen. Darüber hinaus behaupten wir, dass eine sinnvolle und vertrauenswürdige Mensch-Maschine-Interaktion signifikant von expressiven, symbolischen Darstellungen und Argumentationskompetenzen zukünftiger KI-Systeme profitiert bzw. solche sogar voraussetzt.

Die Einbeziehung der Mensch-Maschine-Interaktion zusätzlich zur Integration von symbolischen und subsymbolischen KI-Techniken führt zu einem „doppelt hybriden Ansatz", der neue Möglichkeiten einer ethisch-rechtlichen Kontrolle und Verifikation von KI-Systemen bietet, ohne dazu die „Black-Box öffnen" zu müssen. Unser Forschungsansatz adressiert aus diesem Blickwinkel die folgenden aktuellen Herausforderungen: die Operationalisierung von interpretierbarer und vertrauenswürdiger KI, die (hybride) Integration erfolgreicher Techniken aus subsymbolischer und symbolischer KI und die Transformation von impliziter zu expliziter Wissensrepräsentation, mit dem Ziel, Elemente von Selbstreflexion und Selbstkontrolle in KI-Systemen zu entwickeln und zu studieren und über deren Grenzen und Implikationen, u. a. aus der Perspektive von Mensch-Maschine-Interaktion und Technikphilosophie, zu reflektieren.

„KÜNSTLICHE" RECHTFERTIGUNG

Das Problem der *Black-Box-Governance* hat eine interessante Parallele auf der Ebene der menschlichen Entscheidungsfindung. Die meisten aktuellen Modelle in der Moralpsychologie betrachten die emotionale Intuition als die (oder zumindest eine) initiale Triebkraft des menschlichen Urteilens und Handelns, die erst im Nachhinein (bzw. mit einem wesentlich langsameren zweiten System) mit Hilfe von Gründen rationalisiert wird.[9] In einem sozialen Rahmen des Gebens

8 Garnelo, M; Shanahan, M (2019): Reconciling deep learning with symbolic artificial intelligence: representing objects and relations. In: Current Opinion in Behavioral Sciences 29, S. 17–23.

9 Haidt, J (2001): The emotional dog and its rational tail: a social intuitionist approach to moral judgment. In: Psychology Review 108.4, S. 814–834; Kahneman, D (2013): Thinking, Fast and Slow. New York: Farrar, Straus und Giroux.

und Nehmens von Gründen (z. B. im Hinblick auf moralische Konventionen oder in einem Rechtssystem) kann somit die ursprüngliche Motivation eines einzelnen menschlichen Agenten ignoriert werden, wenn seine Handlungen und seine (nachträgliche) Rechtfertigung den gegebenen sozialen (moralischen oder rechtlichen) Standards entsprechen.[10] Die Kommunikation von Gründen innerhalb eines solchen post-hoc „sozialen Rechtfertigungsmodells" (Social Reasoning Model, SRM) wird nicht überflüssig, sondern umso wichtiger, da nur Gründe die Kohärenz einer moralischen oder rechtlichen Ordnung in einer zunehmend pluralistischen Welt garantieren. Entscheidend für das SRM-Modell ist die relative Unabhängigkeit der rationalen Rechtfertigung vom motivierenden Impuls zum Handeln. Dennoch kann und wird die (regelmäßige) innerlich-subjektive oder soziale Rückkopplung mit rationalen Gründen auf lange Sicht auch die motivationale (intuitiv-emotionale) Disposition der Agenten verändern.

Dieses Post-hoc-SRM ist als „Modell künstlicher sozialer Rechtfertigung" (artificial Social Reasoning Model, aSRM) auf KI-Entscheidungsprozesse übertragbar. Die Black-Box eines opaken subsymbolischen KI-Systems funktioniert wie eine KI-Intuition. Nach dem SRM-Modell ist Transparenz nicht erforderlich, solange das System akzeptable Post-hoc-Gründe für seine Entscheidungen generiert. Moralische und rechtliche Rechenschaftspflichten und Governance könnten durch symbolische (oder sogar subsymbolische) aSRMs ermöglicht werden. Eine symbolische Lösung versucht, die intuitive Entscheidung der Black-Box mit deontischer logischer Argumentation unter Anwendung moralischer oder rechtlicher Standards zu rekonstruieren oder mit einem alternativen Argument zu rechtfertigen. Die Entscheidung eines autonomen Fahrzeugs einem Passanten nicht auszuweichen, ließe sich beispielsweise mit dem Eigeninteresse des Fahrzeughalters am Überleben der Passagiere rekonstruieren, aber ebenso mit dem Argument, dass abrupte Ausweichmanöver unbeteiligte Personen gefährden könnten. Eine pluralistische, expressive „normative Argumentations-Infrastruktur" wie das von uns entwickelte *LogiKEy* unterstützt diesen Prozess , indem es die Exploration, Analyse und Verifikation komplexer normativer Argumente ermöglicht und dazu automatische und interaktive Werkzeuge bereitstellt.[11] LogiKEy verwendet Logik höherer

10 Lomfeld, B (2017): Emotio Iuris. Skizzen zu einer psychologisch aufgeklärten Methodenlehre des Rechts. In: Köhler, S; Müller-Mall, S; Schmidt, F; Schnädelbach, S (Hrsg.): Recht Fühlen. München: Fink, S. 19–32.

11 Benzmüller, C; Parent, X; van der Torre, L (2020): Designing Normative Theories for Ethical and Legal Reasoning: LogiKEy Framework, Methodology, and Tool Support. In: Artificial Intelligence 287, S. 103348. DOI: 10.1016/j. artint.2020.103348.

Ordnung[12] als formale (meta-)logische Grundlage und lässt sich daher leicht für quantifizierte deontische und andere nicht-klassische Logiken erweitern.[13]

Zusätzlich bedarf es einer pluralistischen Ontologie ethisch-rechtlicher Grundwerte, die ein Feld kontroverser, aber sozial akzeptabler „guter Gründe" abstecken.[14] Bei der Ausweich-Entscheidung eines autonomen Fahrzeuges stehen so verschiedene sozial akzeptierte Rechtfertigungen gegeneinander: z. B. eine *utilitaristische* Abwägung von Nutzen (Anzahl von gefährdeten Leben), eine *deontologische* Verantwortungszuweisung eines bewussten Einlassens der Passagiere auf die Gefahr, eine *egalitäre* Gleichbehandlung der Passanten mit Passagieren oder eine *kommunitaristisch-systemische* Stabilität des Vertrauens in die Sicherheit des Straßenverkehrs. Die symbolische Rekonstruktion kann eine Abwägung im Rahmen dieser Gründe darlegen und die Entscheidung als Vorrangrelation für die relevante Umweltsituation rechtfertigen.[15] Eine subsymbolische Lösung könnte ein unabhängiges (zweites) neuronales Netz schaffen, um Gründe für die Ausgabe des (ersten) Entscheidungsnetzes zu produzieren (z. B. im Fall der autonomen Fahrsteuerung). Natürlich ist die Struktur dieses „Rechtfertigungs-Netzes" wieder verborgen. Entscheidend für eine solche zweite Black-Box als Rückkopplung für den ebenfalls verborgenen Entscheidungsprozess ist wieder nicht Transparenz, sondern, dass die ausgegebenen Gründe kohärent den vorgeschriebenen sozialen und ethisch-rechtlichen Standards entsprechen.

Robuste Lösungen für aSRMs könnten auch versuchen, diese beiden Optionen zu integrieren und aufeinander abzustimmen. Darüber hinaus könnte in beiden Szenarien die eingeführte Rückkopplungsschleife des Gebens und Nehmens von Gründen als Lernumgebung (selbstüberwachtes Lernen) in die anfängliche, intuitive Ebene der autonomen Entscheidungsfindung integriert werden, mit dem letztendlichen Effekt, dass sich die Unterschiede auf beiden Ebenen allmählich auflösen könnten.

