Geistes/Sozialwissenschaften

Erschienen in:
Positionen, 35/2021, 50-64
ISSN 1432-4059

Künstliche Intelligenz: natürlich gerecht?

Über Möglichkeiten und Grenzen der Ethik der Künstlichen Intelligenz. Eine Bestandsaufnahme am Beispiel von Sprachverarbeitungssystemen

Elisa Orrù[13], **Universität Freiburg**

Zusammenfassung: Durch die Erfolge der Künstlichen Intelligenz (KI) angetrieben genießt auch die Ethik der KI aktuell Hochkonjunktur. Die Beratung durch Ethikexpert*innen wird zunehmend von Politik und Industrie in Anspruch genommen, um die Risiken, die mit der Erforschung oder Anwendung neuer KI-Technologien einhergehen, proaktiv zu identifizieren und Lösungsvorschläge zu erarbeiten. Doch wie realistisch sind die Erwartungen, die an die KI-Ethik gestellt werden, technologische Entwicklung gerechter und akzeptierbarer zu gestalten? Läuft die Ethik der KI nicht selbst in Gefahr, ein Feigenblatt in den Diensten von im Vorfeld festgelegten wirtschaftlichen und politischen Zielen zu werden? Ausgehend von diesen Fragestellungen präsentiert und diskutiert der Artikel einige der aktuellsten Debatten und Vorschläge für den Umgang mit Chancen und Risiken von KI-Technologien, mit besonderem Fokus auf Sprachtechnologien. Obwohl die Gefahr der Instrumentalisierung präsent ist, so das Fazit, stellt die Ethik der KI aktuell eine wichtige Chance dar, frühzeitig problematische Entwicklungen zu erkennen und gesellschaftliche Debatten, politische Prozesse und rechtliche Regulierungen anzustoßen, welche zukünftige KI-Entwicklungen menschen- und umweltgerechter gestalten können.

Abstract: Driven by the success of artificial intelligence (AI), the ethics of AI is currently enjoying a boom. Advice from ethics experts is increasingly being sought by policymakers and industry to proactively identify the risks associated with new AI technologies and to propose solutions. But how realistic are the expectations placed on AI ethics to make technological development more equitable and acceptable? Doesn't AI ethics itself run the risk of becoming a fig leaf in the service of pre-determined economic and political goals? Based on these questions, this article presents and discusses some of the most current debates and proposals for dealing with the opportunities and risks of AI technologies, with a particular focus on language technologies. Although the danger of instrumentalization is present, the article concludes, the ethics of AI currently represents an important opportunity to identify problematic developments at an early stage and to initiate societal debates, political processes, and legal regulations that can make future AI developments more equitable and environmentally sound.

1. Einführung

1.1. Ethik der KI: ein Forschungsfeld im Aufschwung

Künstliche Intelligenz (KI) findet heute in zahlreichen Lebensbereichen Anwendung. Nach Inhalten im Internet suchen, mit Unterstützung eines Navigationssystems fahren, Sprach-

[13] elisa.orru@philosophie.uni-freiburg.de

nachrichten diktieren, den Saugroboter durch die Wohnung fahren lassen: Diese Beispiele aus dem Alltag zeigen, wie oft und selbstverständlich wir bereits mit KI interagieren. Dass massive Ressourcen in die KI-Entwicklung investiert werden und die hohe Popularität, die viele KI-Systeme genießen, lassen vermuten, dass die Verbreitung von KI in Zukunft weiter zunehmen wird.

Die Aufwärtsentwicklung von KI bietet dabei zahlreiche Chancen, die von der Erleichterung bei der Durchführung täglicher Tätigkeiten und schwerer Arbeiten bis hin zu besserer Diagnostik und Unterstützung bei der Bewältigung von aktuellen Herausforderungen, beispielweise in Verbindung mit dem Klimawandel, reichen. Andererseits bergen diese Technologien und ihr Einsatz auch Risiken und sie rufen Ängste hervor. Eine zunehmend beliebte Strategie, mit Risiken und Bedenken gegenüber KI-Technologien umzugehen, ist die Einbettung von Ethikexpertise in die Entwicklung und Anwendung neuer KI-basierter Systeme. Dementsprechend hat sich das Feld der Ethik der KI in den letzten Jahren zunehmend als zentrale Teildisziplin der Technikethik etabliert, die sich das Ziel setzt, proaktiv Chancen und Risiken der aufsteigenden KI-basierten Technologien zu identifizieren und Lösungen vorzuschlagen.

Besonders sichtbar ist diese Intensivierung der ethischen Beratung bei der Aufstellung von Ethik-Gremien, die damit beauftragt werden, Leitlinien zu generieren, welche idealerweise die konkrete Gestaltung neuer Technologien inspirieren sollen.[14] Auf EU-Ebene fügt sich diese Praxis in eine etwas längere Tradition der ethischen Beaufsichtigung bzw. Begleitung von Forschungsprojekten, die sich mit technologischen Innovationen beschäftigen. Um die Übertragung von ethischen Standards in die Technologienentwicklung zu unterstützen, hat die EU seit einigen Jahren ein Verfahren für alle geförderten Forschungsprojekte eingerichtet, das den Bedarf für eine ethische Aufsicht prüft und gegebenenfalls die Einbettung von ethischer Forschung, Beratung oder Beaufsichtigung in die Projekte sichert. In diesem Feld bin ich seit einigen Jahren in verschiedenen Rollen, nämlich als Ethikforscherin, Leiterin von ethischen Teilprojekten, Begutachterin und *Ethics Advisor*, tätig.[15] Ausgehend von dieser

[14] Im Bereich der KI siehe zum Beispiel in Deutschland den Bericht der parlamentarischen Enquete-Kommission über Künstliche Intelligenz (Deutscher Bundestag Drucksache 19/23700, 19. Wahlperiode 28.10.2020) oder auf globaler Ebene den UNESCO-Entwurf für Leitlinien über die Ethik der KI (Draft text of the recommendation on the ethics of artificial intelligence, Dok. SHS/IGM-AIETHICS/2021/JUN/3 Rev.22 5 June 2021, https://unesdoc.unesco.org/ark:/48223/pf0000377897 (gesehen am 29.10.2021). Über aktuelle Initiativen auf EU-Ebene wird unten näher berichtet.

