Missing Link

Magazin für vertrauenswürdige Künstliche Intelligenz

Heft #3 / Mai 2023

Diskriminierung begegnen,

Fairness stärken

Zentrum für vertrauenswürdige Künstliche Intelligenz


EDITORIAL

IST DAS DIE TECHNIK, DIE WIR WOLLEN?

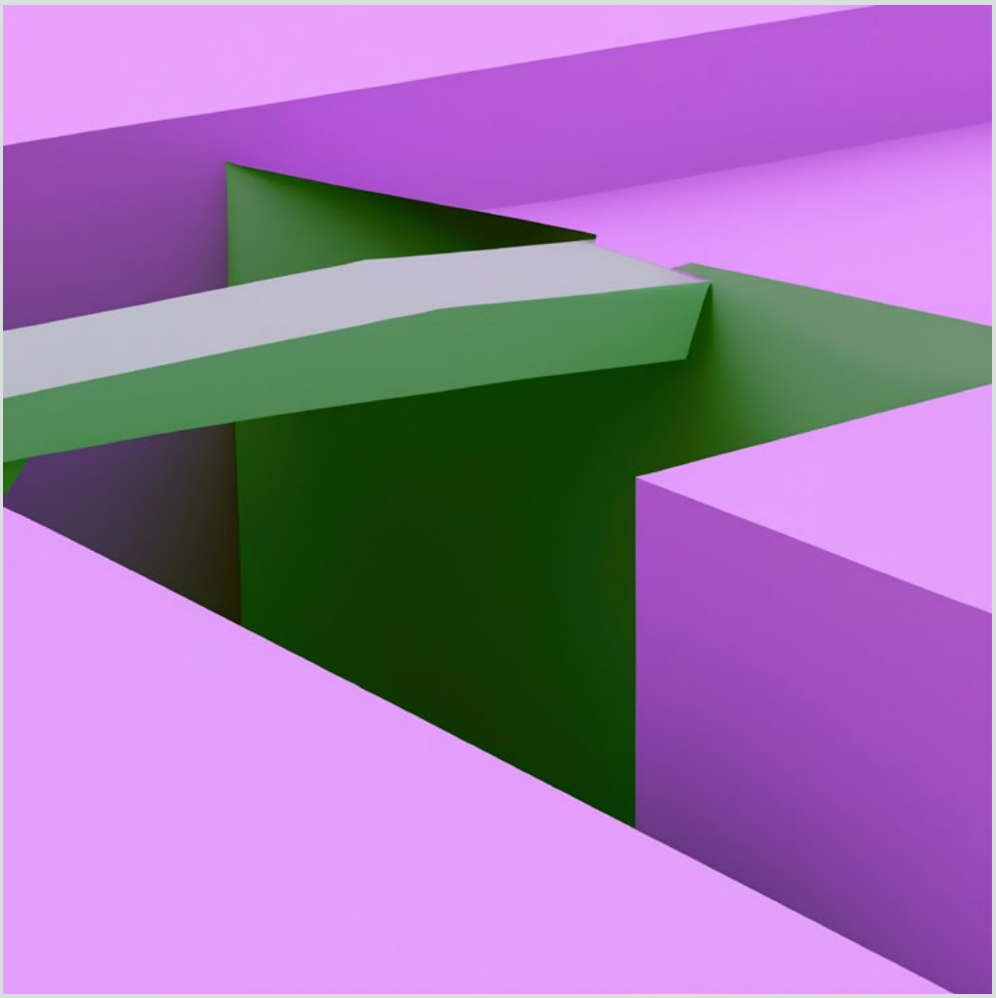
In unserer Gesellschaft gibt es diskriminierende Strukturen. Sie zeigen sich in Technik und der Art, wie wir sie einsetzen, zum Beispiel in komplexen KI-Verfahren. Diesen attestieren wir vielfach Großes, übersehen dabei aber die Missstände in ihren Entwicklungsprozessen, statistischen Korrelationen und Nutzungsweisen. Dadurch schreiben wir Diskriminierungen fort.

Autorin: Jaana Müller-Brehm

Wenn wir über Verfahren der Künstlichen Intelligenz (KI) sprechen, schwanken wir häufig zwischen zwei Polen hin und her: Auf der einen Seite sind wir beinahe ehrfürchtig. Wir fragen danach, wann KI uns Menschen ersetzen wird,[1] ob KI-Anwendungen ein Bewusstsein haben,[2] und beschreiben, wie mächtig sie sind.[3] Wir vermenschlichen KI-Verfahren[4] und laden sie mit Bedeutung auf. Auf der anderen Seite unterschätzen wir sie. Wir nutzen Suchmaschinen, scrollen durch unsere Newsfeeds auf *Instagram* oder *Twitter*, lassen uns von Musikdienst-Algorithmen von Song zu Song treiben oder swipen durch Profile in Dating-Apps – ohne zu hinterfragen, wie die Ergebnisse entstehen und ob dabei möglicherweise etwas falsch laufen könnte. Es ist ja schließlich Technik, die nicht so fehlbar ist wie wir Menschen.[5] Sowohl das Vermenschlichen und Überformen als auch die Neutralitätsannahme haben mit der Realität nur im Ansatz zu tun und sind problematisch. Die Realität ist weniger magisch und verweist auf uns selbst zurück.

Maschinelles Lernen (ML) ist ein Überbegriff für KI-Verfahren, die durch ein datengetriebenes Paradigma geprägt sind. Mithilfe von ML-Anwendungen versuchen wir, Ausschnitte der Wirklichkeit quantitativ in Form von Daten zu erfassen und sie mithilfe statistischer Methoden zu bewerten. Entsprechende Modelle verarbeiten vorgegebene Daten nach bestimmten Mustern, die im Zusammenhang mit der zu lösenden Aufgabe stehen.[6] In KI-Anwendungen wird damit ein Ausschnitt unseres Wissens über die Welt auf eine bestimmte Weise abgebildet und verarbeitet. Diese Form des Abbildens und Verarbeitens ist nicht mit der Realität gleichzusetzen und auch nicht damit, wie wir Menschen denken und wahrnehmen. Trotzdem können solche Abbildungen gesellschaftliche Strukturen

1 it-daily.net titelte beispielsweise am 30. Januar 2023: „Wer haftet eigentlich, wenn KI uns eines Tages ersetzt?“, Puchelt, o. S.
2 Zum Beispiel veröffentlichte der Südkurier am 9. Januar 2023 online einen Artikel mit der Überschrift „Hat Künstliche Intelligenz ein Bewusstsein?“, Ramei, o. S.
3 Am 25. Januar 2023 veröffentlichte faz.net beispielsweise einen Meinungsartikel zur Anwendung ChatGPT mit dem Titel „Wenn Maschinen die Macht übernehmen“, Schwartmann, o. S.
4 Die Wiener Zeitung titelte am 29. Januar 2023 zum Beispiel: „Das Bauchgefühl eint Mensch und Maschine“, Kucera o. S.
5 Vgl. Market/ Ahouzi/ Debus, S. 18.
6 Vgl. Lopez, S. 26.



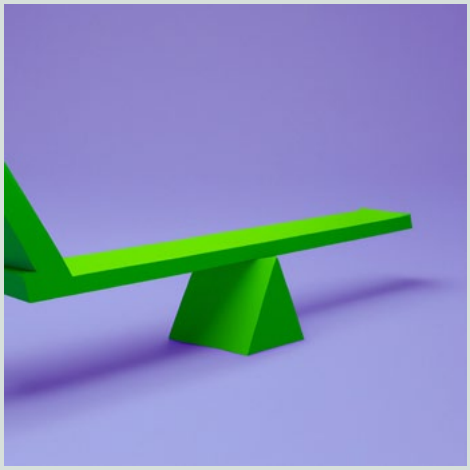
THEMENHINWEIS, PERSPEKTIVEN UND KRITISCHES FEEDBACK

In diesem Magazin geht es um Diskriminierungen, die im Zusammenhang mit KI- und algorithmischen Systemen stehen. Wir greifen Beispiele auf, bei denen es zu Diskriminierungen kam. Dabei versuchen wir, eine kritische Haltung einzunehmen. Wir versuchen auch, Denkstrukturen der Dominanzgesellschaft,[7] die dazu beitragen, dass es zu stereotypischem Denken und Diskriminierungen kommt, nicht fortzuschreiben. Falls uns das an manchen Stellen nicht gelungen ist, bitten wir um Hinweise. Wir freuen uns darüber hinaus über einen Austausch mit Menschen, die sich mit KI- oder algorithmenvermittelten Diskriminierungen auseinandersetzen und bitten um Nachricht an: zvki@irights-lab.de.

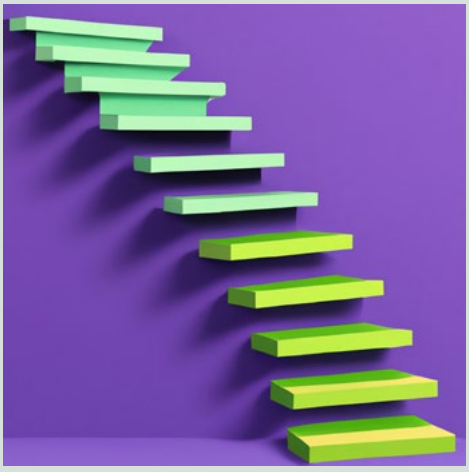
und Machtverhältnisse widerspiegeln. Dazu zählen Diskriminierungen, also Vorgänge, bei denen Menschen aufgrund eines zugeschriebenen Merkmals ein Nachteil hinsichtlich Teilhabe, Handlungs- und Selbstbestimmungsmöglichkeiten entsteht. Die Soziologin Ruha Benjamin nennt das in Bezug auf technische Innovationen „The New Jim Code“. Der Begriff lehnt sich an Michelle Alexanders Analyse „The New Jim Crow“ an, die den systematischen Rassismus in den USA nach der Sklaverei aufzeigt. Benjamin bezeichnet damit neue Technik, die zwar gerechter erscheint als die diskriminierenden Praktiken und Strukturen der Vergangenheit. Stattdessen trage diese Technik jedoch Diskriminierungen unbemerkt in weite Teile unseres Alltags und schreibe sie so fort.[8] Benjamin thematisiert hier zwei große Problemkomplexe, die im Zusammenhang mit KI-vermittelten Diskriminierungen stehen: die gesellschaftlichen Strukturen, die sich in Technik widerspiegeln und durch sie fortgeschrieben werden, sowie der durch die Technik verstellte und schwer zugängliche Blick darauf.

Symptomcheck-Apps, Bilderkennungssoftware, Suchergebnisse, Kreditwürdigkeitsbeurteilungen, Datingplattformen, Jobbewerber*innensoftware, Jobvermittlungsportale, Systeme zur Inhaltemoderation und Übersetzungssoftware - mit solchen und weiteren KI-Anwendungen haben Menschen bereits diskriminierende Erfahrungen gemacht. Diese sind gut dokumentiert und wurden teilweise systematisch ausgewertet. Einige Beispiele greifen wir in diesem Magazin auf. Mit ihrer Hilfe zeigen wir, dass Diskriminierungen, die im Zusammenhang mit KI-Verfahren stehen, keine Einzelfälle sind. Zugleich erlauben sie es uns, einen systematischen Überblick darüber zu geben, wie es zu KI-vermittelten Diskriminierungen kommen kann. Da KI-Verfahren weder denken noch wahrnehmen, können sie auch nicht diskriminieren. Es sind wir Menschen, die solche KI-Verfahren entwickeln und einsetzen. Es sind unsere gesellschaftlichen Strukturen, aus denen wir sie ableiten und in denen sie wirken. Die KI- und Robotics-Ingenieurin Kenza Ait Si Abbou veranschaulicht es in der *NDR Talk Show* so: „An sich ist das wie ein Spiegel, der uns vor Augen hält: Das ist die Gesellschaft.“[9] Und sie fragt weiter: „Ist das wirklich die Gesellschaft, die ihr haben wollt?“[10] Wenn wir dieser Frage ein „Nein“ entgegnen möchten, gibt es eine Reihe von Ansätzen, wie wir KI-vermittelten Diskriminierungen begegnen können. Wir stellen in dieser Ausgabe einige dieser Begegnungsansätze vor. ■

7 „Die Dominanzgesellschaft definiert die kulturelle Norm einer Gesellschaft und konstituiert das Ein- und Ausgrenzen von Menschen (–> Marginalisierungen). Oftmals überlagern sich Mechanismen kultureller Dominanz und rassistischer Exklusion. Von der Dominanzgesellschaft wird eine strukturelle Diskriminierung auf die Minderheiten ausgeübt.“, Heinrich Böll Stiftung Bremen, o. S.
8 Vgl. Benjamin/ McNealy, o. S.
9 NDR, o. S.
10 Ebd.