12 Benzmüller, C; Andrews, P (2019): Church's Type Theory. In: Zalta E N (Hrsg.): The Stanford Encyclopedia of Philosophy. Metaphysics Research Lab, Stanford University, Summer 2019.

13 Eine Reihe neuerer Beispielanwendungen von LogiKEy sind online verfügbar unter logikey.org, darunter auch eine aktuelle Modellierung von Aspekten rechtlicher Abwägung.

14 Lomfeld, B (2019): Grammatik der Rechtfertigung: Eine kritische Rekonstruktion der Rechts(fort)bildung. In: Kritische Justiz 52.4. DOI: 10.5771/9783748905530-516.

15 Benzmüller, C; Fuenmayor, D; Lomfeld, B (2020): Encoding Legal Balancing: Automating an Abstract Ethico-Legal Value Ontology in Preference Logic. In: Workshop on Models of Legal Reasoning, 17th Conference on Principles of Knowledge Representation and Reasoning (KR 2020). https://arxiv.org/abs/2006.12789.

Durch Berücksichtigung verschiedener Arten sozialer (ethischer) Gründe fördern SRMs & aSRMs den normativen Pluralismus und können konkurrierende (maschinen-)ethische Traditionen integrieren: Deontologie, Utilitarismus und Tugendethik. „Rationaler Pluralismus" in der neueren moralischen und politischen Philosophie definiert Rationalität durch Verfahren auf Meta-Ebenen wie „reflektierendes Gleichgewicht" und „überlappender Konsens"[16] oder „rationaler Diskurs".[17] Die zeitgenössische Rechtsphilosophie und -theorie zeigen, wie Recht als demokratische und realweltliche Umsetzung dieser Meta-Verfahren wirken kann, indem es die öffentliche Verhandlung, Abwägung und Argumentation über kontroverse soziale Gründe strukturiert.[18] Die Konstruktion eines pluralistischen aSRM erweitert die meist konsequentialistischen zeitgenössischen Ansätze[19] zur Maschinenethik und moralischen intelligenten Systemen erheblich.

VERNÜNFTIGE MASCHINEN

Unser aSRM-basierter Ansatz stellt einen bedeutenden Schritt in Richtung „Vertrauenswürdigkeit durch Design" dar[20] und schlägt in psychologischer Terminologie eine langsame, rationale (symbolische) *System 2*-Ebene in KI-Systemen vor, die ihre schnellen, „intuitiven", aber opaken (subsymbolischen) Berechnungen auf *System 1*-Ebene rechtfertigen, kontrollieren und (langfristig) sogar trainieren kann.

Die zentralen Innovationen unseres Forschungsansatzes sind:

- **Hybride-reflexive KI-Architektur:** Integration von subsymbolischen und symbolischen KI-Systemen mit reflexiven (autonomen und kollaborativen) Rückkopplungen parallel zur menschlichen Urteilsbildung und Kommunikation als „Modell künstlicher sozialer Rechtfertigung" (artificial Social Reasoning Model, aSRM);

16 Rawls, J (2001): Justice as Fairness: A Restatement. Cambridge: Harvard University Press; Raz, J (1999): Engaging Reason: On the Theory of Value and Action. Oxford: Oxford University Press.

17 Habermas, J (1981): Theorie des kommunikativen Handelns. Frankfurt/M: Suhrkamp.

18 Alexy, R (1978): Theorie der juristischen Argumentation. Frankfurt/M: Suhrkamp. Lomfeld, B (2015): Die Gründe des Vertrages: Eine Diskurstheorie der Vertragsrechte. Tübingen: Mohr Siebeck.

19 Bonnefon, J-F; Shariff, A; Rahwan, I (2016): The social dilemma of autonomous vehicles. In: Science 352.6293, S. 1573–1576. DOI: 10.1126/science.aaf2654; Greene, J; Rossi, F; Tasioulas, J et al. (2016): Embedding Ethical Principles in Collective Decision Support Systems. In: Schuurmans, D; Wellman, M P (Hrsg.): Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12–17, 2016, Phoenix, Arizona, USA, S. 4147–4151.

20 Hamon, R; Junklewitz, H; Sanchez Martin, J (2020): Robustness and Explainability of Artificial Intelligence, JRC technical report, JRC119336, EUR 30040 EN. DOI: 10.2760/57493.

- **Pluralistische Argumentationstools und -modelle:** Holistische Repräsentationen und Tools auf der Ebene der symbolischen (logischen) Verarbeitung werden mit pluralistischen normativen (Werte-)Ontologien kombiniert, um hybride Integration und Rückkopplung mit subsymbolischer KI-Technologie zu ermöglichen;

- **Vertrauen-durch-Kommunikation:** Paradigmenwechsel in der interpretierbaren und vertrauenswürdigen KI von „Erklärbarkeit" hin zu „rationaler Kommunikation" im Sinne von kommunikativer (Inter-)Aktion durch und mit Maschinen.

Wir reagieren mit diesem Forschungsansatz auf zentrale Innovationsherausforderungen, wie sie auf nationaler und europäischer Ebene identifiziert wurden. Anwendungsstudien unseres Modells beispielsweise im Bereich autonomer Mobilität und Scoring-Algorithmen (beispielsweise für Websuche, Reputationsbestimmung oder Kreditvergabe) liefern einen wertvollen Beitrag zur lebhaften, sozialen und politischen Debatte über ethische Herausforderungen von KI. Die Entwicklung erster Fähigkeiten zur (autonomen oder kollaborativen) normativen Argumentation und Kommunikation bei intelligenten Maschinen soll diese perspektivisch zu realem kommunikativen Handeln innerhalb sozialer Systeme befähigen und deren Beziehung zur Gesellschaft als Ganzes verändern. In diesem Sinne widmet sich unser Ansatz der Entwicklung von (und der Reflektion über) praktisch „vernünftige" (reasonable) Maschinen.

In der praktischen Philosophie wird die Fähigkeit über Handlungsgründe nachzudenken, allgemein als „praktische Vernunft" bezeichnet.[21] Wenn künstliche intelligente Systeme normative Gründe hervorbringen und erkennen, die für oder gegen ihre Handlungsentscheidungen sprechen, rechtfertigen sie ihre Handlungen auf in dieser Hinsicht praktisch „vernünftige" Weise. Aus einer pragmatischen gesellschaftlichen Perspektive kann jeder intelligente (menschliche oder künstliche) Agent als praktisch „vernünftig" angesehen werden, wenn er für sein Verhalten „Gründe geben und nehmen" kann,[22] die als „kommunikatives Handeln" sozial akzeptabel sind.[23] Die Erforschung von Methoden und Werkzeugen, die es Maschinen ermöglichen, normative Gründe zu generieren, zu verarbeiten und zu kommunizieren, ebnet den Weg nicht nur für neue Dimensionen der

21 Ratz, J (1999): Practical Reason and Norms. Oxford: Oxford University Press.

22 Brandom, R (1994): Making It Explicit. Cambridge: Harvard University Press.

23 Habermas, J (1981): Theorie des kommunikativen Handelns. Frankfurt/M: Suhrkamp.

Mensch-Maschine-Interaktion, sondern auch für Reflexionen über eine umfassendere künstliche „Vernunft" und die Natur moralischer Handlungsfähigkeit insgesamt.