[15] Aus dieser Tätigkeit sind verschiedene Publikationen entstanden, darunter beispielsweise: Orrù 2015; Orrù 2017, sowie die Projektberichte des Europäischen Projekts TRESSPASS, darunter Orrù 2020 sowie Orrù/Grzondziel 2021 a) und Orrù/Grzondziel 2021 b). Einige im Zusammenhang mit diesen Tätigkeiten entstandenen Erkenntnisse sind in meine neueste Monografie (Orrù 2021) eingeflossen.

Erfahrung und im Rückgriff auf aktuelle Kontroversen und Debatten werde ich im vorliegenden Beitrag Möglichkeiten und Grenzen der Ethik der KI beleuchten.

1.2. Feigenblatt oder wichtiges regulatives Instrument?

Die erwähnte zunehmende Praxis der Formulierung von ethischen Leitlinien im Hinblick auf die KI-Entwicklung wird neuerdings aus vielen Hinsichten kritisiert (Floridi 2019). Im Allgemeinen wird dieser Praxis vorgeworfen, eine „Ethik-Waschmaschine" in Diensten der Industrie zu sein (Metzinger 2019). Laut dieser Kritik wäre die Einsetzung von Ethikgremien eine Strategie, um wirksame und bindende rechtliche Regulierungen zu vermeiden oder mindestens zu verschieben und um die Akzeptanz von gesellschaftlich kontroversen, aber hoch profitablen Technologieanwendungen zu steigern. Insbesondere der 2019 von einer Expert*innengruppe im Auftrag von der EU erarbeiteten „Ethik-Leitlinien für eine vertrauenswürdige KI" (Hochrangige Expertengruppe für Künstliche Intelligenz 2019) wurde vorgeworfen, ein Mittel zur Steigerung der EU-Wettbewerbsfähigkeit auf dem globalen Markt zu sein und kaum echte Steuerungskraft zu besitzen (Metzinger 2019; Gillen 2019).

Bei all dieser Kritik wird aber zumeist auch von den Kritiker*innen anerkannt, dass die Generierung und Anwendung ethischer Richtlinien eine der wenigen aktuell verfügbaren Möglichkeiten darstellen, die zukünftige Entwicklung und den zukünftigen Einsatz der KI gerechter, menschen- und umweltfreundlicher zu gestalten.

Die Ethik der KI scheint daher im Spannungsverhältnis zwischen gezähmter Begleitforschung und aktiver Gestaltungskraft zu stehen. Doch wie hat sich die Ethik der KI bisher innerhalb dieses Feldes bewegt und wie könnte sie den Erwartungen nach effektiver Steuerung gerechter als bisher werden? Um diese Fragen zu beantworten, werde ich zunächst einige Beispiele von KI-Anwendungen präsentieren, die ethische Fragestellungen aufwerfen. Insbesondere werde ich mich auf einer Klasse von KI-Technologien fokussieren, nämlich Sprachmodelle (Kap. 2). Darüber hinaus werde ich zwei philosophische Ansätze präsentieren, die es ermöglichen, den Blick auf ethische Chancen und Risiken der KI zu schärfen und diese in einen breiteren konzeptuellen Kontext zu stellen und zu systematisieren (Kap. 3). Darauffolgend werde ich existierende Lösungsansätze präsentieren, die darauf abzielen, die negativen ethischen Auswirkungen von KI-Anwendungen zu minimieren (Kap. 4). Abschließend werde ich auf die Ausgangsfrage zurückkommen und eine Bilanz über die Funktion und das Potenzial der Ethik der KI ziehen (Kap. 5).

2. Was KI-unterstützte Sprachverarbeitung mit Gerechtigkeit und Klimawandel zu tun hat

Sprachmodelle sind KI-Systeme, die die Wahrscheinlichkeit von Wortsequenzen berechnen und auf dieser Basis Wortvorschläge generieren (Bender u. a. 2021, 611). Sie kommen zum Beispiel bei Websuchen zum Einsatz, etwa wenn der Sucheintrag automatisch vervollständigt wird, sowie bei automatischer Übersetzung oder Spracherkennungssystemen.

2.1. „Reinigungskraft“ oder „Putzfrau“?

Das erste Beispiel bezieht sich auf Programme für die automatische Übersetzung von Texten. Ein in Deutschland besonders beliebtes System ist DeepL. DeepL ist eine kostenlose webbasierte Applikation, in der Nutzer*innen einen Text eingeben können und in Sekundenschnelle eine Übersetzung bekommen, die zwar nicht perfekt, aber in der Regel gut genug ist, um entweder den allgemeinen Sinn etwa eines Zeitungsartikels in einer fremden unbekannten Sprache zu verstehen oder als Basis für eine eigene Übersetzung verwendet zu werden.

Einerseits bietet ein System wie DeepL offensichtliche Vorteile, die auch ethisch relevant sind. Beispielsweise können dadurch Texte in fremden Sprachen zugänglich gemacht werden und somit die Möglichkeiten, sich eine Meinung über bestimmte Themen zu bilden und an öffentlichen Debatten teilzunehmen, gesteigert werden. Andererseits können aber solche Systeme auch zur Verbreitung und Verfestigung von Stereotypen und Benachteiligungsmustern beitragen.

Folgendes Beispiel aus einem im November 2021 selbst durchgeführten Test illustriert diese Risiken. In dem Test habe ich auf DeepL Satzteile mit Berufsbezeichnungen auf Englisch eingegeben. Bekanntlich sind die meisten Substantive auf Englisch nicht genderspezifisch. Um diese Offenheit aufrechtzuhalten, habe ich in dem eingegebenen Text auf Englisch keine Pronomen („she“/„he“) eingegeben. Als Berufsbezeichnungen habe ich „doctor“, „nurse“, „engeneer“ und „cleaner“ eingegeben. Folgender Screenshot visualisiert die Ergebnisse, die DeepL geliefert hat:

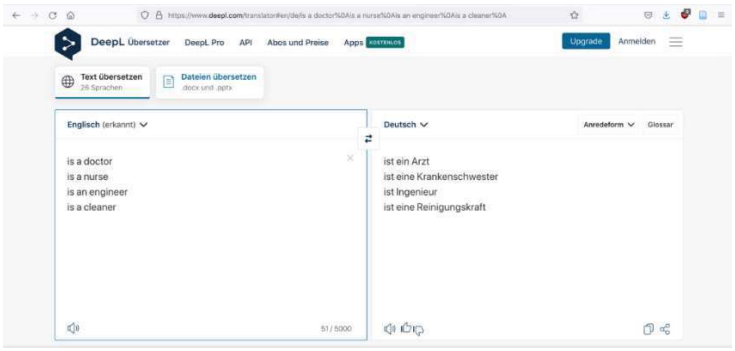

Bild 1. Automatische Übersetzungen von Berufsbezeichnungen durch DeepL

Offensichtlich werden hier Stereotypen in Bezug auf vermeintlich männlich oder weiblich konnotierte Berufe weitergegeben. Ein weiterer Test am selben Tag, aber mit Google Übersetzer durchgeführt, hat die gleichen Übersetzungen geliefert, bis auf die letzte. „Cleaner" wurde nämlich von Google Übersetzter mit „Putzfrau" übersetzt. Diese Unterscheidung in den vorgeschlagenen Ergebnissen zeigt, dass es nicht allein „an den Daten", die zum Trainieren der KI-Systeme verwendet werden, liegt, ob diese Systeme Stereotypen wiedergeben. Einerseits rechnen zwar Sprachmodelle per Definition Wahrscheinlichkeiten aus einem vorgegebenen Datenpool aus, sodass es angesichts der in der allgemeinen Sprache bestehenden Verzerrungen und Stereotypisierungen nicht überrascht, dass es ebenfalls zu „biased" Ergebnissen kommt. Jedoch zeigt das letzte Ergebnis, dass eine Varianz an Ergebnissen möglich ist und dass die Ergebnisse nicht einfach von „den Daten" abhängen. Denn erstens gibt es „die Daten" nicht als solche, sondern es gibt nur Datensammlungen, die von Menschen ausgewählt und als Trainingsdaten eingesetzt werden. Zweitens gibt es Möglichkeiten, auf die Qualität dieser Datensammlungen zu achten sowie Trainingsmechanismen einzusetzen, die dem System „beibringen", stereotypenfreie Optionen zu bevorzugen. Einige dieser Möglichkeiten werden näher im Kapitel 4 unten diskutiert.[16]

2.2. Big Data und „kleine Sprachen"

Im Internet sind verschiedene Applikationen für die automatische Untertitelung von Audio- und Videoinhalten verfügbar. Manche Plattformen, wie YouTube oder Facebook, haben eine eigene integrierte Untertitelungsfunktion für Inhalte, die auf der Plattform abgespielt werden. Für einen weiteren Test habe ich eine eigenständige webbasierte und kostenlose Applikation namens VEED.IO ausgewählt.

Auf VEED.IO können Nutzer*innen ein kurzes Video hochladen, das in wenigen Minuten mit Untertiteln versehen und dann erneut abgespielt wird. Um ein Video mit guter Audioqualität zu verwenden, habe ich das System mit einem kurzen Ausschnitt aus den Tagesschau-Nachrichten getestet.

Auch in diesem Fall sind die Ergebnisse nicht vollständig exakt, aber sie ermöglichen zumindest, den Tenor der Audioinhalte zu verstehen. Aus ethischer Sicht hat dieses System den Vorteil, relativ unkompliziert Zugang zu Audioinhalten für Menschen mit Hörschwierigkeiten anzubieten. Das ist im Hinblick auf gesellschaftliche Teilhabe ein wichtiger Aspekt, der für einen Beitrag solcher Systeme zur gerechteren Gestaltung unserer Gesellschaften spricht. Andererseits unterstützen solche Systeme in der Regel nur eine begrenzte Zahl an

[16] Zudem liefert der Beitrag von Anne Lauscher in dem vorliegenden Band weitere Einsichten in diese Möglichkeiten aus technischer Sicht.

Sprachen. Bei VEED.IO sind es über 100 Sprachen – eine beachtliche Zahl, die aber nur einen Bruchteil der knapp 7000 weltweit gesprochenen Sprachen darstellt (Anderson 2010). Dies führt dazu, dass Menschen, die eine der restlichen über 6900 Sprachen sprechen, nicht von den Vorteilen dieser Technologien profitieren können.

Dass KI-gestützte Sprachsysteme nur für einige Sprachen verfügbar sind, ist teilweise der aktuell meistverfolgten Strategien für die Entwicklung dieser Technologien geschuldet. Denn aktuell setzen die meisten Ansätze auf sehr großen Mengen (in der Größenordnung von Terabytes) von Trainingsdaten, um die Präzision der Ergebnisse der entwickelten Systeme zu verbessern. Dies führt dazu, dass Sprachmodelle aktuell nur für die Sprachen entwickelt werden, in denen der größte Teil der digitalen Inhalte erzeugt wird. Alternative Ansätze in der Informatik, die Sprachmodelle auch auf Basis von kleineren Datenbanken entwickeln und trainieren, existieren zwar, werden aber kaum gefördert und verfolgt (Bender u. a. 2021, 613).

2.3. Alexa & Co: fügsame virtuelle Assistentinnen

Sprachassistenzsysteme wie Alexa, Siri oder Google Assistant sind sehr beliebte KI-gestützte Systeme, die zunehmend im Alltag Anwendung finden. Die Beliebtheit, die sie genießen, beruht auch auf ihre Fähigkeit, die Durchführung bestimmter Aktivitäten zu erleichtern, wie etwa eine Internetrecherche starten, einen Anruf tätigen oder den Wetterbericht checken.

Auch gerade wegen ihrer Beliebtheit und Versatilität können Sprachassistenzsysteme auf besonders schleichende, aber effektive Weise zur Stärkung von Verzerrungen und Stereotypen beitragen. Eine von der UNESCO 2019 veröffentlichte Studie mit dem Titel „I'd blush if I could" untersucht, wie Sprachassistenzsysteme zur Verfestigung von Vorurteilen und Benachteiligungen aufgrund des Geschlechts beitragen können.