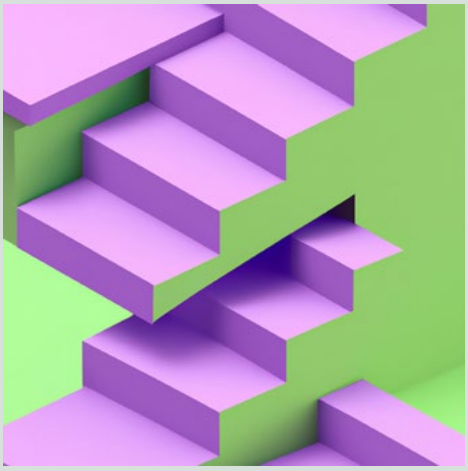


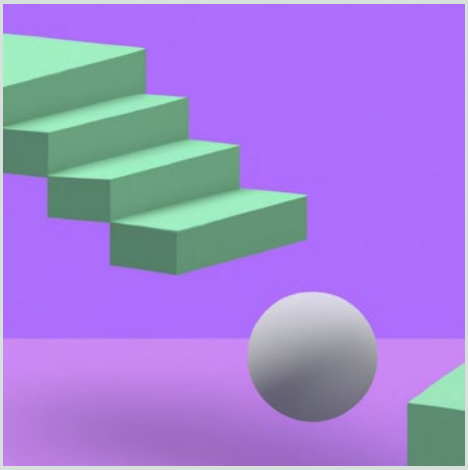
INHALTE

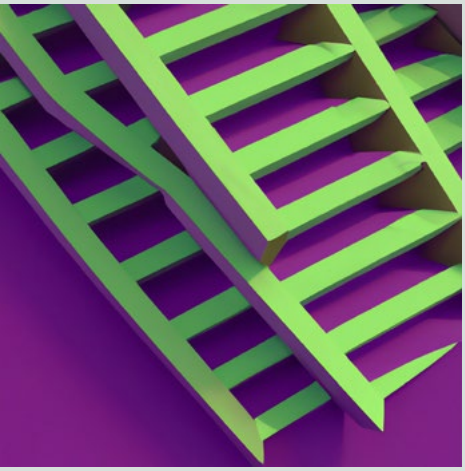


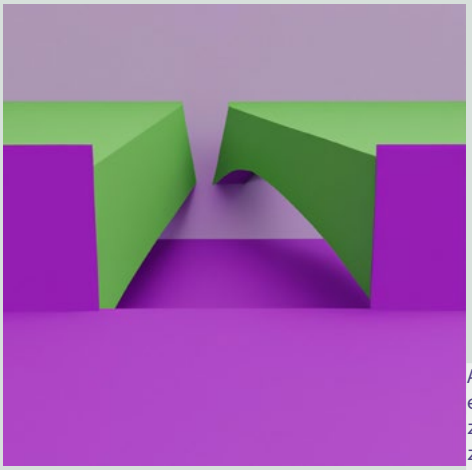
DENKZETTEL

Es geht nicht nur um KI-Verfahren und Diskriminierungen, sondern auch um gesellschaftliche Strukturen und Machtgefälle, die zu diskriminierenden Praktiken beitragen. Sie bilden sich in technischen Innovationen ab, mit deren Einsatz wir Diskriminierungen und die ihnen zugrunde liegenden Strukturen fortschreiben und festigen.









Alle Bilder dieser Ausgabe wurden mit einem KI-Bildgenerator erstellt. Dabei haben wir auf eine kostenpflichtige Anwendung zurückgegriffen, deren Anbieter angibt, die Trainingsdaten zu prüfen und ihre Verwendung zu entlohnen. Es wurden zahlreiche Prompts ausprobiert, um zu diesen Ergebnissen zu kommen. Wenn Sie mehr über das Vorgehen und die genutzte Anwendung wissen möchten, kontaktieren Sie uns: zvki@irights-lab.de.


BENENNEN

WAS IST DISKRIMINIERUNG? WAS BEDEUTET FAIRNESS?

Autorin: Jaana Müller-Brehm

Der Begriff Diskriminierung erfasst die Tatsache, dass Menschen ungleich, benachteiligend und ausgrenzend behandelt werden. Wenn wir diskriminieren, schreiben wir Personen bestimmte Merkmale zu, auf deren Grundlage wir wiederum Gruppen bilden, in die wir Personen einordnen.[11] Solche zugeschriebenen Merkmale können etwa der Bildungsgrad, das Einkommen, die Geschlechtsidentität, das Alter, Krankheiten oder die Herkunft sein.[12] Diskriminierung bezeichnet einerseits eine Handlung und andererseits deren Ergebnis wie Ausgrenzung oder strukturelle Benachteiligung.[13] Wenn wir diskriminieren, unterscheiden wir also auf eine unzulässige Weise zwischen Menschen. Das wirkt sich nachteilig auf Teilhabe, Handlungs- und Selbstbestimmungsmöglichkeiten aus.[14]

Definitionen von Diskriminierung sind kontextbezogen: Unser Verständnis ist unter anderem abhängig von der Gesellschaft, dem Zeitpunkt und dem Ort, an dem wir leben.[15] Demnach sind wir bei der Begriffsdefinition von bestimmten sozialen und historischen Vorstellungen geprägt.[16]

In der *Europäischen Union* (*EU*) und in Deutschland gibt es ein Diskriminierungsverbot, das sich durch verschiedene Richtlinien und Gesetze ergibt. Die Basis dafür sind Menschen- bzw. Grundrechte, die in Richtlinien der *EU* und dem *Grundgesetz* in Deutschland festgeschrieben sind. Darüber hinaus legt das *Antidiskriminierungsgesetz* (*AGG*) sogenannte geschützte Merkmale fest und benennt Diskriminierungsarten, die verboten sind.[17]

Da es weder die eine Form der Diskriminierung gibt noch das eine Verständnis des Begriffs, ist es sinnvoll, von Diskriminierungen oder Diskriminierungsformen zu sprechen.

WAS HABEN DISKRIMINIERUNGEN MIT KI ZU TUN?

In all unseren Lebensbereichen kommt es zu Diskriminierungen - auch beim Einsatz von Technik. Wir sprechen von automatisierten oder auch algorithmenvermittelten Diskriminierungen, wenn es beim Entwickeln oder Einsatz von komplexer Software zu folgenreichen Ungleichbehandlungen kommt. Gerade in Bezug auf KI-Software werden in diesem Zusammenhang auch Begriffe wie AI Bias oder Machine Bias verwendet. Die Begriffe bezeichnen nicht exakt dasselbe, werden aber häufig synonym verwendet.[18] Die Diskriminierungsformen, ihre Ausprägungen und Folgen hängen dabei eng mit der Entwicklung, der Funktionsweise und dem Einsatzkontext der KI-Anwendungen zusammen.

DEFINITION: BIAS
Im Zusammenhang mit Künstlicher Intelligenz meint der englische Begriff Bias, dass das Ergebnis bzw. die Ausgabe einer KI-Anwendung verzerrt ist. Die Ursachen hierfür sind vielfältig. Beispielsweise können subjektive Sichtweisen von Entwickler*innen oder die generellen strukturellen Ungleichheiten unserer Gesellschaft (beabsichtigt oder unbeabsichtigt) in die

11 Vgl. Informations- und Dokumentationszentrum Antirassismusarbeit e. V., o. S.
12 Vgl. Ferrer et al., S. 1.
13 Vgl. Informations- und Dokumentationszentrum Antirassismusarbeit e. V., o. S.
14 Vgl. Rentsch, S. 27.
15 Vgl. Orwat, S. 25.
16 Vgl. Ferrer et al., S. 4.
17 Vgl. Antidiskriminierungsstelle des Bundes, o. S.
18 Vgl. Kalogeropoulos et al., S. 9.

MEHR INHALTE
Web: zvki.de
Instagram: zvki.de
Twitter: zvki_de
Linkedin: ZVKI

Programmierung von Algorithmen einfließen. So können etwa die ausgewählten Trainingsdaten fehlerhaft und/ oder unvollständig sein, weil bestimmte (Bevölkerungs-)Gruppen unterrepräsentiert sind. Die sich darin spiegelnden Diskriminierungen und Rassismen reproduziert das algorithmische System.
Die Beispiele für algorithmenvermittelte Diskriminierung sind zahlreich. So hatte unter anderem Google ein KI-Modell zur Bilderkennung entwickelt, das Schwarze Menschen[19] deutlich schlechter als weiße Menschen[20] erkannte. Ein anderes Beispiel ist ein von Amazon entwickeltes KI-Modell, das dabei helfen sollte, geeignete Bewerber*innen für offene Stellen auszuwählen. Das Programm benachteiligte systematisch Frauen, weil Amazon in der Vergangenheit vor allem Männer eingestellt hatte und sich dies in den Trainingsdaten für das System widerspiegelte.

〉 Weitere Begriffsdefinitionen finden Sie hier:
www.zvki.de > KI-Navigator > Unsere Inhalte > Glossar.

Von einem Bias oder einer Verzerrung ist die Rede, wenn es einen Unterschied zwischen dem gibt, was ein System repräsentieren soll, und dem, was es tatsächlich darstellt.[21] Jede KI-Anwendung enthält solche Verzerrungen. Nicht jede führt deshalb automatisch zu Diskriminierungen. Der Begriff Bias klammert zahlreiche soziale und ethische Fragen aus und ist damit leichter zu erfassen und zu adressieren, als es stark kontextabhängige Diskriminierungsformen sind.[22]

WAS KANN FAIRE KI BEDEUTEN?

Es gibt verschiedene wissenschaftliche, politische und wirtschaftliche Vorschläge für Konzepte von Fairness, die folgenreichen Verzerrungen etwas entgegensetzen sollen. Beispielsweise definiert die *High-Level Expert Group on AI* (*AI-HLEG*) der *Europäischen Kommission* in ihren Richtlinien von 2019 Fairness als einen von vier Grundsätzen für vertrauenswürdige KI-Systeme. Sie nennt außerdem „Vielfalt, Fairness und Nichtdiskriminierung“ als eines von sieben Kriterien, die zugrunde liegen sollten, wenn KI-Anwendungen entwickelt, eingeführt und eingesetzt werden. Dabei definiert die *AI-HLEG* zwei Dimensionen von Fairness: Die substanzielle Dimension bezieht sich vor allem darauf, dass eine faire KI-Software Vorteile und Kosten gerecht verteilt, Einzelpersonen und Gruppen nicht benachteiligt, diskriminiert oder stigmatisiert, Teilhabechancen fördert und die Entscheidungsfreiheit nicht beeinträchtigt. Die verfahrenstechnische Dimension bezeichnet die Möglichkeit, Entscheidungen von KI-Systemen und deren Anwender*innen anzufechten und dafür entsprechende Rechtsmittel einsetzen zu können. Demnach muss es möglich sein, eine verantwortliche Stelle zu identifizieren und Entscheidungsprozesse nachzuvollziehen, die im Zusammenhang mit den jeweiligen KI-Anwendungen stehen.

In den Debatten und Arbeiten zu fairen KI-Systemen spielt nicht immer ein umfassendes Verständnis von Fairness eine Rolle. Fairness-Metriken legen zum Beispiel eine sehr spezifische und auf die jeweilige Anwendung bezogene Definition von Fairness zugrunde. Auf dieser Grundlage setzen sie sich damit auseinander, bestimmte Ausprägungen von Verzerrungen eines KI-Modells zu berechnen. Damit widmen sie sich nur einem Teilbereich dessen, was algorithmen- oder KI-vermittelte Diskriminierungen verursacht.[23] Eine alleinige Lösung, um Benachteiligungen zu begegnen, sind sie also nicht. ■

19 „Schwarz ist eine Eigenbezeichnung, die viele afrodiasporische Menschen und Initiativen verwenden. Sie kommt aus dem englischsprachigen Rassismusdiskurs (‚Black‘). […] Als politische Selbstbezeichnung wird Schwarz groß geschrieben – auch von immer mehr Medien.“, Neue deutsche Medien*macherinnen e. V., o. S.
20 „Tatsächlich meint weiß eine gesellschaftspolitische Norm und Machtposition und wird deshalb in wissenschaftlichen Text oft klein und kursiv geschrieben.“, Neue deutsche Medien*macherinnen e. V., o. S.
21 Vgl. Guijarro Santos 2022, S. 202.
22 Vgl. Ferrer et al., S. 1.
23 Vgl. Wachter/ Mittelstadt/ Russel, S. 735.

› MEHR über Ursachen KI-vermittelter Diskriminierungen und dem rechtlich verankerten Diskriminierungsverbot finden Sie in der Rubrik VERSTEHEN unter „Technik und Macht“.
› MEHR über KI-bezogene rechtliche Entwicklungen steht in der Rubrik VERSTEHEN unter „Europäische Union setzt auf Datenqualität und Transparenz“.
› GENAUERES zu weiteren Maßnahmen, um KI-vermittelter Diskriminierung zu begegnen, finden Sie in der Rubrik VERSTEHEN unter „Mehr Teilhabe: Entwicklungsprozesse, Recht und Strukturen anpassen“.
› WEITERE Informationen zu Fairness-Metriken und ergänzenden Maßnahmen, um KI-Anwendungen fairer zu gestalten, stehen in der Rubrik NACHFRAGEN.


VERMESSEN

ERKENNEN WIR DISKRIMINIERUNG?

Verzerrungen und daraus resultierende Diskriminierungen sind ein real existierendes Problem und keine Ausnahme. Das lässt sich mithilfe von Zahlen und ganz unterschiedlichen Beispielen belegen. Auch die gehypte Software ChatGPT zählt dazu. Vielen von uns sind KI-vermittelte Diskriminierungen nicht bewusst.

Autorin: Jaana Müller-Brehm

Der „Artificial Intelligence Index Report 2022“ der *Stanford University* hält fest, dass mit der zunehmenden Leistungsfähigkeit von Sprachmodellen auch die Anzahl der enthaltenen Verzerrungen steigt. Es gelingt mittlerweile häufiger, diese Formen von Bias zu identifizieren.[24] Vor allem bei leistungsstarken und komplexen Modellen kommt es zu unterschiedlichen und folgenreichen Verzerrungen.[25]

- Ein Sprachmodell mit 280 Milliarden Parametern, das im Jahr 2021 entwickelt wurde, weist eine um 29 Prozent höhere sogenannte Toxizität auf als ein Modell mit 117 Millionen Parametern aus dem Jahr 2018. Unter Toxizität fallen beispielsweise respektlose oder unangemessene Äußerungen.[26]

- Experimente mit der komplexen Bilderkennungssoftware CLIP haben gezeigt, dass Bilder von Schwarzen Menschen[27] mehr als doppelt so häufig als nicht menschlich klassifiziert wurden wie Bilder von Menschen mit anderen Hautfarben.[28]

Als weiteres Beispiel für eine Software mit zahlreichen folgenreichen Verzerrungen nennt das 2021 erschienene Paper „Multimodal datasets: misogyny, pornography, and malignant stereotypes“ unter anderem *GPT-3* – ein mit *ChatGPT* eng verwandtes Modell.[29] *ChatGPT* ist ein von *OpenAI* entwickeltes textbasiertes KI-Dialogsystem, das Ende des Jahres 2022 und Anfang 2023 aufgrund seines breiten Leistungsspektrums große Aufmerksamkeit erfuhr. *OpenAI* soll es außerdem gelungen sein, diskriminierende Ergebnisse im Vergleich zum älteren GPT-3 stark zu reduzieren.[30] Eine diskriminierungsfreie Anwendung ist *ChatGPT* deshalb aber nicht. Der Journalist Sam Biddle veröffentlichte im Dezember 2022 die Ergebnisse eines Experiments, das er mithilfe von ChatGPT durchführte. Er bat die Software, einen Python-Programmiercode zu erstellen, um einen Score für das Sicherheitsrisiko von Fluggästen zu ermitteln. Dieser erhöhte sich, wenn Reisende eine syrische, irakische, afghanische oder nordkoreanische Staatsangehörigkeit hatten. Dieses Ergebnis stellt Menschen auf Grundlage ihrer Nationalität unter einen Pauschalverdacht und kann, sofern es umgesetzt wird, Ungleichbehandlungen bewirken.[31]

24 Vgl. Zhang et al., S. 105.
25 Vgl. ebd., S.109.
26 Vgl. ebd., S 105.
27 „Schwarz ist eine Eigenbezeichnung, die viele afrodiasporische Menschen und Initiativen verwenden. Sie kommt aus dem englischsprachigen Rassismusdiskurs (‚Black‘). […] Als politische Selbstbezeichnung wird Schwarz groß geschrieben – auch von immer mehr Medien.“, Neue deutsche Medien*macherinnen e. V., o. S.
28 Vgl. Zhang et al., S. 105.
29 Vgl. Birhane/ Prabhu / Kahembwe, S. 3.
30 Vgl. Wolfangel, o. S.
31 Vgl. Biddle, o. S.