Neben diesen weitgreifenden und kontroversen Forschungsfragen bleibt das ganz reale Ziel, das Vertrauen in zukünftige KI-Systeme eben dadurch zu stärken, dass autonome Entscheidungen und Handlungen nicht nur erklärt und kontrolliert, sondern kommunikative Interaktionen über deren Gründe ermöglicht werden. Dies bedeutet einen Paradigmenwechsel von Erklärbarkeit und Transparenz zu sozialer Rechtfertigung in der Form „vernünftiger" Kommunikation. Erfolgreiche Kommunikation ist dabei offensichtlich kontextabhängig. Im Bereich autonomer Mobilität etwa macht es für die Generierung sinnvoller Begründungen einen wesentlichen Unterschied, ob die Fehlfunktion einer Fahrzeugkomponente in einer Simulation während des Software-Designs verfolgt und überwacht wird, in Echtzeit dem Fahrer des Fahrzeugs mitgeteilt, später für einen technischen Experten rekonstruiert oder nach einem Unfall zu Zwecken rechtlicher Aufarbeitung auf Gründe befragt wird.

Da normative Kommunikation aber nie rein instrumentell bleibt, sondern gesellschaftliches Miteinander und soziale Systeme konstituiert,[24] könnte bereits ein pragmatischer erster Schritt zu autonomer oder kollaborativer Begründungsfähigkeit die Mensch-Maschine-Beziehung tiefgreifend verändern. Selbst wenn Maschinen so nie von Gründen träumen, könnten sie dann ethisch „vernünftige" Gründe kommunizieren.

24 Luhmann, N (1984): Soziale Systeme: Grundlage einer allgemeinen Theorie. Frankfurt/M: Suhrkamp.

Rainer Mühlhoff

PRÄDIKTIVE PRIVATHEIT: WARUM WIR ALLE „ETWAS ZU VERBERGEN HABEN“

Künstliche Intelligenz (KI) ist seit einigen Jahren wieder ein „gehyptes Thema“. Bestes Indiz dafür ist die öffentliche Aufmerksamkeit, die sich auf die Potentiale von KI und die Zukunftsprognosefähigkeit richtet. Selbstfahrende Autos, humanoide Roboter oder Betriebssysteme, in die man sich verlieben kann, rufen gleichermaßen utopische und dystopische Phantasien auf den Plan, die fließend in Science-Fiction übergehen. Auch ethische Zweifel werden dabei laut, die sich beispielsweise darum drehen, in welcher Weise den autonomen Maschinen moralische Verantwortung für ihr Handeln und ein eigener moralischer Status zugesprochen werden müssen.

Weniger im Zentrum der populären KI-Narrative stehen jene schon jetzt verwendeten datenbasierten KIs, die in die alltäglichen digitalen Medien eingewoben sind: Suchmaschinen; Nachrichtenkuratierung in sozialen Medien; Scoring und Ranking von Menschen im Kontext von Versicherungen und Finanzdienstleistungen; psychologisches Profiling anhand von Verhaltensdaten im Marketing, am Arbeitsplatz, im Bildungsbereich und in den sozialen Netzwerken; KI-unterstützte Polizei- oder Justizarbeit. In diesen Anwendungsfeldern tritt KI in Form von „Prognosesystemen“ auf Grundlage von Machine Learning in Erscheinung. Auch als „prädiktive Analytik“ bezeichnet, bilden solche Prognosesysteme die Grundlage für algorithmisches Entscheiden, Profilbildung und soziale Selektion von Menschen.

Prognosesysteme präsentieren sich allerdings nicht als Roboter, sie sind kein verkörpertes, sprechendes oder handelndes Gegenüber der Menschen in konkreten Interaktionssituationen. Prognosesysteme leben vielmehr in Rechenzentren, entziehen sich der Sichtbarkeit, und haben dennoch schon jetzt enorme Auswirkungen auf unser Denken, Fühlen und Handeln in zahlreichen Lebensbereichen wie Politik, Arbeit, Konsum und sozialer Kommunikation. In diesem Beitrag gebe ich einen kurzen Überblick, worum es sich bei Prognosesystemen genau handelt, worin die Gefahren bestehen und warum demokratische Gesellschaften mit einem neuen Verständnis von Datenschutz und Privatheit darauf reagieren sollten.

KI-BASIERTE PROGNOSESYSTEME

Mit „Prognosesystem" beziehe ich mich auf Machine-Learning-Modelle, die als Input eine Reihe verfügbarer Daten über ein Individuum (oder einen „Fall") erhalten und als Ausgabe die Schätzung einer Zielvariable für dieses Individuum zurückgeben. Die Inputdaten sind dabei typischerweise große Mengen unstrukturierter Hilfsdaten, die leicht zugänglich sind, zum Beispiel Trackingdaten (Browser- oder Standortverläufe) oder Social Media Daten (Likes, Postings, Freund*innen, Gruppenmitgliedschaften), während es sich bei der Zielvariable typischerweise um schwer zugängliche oder besonders sensible Daten handelt, zum Beispiel die Bonität der betroffenen Person, Krankheiten, Suchtverhalten, politische Affinitäten, Geschlecht, sexuelle Orientierung oder psychologische und emotionale Dispositionen.

In der „prädiktiven Analytik" möchte man also anhand leicht zugänglicher Daten schwer zugängliche Daten über Individuen abschätzen. Prädiktive Analytik entsteht überall dort, wo durch alltäglich verwendete digitale Medien massenweise Verhaltens- und Nutzungsdaten anfallen. Das liegt daran, dass prädiktive Modelle den einzelnen Fall anhand von „pattern matching" mit Millionen anderer Fälle abgleichen, ihn einer algorithmisch bestimmten Gruppe besonders ähnlicher Fälle zuordnen und daraus eine Schätzung der unbekannten Zielvariable ableiten. In den meisten Fällen werden solche Modelle mit Verfahren des überwachten Lernens trainiert: Dazu wird eine große Menge sog. „Trainingsdaten" benötigt, ein Datensatz, in dem für eine Kohorte von Individuen beide Datenfelder, also sowohl die Hilfsdaten als auch die sensiblen Zieldaten, erfasst sind. Solche Datensätze fallen regelmäßig im Kontext sozialer Alltagsmedien an, zum Beispiel produziert die Teilmenge aller Facebook-Nutzer*innen, die in ihrem Profil explizit Angaben über ihre sexuelle Orientierung machen, einen Trainingsdatensatz zur Abschätzung der sexuellen Identität; und die Gruppe der Individuen, von denen man gleichzeitig Zugriff auf ihren Browserverlauf und die Daten einer Gesundheits-App hat, produzieren Trainingsdaten für eine KI, die anhand von Browserverläufen Krankheitsdispositionen abzuschätzen lernen kann.

Sobald eine Gruppe von einigen Tausend Individuen freiwillig oder unwissentlich zugleich Hilfsdaten und sensible Daten preisgibt, kann ein Machine-Learning-Modell trainiert werden, welches Korrelationen zwischen den Hilfsdaten und

den sensiblen Daten ermittelt.[1] Solche Modelle werden anschließend dazu verwendet, die sensible Zielvariable auch für Individuen abzuschätzen, über die nur die Hilfsdaten bekannt sind und die selbst keine sensiblen Daten über sich preisgeben würden.