Die bekanntesten Sprachassistenzsysteme weltweit sind, mit wenigen Ausnahmen, entweder ausschließlich oder standardmäßig weiblich. Alexa (Amazon), Cortana (Microsoft) und Siri (Apple) haben sowohl weibliche Namen als auch weibliche Stimmen. Und auch hinter dem genderneutralen Namen „Google Assistant" verbirgt sich eine weibliche Stimme (UNESCO, EQUALS Skills Coalition 2019, 94).[17] Die Wahl weiblicher Stimmen als passender für Assistenzsysteme ist laut Studie der – insbesondere in den überwiegend männlich besetzten Entwicklerteams verbreiteten – Vorstellung geschuldet, Frauen seien besser geeignet für Hilfestellung, Zuarbeiten und Assistenzaufgaben als Männer (UNESCO, EQUALS

[17] Die überwiegend weibliche Charakterisierung von Sprachassistenzsystemen ist typisch auch für wenig verbreitete Systeme (UNESCO, EQUALS Skills Coalition 2019, 95).

Skills Coalition 2019, 97–100). Andererseits verfestigen laut Studie diese Systeme Eigenschaftszuschreibungen, wonach Frauen fügsame, unterwürfige, ständig auf Abruf verfügbare Hilfeleisterinnen seien. Zusätzlich zu diesen allgemeinen Aspekten hebt die Studie hervor, dass die meisten dieser Systeme auf Beleidigung und Belästigung mit Toleranz oder sogar Unterwürfigkeit reagieren (UNESCO, EQUALS Skills Coalition 2019, 204–105).

Eine 2017 durchgeführte und in der UNESCO-Publikation zitierte Studie dokumentierte unter anderem die Reaktionen der benannten Systeme auf die Beleidigung „you are a slut!". Während Siri auf diese Beleidigung die Antwort lieferte, die der UNESCO-Studie den Titel gibt, nämlich „I'd blush if I could", bedankte sich Alexa wie folgt: „Well, thanks for the feedback". Ferner entschuldigte sich Google Assistant („My apologies, I don't understand") und Cortana startete schlicht eine Websuche.

Einige Hersteller haben bereits Maßnahmen ergriffen, um auf die Kritik bezüglich der reproduzierten Genderstereotypen zu reagieren. Während alle benannten Systeme zum Zeitpunkt der ersten Kommerzialisierung standardmäßig weibliche Stimmen hatten, haben beispielsweise Siri und Google Assistant nachträglich jeweils 2013 und 2017 auch voll funktionsfähige männliche Stimmen hinzugefügt (UNESCO, EQUALS Skills Coalition 2019, 115–116). Erst seit Juli 2021 hat Medienberichten zufolge auch Alexa eine männliche Stimme und reagiert optional auf den männlichen Namen „Ziggy" (Erl 2021). All diese Systeme haben jedoch in den meisten Sprachen und Ländern immer noch weibliche Stimmen als Standardoptionen (UNESCO, EQUALS Skills Coalition 2019, 116).

3. Verantwortung und Gerechtigkeit als Eckpunkte der Ethik der KI

Bei der Darstellung der Beispiele sind verschiedene ethische Aspekte angesprochen worden. Ich werde sie nun in Verbindung mit ethischen Konzeptionen bringen, um sie dabei besser systematisieren und ergänzen zu können.

3.1. Das „Prinzip Verantwortung" und der „Fähigkeitenansatz"

Die philosophische Tradition ist reich an ethischen Konzeptionen, die fruchtbar für den Umgang mit Technik gemacht werden können. Die Frage etwa, worin gerechtes Handeln bestünde, hat bereits Platon und Aristoteles intensiv beschäftigt, obwohl zur Zeit der griechischen Polis technisches Handeln noch außerhalb des Horizonts der ethischen Reflexion lag (Jonas 2003, 15-16). Dies hat sich spätestens seit dem 20. Jahrhundert geändert, als die Philosophie die ethische Tragweite neuer technologischer Möglichkeiten erkannte. Seitdem sind sowohl neue ethische Ansätze entwickelt worden, die sich spezifisch mit den Folgen

technologischer Entwicklung auseinandersetzen, als auch ethische Konzeptionen, die sich nicht primär mit technologischen Fragestellungen beschäftigen, jedoch wichtige Anhaltspunkte für die Reflexion über ihre ethischen Implikationen liefern. Unter diesen Theorien sind meines Erachtens zwei Ansätze besonders geeignet, um einen konzeptuellen Hintergrund für die Schärfung und Systematisierung der benannten Chancen und Risiken von KI-Technologien zu liefern. Diese wurden jeweils vom deutsch-amerikanischen Philosophen Hans Jonas und von der US-amerikanischen Philosophin Martha Nussbaum vertreten.

Bereits in den 1970er Jahren entwarf Jonas eine ethische Konzeption, die unmittelbar darauf ausgerichtet ist, die Bedingungen eines ethischen Umgangs mit Technologie, wiewohl nicht primär mit KI, zu formulieren. Laut Jonas hat sich die Reichweite der möglichen Folgen technischen Handelns in der jüngsten Geschichte so ausgedehnt, dass die traditionelle Ethik nicht mehr adäquat scheint, mit diesen umzugehen. Denn diese konzentriere sich auf die unmittelbare zwischenmenschliche Interaktion und auf die direkten Folgen menschlichen Handels. Dagegen sei eine Erweiterung der traditionellen Ethik um eine kollektive und zeitlich ausgedehnte Komponente nötig (Jonas 2003, 22-26).

Jonas' Ethik distanziert sich von der traditionellen Ethik auch dahingehend, dass sie sich „mehr an öffentliche Politik als an privates Verhalten" richtet (Jonas 2003, 37). Spezifisch für Jonas' Konzeption ist auch die Miteinbeziehung der zukünftigen Generationen in die ethische Reflexion. Denn im Mittelpunkt seiner Konzeption, auch „Zukunftsethik" genannt, steht das „Prinzip Verantwortung", das durch folgenden Imperativ zum Ausdruck gebracht werden kann: „Handle so, dass die Wirkungen deiner Handlungen verträglich sind mit der Permanenz echten menschlichen Lebens auf Erden" (Jonas 2003, 36). Wir haben sozusagen eine moralische Pflicht, die Grundlagen der menschlichen Existenz auf der Erde nicht zu gefährden.

Auf den Bereich der KI übertragen lautet dann die zentrale ethische Frage in Anlehnung an Jonas, ob KI-Technologien dazu beitragen, die Möglichkeit echten menschlichen Lebens auf der Erde zu erhalten oder ob sie diese vielmehr gefährden.