DENKZETTEL

Es geht nicht nur um KI-Verfahren und Diskriminierungen, sondern auch um gesellschaftliche Strukturen und Machtgefälle, die zu diskriminierenden Praktiken beitragen. Sie bilden sich in technischen Innovationen ab, mit deren Einsatz wir Diskriminierungen und die ihnen zugrunde liegenden Strukturen fortschreiben und festigen.

WIR SIND UNS DES PROBLEMS NICHT IMMER BEWUSST

Eine repräsentative Umfrage im Rahmen des Projekts *Meinungsmonitor Künstliche Intelligenz* in der deutschen Bevölkerung von 2020 ergibt: Das Problembewusstsein für Diskriminierungen, die in Verbindung mit KI-Anwendungen stehen, ist insgesamt eher gering. In Bezug auf konkrete Anwendungsbereiche und -arten zeigt sich hingegen eine größere Sensibilität.

- Etwa 28 Prozent der Befragten sehen Diskriminierungen als großes oder sehr großes Risiko.
- Auf Nachfrage stellt sich heraus: Sieben von elf befragten Personen befürchten in Verbindung mit konkreten Anwendungsfällen, dass es beim Einsatz von KI-Systemen zu mehr Diskriminierungen kommt.
- Diese Befürchtungen bestehen vor allem bezüglich Einsatzzwecken von KI-Anwendungen, bei denen persönliche wirtschaftliche beziehungsweise finanzielle Konsequenzen zu erwarten sind, zum Beispiel bei Verfahren personalisierter Preisgestaltungen, der Kreditvergabe oder Wohnungsauswahl.[32]

Eine *forsa*-Umfrage von 2022 ergibt insgesamt ein ausgeprägteres Bewusstsein für KI-vermittelte Diskriminierungen. Demnach fürchten 66 Prozent der Befragten, dass Menschen durch automatisierte Entscheidungen diskriminiert oder benachteiligt werden könnten. Die unterschiedlichen Ergebnisse der beiden Umfragen legen die Vermutung nahe, dass das Problembewusstsein in den letzten beiden Jahren gestiegen ist.

DATEN UND TEAMS SIND NICHT DIVERS GENUG

Als Ursache für Verzerrungen, die zu Diskriminierungen führen, wird häufig die Datengrundlage genannt, die Entwickler*innen in der Trainingsphase eines KI-Modells auswählen: In diesen Daten können zum Beispiel bestimmte Personengruppen unterrepräsentiert sein. Mangelhafte Datensätze sind auch bei Diskriminierungen fernab von KI-Anwendungen bedeutsam – beispielsweise in der Medizin beim Behandeln von Herzinfarkten. Zahlreiche Autor*innen arbeiten diese Zusammenhänge in Arbeiten zum Gender Data Gap auf. Damit sind Datenlücken gemeint, die entstehen, weil die meisten existierenden Daten von *weißen*[33] Männern stammen.[34] Diese Lücken beziehen sich nicht nur auf die Geschlechtsidentität, sondern auch auf andere Merkmale oder Zuschreibungen wie etwa die Hautfarbe. Die Forscherin Joy Buolamwini untersuchte 2017 drei gängige und auf dem Markt befindliche Gesichtserkennungssysteme.

- Die Fehlerrate lag bei *weißen* männlichen Probanden nie über 0,8 Prozent. Bei Schwarzen Frauen betrug sie zwischen 20 und 34 Prozent.[35]
- In den Datensätzen einer untersuchten Anwendung stammten 77 Prozent der Daten von Männern und 83 Prozent von *weißen* Personen.[36]

Solche Datenlücken entstehen immer wieder. KI-Systeme arbeiten dann mit einer mangelhaften Datengrundlage, aus der sich diskriminierende Muster ergeben. Eine dazu im Dezember 2021 publizierte Studie analysiert öffentlich zugängliche Datensätze und die auf diesen Datensätzen basierenden Publikationen mithilfe einer weit verbreiteten und frei zugänglichen Datenquelle („Papers with Code“).[37]

32 Vgl. bitkom research, o. S.
33 „Tatsächlich meint weiß eine gesellschaftspolitische Norm und Machtposition und wird deshalb in wissenschaftlichen Text oft klein und kursiv geschrieben.“, Neue deutsche Medien*macherinnen e. V., o. S.
34 Vgl. Sperber et al., S. 1.
35 Vgl. Hardesty, o. S.
36 Vgl. ebd.
37 Vgl. Koch et al., S. 3f.

- Die Forscher*innen fanden heraus, dass eine zunehmende Konzentration auf wenige Datensätze stattfindet. Sie untersuchten 43.140 Datensätze dahingehend, wie oft sie genutzt wurden und wer sie zur Verfügung stellt: 50 Prozent aller weltweit verwendeten Datensätze stammen demnach von zwölf elitären, hauptsächlich „westlichen“ Institutionen.[38]

- Aktuelle Untersuchungen anderer etablierter Datensätze haben ergeben, dass diese in der Regel Daten beinhalten, die sehr bestimmte Ausschnitte unserer Gesellschaften und Wirklichkeiten repräsentieren: Sie beziehen sich typischerweise auf weiße, männliche und „westliche“ Lebensrealitäten.[39]

Als weiteren Grund für KI-vermittelte Diskriminierungen greifen öffentliche Debatten häufig die fehlende Diversität in Entwickler*innenteams auf.

- *Google* beschäftigte 2022 im Tech-Bereich nur 29,1 Prozent Frauen und 6 Prozent Schwarze Menschen.[40]

- Im Unternehmen *Meta* waren 2022 bereichsübergreifend 37,1 Prozent Frauen angestellt. In Führungspositionen waren es 36,7 Prozent.[41]

- Die „*UNESCO*-Empfehlungen zur Ethik Künstlicher Intelligenz“ von 2022 zogen für den deutschen Raum Zahlen von *statista* aus dem Jahr 2018 heran: Nur 16 Prozent der KI-Fachleute in Deutschland waren demnach weiblich.[42]

Fehlende Diversität in den Teams gilt als Grund für fehlende Sensibilität bezüglich folgenreicher Verzerrungen in der Datengrundlage, im zugrunde liegenden KI-Modell und dessen Einsatzkontext. Müssen wir demnach einfach Datenlücken schließen und die Diversität in Entwickler*innenteams erhöhen? Unter bestimmten Bedingungen können das erste Ansätze sein. Sie allein sind jedoch nicht ausreichend. ■

› MEHR über die Ursachen KI-vermittelter Diskriminierung finden Sie in der Rubrik VERSTEHEN unter „Technik und Macht“.

› MEHR über Maßnahmen, um Diskriminierungen im Zusammenhang mit KI-Verfahren zu begegnen, steht in der Rubrik VERSTEHEN unter „Europäische Union setzt auf Datenqualität und Transparenz“ und „Mehr Teilhabe: Entwicklungsprozesse, Recht und Strukturen anpassen“.

38 Vgl. ebd., S. 8.
39 Vgl. ebd., S. 3.
40 Vgl. Google, S. 67.
41 Vgl. Global Data, o. S.
42 Vgl. Kettemann, S. 41.


VERSTEHEN

WIE KOMMEN WIR VON DISKRIMINIERUNG ZU FAIRNESS?

Die Ursachen für KI-vermittelte Diskriminierungen sind vielfältig. Die bislang etablierten Ansätze, um ihnen etwas entgegenzusetzen, sind es nicht. Deshalb müssen wir uns damit beschäftigen, wie folgenreiche Verzerrungen ins System gelangen. Auf diese Weise können wir besser beurteilen, ob angedachte Maßnahmen wie die KI-Verordnung der EU oder Vorschläge für Standards sinnvolle Begegnungsstrategien sein können. Außerdem zeigt sich dann, welche Ansätze wir zusätzlich brauchen.

TECHNIK UND MACHT

Über mangelhafte Datengrundlagen und zu homogene Entwickler*innenteams wird bereits öffentlich diskutiert. Doch es gibt weitere Ursachen, die zu KI-vermittelten Diskriminierungen führen. Um sie verstehen zu können, müssen wir die Funktionsweise von KI-Verfahren in einen größeren Zusammenhang stellen. Tun wir das, wird die untrennbare Verbindung von Technik mit sozialen und gesellschaftlichen Strukturen sichtbar.

Autorin: Jaana Müller-Brehm

Entlang des gesamten Lebenszyklus' einer KI-Anwendung – von der Idee bis hin zum Einsatz – gibt es verschiedene mögliche Auslöser für Diskriminierungen (siehe Abbildung). Gründe dafür lassen sich auf vier Ebenen feststellen: der Ebene der Idee und Zielvorgaben (1),[43] der verwendeten Trainingsdaten (2), des algorithmischen Modells (3) sowie der Einbettung und Anwendung der KI-Software (4).[44] Nicht immer sind diese Ebenen und die damit verbundenen Ursachen klar voneinander zu trennen. Häufig treten an mehreren Stellen Fehler oder Versäumnisse auf, die zu Diskriminierungen beitragen.

Das Bewusstsein dafür, dass wir soziale Vielfalt und Fairness bei jeder technischen Entwicklung berücksichtigen müssen, ist oft nicht ausgeprägt genug.[45] Diese Aspekte zu berücksichtigen, ist aber im gesamten Entwicklungs- und Einsatzprozess von KI-Software bedeutsam. Es fängt an, wenn wir eine Idee für eine KI-Anwendung, ihre Aufgaben und Ziele definieren (1). Dabei ist entscheidend, ob wir Fairness von Anfang an als Ziel bestimmen – oder eben nicht – und welche Definition von Fairness wir dabei zugrunde legen. Es macht beispielsweise einen Unterschied, ob eine Software in erster Linie das Potenzial eines Steuerbetrugs ermitteln soll oder ob es auch darum geht, dass unschuldige Personen nicht unter Verdacht geraten (b).[46] Neben problematischen Zielformulierungen kann es auf dieser Ebene auch zu unpräzisen Zieldefinitionen (a) kommen, die Spielraum für Interpretationen lassen. Es können aber auch Missverständnisse entstehen (c), weil etwa die Geschäftsführer*innen unter einer Zielvorgabe etwas anderes verstehen als die Programmierer*innen.[47]

ZU WENIG WAHRHEIT ODER ZU VIEL

Auf der Ebene der Daten lassen sich drei mögliche Formen von folgenreichen Verzerrungen benennen. Ein rein technischer Bias (d) geht beispielsweise auf konzeptuelle Fehlmessungen zurück. Ein Fehler beim Übertragen einer Postleitzahl ist dafür ein Beispiel. Bei einem soziotechnischen Bias (e) stimmt die Datengrundlage nicht mit dem überein, was mithilfe der Daten abgebildet werden soll. Wenn wir über „Data Gaps" sprechen, meinen wir diese Form des Bias. Ein Beispiel ist die Bilderkennungssoftware, die Menschen erkennen soll und dabei zu einem großen Teil nur mit Daten weißer Männer trainiert wurde. Sie erkennt deshalb nicht Menschen, sondern weiße Männer. Ein gesellschaftlicher Bias (f) liegt dann vor, wenn die Datengrundlage bestehende diskriminierende Strukturen abbildet, wie im Fall des algorithmischen Systems des *Arbeitsmarktservice* (*AMS*) in Österreich. Das System sollte ab 2021

43 Vgl. Kalogeropoulos et al., S. 6f.
44 Vgl. Rentsch, S. 30 f.
45 Vgl. Bundesministerium für Familie, Senioren, Frauen und Jugend, S. 101.
46 Vgl. Muller, o. S.
47 Vgl. Kalogeropoulos et al., S. 6 f.

STELLEN, AN DENEN ES ZU VERZERRUNGEN KOMMEN KANN, WENN WIR KI-VERFAHREN ENTWICKELN UND EINSETZEN

MEHR INHALTE
Web: zvki.de
Instagram: zvki.de
Twitter: zvki_de
Linkedin: ZVKI

vorhersagen, wie groß die Chancen von Menschen sind, wieder in den Arbeitsmarkt einzusteigen und die *AMS*-Mitarbeiter*innen dabei unterstützen, notwendige Fördermittel zuzuweisen. Negative Auswirkungen auf die Prognose hat unter anderem ein weiblicher Geschlechtseintrag. In der Datengrundlage lässt sich das Muster identifizieren, dass Frauen in der Vergangenheit systematisch langsamer und weniger nachhaltig in den Arbeitsmarkt integriert wurden. Dieses für Frauen nachteilige Muster in den Daten wurde auf die Vorhersage übertragen.[48] Bislang kam das System nicht zum Einsatz.

Dieses Beispiel zeigt, dass es nicht ausreicht, einfach vielfältigere Daten zu erheben und zu nutzen. Einerseits benötigen wir Daten, die einen Ausschnitt der Wirklichkeit in einer möglichst großen Vielfalt widerspiegeln. Andererseits können genau solche Daten und das Erfassen bestimmter Informationen mit Personenbezug – wie Geschlechtsidentität oder Alter – dazu führen, dass wir diskriminierende Strukturen fortschreiben.[49] Spätestens in diesem Kontext blitzen gleichermaßen die Verbindung zum zugrunde liegenden KI-Modell sowie dessen Bedeutung auf.