Mediziner*innen von der University of Pennsylvania haben gezeigt, dass sich mit dieser Vorgehensweise anhand von Facebook-Daten vorhersagen lässt, ob ein*e Nutzer*in an Krankheiten wie Depression, Psychosen, Diabetes oder Bluthochdruck leidet.[2] Facebook selbst hat bekannt gegeben, mittels Künstlicher Intelligenz suizidale Nutzer*innen anhand ihrer Postings erkennen zu können. Eine weitere Studie zeigt, dass die Daten über Facebook-Likes dazu verwendet werden können, „eine Reihe höchst sensibler persönlicher Attribute vorherzusagen, darunter sexuelle Orientierung, Ethnie, religiöse und politische Ansichten, Persönlichkeitseigenschaften, Intelligenz, happiness, Suchtverhalten, Trennung der Eltern, Alter und Geschlecht".[3]

Solche prädiktiven Analysen stoßen zum Beispiel bei Versicherungskonzernen auf großes Interesse, weil sie eine individuelle Risikobemessung erlauben. Krankenversicherungen können ihre Kund*innen mit Rabatten zum Gebrauch eines Fitness-Trackers motivieren, dessen Daten zentral gespeichert werden und so mit den Behandlungsdaten der Krankenkassen korreliert werden können, um individuelle Risikoprofile zu bestimmen. So genannte „Pay-as-you-Drive"-Tarife von KFZ-Versicherungen verwenden Positions-Tracking und Beschleunigungssensoren in den Fahrzeugen, um mittels prädiktiver Analytik individuelle Versicherungsprämien in Abhängigkeit vom Fahrstil und Aufenthaltsort zu bestimmen. Im Human-Resource-Management werden Bewerber*innen bei Jobausschreibungen mittels „hiring algorithms" vorsortiert, ein Verfahren, das in den USA bereits heute bei einer Mehrheit der Einstellungsverfahren verwendet wird.[4]

1 Ob solche Modelle statistisch valide sind, ist hiermit nicht ausgesagt; in vielen Fällen sind sie es nicht, zum Beispiel, weil die Trainingsdaten nicht repräsentativ für die relevante Grundgesamtheit sind. Entscheidend ist an dieser Stelle, dass diese Verfahren häufig ungeachtet statistischer Erwägungen trotzdem angewendet werden; siehe Mühlhoff, R (2020): Automatisierte Ungleichheit: Ethik der Künstlichen Intelligenz in der biopolitischen Wende des Digitalen Kapitalismus. In: Deutsche Zeitschrift für Philosophie, 68(6), S. 867–890. DOI: 10.1515/dzph-2020-0059.

2 Merchant, R M et al. (2019): Evaluating the Predictability of Medical Conditions from Social Media Posts. In: PLOS ONE 14, Nr. 6. DOI: 10.1371/journal.pone.0215476.

3 Kosinski, M et al. (2013): Private Traits and Attributes are Predictable from Digital Records of Human Behavior. In: Proceedings of the National Academy of Sciences 110, Nr. 15. DOI: 10.1073/pnas.1218772110.

4 O'Neil, C (2016): Weapons of Math Destruction. New York: Crown, S. 108, 148.

Zu den ersten Anwendungen prädiktiver Analytik gehört die gezielte Werbung. So ist es einer US-amerikanischen Supermarktkette im Jahr 2011 gelungen, anhand der Einkaufsdaten, die über Rabattprogramme (customer loyalty cards) gesammelt werden, schwangere Kundinnen zu identifizieren.[5] Im Credit Scoring wird ebenfalls schon lange auf prädiktive Modelle zurückgegriffen, die durch Machine Learning eine weitere Verfeinerung erhalten haben. „All data is credit data" lautet der Leitspruch jener Sparte der Finanzindustrie, die mit alternativen Kreditrisikomodellen auf Grundlage von Verhaltens- und Nutzungsdaten auch noch diejenigen mit Krediten versorgen möchte, die nach klassischem Ermessen nicht kreditwürdig sind: So genannte „payday lending"-Anbieter wie das vom Ex-Google-Mitarbeiter Douglas Merrill gegründete Fintec-Unternehmen ZestFinance oder die deutsche Firma Kreditec nutzen mit KI-basierten Kreditangeboten die Ausweglosigkeit sozioökonomisch schlechter gestellter Menschen aus – mit Jahreszinssätzen, die nicht selten 300 Prozent übersteigen.[6]

BIOPOLITIK: VON DER KUNDENVERWALTUNG ZUM BEVÖLKERUNGSMANAGEMENT

Auch wenn die Finanz- und Versicherungsbranche und das individualisierte Marketing der sozialen Medien die Entwicklung prädiktiver Analytik hauptsächlich angetrieben haben, kommt es für eine aktuelle ethische und politische Besprechung dieses Themas darauf an, dass das Anwendungsfeld dieser Technologie allmählich den Bereich der B2C-Operationen („Business-to-Costumer") überschreitet und zunehmend die Relationen zwischen Staat und Staatsbürger*innen („Government-to-Citizen", G2C) erfasst. Profilbildung, Risiko-Scoring und automatisierte Entscheidungsfindung werden nicht mehr nur für selektive Information, differenzielle Preisgestaltung und andere Strategien eingesetzt, die lediglich das wirtschaftliche Verhalten von Marktteilnehmern beeinflussen. Vielmehr werden diese Techniken nun dafür verwendet, die Beziehung des Einzelnen zum Staat, einschließlich der wohlfahrtsstaatlichen Institutionen, der Bildungs- und Gesundheitseinrichtungen, des Sicherheitsapparats und der Politik zu prägen. Das zeigt sich beispielsweise im Feld der prädiktiven Polizeiarbeit

5 Duhigg, C (2012): How Companies Learn Your Secrets. In: The New York Times, 16. Februar 2012, https://www.nytimes.com/2012/02/19/magazine/shopping-habits.html [20.11.2020].

6 Lippert, J (2014): ZestFinance Issues Small, High-Rate Loans, Uses Big Data to Weed out Deadbeats. In: Washington Post, 11. Oktober 2014.

und Kriminologie,[7] in der Gesundheitsversorgung, im Bildungsbereich sowie im Jugendschutz.[8]

Die Bereichserweiterung der KI von den B2C- zu den G2C-Relationen vollzieht sich dabei *nicht*, wie man meinen könnte, indem KI-Technologie vom privaten in den öffentlichen Sektor übertragen wird. Zu beobachten ist hingegen eine tendenzielle Integration staatlicher Institutionen in private Plattformen und Datennetzwerke, denn Machine Learning-Technologie ist so tief mit privaten Daten-Infrastrukturen verstrickt, dass sie nicht anderswo und getrennt von den Wirtschaftsakteuren repliziert werden kann. Die Verwendung von KI im G2C-Bereich bedeutet deshalb eine Privatisierung öffentlicher Fürsorgeinfrastruktur; staatliche Institutionen und G2C-Kommunikation werden schrittweise in die Wertschöpfungslogiken des digitalen Kapitalismus integriert.

Diese Entwicklung bezeichne ich als „Biopolitik der KI".[9] Damit ist ein Staats- und Wirtschaftswesen des algorithmischen Bevölkerungsmanagements gemeint, das tief in die biologischen, ökonomischen und sozialen Prozesse integriert ist und sich als Kontrollapparat manifestiert. Die Biopolitik der KI verfährt dabei nach dem Prinzip der „Gruppenhaft", das dem oben erwähnten „pattern matching" inhärent ist und eine prognostische Verwaltung großer Kohorten und Menschenmengen im Stil eines Risiko-Controllings ermöglicht: Jedes Individuum wird individuell (als Risikofaktor) erfasst, angesprochen und behandelt, dabei aber doch nicht aus dem Glaskäfig einer virtuellen Vergleichsgruppe entlassen.

Die KI-basierte algorithmische Bevölkerungsverwaltung hat deshalb den Effekt, soziale Ungleichheit zu zementieren oder sogar weiter anzufachen. Virginia Eubanks hat dafür den Begriff der „Automatisierung von Ungleichheit" ins Spiel gebracht:[10] durch den Einsatz der Algorithmen entsteht – teils entlang neuer, teils entlang bestehender Strukturen – ein unbemerktes soziales Gefälle, das dadurch geprägt ist, dass Individuen in Bezug auf ihren Zugriff auf staatliche Ressourcen, ihrer Belegung mit Beschränkungen (Sicherheit, Polizeiarbeit, Jugendschutz, öffentliche Gesundheit) und ihren ökonomischen Chancen unterschiedlich behandelt werden. Damit kommt es zu einer computerisierten Form der sozialen

7 Hao, K (2019): AI Is Sending People to Jail – and Getting it Wrong. In: MIT Technology Review, 21. Januar 2019, https://www.technologyreview.com/2019/01/21/137783/algorithms-criminal-justice-ai [20.11.2020].