Zusammen mit dem Ökonomen und Nobelpreisträger Amartya Sen hat Nussbaum seit den 1980er Jahren den sog. Fähigkeitenansatz entwickelt (Nussbaum 2015, 53–68). Anders als Jonas beschäftigt sich Nussbaum nicht primär mit technologischer Entwicklung. Im Mittelpunkt ihrer Konzeption steht vielmehr die Frage, was eine Gesellschaft als „gerecht" charakterisiert. Laut Nussbaum hängt der Grad von Gerechtigkeit, der in einer gegebenen Gesellschaft

erreicht wird, eng mit der Möglichkeit aller Bürger*innen[18] zusammen, ein der Menschenwürde entsprechendes Leben zu führen. Dafür muss ein Minimum an zehn zentralen menschlichen Fähigkeiten gegeben werden (Nussbaum 2015, 41). Diese Fähigkeiten umfassen: 1) Leben; 2) körperliche Gesundheit; 3) körperliche Unversehrtheit, Sinne, Vorstellungskraft; 4) Denken; 5) Gefühle; 6) praktische Vernunft; 7) Zugehörigkeit; 8) „andere Gattungen"; 9) Spiel und 10) Kontrolle über die eigene Umwelt (Nussbaum 2015, 41–42). Für unsere Diskussion in Bezug auf KI sind zwei dieser Fähigkeiten besonders relevant, nämlich Zugehörigkeit und Kontrolle über die eigene Umwelt. Zugehörigkeit impliziert unter anderem die Fähigkeit, an gesellschaftlicher Interaktion teilnehmen zu können und nicht diskriminiert zu werden. Zur Kontrolle der eigenen Umwelt gehört die Fähigkeit, an politischen Entscheidungsprozessen teilnehmen zu können.

Es ist wichtig anzumerken, dass für Nussbaum die ersten Adressaten der Forderungen, die sich aus ihrer Konzeption der Gerechtigkeit ergeben, Regierungen sind (Nussbaum 2015, 40–41). Es wäre daher reduktiv, die Forderungen nach diesen Fähigkeiten von einzelnen KI-Technologien abhängig zu machen. Jedoch sind einzelne Technologien immer in sozialen Kontexten eingebettet, die maßgeblich beeinflussen, wie diese Technologien zum Einsatz kommen. Es scheint daher plausibel, den Fähigkeitenansatz als Bezugspunkt für eine Diskussion über KI-Technologien und Gerechtigkeit zu verwenden, und dies umso mehr, weil die Lösungen, die zur Steigerung der Gerechtigkeit in Bezug auf KI-Anwendungen beitragen können, keinen rein technischen Charakter haben, sondern immer auch gesellschaftlicher und oft auch rechtlicher und politischer Natur sind. Auf KI bezogen könnte daher die Frage der Gerechtigkeit im Sinne Nussbaums so umformuliert werden, dass KI-Technologien zur gerechteren Gestaltung einer Gesellschaft beitragen, wenn sie auf eine Art und Weise zum Einsatz kommen, welche die Steigerung der Fähigkeiten fördert. Dagegen trügen KI-Technologien zur Ungerechtigkeit bei, wenn ihre Anwendung Zustände oder Faktoren begünstigt, die sich hemmend auf diese Fähigkeiten auswirken.

3.2. Chancen und Risiken von KI-gestützten Sprachverarbeitungssystemen

Wenn wir nun zu den dargestellten KI-Beispiele zurückkehren, können wir im Lichte der zwei ethischen Konzeptionen von Nussbaum und Jonas festhalten, dass einerseits Sprachunterstützungssysteme zur gerechteren Gestaltung unserer Gesellschaften maßgeblich beitragen können, denn:

[18] Dass Nussbaum den Fokus auf Bürger*innen beschränkt, hat damit zu tun, dass die Gerechtigkeitsforderungen, die sich aus ihrer Theorie ergeben, hauptsächlich an (nationale) Regierungen gerichtet sind (s. unten). Ich werde im Folgenden annehmen, dass sich die Forderungen, die sich aus ihrer Theorie ergeben, auch auf Nicht-Bürger*innen erstrecken können.

- Erstens können sie den Zugang zu kulturellen, politischen oder wissenschaftlichen Inhalten für alle erweitern, indem Texte in unbekannten Sprachen erschlossen werden;
- zweitens können diese Systeme auch Inklusion fördern, indem sie kulturelle, politische sowie gesellschaftlich relevante Inhalte für Gehörlose und Schwerhörige unkompliziert zugänglich machen.

Mit den Begriffen Nussbaums formuliert: Wenn diese Systeme in einem entsprechenden gesellschaftlichen, politischen und rechtlichen Umfeld eingesetzt werden, können sie zur Gerechtigkeit beitragen, weil sie die Fähigkeiten der Nutzer*innen erhöhen können, an gesellschaftlicher Interaktion teilzunehmen und sich politisch zu beteiligen.

Andererseits können diese Technologien auch negative Folgen mit sich bringen, insbesondere in den folgenden Hinsichten:

- Erstens können diese Systeme leicht zur Amplifizierung von existierenden Verzerrungen, Vorurteilen und Diskriminierungsmustern führen. Denn sie sind „stochastische Papageie" (Bender u. a. 2021 616–618), nämlich mathematische Systeme, die weder die Bedeutung von den Texten „verstehen", mit denen sie arbeiten, noch verbinden sie die produzierten Texte mit Bedeutung. Weil Inhalte, die von gesellschaftlich privilegierten Gruppen hergestellt werden – aber auch offensichtlich rassistische und sexistische Inhalte –, größere Wahrscheinlichkeiten haben, im Internet überrepräsentiert oder sichtbarer zu sein, haben diese Inhalte auch eine höhere Wahrscheinlichkeit, die Vorschläge der Sprachmodelle stärker zu beeinflussen. Hinzu kommt, dass die Teams, die aktuell an der Konzeption und Entwicklung von KI-Technologien beteiligt sind, in der Regel wenig Diversität aufweisen. Etwa sind heute weltweit über 90 % der Softwareentwickler*innen noch immer Männer.[19]
- Zweitens sind die Vor- und Nachteile dieser Systeme ungerecht verteilt. Weil Sprachmodelle aktuell große Mengen an Trainingsdaten benötigen, werden sie meistens nur für diejenigen Sprachen entwickelt, in denen der größte Teil der digitalen Inhalte erzeugt wird. Dies sind typischerweise Englisch und Sprachen, die in reicheren, hochdigitalisierten Ländern gesprochen werden. Noch 2020 hatten 90 % der weltweiten Sprachen, die insgesamt von über einer Milliarde Menschen gesprochen werden, kein oder kaum Unterstützung durch Sprachtechnologien (Bender u. a. 2021, 612; Joshi u. a. 2020, 6285).
- Drittens wirken sich Sprachmodelle erschwerend auf die Ursachen der Klimakrise aus. Weil diese Systeme aktuell meistens auf die Verfügbarkeit von großen Datenbanken und