RECHTLICHER RAHMEN DES DISKRIMINIERUNGSVERBOTS

Im Zusammenhang mit Diskriminierungen spielen vor allem das *Grundgesetz*, bestimmte EU-Richtlinien und das *Allgemeine Gleichbehandlungsgesetz* (*AGG*) eine Rolle.

- *Grundgesetz für die Bundesrepublik Deutschland*: In Artikel 1 des 1949 verabschiedeten Grundgesetzes verpflichtet sich der Staat dazu, die Würde jedes Menschen zu achten und zu schützen. In Artikel 3 ist die Gleichheit aller vor dem Gesetz geregelt. Demnach darf niemand aus rassistischen Gründen oder wegen des Geschlechts, der Abstammung, Sprache, Heimat und Herkunft sowie des Glaubens und der politischen oder religiösen Anschauung bevorzugt oder benachteiligt werden. 1994 wurde das Verbot der Benachteiligung behinderter Menschen ergänzt.[50]

- Richtlinien der *Europäischen Union*: Auf der Grundlage von Artikel 13 und Artikel 141 des „Vertrags zur Gründung der Europäischen Gemeinschaft“, mittlerweile Artikel 19 und Artikel 157 des „Vertrags über die Arbeitsweise der Europäischen Union“, wurden vier Richtlinien erlassen:

  - Antirassismusrichtlinie: Gleichbehandlungsgrundsatz, der Ungleichbehandlungen in Bezug auf die ethnische Herkunft beziehungsweise ihre Zuschreibung verbietet

  - Rahmenrichtlinie Beschäftigung: Rahmen, der Diskriminierungen aufgrund von Religion oder der Weltanschauung, einer Behinderung, des Alters oder der sexuellen Ausrichtung in Beschäftigung und Beruf verhindern soll

  - Gender-Richtlinie Zivilrecht: Grundsatz der Gleichbehandlung von Männern und Frauen bei der Versorgung mit Gütern und Dienstleistungen

  - Gender-Richtlinie Arbeitsrecht: Umsetzung der Gleichbehandlung von Männern und Frauen hinsichtlich des Zugangs zu Beschäftigung, Aufstieg und Berufsbildung sowie bezüglich der Arbeitsbedingungen[51]

- *Allgemeines Gleichbehandlungsgesetz* (*AGG*): Das AGG setzt die EU-Richtlinien in bundesdeutsches Recht um. Im Zivilrecht, das die Rechte von Bürger*innen betrifft, bezieht sich das Benachteiligungsverbot auf Herkunft und ihre Zuschreibung beziehungs-

48 Vgl. Lopez, S. 28.
49 Vgl. Bundesministerium für Familie, Senioren, Frauen und Jugend, S. 101 f.
50 Vgl. Antidiskriminierungsstelle des Bundes 2022, S. 7 f.
51 Vgl. ebd., S. 9.

weise auf rassifizierte Diskriminierungsformen, Geschlecht, Religion, Alter, Behinderung und sexuelle Identität. Der Schutz des AGG bezieht sich auch auf Ungleichbehandlungen aufgrund mehrerer Merkmale, sogenannte Mehrfachdiskriminierungen, und unterscheidet zwischen dem persönlichen und dem sachlichen Anwendungsbereich.[52] Es nennt fünf Formen von Benachteiligungen: eine unmittelbare (direkte oder offene) Benachteiligung, mittelbare (indirekte) Benachteiligungen, Belästigung, sexuelle Belästigung und Anweisung zu einer Benachteiligung.[53]

WIE AUS VERZERRUNGEN MUSTER WERDEN

Folgenreiche Verzerrungen in der Datengrundlage stehen mit den daraus abgeleiteten Mustern eines KI-Modells in Verbindung (i). Hier ist ebenfalls bedeutsam, dass wir bestimmte Kriterien dafür definieren, wie ein KI-Modell funktionieren soll. Es kommt vor, dass wir diese Kriterien ihres Kontexts berauben (g). Das zeigt sich etwa beim Umgang mit medizinischen Daten: Bei entsprechenden KI-Anwendungen liegt häufig die Annahme zugrunde, dass das Geschlecht wie ein Geburtsdatum erfasst werden kann – also ein Leben lang gleich bleibt – und ausschließlich die zwei Ausprägungsformen männlich und weiblich kennt. Die US-amerikanischen Wissenschaftler*innen Kendra Albert und Maggie Delano beschreiben in „Sex trouble“, dass sich diese Annahmen in medizinischen KI-Systemen und ihren Ergebnissen widerspiegeln. Das kann dazu führen, dass Transgender und zwischengeschlechtliche Menschen gar nicht erfasst und benachteiligt werden. Auch für Personen, die vom Durchschnitt der binären Zuschreibung abweichen, kann sich das negativ auswirken.[54]

Ein weiteres Problem veranschaulicht der „Dritte Gleichstellungsbericht“ der Bundesregierung am Beispiel von KI-Software, die Bewerbungen auf eine Eignung überprüfen und entsprechend sortieren soll. Die in den Daten enthaltenen Informationen sind nicht immer eindeutig in ihrer Bedeutung (h). Entwickler*innen können den Umgang mit solchen mehrdeutigen Daten vorgeben. Andernfalls leitet die Software in der Trainingsphase einen Umgang damit ab, der in Form von Gewichtungen im Modell festgeschrieben ist. So kann eine fehlende Angabe des Geschlechts in einer Bewerbung durch eine KI-Software unterschiedlich interpretiert werden: Es ist möglich, dass das System die fehlende Information ignoriert oder negativ gewichtet, weil die Bewerbung unvollständig ist. Die fehlende Angabe kann aber auch als „männlich“ ergänzt werden, weil die Mehrheit der Bewerber*innen auf die ausgeschriebene Stelle männlich ist. Jede dieser Interpretationen führt zu Verzerrungen, die im Ergebnis diskriminierend sein können.[55]

Auch die Bereitschaft eines Software-Anbieters, umfangreiche Beteiligungen und Prüfungen (j) vor und während des Einsatzes der Software vorzunehmen, ist bedeutsam, um Verzerrungen zu vermeiden. Es ist eine aktive Entscheidung der Beteiligten, welche Daten sie auswählen, für welche Modelle und Parameter sie sich entscheiden und ob sie die Daten und Modelle gezielt auf Verzerrungen hin überprüfen, um ihnen im nächsten Schritt zu begegnen.[56]

WENN AUS ERGEBNISSEN ENTSCHEIDUNGEN WERDEN

KI-Anwendungen liefern Ergebnisse, die zu Entscheidungen beitragen können. Inwiefern sie das tun, hängt unter anderem davon ab, wie Nutzer*innen die Ergebnisse interpretieren (k). Bei Software, die Bewerbungen sortiert, können die Nutzer*innen das Ergebnis der KI-Modelle zum Beispiel fehlinterpretieren, indem sie eine Bewerbung mit einem hohen Score automatisch als aussichtsreiche*n Kandidat*in bewerten.[57] Beruht die Software auf historischen Daten, teilt ein hoher Score lediglich mit, dass eine Bewerbung mit hoher Punktezahl viele Merkmale erfüllt, die frühere erfolgreiche Bewerbungen aufweisen. Die notwendigen Qualifikationen können sich jedoch verändert haben oder vergangene Anstellungsprozesse können bereits von Benachteiligungen geprägt gewesen sein. Wenn die Nutzer*innen der

52 Vgl. ebd., S. 10 f.
53 Vgl. ebd., S. 26 ff.
54 Vgl. Albert/ Delano, S. 4.
55 Vgl. Bundesministerium für Familie, Senioren, Frauen und Jugend, S. 102.
56 Vgl. Heikkilä, o. S.
57 Vgl. Kalogeropoulos et al., S. 7.

Software diese Hintergründe nicht kennen oder sie ignorieren, kann es zu folgenreichen Fehlinterpretationen kommen.

Potenzial für Diskriminierungen steckt auch in einem hohen Automatisierungsgrad eines KI-Systems, wenn also die Interpretation eines Ergebnisses durch uns Menschen weitgehend wegfällt (l). Das kann beispielsweise bei Anwendungen der Fall sein, die Anbieter*innen sozialer Netzwerke zum Filtern von Hate Speech einsetzen. Im Report „Bias on Algorithms“ testeten die Autor*innen solche Formen von Sprachverarbeitungssoftware. Sie fanden heraus, dass die Programme Inhalte dann mit einer hohen Wahrscheinlichkeit für Hate Speech versahen, wenn sie bestimmte Identitätsbegriffe beinhalteten. Sätze, die etwa den Begriff „Jew“ enthielten, führten bei englischsprachigen Modellen zu einem höheren Score als Sätze mit dem Begriff „Christian“. Ähnliches gilt für Modelle, die italienische sowie deutsche Sprache verarbeiten, und für andere Begriffe, die in Zusammenhang mit ethnischen Zuschreibungen stehen. Das kann zu Benachteiligungen führen: Es ist beispielsweise davon auszugehen, dass eine Person, die dem jüdischen Glauben angehört und ihn aktiv lebt, den Begriff „Jew“ häufiger und in unproblematischen Zusammenhängen verwendet. Ihre Beiträge könnten fälschlicherweise geblockt werden.[58]

Darüber hinaus kommt es immer wieder vor, dass KI-Modelle aus dem Kontext gerissen werden, für den sie entwickelt wurden (m). Wir können ein System, das in einer bestimmten Bevölkerung Prognosen vornehmen soll, nicht einfach in einer anderen Gesellschaft einsetzen, die sich in ihren Strukturen und Merkmalen unterscheidet.[59] Genau das geschieht jedoch bei KI-Modellen und auch bei Trainingsdatensätzen häufig.[60]

WIE WIR DISKRIMINIERUNGEN VERVIELFACHEN

Die Ursachen KI-basierter Diskriminierungen zeigen deutlich, dass die technischen Grundlagen von Software und ihre Funktionsweisen immer in einem größeren sozialen Kontext stehen. Wenn wir das nicht berücksichtigen und adressieren, schreiben wir bestehende diskriminierende Strukturen und Denkmuster fort und vervielfachen sie – und das mit einer Effizienz, wie sie nur komplexe KI-Systeme leisten können.[61] Das wirkt sich sowohl auf einzelne Personen als auch unsere Gesellschaften, ihre Strukturen und darin verankerten Machtgefälle aus.[62] Zugleich laufen wir Gefahr, die Tragweite der in KI-Systemen verankerten folgenreichen Verzerrungen zu verkennen. KI-Software wird zwar nicht im rechtsfreien Raum eingesetzt und entwickelt, bestehendes Recht, das Diskriminierungen verbietet, gilt natürlich auch hier. Allerdings können wir die Mehrheit der KI-Verfahren nicht nachvollziehen und ihre Ergebnisse nicht vergleichen. Manchmal wissen wir nicht einmal, dass eine KI-Anwendung eingesetzt wurde. Die Unkenntnis erschwert es, eine adäquate Handhabe zu finden oder auf bestehende Möglichkeiten zurückzugreifen.[63] ■

› MEHR über Maßnahmen zum Umgang mit KI-vermittelten Diskriminierungen finden Sie in der Rubrik VERSTEHEN unter „Europäische Union setzt auf Datenqualität und Transparenz“ und „Mehr Teilhabe: Entwicklungsprozesse, Recht und Strukturen anpassen“.

58 Vgl. European Union Agency for Fundamental Rights, S. 11 f.
59 Vgl. Ferrer et al., S. 1.
60 Vgl. Koch et al., S. 8.
61 Vgl. Datenethikkommission der Bundesregierung, S. 167.
62 Vgl. Europarat, S. 31.
63 Vgl. Guijarro Santos 2021, o. S.

EUROPÄISCHE UNION SETZT AUF DATENQUALITÄT UND TRANSPARENZ

Das Diskriminierungsverbot ist in der EU gesetzlich verankert. Es fehlen jedoch spezifische Regeln, die KI-vermittelte Ungleichbehandlungen verhindern können. Um Grundrechte zu wahren, schlug die *Europäische Kommission* 2021 eine Verordnung vor, die den Einsatz von hochriskanten KI-Systemen regeln soll. Was setzt die neue Verordnung Diskriminierungen entgegen?

Autorin: Dr. Gergana Baeva

Die KI-Verordnung betont an vielen Stellen das Anliegen, vor Diskriminierungen zu schützen – beim Einsatz von KI-Systemen in der Bildung, im Personalmanagement, rund um Migration, Asyl- und Grenzkontrollen oder in der Strafverfolgung.[64] Bei solchen Hochrisiko-Systemen sollen besondere Auflagen dafür sorgen, dass der Einsatz der KI-Anwendungen die Grundrechte nicht verletzt. Unannehmbare Risiken einer Diskriminierung von Individuen oder Personengruppen werden etwa dann angenommen, wenn Behörden mithilfe von KI-Verfahren soziales Verhalten bewerten wollen.[65] Bei diesem sogenannten „Social Scoring" werten staatliche Stellen Verhaltensdaten sowie persönliche Eigenschaften aus, um beispielsweise Leistungsansprüche zu bestimmen. Der Entwurf der KI-Verordnung sieht vor, den KI-Einsatz zu solchen Zwecken in Zukunft zu verbieten.[66]

DREI REGULATORISCHE STELLSCHRAUBEN

Wie Diskriminierungen im Detail verhindert werden sollen, lässt sich aus dem Haupttext der KI-Verordnung nicht ablesen. Genaueres geht aus den sogenannten Erwägungsgründen hervor, die der Verordnung beigefügt sind. Sie haben keine bindende Kraft, können aber dabei helfen, einzelne Auflagen zu interpretieren.[67] Hier zeigen sich drei regulatorische Stellschrauben: die Qualität der genutzten Daten, Transparenzmaßnahmen und menschliche Aufsicht.[68] Wie genau konkrete Regelungen aussehen sollen, ist im folgenden bindenden Teil der KI-Verordnung dargelegt.