8 Eubanks, V (2017): Automating inequality. New York: St. Martin's Press.

9 Mühlhoff, R (2020): Automatisierte Ungleichheit, a. a. O.

10 Ebd.

Selektion, die die Gesellschaft in unsichtbare soziale Klassen unterteilt, z. B. in solche Menschen, die mutmaßlich ein Sicherheits- oder Gesundheitsrisiko darstellen, besseren oder schlechteren Zugang zu medizinischer Versorgung erhalten, aufgrund ihres Lernverhaltens in der Schule oder an der Universität mit einem privilegierten Zugang zu bestimmten Berufen rechnen können oder eher Opfer häuslicher Gewalt werden und deshalb präventiv vom Jugendschutz überwacht werden sollten.[11]

PRÄDIKTIVE PRIVATHEIT

Angesichts der skizzierten Entwicklung, in der KI in die Bereiche der sozialen Beziehungen, der Politik, des Sicherheits- und Justizwesens und der öffentlichen Verwaltung vordringt, stehen unsere Gesellschaften vor einer grundlegend neuen Herausforderung des Datenschutzes. Im ersten Abschnitt wurde gezeigt, dass mittels prädiktiver Analytik anhand von leicht zugänglichen und oft anonym erhobenen Hilfsdaten (z. B. Trackingdaten, Bewegungsprofile, Social Media Daten) sensible Informationen über beliebige Individuen abgeleitet werden können – auch ohne deren Wissen oder Zustimmung. In dieser Konstellation tritt nun ein neuer Typus der Verletzung von Privatsphäre zutage, mit dem viele Menschen im Alltag noch gar nicht rechnen, denn klassischerweise stellt man sich die Verletzung der Privatsphäre als intrusiven Akt vor, in dem sensible Daten gezielt entwendet oder zweckentfremdet werden. Durch prädiktive Analytik kann die Privatsphäre eines Individuums aber verletzt werden, indem sensible Informationen aus anderen Daten *abgeleitet* oder *vorhergesagt* werden. Wir stehen deshalb angesichts prädiktiver Analytik vor einem neuen ethischen und politischen Problem: Sensible Informationen werden hier nicht durch ein Datenleck oder unerlaubte Weitergabe zuvor erhobener Daten preisgegeben, sondern durch Abschätzung von Verhaltensähnlichkeiten in einem kollektiv produzierten Datenpool.

Um den möglichen Missbrauch dieser Technologie zu problematisieren, benötigen wir ein erweitertes Verständnis von Privatsphäre, das sich auch auf *abgeschätzte*, nicht nur auf explizit *erhobene* Informationen erstreckt. Die Privatheit einer Person ist auch dann verletzt, wenn sensible Informationen ohne ihr Wissen und gegen ihren Willen durch Vergleich mit vielen anderen Personen abgeschätzt werden. Ich bezeichne diese Herangehensweise an Datenschutz als „prädiktive Privatheit".[12]

11 Ebd., S. 127ff.

12 Mühlhoff, R (2020): Predictive Privacy: Towards an Applied Ethics of Data Analytics. SSRN preprint, https://ssrn.com/abstract=3724185.

Dieser Begriff kann den Ausgangspunkt für eine effizientere gesetzliche Regulierung von Big-Data- und KI-Anwendungen bilden und helfen, eine Lücke zu schließen, die durch bestehende Datenschutzgesetzgebung teilweise mit produziert wird: Während die Verarbeitung von Daten über geschützte Attribute wie Geschlecht, sexuelle Orientierung, Religionszugehörigkeit, Ethnie etc. in den meisten Datenschutzgesetzgebungen streng geschützt ist und deshalb mit hohen rechtlichen Risiken und operativen Herausforderungen (z. B. Informations- und Einwilligungsverfahren oder sichere Datenspeicherung) verbunden ist, gehen Unternehmen dazu über, statt dieser Datenfelder mit sogenannten „proxies" zu arbeiten. Dabei handelt es sich um algorithmisch bestimmte Kombinationen von (ungeschützten) Hilfsdaten, die mittels prädiktiver Analytik eine Vorhersage über die geschützten Attribute zulassen. Prädiktive Analytik birgt also nicht nur neue Möglichkeiten der Privatsphäreverletzung, sondern kann auch dafür eingesetzt werden, das bestehende Datenschutzniveau zu unterlaufen.

Um den Schutz prädiktiver Privatheit wirkungsvoll in Regulierung umsetzen zu können, benötigen wir zuerst ein breites gesellschaftliches Bewusstsein für den Mechanismus prädiktiver Privatsphäreverletzungen. KI-basierte Prognosen sind nur möglich, wenn und weil viele Bürger*innen keine Bedenken haben, ihre Daten freiwillig (und ggfs. anonymisiert) zur Verfügung zu stellen. „Ich habe doch nichts zu verbergen" ist eine weit verbreitete moralische Haltung, die es großen Plattformunternehmen erst ermöglicht, umfassende Trainingsdatensätze zu generieren. Wie weiter oben ausgeführt, wird zum Training eines prädiktiven Modells eine Gruppe von Nutzer*innen benötigt, die sowohl die Hilfsdaten *als auch* die sensiblen Zieldaten über sich preisgeben. Auf gesellschaftlichem Maßstab betrachtet, reicht dafür in vielen Fällen eine Minderheit aus, die für sich keine Probleme bei der Preisgabe dieser Daten sieht oder sich ihrer nicht bewusst ist. Oft ist diese Einstellung bei Personen zu finden, die sich selbst für „normal" halten, nicht davon ausgehen, dass sie in den Fokus von „Überwachung" gelangen könnten, und die selbst in ihrer gesellschaftlichen Position keine negativen Auswirkungen durch prädiktive Analytik erlebt haben. Doch nur anhand der Daten vieler „normaler" Nutzer*innen, die meinen „nichts zu verbergen zu haben", lassen sich die prädiktiven Algorithmen trainieren, die *andere* Individuen als Abweichler*innen erkennen können. Sich der Auswirkungen des eigenen Umgangs mit sensiblen Daten für Dritte bewusst zu werden, ist die ethische Dimension prädiktiver Privatheit. Denn auch die weltweit fortschrittliche europäische Datenschutzgrundverordnung ist gegen die Gefahren von Big Data und KI weitestgehend wirkungslos, insbesondere, wenn sie dadurch zustande

kommen, dass zahlreiche Nutzer*innen in die Verarbeitung ihrer sensiblen Daten einwilligen.[13]

Dass KI-basierte Prognosesysteme zurzeit noch weitestgehend ohne wirksame regulatorische Hürden betrieben werden können, liegt auch an der weit verbreiteten, jedoch zu eng gefassten liberalistischen Auslegung von Datenschutz: Jede Person soll nach eigenem Ermessen entscheiden können, was mit den eigenen Daten passiert. Diese Herangehensweise ist eng mit dem klassischen bürgerlichen Verständnis von Privatheit verknüpft, das auf die Kontrolle des Zugangs zur eigenen privaten Sphäre zugerichtet ist. Doch im Zeitalter von Big Data und prädiktiver Analytik kommt dieses individualistische Prinzip an seine Grenzen: Daten, die man selbst freiwillig weitergibt, können dazu verwendet werden, sensible Informationen über *andere* Menschen abzuschätzen; und umgekehrt kann man selbst aufgrund der Daten, die *andere* über sich preisgeben, unterschiedlich behandelt werden. Es kann uns also nicht egal sein, wie unsere Mitmenschen mit ihren Daten umgehen. Und weil die negativen Auswirkungen prädiktiver Analytik nicht auf alle Gesellschaftsmitglieder gleich verteilt sind, sondern überproportional die Armen, weniger Gebildeten, Schwachen, Kranken und sozioökonomisch Benachteiligten treffen, stehen demokratische Gesellschaften hier in einer *kollektiven Verantwortung:* Wir alle müssen dafür sorgen, dass mit unseren Daten kein Missbrauch getrieben werden kann – und es ist eine gute Faustregel, davon auszugehen, dass ein solcher Missbrauch in den meisten Fällen *nicht* uns selbst trifft.