[19] Software developers: distribution by gender 2021, https://www.statista.com/statistics/1126823/worldwide-developer-gender/ (gesehen am 30.11.2021).

viel Rechenkapazität angewiesen sind, ist ihre Entwicklung sehr ressourcenintensiv. Es wurde berechnet, dass das Trainieren eines großen Sprachmodells (das nur einer von vielen Schritten der Entwicklung eines KI-Systems ist) ca. 284 Tonnen CO_2 produziert. Zum Vergleich: Im Durchschnitt ist jeder Mensch für die Produktion von 5 Tonnen CO_2 pro Jahr verantwortlich (Bender u. a. 2021, 612; Strubell u. a. 2019).

Die ersten beiden Nachteile dieser Technologien können in Anlehnung an Nussbaum als problematisch im Hinblick auf Gerechtigkeit angesehen werden, weil sie dazu beitragen, bestehende Benachteiligungen zu verstärken, entweder für bestimmte Gruppen innerhalb einer Gesellschaft oder für Menschen, die Sprachen verwenden, die nicht zu den in der KI überrepräsentierten Sprachen gehören. Die ethische Bedeutung des dritten Nachteils wird in Anlehnung an Jonas' „Prinzip Verantwortung" deutlich: Diese Technologien, so wie sie heute entwickelt und eingesetzt werden, tragen nicht zum nachhaltigen Verbrauch der Ressourcen unseres Planeten bei. Global und intergenerationell betrachtet sind es prinzipiell die Bewohner*innen der reicheren, hochdigitalisierten Länder, die von den Vorteilen der aktuellen Sprachverarbeitungssysteme profitieren. Dagegen treffen die Klimaschäden, die durch diese Systeme mitverursacht werden, Länder des globalen Südens und zukünftige Generationen am härtesten, die zudem wenige Chancen haben, von den Vorteilen dieser Systeme zu profitieren.

4. Lösungsansätze

Im vielfältigen Forschungsfeld der Ethik der KI wurden bereits Lösungsansätze entwickelt, um die Nachteile dieser Systeme zu minimieren.

Als Option für die Milderung vieler der benannten Unzulänglichkeiten aktueller Sprachmodelle hat Timnit Gebru, ehemalige Co-Leiterin der Ethik-Abteilung von Google, zusammen mit anderem Kolleg*innen vorgeschlagen, auf kleinere und sorgfältiger sortierte Datenbanken zu setzen (Bender u. a. 2021, 623). Die Verwendung kleinerer Datasets würde es erstens ermöglichen, besser auf die Qualität und Diversität der Trainingsdaten zu achten. Das würde wiederum die Neigung automatisierter Systeme mildern, bereits existierende Verzerrungen oder Diskriminierungen zu amplifizieren. Darüber hinaus würde es die Bevorzugung kleinerer und genauerer Trainingsdaten ermöglichen, auf ressourcenschonende Weise Sprachmodelle zu entwickeln. Dies würde wiederum auch eine gerechtere Verteilung der Vor- und Nachteile von Sprachsystemen im Allgemeinen ermöglichen. Ebenfalls positiv auf die Gerechtigkeit würde sich die Möglichkeit auswirken, durch die Verwendung kleinerer Datasets Sprachmodelle auch für Sprachen zu entwickeln, für die keine Terabytes an Trainingsdaten zur Verfügung stehen.

Aus der UNESCO-Studie über Sprachassistenzsysteme entsprang ferner die Forderung, durch Bildung eine erhöhte Teilhabe von Frauen an der Gestaltung und Entwicklung von KI-Systemen zu bewirken. Die Studie selbst warnt davor, einen Automatismus zwischen Mädchen- und Frauenbildung und der Entstehung diverserer Teams bei der Entwicklung von KI-Systemen einerseits und zwischen letzterer und der Reduzierung von Bias in KI-Systemen andererseits einfach vorauszusetzen. Dennoch scheint mehr Frauenbeteiligung eine nötige, wenn auch nicht ausreichende Bedingung für eine gerechtere Gestaltung der KI-Technologien zu sein.

In der Folge der UNESCO-Studie wurde 2019 zudem die Kampagne „Hey update my voice!" gestartet, die für den Zusammenhang zwischen Stereotypisierungen in Sprachassistenzsystemen und Gewalt gegen Frauen sensibilisieren will.[20] Die Kampagne ruft dazu auf, „Antworten" zu „spenden", die von den Sprachassistenzsystemen übernommen werden können – als Alternativen zu den aktuellen Antworten, die wie erwähnt oft unterwürfig, tolerierend oder bagatellisierend auf sexuelle Beleidigungen oder Belästigungen reagieren.

5. Conclusio: was kann und was soll die Ethik der KI leisten?

Die präsentierten Lösungsansätze zeigen meines Erachtens eine Reihe von möglichen Nuancen im Umgang mit ethischen Risiken und Chancen der KI. Sie verweisen auf praktikable Lösungen, die zumindest einige der schwerwiegendsten Nachteile aktueller KI-Applikationen mindern könnten.

Die Frage, ob diese ethischen Ansätze aber auch in der Lage sind, grundsätzliche Kritik an bestimmten technologischen Entwicklungen zu formulieren, bedarf einer nuancierten Antwort.

Einerseits kann das Risiko nicht ausgeschlossen werden, dass solche Vorschläge zu leeren Etiketten werden können, die nur den Anschein von *„ethics-compliance"* geben, die Substanz der Nachteile aber unberührt lassen. Es ist daher eine zentrale Aufgabe der Technikethik selbst, die eigenen Vorschläge kontinuierlich zu überprüfen, gegebenenfalls zu ergänzen oder zu revidieren und gegen missbräuchliche Aneignungen zu verteidigen.