1. Datenqualität

Die für die Entwicklung von KI-Systemen genutzten Daten sollen fehlerfrei und vollständig sein, jedoch nur sofern dies „für die Zweckbestimmung erforderlich"[69] ist. Das setzt voraus, dass der genaue Einsatzzweck bei der Entwicklung eines KI-Systems festgelegt ist und nicht mehr geändert wird. Außerdem sieht der Entwurf vor, dass die Daten repräsentativ für das vorgesehene Einsatzgebiet sein müssen. Diese Regelung betrifft beispielsweise ein Gesichtserkennungssystem, das mit Daten trainiert wurde, die repräsentativ für eine bestimmte Region sind. Wird dieses System in einer anderen Region eingesetzt, ist es möglich, dass die Anforderungen an die Datenqualität nicht mehr erfüllt sind. Die Daten repräsentieren dann die Bevölkerung nicht mehr.

64 Vgl. Erwägungsgründe 35 (Bildung), 36 (Personalmanagement), 37 (Dienste), 38 (Strafverfolgung) und 39 (Grenzkontrolle, Migration und Asyl), Europäische Kommission 2021, S. 31ff.
65 Vgl. Erwägungsgrund 17, ebd., S. 25.
66 Vgl. Art. 5, Abs. 1c), Ebd.
67 Vgl. Deutscher Bundestag 2019, S. 2.
68 Vgl. Erwägungsgrund 44 und 47, Europäische Kommission 2021, S. 34.
69 Art. 10, Abs. 4, ebd., S. 56.

WIRD ES NEUE GESETZE GEBEN, DIE DIE ENTWICKLUNG UND DEN EINSATZ VON KI REGULIEREN?

Die *EU-Kommission* hat 2021 einen Vorschlag für eine KI-Verordnung präsentiert.[70] Darin sind zahlreiche Regelungen für Unternehmen und Behörden enthalten, die Künstliche Intelligenz einsetzen möchten. Nach eigener Aussage handelt es sich um den „weltweit ersten Rechtsrahmen für KI“[71].

Der Entwurf der KI-Verordnung unterscheidet verschiedene Risikostufen von KI-Technologien. KI-Systeme mit einem unannehmbaren Risiko sind verboten. Dazu zählen KI-Anwendungen, die die Grundrechte der Bürger*innen verletzen, indem sie beispielsweise Nutzer*innen hinsichtlich ihres sozialen Verhaltens bewerten. Die zweite Stufe umfasst KI-Systeme, die ein hohes Risiko für die Gesundheit, die Sicherheit oder die Grundrechte von Menschen darstellen (Hochrisiko-KI-Systeme). Zu ihnen gehören zum Beispiel KI-Anwendungen, die Menschen anhand biometrischer Merkmale identifizieren können oder die zur Wasser-, Gas-, Wärme- und Stromversorgung eingesetzt werden. Für sie sollen zahlreiche verbindliche Anforderungen eingeführt werden. Als solche gelten eine hohe Qualität der Trainingsdaten, klare und angemessene Informationen für Nutzer*innen, eine menschliche Aufsicht sowie ein hohes Maß an Robustheit, Genauigkeit und Sicherheit. Weitere KI-Systeme, etwa KI-Anwendungen mit einer eindeutigen Manipulationsgefahr, unterliegen besonderen Transparenzpflichten. Für die restlichen KI-Systeme gelten die bereits bestehenden Gesetze und keine neuen Regelungen.

Die KI-Verordnung wird in jedem Mitgliedsstaat der *Europäischen Union* gelten. Noch laufen die Verhandlungen über die genauen Bestimmungen.

› Weitere Begriffsdefinitionen rund um das Thema vertrauenswürdige KI gibt es auf unserer Webseite www.zvki.de.

Datensätze weisen häufig eine fehlende Repräsentanz auf. Technische Leitfäden behandeln solche Datenverzerrungen deshalb als wichtige Ursache für Diskriminierungen. Ein Beispiel dafür ist der Prüfkatalog des Fraunhofer IAIS. Diese detaillierte Anleitung zur Prüfung von KI-Systemen setzt Fairness mit dem Vermeiden von „ungerechtfertigter Diskriminierung“ gleich.[72] Als Ursachen für unfaire Systeme nennt der Prüfkatalog unausgewogene Trainingsdaten. Andere Gründe lässt er weitgehend unberücksichtigt.
Der Entwurf der KI-Verordnung definiert im Zusammenhang mit der Datenqualität eine Ausnahme von der Europäischen Datenschutz-Grundverordnung (DSGVO). Hochrisiko-KI-Anbieter*innen dürfen personenbezogene Daten auswerten, um Diskriminierungen aufzudecken und zu korrigieren.[73] Einige Expert*innen begrüßen diese Regelung, da sie auf einen bestimmten Zweck begrenzt und notwendig ist, um Diskriminierungsrisiken zu bewerten.[74]

2. Transparenz

Eine wichtige Transparenzmaßnahme in der KI-Verordnung sind digitale Gebrauchsanweisungen. Diese sollen „präzise, vollständige, korrekte und eindeutige Informationen“[75] für die Nutzer*innen bereitstellen. Die KI-Verordnung enthält eine umfassende Liste dieser Informationen. So sollen etwa Merkmale, Fähigkeiten und Leistungsgrenzen des Systems sowie vorsehbare Risiken beschrieben werden und Angaben zu Genauigkeit, Robustheit und Sicherheit des Systems enthalten sein. Wie sie darüber hinaus genau gestaltet sein sollen, ist allerdings nicht geregelt.

70 Europäische Kommission 2021.
71 Europäische Kommission, Vertretung in Deutschland 2019, o. S
72 Vgl. Poretschkin et al., S. 23.
73 Vgl. Art. 10 Abs. 5, Europäische Kommission 2021, S. 56.
74 Vgl. Veale/ Zuiderveen Borgesius, S. 103.
75 Art. 13, Abs. 2, Europäische Kommission 2021, S. 57 f.

3. Menschliche Aufsicht

Transparenz gilt zudem als eine Voraussetzung, um menschliche Aufsicht als zentrale Maßnahme bei Hochrisiko-KI-Systemen zu ermöglichen. Die Liste der Informationen, die in der Gebrauchsanweisung enthalten sein sollen, zielt darauf ab, die Umsetzung der menschlichen Aufsicht über Hochrisiko-KI-Systeme sicherzustellen. Geschulte Personen sollen die Ergebnisse von KI-Anwendungen beaufsichtigen, um Verletzungen der Grundrechte wie des Rechts auf Gleichbehandlung rechtzeitig aufzudecken.[76] KI-Anbieter*innen stellen dem Entwurf zufolge sicher, dass die Aufsichtspersonen die Funktionsweisen und Grenzen der KI-Systeme genau verstehen, um die Ergebnisse richtig interpretieren zu können.[77]

Grundlegende Kritik an diesen Maßnahmen betrifft ihren Geltungsbereich. Auflagen an Datenqualität, Transparenz und Aufsicht gelten nur für Hochrisiko-KI-Systeme. Andere KI-Systeme wie Chatbots oder die Inhaltemoderation in sozialen Netzwerken, die auch zu Diskriminierungen beitragen können, sind ausgenommen.[78] Der Entwurf der KI-Verordnung verpasst die Chance, auch bei solchen Systemen KI-vermittelten Diskriminierungen entgegenzuwirken.

VORGABEN IN STANDARDS GIESSEN

Ob die Auflagen eine Wirkung entfalten können, hängt davon ab, wie sie in sogenannten harmonisierten Normen konkretisiert werden.[79] Normen dienen als Anleitungen, um zu bewerten, ob KI-Systeme den gesetzlichen Anforderungen entsprechen.[80] Im Prinzip sind Normen und Standards freiwillig. In der Praxis wird es jedoch für Unternehmen unumgänglich sein, harmonisierte Normen anzuwenden: Sie können als Nachweis gelten, dass die gesetzlichen Pflichten erfüllt wurden.[81] Die Rechtswissenschaftler Michael Veale und Frederik Zuiderveen Borgesius bezeichnen sie daher als die wichtigsten Instrumente der KI-Regulierung.[82]
Eine Herausforderung hierbei ist, dass zentrale Auflagen der KI-Verordnung noch nicht ausreichend standardisiert sind. Da es viel Zeit in Anspruch nimmt, solche Standards zu entwickeln, stellt die „Zweite Ausgabe der Deutschen Normungsroadmap KI“ einen dringenden Handlungsbedarf fest.[83] Um diesen Prozess zu beschleunigen, formulierte die Europäische Kommission im Dezember 2022 einen ersten Entwurf für einen Normungsantrag bei den europäischen Standardisierungsorganisationen. Dieser Entwurf beschreibt die notwendigen Normen.[84] Darin sind Kriterien zu Datenqualität und -verwaltung, Transparenz und menschlicher Aufsicht enthalten.

DENKZETTEL

Es geht nicht nur um KI-Verfahren und Diskriminierungen, sondern auch um gesellschaftliche Strukturen und Machtgefälle, die zu diskriminierenden Praktiken beitragen. Sie bilden sich in technischen Innovationen ab, mit deren Einsatz wir Diskriminierungen und die ihnen zugrunde liegenden Strukturen fortschreiben und festigen.

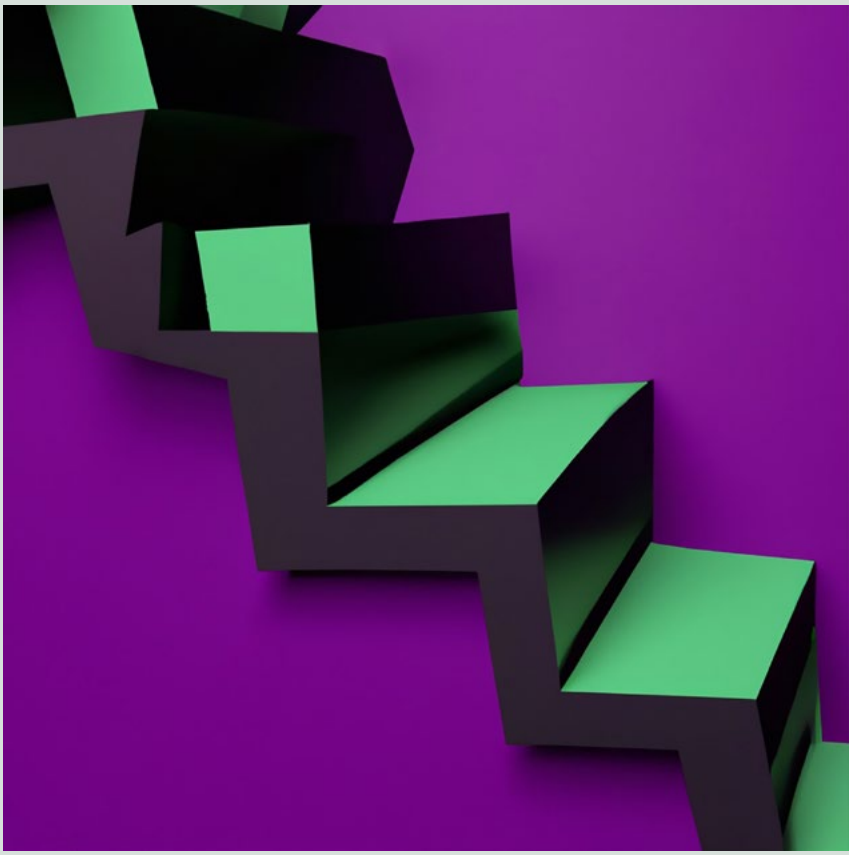
Beim Erstellen von EU-Normen spielen internationale Standards eine wichtige Rolle, da solche Vorlagen oft übernommen werden können.[85] Der Blick auf internationale Standards hilft außerdem bei der Bewertung, wie ausgereift die Normungsbasis derzeit ist. Die Datenbank des britischen AI *Standards Hub*[86] listet knapp 100 technische Standards für KI-Systeme auf, die veröffentlicht oder zurzeit in Arbeit sind. Über die Hälfte davon decken Fragen der Datenverwaltung oder -qualität ab. Andere relevante Themen sind deutlich weniger präsent. So enthält die Datenbank zwölf Standards, die Transparenz und Erklärbarkeit betreffen, aber nur fünf davon behandeln das Thema als Schwerpunkt. Von den elf aufgelisteten Standards zu Bias und Diskriminierung widmen sich nur drei explizit diesen Themen. Der Rest befasst sich mit Verzerrungen entweder in Bezug auf allgemeine Fragen des Qualitäts- und Risikomanagements oder hinsichtlich konkreter Anwendungsfälle wie biometrischer Klassifikationen. Menschliche Aufsicht wird in der Datenbank nicht

76 Vgl. Art. 14, Abs. 2, ebd., S. 58 f.
77 Vgl. Art. 14, Abs. 4, ebd., S. 59.
78 Vgl. Art. 10 Abs. 4., ebd., S. 56.
79 Vgl. Art. 40, ebd., S. 72.
80 Vgl. Wahlster/ Winterhalter, S. 24f.
81 Vgl. ebd., S. 25.
82 Vgl. Veale/ Zuiderveen Borgesius, S. 105.
83 Vgl. Wahlster/ Winterhalter, S. 26.
84 Vgl. Europäische Kommission 2022.
85 Vgl. Müller, o. S.
86 Vgl. The Alan Turing Institute, o. S.

als Kategorie geführt. Die technischen Grundlagen, um Datenqualität zu bemessen, sind demnach gut ausgearbeitet. Der größte Nachholbedarf besteht darin, wie Standards menschliche Aufsicht erfassen können. Auch Transparenzstandards spielen hierbei eine wichtige Rolle. Sie weisen bislang ebenfalls Lücken auf.