13 Wachter, S (2019): Data Protection in the Age of Big Data. In: Nature Electronics 2 (1), S. 6–7.

Simone Kaiser, Marie Lena Heidingsfelder und Martina Schraudner

IF YOU WANT TO GO FAR, GO TOGETHER: GESELLSCHAFTS-FORESIGHT UND ZUKUNFTSBILDER ALS SCHLÜSSEL FÜR VERANTWORTLICHE KI-GESTALTUNG

Verantwortung als Standortfaktor für deutsche und europäische KI-Entwicklung

Für die Europäische Kommission und die deutsche Bundesregierung ist die Sache klar: *AI made in Europe* kombiniert technologische Exzellenz mit Vertrauen und sichert so eine gemeinwohlorientierte Entwicklung von Künstlicher Intelligenz.[1] AI made in Europe ist die „europäische Antwort auf datenbasierte Geschäftsmodelle […], die unserer Wirtschafts-, Werte- und Sozialstruktur entspricht".[2] Aber nicht nur Regierungen haben inzwischen Anforderungen an KI-Technologien formuliert, auch die großen Tech-Unternehmen entwerfen Leitlinien zur verantwortlichen Gestaltung von KI-Technologien und deren Anwendungen.[3] Die Anforderung, KI verantwortlich zu gestalten, ist längst mehr als nur eine warnende Stimme von Datenschützer*innen – sie ist zum Standortfaktor für europäische und deutsche KI-Entwicklung und die darauf aufbauenden Geschäftsmodelle geworden. KI-Technologien brauchen Daten. Angesichts einer immer stärkeren Verschränkung des *Real Life* mit unseren virtuellen Identitäten und Datenspuren wird deutlich: Die verantwortliche Gestaltung von neuen KI-Technologien kann nicht allein in Forschungs- und Entwicklungsabteilungen realisiert werden. Verantwortliche KI-Gestaltung ist keine Fingerübung für Technologiespezialisten, sie betrifft uns als Gesellschaft insgesamt: Denn weit mehr als uns Filmempfehlungen auf Netflix zu geben, kommt KI die Aufgabe zu, Pluralität, Selbstbestimmung und Teilhabe – zentrale Funktionsmechanismen demokratischer Gesellschaften – auch zukünftig sicherzustellen. Diesem Anspruch werden wir nur gerecht, wenn wir die Akzeptanzbedingungen und ethischen Fragen rund um KI frühzeitig unter umfassenden Einbezug gesellschaftlicher Akteure erörtern.

In diesem Beitrag gehen wir der Frage nach, wie ein Partizipationsprozess gestaltet werden kann, der einen breiten und chancenorientierten

1 Europäische Kommission (2020): Weißbuch zur Künstlichen Intelligenz – ein europäisches Konzept für Exzellenz und Vertrauen, Brüssel.

2 Deutsche Bundesregierung (2018): Strategie Künstliche Intelligenz der Bundesregierung, Berlin, S. 9, online unter: https://www.bmbf.de/files/Nationale_KI-Strategie.pdf [20.10.2020].

3 Fjeld, J; Achten, N; Hilligoss, H et al. (2020): Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI, Berkman Klein Center Research Publication No. 2020-1.

gesellschaftlichen Dialog über wünschenswerte KI-Zukünfte initiiert und die Ergebnisse dieses Dialogs anschließend wieder für die Technologieentwicklung nutzbar macht. Dabei plädieren wir für die Umsetzung eines *Gesellschafts-Foresights* und die Gestaltung von *Zukunftsbildern*, um gesellschaftliche Wünsche und Ansprüche an neue KI-Technologien zu identifizieren sowie gemeinwohlorientierte technologische Zukünfte zu gestalten.

GESTALTUNGSORIENTIERTER GESELLSCHAFTS-FORESIGHT ALS GRUNDLAGE VERANTWORTLICHER KI-GESTALTUNG

Vertrauen und gesellschaftliche Akzeptanz sind der Schlüssel für eine erfolgreiche Einführung neuer KI-Technologien.[4] Um diesen Anspruch umzusetzen, müssen neue Strategien in Forschung und Technologieentwicklung ethische Konfliktfelder und Akzeptanzbedingungen, beispielsweise in Bezug auf die informationelle Selbstbestimmung, möglichst frühzeitig in den Blick nehmen, um diese proaktiv berücksichtigen zu können. Am Fraunhofer Center for Responsible Research and Innovation (CeRRI) verstehen wir *verantwortliche KI* als den partizipativen Gestaltungsprozess und die ethischen Rahmenbedingungen, die das Design, den Einsatz, die Verwendung sowie die Kontrolle gemeinwohl- und menschenzentrierter AI-Systeme ermöglichen. Diese Definition verweist auf zwei relevante Dimensionen: Einmal auf den partizipativen Gestaltungsprozess, über den Vertrauen und Akzeptanz hergestellt werden können, zum anderen auf die gleichzeitige inhaltliche Verhandlung und Verständigung im Rahmen dieses partizipativen Prozesses über die ethischen Akzeptanzbedingungen.

Ein strategisches Werkzeug für Unternehmen, Forschungseinrichtungen und politische Akteure, um verantwortliche KI umzusetzen, kann ein gestaltungsorientierter Gesellschafts-Foresight sein.[5] Die traditionelle Zukunftsforschung hat den Anspruch, sich auf mögliche Zukunftsszenarien möglichst gut vorzubereiten. Die Stichworte sind hier oft „preparedness" oder „resilience". Im Unterschied zu dieser

4 Europäische Kommission (2020) a. a. O.; Weber, M; Buschbacher, F (2017): Künstliche Intelligenz–Wirtschaftliche Bedeutung, gesellschaftliche Herausforderungen, menschliche Verantwortung. Bitkom eV. und DFKI, Berlin/Kaiserslautern.

5 Kaiser, S; Glatte, H; Schroth, F et al. (2019): Vorausschau². Neue Impulse für Strategie und Innovation durch Verknüpfung von Technologie- und Gesellschaftsvorausschau, Fraunhofer Verlag, Stuttgart.; Kaiser, S; Glatte, H; Bitter, F et al. (2018): Zukunftsgestaltung als kollaborativer Prozess. Designbasierte Zukunftsszenarien als Strategietool in komplexen Ökosystemen. In: Gausemeier, J; Bauer, W; Dumitrescu, R (Hrsg.) (2018): 14. Symposium für Vorausschau und Technologieplanung, 8. und 9. November 2018, Universität Paderborn, Berlin.

traditionellen Mission der Zukunftsforschung fokussieren gestaltungsorientierte Foresight-Ansätze auf Handlungsmöglichkeiten in der Gegenwart. Dem Foresight-Prozess geht es in diesem Verständnis weniger darum, mögliche Entwicklungen zu *kennen* oder die Wahrscheinlichkeit von Entwicklungen zu *bewerten*, sondern über die Identifikation erwünschter Entwicklungen Zukunft aktiv zu *gestalten.*[6] Damit liefern gestaltungsorientierte Foresight-Ansätze die methodische Grundlage für die Umsetzung eines Gesellschafts-Foresight, bei dem die Wissensbestände, Bedarfe und Perspektiven möglichst vieler verschiedener Stakeholder einbezogen werden, um den zukünftigen Möglichkeitsraum möglichst breit aufzuspannen. Auf dieser Grundlage lassen sich wünschenswerte Zukünfte verhandeln und identifizieren und darauf aufbauend entsprechende gemeinwohlorientierte Transformationsprozesse initiieren. So werden Bürgerinnen und Bürger befähigt, ihre Ansprüche an eine technologische Zukunft zu formulieren und wünschbare Zukunftsbilder zu beschreiben. Mit einem Gesellschafts-Foresight können unterschiedliche Ziele erreicht werden:

- **Inklusion und Partizipation:** Der Gesellschafts-Foresight zielt auf den Einbezug *breiter gesellschaftlicher Gruppen.* Indem visuelle und haptische Elemente in Zukunftsbildern genutzt werden, gelingt es, über Sprache als Dialogmedium hinauszugehen. Damit wird eine niedrigschwellige und intuitive Formulierung von individuellen Wünschen ermöglicht. Eine Einbindung von in Beteiligungsprozessen regelmäßig unterrepräsentierten Bevölkerungsgruppen, wie bspw. aus migrantischen oder bildungsfernen Milieus, wird dadurch erleichtert. Darüber hinaus fördern die Methoden gemeinsame Diskussionsprozesse und gestaltende, konstruktive Ergebnisse.