Andererseits waren einige der präsentierten Lösungsansätze in der Lage, implizite Annahmen und einigermaßen „grundsätzliche" Aspekte aktueller KI-Systeme infrage zu stellen. So hat etwa die Debatte um die Sprachassistenzsysteme auf grundsätzliche Mechanismen von

[20] https://en.unesco.org/news/hey-update-my-voice-movement-exposes-cyber-harassment (gesehen am 30.11.2021).

Macht- und Rollenverteilung zwischen den Geschlechtern hingewiesen und ist meines Erachtens schwer als „ethische Waschmaschine“ abzutun. Auch die Tatsache, dass Gebru laut übereinstimmenden Medienberichten wegen ihres Vorschlags ihre Stelle bei Google aufgeben musste und dass einige der Mitautor*innen ihres Artikels auf Anweisung der jeweiligen Arbeitgeber*innen auf die Unterzeichnung des Artikels verzichten mussten, zeigt, dass auf den ersten Blick harmlose Vorschläge, wie die Größe der Trainingsdatenbanken zu reduzieren, „grundsätzlich“ genug sind, um Irritationen und ernsthafte Konsequenzen für die Beteiligten nach sich zu ziehen.

In einigen weiteren Fällen wurde innerhalb der Ethik der KI auch der Versuch unternommen, ein grundsätzliches Veto gegen bestimmte technologische Entwicklungen zu formulieren. So wurde etwa im Rahmen der Aufstellung der benannten EU-Leitlinien für vertrauenswürdige KI eine Untergruppe damit beauftragt, sogenannte „rote Linien“ zu erarbeiten, nämlich „nichtverhandelbare ethische Prinzipien, die festlegen, was in Europa mit KI nicht gemacht werden darf“ (Metzinger 2019). Der Versuch ist zunächst gescheitert, was unter anderem einer der Anlässe für die scharfe Kritik an den Leitlinien war, die anfangs erwähnt wurde.

Dabei ist eine Ambivalenz zum Tragen gekommen, die die von der EU-beauftragte Ethikberatung öfters prägt. Diese besteht in der Einbettung der Ethikarbeit in einen Kontext, in dem die Ziele der technologischen Entwicklung bereits gesetzt sind. Denn der allgemeine Rahmen, in dem die ethische Tätigkeit sich bewegen soll, ist oft schon im Vorfeld definiert und an strategischen Zielen orientiert (etwa der Stärkung der Wettbewerbsfähigkeit der EU), die nicht primär ethisch konnotiert sind oder sogar mit ethischen Prinzipien in einem Spannungsverhältnis stehen (Europäischen Kommission 2018). Solche vorgegebenen Ziele stellen tatsächliche Herausforderungen an eine Ethik dar, die sich nicht nur als „Begleitung“ von bereits im Vorfeld festgelegten Zielen versteht. Andererseits sind die von der EU-Kommission oder anderen Auftraggebern definierten Rahmenbedingungen nicht „in Stein gemeißelt“. So hat etwa die Arbeit der Untergruppe zur Verbietung bestimmter Technologien und die daraus resultierende Debatte einen Prozess in Gang gesetzt, der vom EU-Parlament aufgegriffen wurde und aktuelle legislative Initiativen prägt. Die EU-Kommission hat beispielsweise im April 2021 einen Entwurf für ein Gesetz über Künstliche Intelligenz vorgelegt, der einige KI-Anwendungen grundsätzlich verbietet (Europäische Kommission 2021). Auch das EU-Parlament diskutiert aktuell die Einführung eines Verbotes einiger KI-Systeme, etwa von *social scoring systems* und privaten Gesichtserkennungsdatenbanken sowie von automatischen Erkennungssystemen in öffentlichen Räumen und bei Grenzkontrollen (European Parliament 2021).

Genau hier und trotz des anfänglichen Scheiterns zeigt sich meines Erachtens die mögliche Stärke einer Ethik der KI. Denn Ethik darf keine rechtliche Regulierung ersetzen, sie kann aber eine Chance darstellen, relevante Probleme zu fokussieren und dafür zu sensibilisieren, bevor eine rechtliche Regulierung einsetzen kann, oder auch dort, wo rechtliche Grauzonen bestehen. Auf bestimmte technische Anwendungen hinzuweisen, die angesichts ethischer Risiken verboten werden sollten, sollte meines Erachtens zu den Möglichkeiten der Ethik der KI gehören. Jonas' Philosophie bietet hierfür Anhaltspunkte und einen theoretischen Rahmen, um diese einzubetten. Ein solches Verbot wäre natürlich nicht von der Ethik selbst durchsetzbar, es könnte aber erstens gesellschaftliche Debatten und politische Prozesse anstoßen und zweitens eine Vorlage für verbindliche rechtliche Regulierungen liefern.

Die ethische Reflexion über KI sowie über neue, disruptive Technologien im Allgemeinen kann sogar den Anlass bieten, um bestehende und historische Ungerechtigkeiten hervorzuheben und neu zu diskutieren. Wenn das gelingt, kann Technikethik breitere gesellschaftliche Debatten auslösen und dabei einen Beitrag dazu leisten, bestimmte Werte und ethische Prinzipien in unseren Gesellschaften allgemein zu stärken. Dabei sollte neben der Diskussion von Chancen und Risiken spezifischer Technologien auch die Frage, was ein wünschenswertes (Zusammen-)Leben darstellt und welche Rolle Technologien in dieser Vision spielen sollen, stärker in den Fokus rücken.

Wir befinden uns an der Schwelle eines historischen Durchbruchs und erleben einen technisch-gesellschaftlichen Wandel, der unsere Gesellschaft für die nächsten Jahrzehnte prägen wird. Die ethische Werkzeugkiste, die uns zur Verfügung steht, ist nicht perfekt. Sie liefert aber wichtige Instrumente, um diesen Wandel zu steuern und ihn gerecht und ethisch verträglich zu gestalten. Wenn uns dies gelingt, werden wir wichtige Voraussetzungen für eine Gesellschaft schaffen, die menschen- und umweltgerechter ist.