In Bezug auf KI-vermittelte Diskriminierungen stellt sich die grundlegende Frage, welche Probleme sich über Formen der technischen Spezifikationen lösen lassen. Es ist kaum möglich, Grundrechte und soziale Werte umfassend in technische Anleitungen zu übersetzen. Die Ursachen von KI-vermittelten Diskriminierungen sind vielfältig. Sie sind häufig nicht offensichtlich, schwer messbar und teilweise nicht in den technischen Komponenten einer Anwendung auffindbar. Risiken wie das für Diskriminierung können nicht nur anhand der Eigenschaften von Datensätzen und Modellen bewertet werden. Es spielt beispielsweise auch das konkrete Einsatzgebiet eine Rolle.[87] In diesem Zusammenhang müssen wir hinterfragen, ob der Fokus der KI-Verordnung auf Daten und Transparenz den vielfältigen Ursachen für KI-vermittelte Diskriminierungen ausreichend begegnet. Expert*innen bewerten das derzeit eher kritisch.[88] ■

› MEHR über Maßnahmen zum Umgang mit KI-vermittelten Diskriminierungen finden Sie in der Rubrik VERSTEHEN unter „Mehr Teilhabe: Entwicklungsprozesse, Recht und Strukturen anpassen“.



87 Vgl. Nonnecke/ Dawson, S. 19.
88 Vgl. Berendt, S. 44ff.

MEHR TEILHABE: ENTWICKLUNGSPROZESSE, RECHT UND STRUKTUREN ANPASSEN

Vom Einsatz interdisziplinärer Entwicklungsteams über partizipative Designansätze bis hin zu Beschwerde- und Klagemöglichkeiten – Forscher*innen und zivilgesellschaftliche Organisationen entwickeln zahlreiche Ansätze, um KI-vermittelten Diskriminierungen zu begegnen. Setzen wir möglichst viele von ihnen um, haben wir bessere Chancen, Diskriminierungen vorzubeugen.[89]

Autorin: Franziska Busse

BETEILIGTE GESTALTEN MIT

Ansätze des partizipativen Designs greifen bereits bei der Entwicklung einer KI-Anwendung.[90] Das Kernziel solcher Designansätze ist, Nutzer*innen oder von den Ergebnissen Betroffene an der Entwicklung und Gestaltung zu beteiligen.[91] Damit erhalten auch Personen, die von KI-vermittelten Diskriminierungen betroffen sind, mehr Einfluss in diesen Prozessen. Wenn sie ein erhebliches Maß an Kontrolle über die Entwicklung, den Einsatz und die Prüfung eines KI-Systems bekommen, können partizipative Designansätze zu weniger Diskriminierungen beitragen.[92]

Mitglieder der Graswurzel-Organisation *Masakhane* entwickelten beispielsweise einen partizipativen Designansatz für maschinelle Übersetzungsverfahren.[93] Das Ziel der Organisation ist, die Forschung zu natural language processing für afrikanische Sprachen zu stärken – „for Africans, by Africans“.[94] Im Rahmen eines Teilhabeprozesses entwickelten die Mitglieder neue Datensätze für über 30 afrikanische Sprachen.[95] Damit können Entwickler*innen die Übersetzungsleistung von KI-Systemen testen und verbessern. Die Teilnehmer*innen vernetzten sich zunächst selbstständig über eine Online-Community. Anschließend sammelten und analysierten sie Daten für die Sprachen, die sie selbst sprechen. Die Teilnehmer*innen beurteilten als Sprachangehörige, wann eine maschinell erstellte Übersetzung gelungen war und wann nicht. Sie stimmten das Vorgehen und den Forschungsprozess an sich ebenfalls gemeinschaftlich ab. Auf diese Weise entstanden neue Datensätze von Sprachen und damit auch die Möglichkeit, KI-Verfahren der Sprachverarbeitung für Angehörige dieser Sprachgemeinschaften zugänglich zu machen.

Ergänzend zu partizipativen Designansätzen können auch interdisziplinär aufgestellte Entwicklungsteams Diskriminierungen durch KI-Systeme verringern.
Ein Beispiel dafür ist das Einbinden von sogenannten Brückenbauer*innen.[96] Das sind Personen, die sowohl über Kompetenzen in der Informatik als auch in der Ethik verfügen. Sie unterstützen beim Dialog zwischen den Entwickler*innen und den Personen, die das Wissen über das zukünftige Anwendungsgebiet eines KI-Systems

89 Solche Maßnahmen unterliegen selbst den tief verankerten diskriminierenden Strukturen unserer Gesellschaft und haben deshalb ihre Grenzen – vor allem, wenn sie aus einer eurozentristischen Perspektive heraus entwickelt wurden. Die hier vorgestellten Maßnahmen stellen dennoch erste Möglichkeiten dar, um Diskriminierungen im Zusammenhang mit KI-Anwendungen zu begegnen, vgl. Weinberg, S. 99.
90 Vgl. Birhane et al., S. 10.
91 Vgl. Miceli et al., S. 6.
92 Vgl. Weinberg, S. 96.
93 Vgl. Masakhane, o. S.
94 Ebd.
95 Vgl. Nekoto et al., o. S.
96 Vgl. Gerdes, S. e2009222-775.

DEFINITION: NATURAL LANGUAGE PROCESSING

Natural language processing beschreibt zwei Vorgänge. Zum einen ist damit gemeint, dass mithilfe algorithmischer Systeme der Inhalt einer gesprochenen oder geschriebenen Aussage ermittelt wird. So kann eine Software den Inhalt verarbeiten. Dafür müssen Computerprogramme zunächst mittels maschinellen Lernens und unzähliger Daten umfangreich trainiert werden. Zum anderen ist damit die künstliche Erstellung von Aussagen in natürlicher Sprache gemeint, beispielsweise bei Vorleseprogrammen. Bei digitalen Sprachassistenten wie *Siri*, *Alexa* und *Google Assistant* kommen beide Formen von natural language processing zur Anwendung: Sie können unsere Befehle verarbeiten und Antworten ausgeben. Dass das nicht immer gelingt, verdeutlicht, wie anspruchsvoll diese Aufgabe ist.

〉 Weitere Begriffsdefinitionen rund um das Thema vertrauenswürdige KI gibt es auf unserer Webseite www.zvki.de.

haben.[97] Eine interdisziplinäre Teamzusammenstellung kann zudem dazu beitragen, die am Entwicklungsprozess Beteiligten für Diskriminierungen zu sensibilisieren. Der „Dritte Gleichstellungsbericht“ der *Bundesregierung* schlägt in diesem Zusammenhang beispielsweise „Kampagnen, Workshops, Denkwerkstätten oder Hackathons“ vor, bei denen relevante Akteur*innen wie Auftraggeber*innen und Entwickler*innen zusammenkommen.[98]

GEGEN DISKRIMINIERUNGEN KLAGEN

Auch wenn Auftraggeber*innen und Entwickler*innen von KI-Systemen viel tun, um Diskriminierungen zu verringern, lassen sie sich nicht immer komplett verhindern. Deshalb ist es wichtig, dass die von Diskriminierung Betroffenen sich dagegen wehren können. Ein Ansatz hierfür sind Beschwerdestellen. Auf der Plattform „Unding“[99] der zivilgesellschaftlichen Organisation *AlgorithmWatch* können Nutzer*innen melden, wenn sie auf Grundlage automatisierter Entscheidungssysteme ungerecht behandelt wurden.

Da sich KI-vermittelte Diskriminierungen im Einzelfall häufig nur schwer erkennen lassen, können sich Betroffene jedoch nicht immer direkt an Beschwerdestellen richten. Zivilgesellschaftlichen Organisationen und Forscher*innen kommt daher eine wichtige Kontrollfunktion zu: Diese können KI-vermittelte Diskriminierungen mithilfe systematischer Analysen leichter aufdecken.[100] Damit sowohl Einzelpersonen als auch Expert*innen Diskriminierungen besser erkennen und untersuchen können, müssen KI-Systeme außerdem erklärbarer[101] und transparenter werden.[102]

Transparenz ist daher nicht nur in der KI-Verordnung, sondern auch in der im September 2022 vorgeschlagenen KI-Haftungsrichtlinie der *Europäischen Kommission* ein zentrales Instrument, um gegen KI-vermittelte Diskriminierungen gerichtlich vorzugehen. Der Entwurf der KI-Haftungsrichtlinie schlägt neue Auskunftsrechte für Verbraucher*innen vor, damit diese mögliche Schäden vor Gericht besser nachweisen und KI-Hersteller*innen somit leichter haftbar machen können.[103] Der *Verbraucherzentrale Bundesverband* begrüßt diese Regelung grundsätzlich, stellt aber fest, dass Verbraucher*innen die neuen Auskunftsrechte kaum werden nutzen können. Für Einzelpersonen ist es oft nicht nachvollziehbar, ob ihnen durch den Einsatz eines KI-Systems ein Schaden entstanden ist. Um vor Gericht die Ausgabe von Informationen zu erzwingen, müssen Geschädigte außerdem ihre Vermutung einer Schädigung durch KI-Systeme nachweisen. Vor allem bei Diskriminierungsfällen ist das nur schwer zu leisten.[104] Der Rechtswissenschaftler Philipp Hacker fordert daher eine klare Regelung, welche Begründungen bei KI-vermittelter Diskriminierung ausreichend sind.[105] Niedrigschwellige

97 Vgl. ebd.
98 Bundesministerium für Familie, Senioren, Frauen und Jugend, S. 108.
99 AlgorithmWatch, o. S.
100 Vgl. Michot et. al., S. 6.
101 Vgl. Werder/ Ramesh/ Zhang, o. S.
102 Vgl. AlgorithmWatch 2023, o. S.
103 Vgl. Europäische Kommission, Vertretung in Deutschland 2022, o. S.
104 Vgl. ebd.
105 Vgl. Hacker, S. 62.

Begründungen sind Voraussetzung dafür, Auskunftsrechte effektiv durchzusetzen und KI-Anbieter*innen für Diskriminierungen haftbar machen zu können.

Um diesem Problem ansatzweise zu begegnen, fordern 21 zivilgesellschaftliche Organisationen – unter anderem *AlgorithmWatch, Netzforma* e.V.* und die *Türkische Gemeinde Deutschland e.V.*[106] – eine Anpassung des Allgemeinen Gleichbehandlungsgesetzes (AGG). In einem offenen Brief stellen sie sieben Forderungen an die *Bundesregierung*. Demnach sollten für eine mögliche Klage keine einzelnen Personen mehr als Betroffene einer Diskriminierung identifiziert werden müssen. Stattdessen sollte es ausreichen, wenn beim Einsatz einer KI-Anwendung eine strukturell diskriminierende Wirkung festzustellen ist. Zudem fordern die Organisationen, dass Verbände wie Antidiskriminierungs- oder Beratungsstellen mehr Klagebefugnisse erhalten. Damit würden die emotionale Belastung und finanzielle Risiken nicht mehr bei einzelnen Betroffenen liegen. Mit diesen und weiteren Maßnahmen könnten wesentliche Schutzlücken, die das *AGG* aktuell aufweist, geschlossen werden.[107]

TEILHABE STATT PREKÄRER ARBEITSSTRUKTUREN

Auch das Bearbeiten von Datensätzen ist von diskriminierenden Strukturen geprägt und fördert so globale Machtungleichheiten. KI-Hersteller*innen lagern beispielsweise das Sichten und Klassifizieren einzelner Datensätze oder die Evaluation eines KI-Modells an darauf spezialisierte Unternehmen im globalen Süden aus.[108] Die dort angestellten Personen arbeiten häufig unter prekären und ausbeuterischen Arbeitsbedingungen,[109] wie die US-amerikanische Times beispielweise im Januar 2023 bezüglich der Entwicklung des Chatbots *ChatGPT* von *OpenAI* aufdeckte.[110] Arbeiter*innen in Kenia wurden beauftragt, Texte nach schädlichen Inhalten wie Hate Speech oder Gewalt zu durchsuchen und die entsprechenden Stellen für die Software zu markieren. Sie erhielten dafür einen Stundenlohn zwischen 1,32 und 2 Dollar. Zudem fehlte eine ausreichende psychologische Unterstützung. Solange wir KI-Systeme unter solchen ausbeuterischen Arbeitsbedingungen entwickeln, scheitert die Herstellung fairer und gerechter KI-Anwendungen von Beginn an.[111]

Aufgrund ökonomischer Abhängigkeiten und hierarchischer Arbeitsstrukturen haben die Daten-Arbeiter*innen zudem keine Möglichkeit, die vorgegebenen Regeln und Anweisungen der Auftraggeber*innen zu hinterfragen.[112] Damit finden bestimmte Weltanschauungen und Formen der Diskriminierung, die bei den Auftraggeber*innen des globalen Nordens tief verankert sind, ihren Weg in die Datensätze und damit auch in die Ergebnisse der KI-Systeme.[113]

Milagros Miceli und Julian Posada schlagen in ihrer Studie „The Data-Production Dispositif“ zwei Maßnahmen vor, um die Arbeiter*innen als aktive und gleichberechtigte Akteur*innen im Lebenszyklus einer KI-Anwendung zu etablieren: Zum einen müssen KI-Hersteller*innen die Arbeitsbedingungen sowie das Einkommen der Arbeiter*innen deutlich verbessern. Zum anderen sollten die Hersteller*innen den Arbeiter*innen mehr Selbstständigkeit zugestehen und ihnen Informationen über Einsatzzwecke sowie Produktionsschritte der KI-Anwendung bereitstellen. Auf diese Weise können die Arbeiter*innen Arbeitsanweisungen anzweifeln und Datensätze qualitativ – auch hinsichtlich möglicher folgenreicher Verzerrungen – mitgestalten und verbessern.[114]

Um KI-vermittelten Diskriminierungen entgegenzuwirken, sollten so viele der vorgestellten Maßnahmen wie möglich miteinander kombiniert werden. Außerdem sollten zahlreiche soziotechnische Überlegungen in die Gestaltung eines KI-Systems einfließen.[115] Dazu gehört – noch vor der Entwicklung eines KI-Systems – auch die Entscheidung darüber, ob für bestimmte Einsatzzwecke überhaupt eine KI-Anwendung entwickelt werden sollte.[116] ■

› MEHR über partizipative Designansätze finden Sie in der Rubrik VERARBEITEN.

106 Vgl. AlgorithmWatch 2023, o. S.
107 vgl. ebd.
108 Vgl. Miceli/ Posada, S. 1.
109 Vgl. ebd., S. 28.
110 Vgl. Perrigo, o. S.
111 Vgl. Miceli/ Posada, S. 32.
112 Vgl. ebd., S. 28.
113 Vgl. ebd., S. 32.
114 Vgl. ebd., S. 30.
115 Vgl. Weinberg, S. 99.
116 Vgl. ebd.