- **Frühzeitiger Einbezug:** Der Gesellschafts-Foresight ermöglicht eine aktive und partizipative Technologiegestaltung zu einem Zeitpunkt, an dem noch wenige technologische Pfadabhängigkeiten bestehen. Im Mittelpunkt der Gestaltung von Zukunftsbildern steht die Identifikation wünschbarer Zukünfte. Mit dieser *langfristigen Gestaltungsperspektive* und *normativen Orientierung* unterscheidet sich der Gesellschafts-Foresight von Beteiligungsprozessen, die auf Informationsvermittlung oder Akzeptanzbeschaffung für neue Technologien fokussieren.

6 Kaiser, S; Glatte, H; Schroth, F et al. (2019) a. a. O.; Tuomi, I (2013): Next-generation foresight in anticipatory organizations: background study for the European forum on forward-looking activities (EFFLA). European Commission, Brüssel.; Tully, C; Rhydderch, A; Glenday, P (2017): With foresight the frog might not croak. Online unter: https://www.chathamhouse.org/system/files/publications/twt/With%20foresight%20the%20frog%20might%20not%20croak.pdf.

- **Gemeinwohlorientierung:** Der Gesellschafts-Foresight betont und verankert eine Perspektive, die Technologieentwicklung nicht als linearen oder gar naturgesetzlichen Fortschrittsprozess versteht, sondern als gesellschaftlichen, sozialen und iterativen *Gestaltungs*prozess, der mit und für den Menschen erfolgen soll. Dies stärkt Akzeptanz und Vertrauen in die Technologieentwicklung.

- **Innovation:** Partizipation und Inklusion sichern nicht nur Vertrauen und Akzeptanz. Über die Einbeziehung unterschiedlicher Wissensbestände und Erfahrungshintergründe generieren sie neue Innovationsimpulse für Forschung und Entwicklung.

Damit der Anspruch des breiten Einbezugs gesellschaftlicher Akteure im Rahmen eines Gesellschafts-Foresights gelingen kann, kommen Zukunftsbilder als eine Methode aus der Designforschung („Design for debate") zum Einsatz.

Zukunftsbilder als Basis für einen breiten Dialog zu wünschenswerten KI-Zukünften werden auf Basis wissenschaftlicher Projektionen oder partizipativer Gestaltungsprozesse durch Wissenschaftler*innen in Zusammenarbeit mit Texter*innen und Designer*innen entwickelt. Sie ermöglichen eine Diskussion von erwünschten technologischen Zukünften, ohne dass die Rezipient*innen auf spezifisches technologisches Wissen und Fachtermini zurückgreifen müssen.[7] Sie bringen die Ergebnisse technologischer und gesellschaftlicher Foresight-Prozesse anhand von Bildern, Objekten und Narrationen (Texten, Filmen, Audios) auf den Punkt und zeigen dabei insbesondere die kontroversen und ambivalenten Aspekte einer möglichen Zukunft. Sie stellen also weder Utopien noch Dystopien dar, sondern leuchten die Grauzonen technologischer Innovationen aus. Bei der Entwicklung von Zukunftsbildern ist sowohl der visuelle als auch der narrative Aspekt von Bedeutung:[8] Der Einsatz von Bildern und Objekten eignet sich gut, um Interesse zu wecken und eine konkrete Vorstellung der möglichen Zukunft zu vermitteln, wohingegen der Einsatz von Narrationen es ermöglicht, komplexe Zusammenhänge zwischen Mikro-, Meso- und Makro-Ebene in soziotechnischen

7 Schraudner, M; Kaiser, S; Heidingsfelder, M et al. (2017): Shaping Future. Neue Methoden für Partizipation in Forschung und Innovation, Fraunhofer Verlag, Stuttgart.; Kaiser, S; Glatte, H; Schroth, F et al. (2019).

8 Heidingsfelder, M (2018): Zukunft gestalten. Design Fiction als Methode für partizipative Foresight-Prozesse und bidirektionale Wissenschaftskommunikation; Sterling, B (2009): Design Fiction. In: Interactions, 16(3), S. 20–24.; Wakkary, R; Odom, W; Hauser, S et al. (2016): A short guide to material speculation: Actual artifacts for critical inquiry. In: Interactions, 23(2), S. 44–48.

Systemen verständlich darzustellen. Insbesondere im Bereich von KI, der zum einen sehr abstrakt und zum anderen stark durch eine Aufspaltung des Diskurses in Utopien und Dystopien geprägt ist, bietet die Entwicklung ambivalenter und anschaulicher Zukunftsbilder eine gute Möglichkeit, in einen breiten gesellschaftlichen Dialog zu treten. So können aus einem partizipativen Prozess zur verantwortlichen Gestaltung von KI bestimmte Anwendungsszenarien und Touchpoints sichtbar werden.

Die Zahl der Menschen, die in Foresight-Prozesse einbezogen werden können, ist durch das Format selbst begrenzt: Die explorative Erarbeitung von Zukunftsbildern und der direkte Dialog mit Expert*innen und Stakeholdern erfordern ein hohes Maß an Zeit und Engagement, so dass eine gleichzeitig tiefe, aber auch repräsentative Beteiligung nur jeweils mit einem sehr hohen Aufwand möglich ist. Dies ist das sog. „Repräsentativitätsdilemma" partizipativer Prozesse:[9] Da eine vollständige Beteiligung aller Menschen aus den jeweiligen Zielgruppen in partizipativen Prozessen kaum möglich ist, muss eine Auswahl der relevanten Akteure getroffen werden – insbesondere bei einem Thema wie KI, das transformativ auf die gesamte Gesellschaft wirkt. Dieses Dilemma zieht praktische Fragen nach der notwendigen Größe partizipativer Verfahren nach sich, insbesondere wenn politische Steuerungsprozesse daraus abgeleitet werden sollen, denn: „In essence, legitimization follows representation, and ideally, authority follows legitimacy".[10] Für die Legitimität partizipativer Prozesse ist es somit wichtig, sowohl über Diversität als auch über die Anzahl der beteiligten Menschen ein möglichst breites und repräsentatives Bild gesellschaftlicher Bedarfe, Akzeptanzhaltungen und Wünsche zu erhalten. Um die Ergebnisse der „tiefen" Partizipation in einen breiten Dialog zu wünschenswerten Zukünften zu übertragen, eignen sich Zukunftsbilder im Besonderen, da sie als „Scharniere" zwischen der tiefen und explorativen Partizipation ausgewählter gesellschaftlicher Akteure und dem breiten Dialog fungieren: Zukunftsbilder bringen die relevanten und vor allem die kontroversen Ergebnisse der explorativen Phase visuell und narrativ auf den Punkt und bieten so die Basis für einen repräsentativen Dialog.

Um diesen Dialog produktiv zu führen, müssen die objekthaften und visuellen Bestandteile der Zukunftsbilder einerseits konkrete Möglichkeitsräume aufspannen, andererseits dürfen diese nicht schon klar vermessen sein. Nur so können

9 van der Helm, R (2007): Ten insolvable dilemmas of participation and why foresight has to deal with them. In: Foresight, 9(3), S. 3–17.