Literatur

Anderson, Stephen, Languages: A Very Short Introduction, Linguistic Society of America, Brochure Series: Frequently Asked Questions, 2010, https://www.linguisticsociety.org/content/how-many-languages-are-there-world (gesehen am 29.10.2021).

Bender, Emily M. u. a., On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?, in: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 2021, 610–623.

Deutscher Bundestag, Bericht der Enquete-Kommission Künstliche Intelligenz – Gesellschaftliche Verantwortung und wirtschaftliche, soziale und ökologische Potenziale, Drucksache 19/23700, 19. Wahlperiode 28.10.2020.

Erl, Josef, Alexa bekommt männliche Stimme und neues Aktivierungswort, 21.07.2021, https://mixed.de/alexa-bekommt-maennliche-stimme-und-neues-aktivierungswort/ (gesehen am 30.11.2021).

Europäische Kommission, Mitteilung, Künstliche Intelligenz für Europa, COM/2018/237, 25.04.2018.

---, Vorschlag für eine Verordnung des Europäischen Parlaments und des Rates zur Festlegung harmonisierter Vorschriften für Künstliche Intelligenz (Gesetz über Künstliche Intelligenz) und zur Änderung bestimmter Rechtsakte der Union, COM/2021/206 final, 21.04.2021.

European Parliament News 2021, Use of artificial intelligence by the police: MEPs oppose mass surveillance, https://www.europarl.europa.eu/news/en/press-room/20210930IPR13925/use-of-artificial-intelligence-by-the-police-meps-oppose-mass-surveillance (06.10.2021, gesehen am 29.10.2021).

Floridi, Luciano, Translating Principles into Practices of Digital Ethics: Five Risks of Being Unethical, in: Philosophy & Technology, 32/2, 2019, 185–193.

Gillen, Erny, Umgang mit KI: Wollen wir Sicherheit oder Wettbewerb?, in: FAZ.NET, 10.01.2019, https://www.faz.net/aktuell/feuilleton/debatten/ethik-und-ki-wollen-wir-sicherheit-oder-wettbewerb-15980382.html (gesehen am 29.10.2021).

Hochrangige Expertengruppe für Künstliche Intelligenz, Ethik-leitlinien für eine vertrauenswürdige KI. LU: Amt für Veröffentlichungen der Europäischen Union 2019, https://data.europa.eu/doi/10.2759/22710 (gesehen am 30.11.2021).

Jonas, Hans, *Das Prinzip Verantwortung: Versuch einer Ethik für die technologische Zivilisation*. Frankfurt am Main: Suhrkamp Verlag 2003.

Joshi, Pratik u. a., The State and Fate of Linguistic Diversity and Inclusion in the NLP World, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, 6282–6293, https://aclanthology.org/2020.acl-main.560 (gesehen am 04.11.2021).

Metzinger, Thomas, Nehmt der Industrie die Ethik weg!, in: Der Tagesspiegel, 08.04.2019, 2019, https://www.tagesspiegel.de/politik/eu-ethikrichtlinien-fuer-kuenstliche-intelligenz-nehmt-der-industrie-die-ethik-weg/24195388.html (gesehen am 30.11.2021).

Nussbaum, Martha, *Fähigkeiten schaffen: Neue Wege zur Verbesserung menschlicher Lebensqualität*. Freiburg/München: Verlag Karl Alber 2015.

Orrù, Elisa, Effects and effectiveness of surveillance technologies: mapping perceptions, reducing harm, European University Institute Working Papers, LAW 2015/39 - SURVEILLE.

---, Minimum Harm by Design. Reworking Privacy by Design to mitigate the risks of surveillance, in: Leenes, Ronald u. a. (Hg.), *Computers, Privacy and Data Protection: Invisibilities & Infrastructures*. Dordrecht: Springer 2017, 107–137.

---, TRESSPASS D9.2 „Project Baseline for Research Ethics“, 2020, https://cordis.europa.eu/project/id/787120/results (gesehen am 29.10.2021).

--- , *Legitimität, Sicherheit, Autonomie*. Baden-Baden: Nomos 2021.

Orrù, Elisa, Grzondziel, Marianne, TRESSPASS D9.8 „Updated framework for assessing direct ethical, legal and societal impact of risk based screening concepts“, 2021 a), https://cordis.europa.eu/project/id/787120/results (im Erscheinen).

---, D9.9 „Guidelines for decision-makers“, 2021 b), https://cordis.europa.eu/project/id/787120/results (im Erscheinen).

Software developers: distribution by gender 2021, https://www.statista.com/statistics/1126823/ worldwide-developer-gender/ (gesehen am 30.11.2021).

Strubell, Emma u. a., Energy and Policy Considerations for Deep Learning in NLP, 2019, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 3645–3650, https://aclanthology.org/P19-1355 (gesehen am 30.11.2021).

UNESCO, EQUALS Skills Coalition, I'd blush if I could: closing gender divides in digital skills through education, 2019, https://unesdoc.unesco.org/ark:/48223/pf0000367416.page=1 (gesehen am 29.10.2021).

UNESCO, Draft text of the recommendation on the ethics of artificial intelligence, Dok. SHS/IGM-AIETHICS/2021/JUN/3 Rev.22 5 June 2021, 2021, https://unesdoc.unesco.org/ark:/48223/pf0000377897 (gesehen am 29.10.2021).

Elisa Orrù (PD Dr.) ist seit 2013 als wissenschaftliche Mitarbeiterin und Projektleiterin an der Albert-Ludwigs-Universität Freiburg tätig. Dort schloss sie 2020 ihre Habilitation im Fach Philosophie über die Legitimität der EU-Sicherheitspolitik im Kontext der Digitalisierung ab. Das Studium der Philosophie absolvierte sie im Jahr 2004 an der Universität Mailand. 2008 promovierte sie an der rechtswissenschaftlichen Fakultät der Universität Pisa. Für ihre Dissertation wurde sie im Rahmen des Forschungspreises des sizilianischen Kollegs für Philosophie (Syrakus) und durch das italienische Institut für philosophische Studien (Neapel) ausgezeichnet. Sie forschte unter anderem am Max-Planck-Institut zur Erforschung von Kriminalität, Sicherheit und Recht in Freiburg und am Centre for Human Values der Princeton University.