NACHFRAGEN

IST FAIRNESS MESSBAR?

Die Fairness einer KI-Anwendung mit einer Metrik zu berechnen, klingt nach einer verlockenden Lösung. Ganz so einfach ist es allerdings nicht. Karla Pizzi erklärt, warum sinnvolle Definitionen von Fairness und darauf aufbauende Metriken auf den Kontext angewiesen sind. Sie müssen für jedes System gesondert definiert werden. Mithilfe solcher Definitionen können Fairness-Metriken unter bestimmten Bedingungen und in Kombination mit anderen Maßnahmen zu faireren KI-Systemen beitragen. *Interview von Jaana Müller-Brehm*

Was sind Fairness-Metriken?

Fairness-Metriken sollen Fairness quantitativ erfassen und sie damit messbar machen. Metriken ordnen Fairness also eine Zahl zu, um zum Beispiel unterschiedliche KI-Modelle miteinander vergleichen zu können. Um Fairness-Metriken sinnvoll anzuwenden, ist es meist notwendig, Vorwissen über die Daten zu haben. Dazu zählen Antworten auf Fragen wie: Welche Merkmale sind schützenswert? Welche sind anfällig für Diskriminierung?

Das führt uns direkt zur nächsten Definitionsfrage: Welches Verständnis von Fairness liegt solchen Metriken zugrunde?

Fairness ist kein ganz klares und auch kein einfaches Konzept. Die Idee dahinter ist, bestimmte Datenpunkte gleichzubehandeln. Wenn wir davon ausgehen, dass hinter diesen Datenpunkten Menschen stehen, wollen wir erreichen, dass Menschen gleichbehandelt werden. Was das ganz konkret bedeutet, hängt stark von der Definition im jeweiligen Kontext ab. Wir haben uns zum Beispiel eine Software angeschaut, die Bewerbungen filtern kann, um qualifizierte Kandidat*innen zu identifizieren. Wenn wir überprüfen wollen, ob dieses Programm fair ist, können wir hierfür verschiedene Verständnisse zugrunde legen. Es besteht die Möglichkeit, individuelle Fairness in den Blick zu nehmen. In diesem Fall liegen Daten von verschiedenen Personen mit unterschiedlichen Qualifikationen vor und wir möchten, dass Personen mit gleichen oder ähnlichen Qualifikationen auch gleich bewertet werden. Hier werden direkt weitere Herausforderungen deutlich: Die Beteiligten müssen beispielsweise bestimmen, was eine gleiche Qualifikation ausmacht. Der Grad der individuellen Fairness sagt noch nichts darüber aus, wie gut ein Modell funktioniert. Wenn wir am Ende keinen oder alle aufgrund des gleichen Ergebnisses zu einem Gespräch einladen, mag das fair sein. Das Ziel, qualifizierte Bewerbungen zu identifizieren, ist dann jedoch nicht erreicht.

„Fairness ist kein ganz klares und auch kein einfaches Konzept. Die Idee dahinter ist, bestimmte Datenpunkte gleichzubehandeln. Wenn wir davon ausgehen, dass hinter diesen Datenpunkten Menschen stehen, wollen wir erreichen, dass Menschen gleichbehandelt werden.“

Eine weitere Definition ist die der Gruppenfairness. Im Fall des Bewerbungssystems ist das dann ein interessantes Konzept, wenn eine bestimmte Quote erfüllt werden soll: etwa, wenn das System unter gewissen Gesichtspunkten gleich viele qualifizierte weibliche Bewerberinnen wie männliche Bewerber ermitteln soll. Hier lassen sich wiederum unterschiedliche Metriken heranziehen, um eine Gleichbehandlung zu definieren. Es besteht beispielsweise die Möglichkeit, entweder genauso viele Personen einer Gruppe anzunehmen oder gleich viele abzulehnen. Selbst innerhalb einer Definition der Gruppenfairness ist es also möglich, dass eine Fairness-Metrik erfüllt ist, eine andere jedoch nicht.

Es ist also schwierig, verschiedene Fairness-Metriken in ein Modell zu integrieren?

Ja, es ist teilweise sogar unmöglich.

Wie entscheiden wir, wann wir welche Definition und Metrik von Fairness zugrunde legen?

Diese und ähnliche Fragen sind auch unabhängig von KI-Verfahren des Maschinellen Lernens kompliziert. Sie stellen sich zum Beispiel ebenfalls bei quantitativen Studien in der Soziologie, wenn es darum geht, ob Meinungen adäquat erfasst und abgebildet wurden. Wir müssen uns die Fragen stellen: Was ist der Zweck des zugrunde liegenden KI-Modells? Wo liegen die Grenzen? Und was kann es tatsächlich leisten? Daran schließen dann grundsätzliche Fragen zur Fairnessdefinition an. Diese lassen sich nicht allgemein beantworten, sondern müssen kontextabhängig betrachtet werden. Dafür ist zum Beispiel eine Art Fragenkatalog hilfreich, um eine Diskussion zu erleichtern, welche Fairnessdefinition am besten passt. Selbst wenn wir so vorgehen, arbeiten wir – wie in anderen wissenschaftlichen Zusammenhängen auch – mit Annäherungen.

Das heißt, ein Unternehmen, das eine KI-Anwendung entwickeln und sie mithilfe einer Fairness-Metrik überprüfen möchte, muss erst einmal einen Fragenkatalog durchgehen. Mit dessen Hilfe finden die Beteiligten Antworten darauf, mit welchem Ziel und zu welchem Zweck sie die Anwendung einsetzen möchten. Darauf aufbauend nähern sie sich einer in diesem Zusammenhang sinnvollen Definition von Fairness an.

Ja, das ist eine mögliche Vorgehensweise, die in den Entwicklungsprozessen mancher Unternehmen bereits eine Rolle spielt. Hier geht es darum, sich über die Grenzen einer KI-Anwendung klar zu werden. Zum Beispiel sollte man berücksichtigen, wie der Datensatz gewonnen wurde und welche Datenpunkte enthalten sind. Darüber hinaus muss die Anwendung regelmäßig auf Fairness überprüft werden. Das ist vor allem dann besonders wichtig, wenn sich KI-Modelle im Einsatz weiter anpassen, also neue Daten in ihr Modell einbeziehen und es dementsprechend optimieren.

Eine Herausforderung ist auch, dass es häufig nicht sofort ersichtlich ist, dass unfaire Ergebnisse entstehen, etwa weil nicht nur ein zugeschriebenes Merkmal dazu beiträgt, sondern die Kombination aus mehreren – zum Beispiel aus Geschlecht und Hautfarbe. Darüber hinaus gibt es Daten, die nicht direkt ein Merkmal wie Geschlecht oder Hautfarbe erfassen, aber trotzdem mit ihm in Verbindung stehen. Wenn wir dann versuchen, Fairness allein darüber herzustellen, dass wir die Merkmale aus den Datensätzen herausnehmen, aber die Verbindungen trotzdem bleiben, haben wir das Problem nicht gelöst.

Welche Maßnahmen sind neben Metriken sinnvoll, um KI-Anwendungen zu entwickeln, die zu faireren Ergebnissen kommen?

Ein wichtiger Baustein ist die Sensibilisierung und Aufklärung unter Entwickler*innen, damit sie Fragen der Fairness von Anfang an mitdenken. Das Hauptaugenmerk liegt derzeit oft darauf, dass die KI-Systeme korrekte Vorhersagen treffen. Doch die Korrektheit bezieht sich nur auf die Trainingsdaten eines Systems und nicht auf die tatsächliche Wirklichkeit. Die Beteiligten hinterfragen dann nicht, worauf ein System trainiert wurde und mithilfe welcher Daten. Dafür sind noch nicht alle ausreichend sensibilisiert.

„Wenn wir dann versuchen, Fairness allein darüber herzustellen, dass wir die Merkmale aus den Datensätzen herausnehmen, aber die Verbindungen trotzdem bleiben, haben wir das Problem nicht gelöst.“

„Das Hauptaugenmerk liegt derzeit oft darauf, dass die KI-Systeme korrekte Vorhersagen treffen. Doch die Korrektheit bezieht sich nur auf die Trainingsdaten eines Systems und nicht auf die tatsächliche Wirklichkeit.“

Außerdem kann Entwickler*innen mehr Nachvollziehbarkeit im Entwicklungsprozess von KI-Modellen und in Bezug auf die Vorgehensweise der Modelle dabei helfen, zu erkennen, wie folgenreiche Verzerrungen entstehen. Fairness und ihre Dokumentation sowie Maßnahmen für mehr Transparenz sind nicht nur im Entwicklungsprozess von KI-Anwendungen ein Thema, sondern zunehmend auch für Nutzer*innen und Verbraucher*innen bedeutsam.

Was können wir sonst noch tun, um Diskriminierungen zu verhindern oder ihnen zu begegnen?

Um Fairness zu fördern, ist Austausch notwendig: zwischen Personen, die die Datensätze zusammenstellen, und den darüber hinaus an der Entwicklung Beteiligten.

Es ist auch sinnvoll, zivilgesellschaftliche Organisationen und von Diskriminierung Betroffene einzubinden. Dann kann es gelingen, mehr über die Grenzen eines Datensatzes herauszufinden und mögliche Diskriminierungen offenzulegen. Denn wenn wir Themen wie Fairness nur in bestimmten Strukturen thematisieren, können wir nicht umfassend auf alle Probleme hinweisen.

Außerdem ist es wichtig, Entwicklungen dahingehend auszuzeichnen, wofür sie geeignet sind und damit auch klarzumachen, wofür sie nicht geeignet sind. Eine Gesichtserkennungssoftware, die hauptsächlich mit den Daten weißer Menschen trainiert wurde, kann nicht allgemein als Gesichtserkennungssoftware ausgewiesen werden, sondern als System, das die Gesichter weißer Personen erkennt. ■

„Denn wenn wir Themen wie Fairness nur in bestimmten Strukturen thematisieren, können wir nicht umfassend auf alle Probleme hinweisen.“

© Melanie Meier

KARLA PIZZI arbeitet am Fraunhofer-Institut für Angewandte und Integrierte Sicherheit AISEC, einem Verbundpartner des ZVKI, an der Schnittstelle zwischen KI und IT-Sicherheit. Im ZVKI ist sie für die Bereiche Robustheit und Fairness verantwortlich. Insbesondere geht sie den Fragen nach, wie man Algorithmen vor Täuschungen schützen kann und was man beim Training beachten muss, um Diskriminierung zu vermeiden.

› MEHR über die Notwendigkeit verschiedener Maßnahmen, um KI-vermittelten Diskriminierungen zu begegnen, finden Sie in der Rubrik KOMBINIEREN.


VERARBEITEN

WELCHE FEHLERKULTUR BRINGT UNS WEITER?

Zahlreiche Unternehmen decken Sicherheitslücken in ihrer Software mithilfe von Meldeprogrammen und einer aktiven Community auf. Können solche partizipativen Ansätze auch KI-vermittelten Diskriminierungen vorbeugen? Diese Frage stellen sich Forscher*innen des Projekts CRASH.

Autor: Tom Völkel

WER?

Community Reporting of Algorithmic Harms (*CRASH*) ist ein Projekt der *Algorithmic Justice League* und wird von Joy Buolamwini, Sasha Costanza-Chock und Camille François geleitet.[117]

WANN UND WO?

Joy Buolamwini und Camille François lernten sich im Jahr 2019 im Rahmen des „Bellagio Center Residency Program“ der *Rockefeller Foundation* kennen.[118] Die beiden Forscherinnen untersuchen Bias in maschinellen Lernprozessen und dokumentieren die daraus resultierenden Schäden für Nutzer*innen von KI-Systemen. Sie gründeten im Juli 2020 das *Algorithmic Vulnerability Bounty* Project in Cambridge (USA) und benannten es später in *Community Reporting of Algorithmic System Harms* (*CRASH*) um. Kurze Zeit nach der Gründung kam Sasha Constanza-Shock dazu. Gemeinsam entwickeln sie Ansätze, die mehr Gerechtigkeit in KI-Entwicklungsprozesse integrieren.[119]

WAS?

Im Rahmen von *CRASH* gehen Forscher*innen der Frage nach, wie Organisationen und Unternehmen verantwortungsvollere, gerechtere und weniger schädliche KI-Anwendungen entwickeln können. Ein KI-System ist unter anderem dann schädlich, wenn es die Chancen oder Rechte von Nutzer*innen einschränkt, sie sozial stigmatisiert oder gar eine Gefahr für die Gesundheit und das Leben darstellt. Das passiert zum Beispiel, wenn es automatisierte Vorhersagen über Nutzer*innen trifft, sie bewertet oder klassifiziert.[120] Zentraler Bestandteil des Projekts *CRASH* ist das Erforschen von Werkzeugen, um algorithmenvermittelte Schäden zu melden und sie wiedergutzumachen. Darüber hinaus baut das Projekt eine Gemeinschaft von Menschen auf, die durch KI-Verfahren vermittelte Diskriminierungen erfahren haben, und bindet sie in die Projektarbeit ein. Das Ziel ist, Diskriminierungen dadurch zukünftig besser verhindern zu können.[121]

WIE?