10 Ebd., S.9.

Diskussionsräume geöffnet werden und die Beteiligten ihre Perspektiven einbringen. Um dieses Ziel zu erreichen, sind mehrere Faktoren zu beachten: Der Ort (lokal oder virtuell), die Größe und die Dauer der Ausstellung der Zukunftsbilder, die Zielgruppe, der Einsatz von Medien (visuell, auditiv, filmisch), die Objekte selbst sowie die Interaktionsmöglichkeiten bestimmen, ob die Zukunftsbilder nicht nur einen (räumlichen) Zugang zu den Objekten bieten, sondern auch eine sinnstiftende Beschäftigung mit ihnen ermöglicht und so Partizipation fördern.

In der Gestaltung der Interaktionsmöglichkeiten bieten sich sozialwissenschaftliche Methoden wie quantitative und qualitative Umfragen an. Damit kann ein repräsentatives Bild der Anforderungen und Akzeptanzbedingungen erhoben werden. Um darüber hinaus den Austausch und Dialog zu fördern und gleichzeitig eine besonders hohe Reichweite für die Zukunftsbilder zu erlangen, haben moderierte und auswertbare Online-Plattformen besonderes Potential für die Dialoggestaltung. Die hier genannten Partizipations- und Feedbackoptionen zeichnen sich durch eine relativ kurze Interaktionsdauer aus, um breite Beteiligung zu ermöglichen. Die Ergebnisse der Zukunftsbild-Dialoge werden für die Technologieentwicklung nutzbar, wenn sie an KI-Akteure aus Forschung, Politik und Wirtschaft zurückgespielt werden. Zu diesem Zweck müssen die Ergebnisse dokumentiert und sozialwissenschaftlich ausgewertet werden, bevor sie in Gestaltungsempfehlungen und konkrete Technologie-Roadmaps übersetzt werden können. Gemeinsam mit Technologie-Expert*innen, Politiker*innen und Unternehmensvertreter*innen können auf dieser Basis Transformationsprozesse gestaltet werden.

FAZIT: PARTIZIPATION ALS SCHLÜSSEL FÜR DIE PRAKTISCHE UMSETZUNG VON VERANTWORTLICHER KI UND GEMEINWOHLORIENTIERTER TRANSFORMATION

Der Anspruch verantwortlicher KI-Gestaltung kann über einen gestaltungsorientierten Gesellschafts-Foresight in Kombination mit Zukunftsbildern als Basis für breite gesellschaftliche Debatten praktisch umgesetzt werden. Ein solch gestaltungsorientierter Partizipationsprozess schafft Räume, um offene ethische Fragen, Zielkonflikte und Risiken gesellschaftlich zu verhandeln sowie chancenorientierte und breit getragene KI-Zukünfte zu gestalten. Er liefert damit auch die Grundlage für einen gemeinwohlorientierten Transformationsprozess. Dabei liegt ein besonderes Potenzial in der Kombination eines gestaltungsorientierten Gesellschafts-Foresight mit der Methode der Zukunftsbilder, denn sie

erweitert die Möglichkeiten einer zahlenmäßig notwendigerweise begrenzten Gruppe um eine tiefe gesellschaftliche Auseinandersetzung. Hier lässt sich an den Potentialen und der Reichweite von Online-Dialogen anknüpfen. Darüber hinaus ist der Partizipationsprozess bidirektional: Einerseits erhalten gesellschaftliche Akteure Informationen und Wissen zu technologischen Entwicklungen, andererseits fließen ihre Bedarfe und Impulse zurück an die beteiligten Expertinnen und Experten. Auf diese Weise tragen die Ansätze dazu bei, eine gemeinsame Sprache und ein gemeinsames Verständnis zwischen unterschiedlichen Stakeholdern zu entwickeln und so vielfältige Wissensbestände und Perspektiven einzubeziehen. Ohne Frage ist ein solcher Partizipationsprozess zur Gestaltung verantwortlicher KI methodisch voraussetzungsvoll, zeitaufwendig und erfordert neue Kompetenzen in der Technologieplanung und der Foresightmethode. Diese Tatsache in einem schnellen globalen Wettbewerb als Gegenargument für solche Prozesse heranzuziehen, ist aber zu kurz gegriffen. Denn die Auswirkungen, die KI auf unsere Gesellschaft hat und haben wird, sind komplex und vielschichtig. Je frühzeitiger Akzeptanzbedingungen und ethische Fragen adressiert werden, desto eher können sie als Innovationsquelle genutzt und Lösungen entwickelt werden, die nicht nur technologisch funktionieren, sondern auch gesellschaftlich akzeptiert werden und damit auf dem Markt erfolgreich und rechtlich legitimiert sind. Ein gestaltungsorientierter Gesellschafts-Foresight ermöglicht dabei, Kriterien für verantwortliche KI wie bspw. Kontrolle, Fairness, Erklärbarkeit, Sicherheit, Rechenschaftspflicht oder Schutz der Privatsphäre in die Umsetzung zu bringen, zu konkretisieren und so Handlungssicherheit und einen wertebasierten Orientierungsrahmen für Entwickler*innen zu schaffen. Ein solcher Prozess kann die bestehende Technologieplanungs- und Foresightprozesse in Ministerien, in Forschungseinrichtungen und in Unternehmen erweitern und ergänzen; insbesondere dann, wenn europäische Werte frühzeitig in die Technologieentwicklung einzogen werden sollen. Damit können vorhandene Kompetenzen genutzt, Kapazitäten gebündelt und Synergieeffekte bspw. bezüglich ähnlicher ethischer Fragestellungen in unterschiedlichen KI-Anwendungsfeldern gebündelt werden. Wir Europäer*innen haben uns festgelegt: Wir möchten KI-Technologien auf Basis eines europäischen Wertefundaments entwickeln und nutzen. Wir möchten einen gemeinwohlorientierten technologischen Transformationsprozess gestalten. Dies erfordert die Zeit und den Mut, dieses Wertefundament in Zweifelsfällen auszuloten und zu ergründen. Genau darin liegt der Standortvorteil deutscher und europäischer Technologieentwicklung.

TITEL DER REIHE »#VERANTWORTUNGKI – KÜNSTLICHE INTELLIGENZ UND GESELLSCHAFTLICHE FOLGEN«

Heft 1/2020
Isabella Hermann, Georgios Kolliarakis, Fruzsina Molnár-Gábor, Timo Rademacher, Frauke Rostalski
VERTRAUENSWÜRDIGE KI?
VORAUSSCHAUENDE POLITIK!

Heft 2/2020
Isabella Hermann, Frauke Rostalski, Günter Stock
KOMPETENT EIGENE ENTSCHEIDUNGEN TREFFEN?
AUCH MIT KÜNSTLICHER INTELLIGENZ!

Der zunehmende Einsatz von sogenannter „Künstlicher Intelligenz“ (KI) verspricht viele Verbesserungen, beispielsweise durch Bilderkennung in der Medizindiagnose. Er birgt aber auch das Risiko, dass Menschen durch irrtümliche Vorhersagen von KI-Systemen zu Schaden kommen können. In solchen Fällen wird es immer schwieriger zu bestimmen, wer die Verantwortung trägt. Die Reihe #VerantwortungKI – Künstliche Intelligenz und gesellschaftliche Folgen bietet ein Forum für Beiträge über die ethischen, rechtlichen und gesellschaftspolitischen Chancen und Risiken des Einsatzes von KI mit einem besonderen Blick auf den Verantwortungsbegriff. Die Beitragsreihe wird von der interdisziplinären Arbeitsgruppe *Verantwortung: Maschinelles Lernen und Künstliche Intelligenz* betreut.

ISBN 978-3-939818-93-9