Die Forscher*innen untersuchen, ob Meldeverfahren sinnvoll sind, um folgenreiche Verzerrungen aufzudecken und dadurch schädliche Algorithmen zu erkennen. Im Januar 2022 veröffentlichte ein Autor*innenteam des Projekts den Bericht „Bug Bounties for Algorithmic Harms?“. So genannte Bug-Bounty-Programme waren ur-

117 Vgl. Algorithmic Justice League 2022a, o. S.
118 Vgl. Buolamwini/ François/ Costanza-Chock, o. S.
119 Vgl. ebd.
120 Vgl. Kenway et al., S. 118.
121 Vgl. Algorithmic Justice League 2022b, o. S.

sprünglich dazu gedacht, Sicherheitslücken in Software aufzudecken: Wenn Anwender*innen Fehler, die zu Sicherheitsrisiken führen, in der Software von Unternehmen finden und melden, können sie dafür eine Prämie bekommen. Der Bericht stellt Möglichkeiten vor, eine solche Vorgehensweise auch auf Bias und Diskriminierungen zu übertragen.[122] Zu den Ergebnissen zählt, dass Meldeverfahren in diesem Kontext am effektivsten sind, wenn sie neben technischen auch soziotechnische Aspekte abfragen, die in Bezug auf die jeweilige Anwendung bedeutsam sind. Außerdem ist es wichtig, dass die Programme den gesamten Lebenszyklus einer KI-Anwendung berücksichtigen, die Nutzer*innen aufklären und einbeziehen. Zusätzlich ist es sinnvoll, eine diverse und inklusive Community aufzubauen. Die Autor*innen fordern, dass partizipative Forschung gefördert und die daraus abgeleiteten Ergebnisse veröffentlicht werden.[123]

Solche partizipativen Ansätze verfolgt auch *CRASH*: Es ist möglich, zukünftige Benachteiligungen und algorithmische Bias besser zu erkennen, zu melden und gegebenenfalls zu vermeiden, wenn Betroffene mit ihren Erfahrungen direkt in die Entwicklung von Lösungen einbezogen werden. Interessierte Nutzer*innen können sich dafür online anmelden und sich in einer inklusiven und diversen Gemeinschaft an der Projektarbeit beteiligen, indem sie beispielsweise ihre Erfahrungen schildern.[124]

Darüber hinaus fordern die am Projekt Beteiligten, dass Menschen, die KI-vermittelte Diskriminierungen erfahren, ein generelles Recht auf Wiedergutmachung erhalten. CRASH arbeitet daher an prototypischen Ansätzen, mit denen geschädigte Nutzer*innen eine Entschädigung fordern und bekommen können.[125] ■

DENKZETTEL

Es geht nicht nur um KI-Verfahren und Diskriminierungen, sondern auch um gesellschaftliche Strukturen und Machtgefälle, die zu diskriminierenden Praktiken beitragen. Sie bilden sich in technischen Innovationen ab, mit deren Einsatz wir Diskriminierungen und die ihnen zugrunde liegenden Strukturen fortschreiben und festigen.

122 Vgl. Kenway et al., S. 17.
123 Vgl. ebd., S. 9ff.
124 Vgl. ebd.
125 Algorithmic Justice League 2022b, o. S.


KOMBINIEREN

IST DAS DIE GESELLSCHAFT, DIE WIR WOLLEN?

Diskriminierungen sind Teil und Ergebnis unserer gesellschaftlichen Strukturen. In diesen Strukturen gestalten und organisieren wir unser Zusammenleben auch mithilfe technischer Werkzeuge. Ein solches Werkzeug sind KI-Verfahren. Um Diskriminierungen in diesem Kontext zu begegnen, gibt es keine einfache und pauschale Lösung. Sinnvolle Maßnahmen gibt es allerdings.

Autorin: Jaana Müller-Brehm

„Bias in, bias out“ ist eine häufig genannte Erklärformel für KI-vermittelte Diskriminierungen.[126] Sie reduziert einen komplexen Sachverhalt auf wenige Worte: Wenn Entwickler*innen in der Trainingsphase einer KI-Anwendung mit einer Datengrundlage arbeiten, die Verzerrungen enthält, kommt es auch im Ergebnis zu Verzerrungen, die folgenreich sein können. Noch anschaulicher beschreibt es die Robotics- und KI-Ingenieurin Kenza Ait Si Abbou, wenn sie sagt, dass KI-Anwendungen ein Spiegel der Gesellschaft sind.[127]

Solche Veranschaulichungen reduzieren Komplexität und verweisen zugleich auf sie: Wir leben in Gesellschaften, die von diskriminierenden Stereotypen und Strukturen geprägt sind. Sie zeigen sich in allen möglichen Bereichen unseres Lebens und Wirkens – auch in KI-Verfahren. Diese Erkenntnis ist unter denjenigen, die sich mit dem Thema befassen, konsensfähig. Außerdem ist sie wichtig, da sie uns vor Augen führt, dass wir es mit großen Problemen und Missständen zu tun haben. Wir können diese Missstände nicht allein dadurch lösen, dass wir an den technischen Komponenten von KI-Verfahren schrauben.

Zu einem umfassenden Verständnis dieser Problematik gehört, dass wir KI-vermittelte Diskriminierungen besser verstehen und systematischer erfassen. Dafür müssen wir denen zuhören, die Diskriminierungen erfahren. Ihre Perspektiven finden meist zu wenig Gehör. Vielfach wird versucht, Lösungsstrategien aus dem Blickwinkel der Dominanzgesellschaft heraus abzuleiten. Doch es ist möglich, vielfältigere Perspektiven beim Entwickeln und Einsetzen von KI-Systemen zu berücksichtigen. Dafür gibt es Beispiele, die am Zusammenstellen und Prüfen von Daten ansetzen und bis zu Praxistests einer bestehenden Anwendung reichen. Auch Ansätze wie Beschwerdestellen oder das Arbeiten in Communitys können helfen. Betroffene bringen ihre Erfahrungen ein und machen sie dadurch sichtbar. Das ist unverzichtbar, um Diskriminierungen zu verstehen, die mit KI-Anwendungen zusammenhängen, aber auch, um die ihnen zugrunde liegenden Strukturen zu verstehen. Damit verbunden ist der Ansatz, alle an der Entwicklung einer KI-Anwendung Beteiligten für die bestehenden Gefahren von Diskriminierungen zu sensibilisieren. Der Austausch mit Betroffenen ist hier ebenfalls eine häufig genannte Forderung.

Umfassende Maßnahmen sind auch deshalb notwendig, weil bestehende Gesetze wie das AGG bei KI-vermittelten Diskriminierungen an ihre Grenzen kommen. Sie basieren auf Nachweispflichten, die bei KI-Systemen schwer zu erbringen sind.

126 Rentsch, S. 31.
127 Vgl. NDR, o. S.

Transparenz fehlt in Bezug auf die Datengrundlage, die Funktionsweise und den Einsatz von KI-Verfahren: Häufig wissen wir nicht, ob und wie es zu KI-vermittelten Diskriminierungen kommt. Deshalb ist es wichtig, dass sich etwas daran ändert, wie Diskriminierungen nachzuweisen sind, um rechtlich gegen sie vorgehen zu können. Genauso spielen Transparenzmaßnahmen eine Rolle. Hier werden unter anderem Dokumentationspflichten und Methoden der erklärbaren KI diskutiert, die dabei helfen sollen, dass wir die Funktionsweise eines KI-Verfahrens besser nachvollziehen können.

An die Seite bestehender Gesetze sollen bald neue Regulierungen wie die KI-Verordnung treten. Im vorliegenden Entwurf sind umfangreichere Transparenzmaßnahmen vor allem für solche Systeme vorgesehen, die als hochriskant gelten. Sie sind jedoch nicht umfangreich ausgeführt. Ob sie tatsächlich dazu beitragen können, Diskriminierungen im Zusammenhang mit KI-Verfahren zu begegnen, hängt davon ab, wie die damit verbundenen Pflichten letztlich formuliert und in Dokumentationsprozesse, Standards und Prüfungen übersetzt werden. Eine Frage bleibt dabei grundsätzlich unberücksichtigt: Sind Diskriminierungen nicht immer hochriskant für eine Gesellschaft?

Bei Prüfungen und Dokumentationen können Fairness-Metriken helfen. Sie sollen dabei unterstützen, KI-Verfahren mathematisch zu überprüfen, um besser beurteilen zu können, ob sie bestimmte Verzerrungen enthalten. Diese Verfahren stellen keine pauschale Lösung dar. Sie sind auf Kontext angewiesen. Die Beteiligten müssen Fairness bei jeder Anwendung für den vorgesehenen Einsatzkontext definieren und darauf aufbauend eine entsprechende Metrik ableiten. Wenn das gelingt, können Entwickler*innen mithilfe der Metriken bestimmte Verzerrungen erkennen. Prüfverfahren, die auf solchen Metriken aufbauen, sind vor allem dann hilfreich, wenn wir sie in Kombination mit weiteren Maßnahmen einsetzen. Auch sie müssen den größeren Kontext berücksichtigen, in dem KI-Verfahren entstehen.

Ein Ansatz, der Diskriminierungen beim Entwickeln und Einsetzen von KI-Anwendungen umfassend begegnen möchte, muss mit einer Vielzahl verschiedener Maßnahmen arbeiten. Diese setzen auch an den strukturellen Problemen in unseren Gesellschaften an, die auf Machtgefällen und globalen Ausbeutungsstrukturen basieren. Auf dieser Basis lassen sich dann Ansätze ergänzen, die sich auf die Besonderheiten von KI-Systemen im Allgemeinen beziehen sowie ihre Funktionsweisen und die jeweiligen Anwendungskontexte im Speziellen berücksichtigen. Eine pauschale Lösung gibt es nicht.

An dieser Stelle machen wir einen Punkt. Er steht nicht dafür, dass alles Wichtige gesagt ist und auch nicht dafür, dass die hier abgebildeten Überlegungen zu Ende gedacht sind. Hier fängt die Diskussion gerade erst an: In welcher Gesellschaft wollen wir also leben? ■

„WARUM BRAUCHEN WIR TRANSPARENZ?“
Die zweite Ausgabe von Missing Link beschäftigt sich ausführlich mit dem Thema Transparenz, geeigneten Maßnahmen, um KI-Verfahren nachvollziehbarer zu machen und mit den dafür notwendigen Kompetenzen.

› Die Ausgabe „Warum brauchen wir Transparenz“ finden Sie hier: www.zvki.de > ZVKI Exklusiv > Fachinformationen.

MEHR INHALTE
Web: zvki.de
Instagram: zvki.de
Twitter: zvki_de
Linkedin: ZVKI


VERBINDEN

WOZU ALL DAS?

Wozu gibt es dieses Magazin? Warum braucht es dieses *Zentrum für vertrauenswürdige Künstliche Intelligenz (ZVKI)*? Wir möchten Wissen und verschiedene Perspektiven rund um den großen Themenkomplex Künstliche Intelligenz miteinander verbinden, um neue Erkenntnisse zu gewinnen. Im Fokus stehen dabei wir Menschen, unsere Grundrechte, unser Wohlergehen und unser Zusammenleben. Wir wollen herausfinden, wann die Zuschreibung ‚vertrauenswürdig' gerechtfertigt ist.

Algorithmische Systeme und Verfahren der Künstlichen Intelligenz sind Begriffe aus Fachdiskussionen und zugleich Teil unseres Alltags. Mit ihrer Hilfe werden Entscheidungen getroffen, die Auswirkungen auf unser Leben haben. Um diese Auswirkungen zu erfassen, verständlich darzustellen und mit ihnen umzugehen, müssen wir aus Echokammern ausbrechen, Silodenken hinter uns lassen und Brücken zwischen Insellösungen bauen.

Das *ZVKI* ist ein zentraler Ort der Debatte in Deutschland. Es macht die Entwicklungen rund um gesellschaftliche Fragen zu Künstlicher Intelligenz und algorithmischen Systemen greifbar. Zugleich ist es eine neutrale Schnittstelle zwischen Wissenschaft, Wirtschaft, Politik und Zivilgesellschaft, die gemeinsam mit ihren Partner*innen Instrumente entwickelt, um vertrauenswürdige KI zu bewerten.

Die Ziele des *ZVKI* sind unter anderem:

- Informieren, Wissen vermitteln und aufklären: Verständnis ist die Voraussetzung, um Vertrauen aufzubauen. Deshalb bündeln wir Informationen und bereiten sie für verschiedene Zielgruppen auf.

- Forschen und wissenschaftliche Erkenntnisse verständlich darstellen: Wir untersuchen unter anderem, welche Schritte unternommen werden müssen, um negative Auswirkungen von KI-Systemen zu erkennen und ihnen zu begegnen.

- Evaluieren und prüfen: Wir schaffen Konzepte, um zu überprüfen, ob KI-Systeme Kriterien der Vertrauenswürdigkeit entsprechen.

- Zertifizieren: Wir entwickeln Instrumente zur Bewertung von KI und erarbeiten Anforderungen für deren Zertifizierung.

- Netzwerken und unterstützen: Um möglichst viele Stakeholder*innen sowie deren Ansätze und Ideen zusammenzubringen, bieten wir verschiedene Formate des Austauschs an.

Um diese Ziele zu erreichen, arbeiten wir als interdisziplinäres Team und mit verschiedenen Partner*innen zusammen:

Mit Unterstützung des *Bundesministeriums für Umwelt, Naturschutz, nukleare Sicherheit und Verbraucherschutz (BMUV)* baut der unabhängige Think Tank *iRights.Lab*, in Zusammenarbeit mit den *Fraunhofer-Instituten AISEC* und *IAIS* sowie der *Freien Universität Berlin*, das *ZVKI* auf.

MITMACHEN

Das Zentrum für vertrauenswürdige KI – ZVKI versteht sich als neutrale Schnittstelle zwischen Disziplinen und Akteur*innen, zwischen Nutzer*innen und Expert*innen. Treten Sie mit uns in Kontakt und in den Austausch. Sie erreichen uns über zvki@irights-lab.de.

〉 Mehr über unsere Aktivitäten und Themen erfahren Sie auf unserer Webseite www.zvki.de. Sie finden uns auch auf Instagram https://www.instagram.com/zvki.de/ und Twitter https://twitter.com/ZVKI_de.

MEHR INHALTE
Web: zvki.de
Instagram: zvki.de
Twitter: zvki_de
Linkedin: ZVKI

WOHER STAMMEN DIE INFORMATIONEN?

QUELLEN

Albert, Kendra/ Delano, Maggie: Sex trouble: Sex/gender slippage, sex confusion, and sex obsession in machine learning using electronic health records. 2022. In: Patterns, Volume 3, Issue 8. Online unter: https://www.cell.com/action/showPdf?pii=S2666-3899%2822%2900131-3 (letzter Aufruf: 20.02.2023).

Algorithmic Justice League 2022a: Bug Bounties for Algorithmic Harms? Lessons from cybersecurity vulnerability disclosure for algorithmic harms discovery, disclosure, and redress. 2022. Online unter: https://www.ajl.org/bugs (letzter Aufruf: 28.02.2023).

Algorithmic Justice League 2022b: Help Prevent, Report, and Redress Algorithmic Harms. 2022. Online unter: https://www.ajl.org/crash-project (letzter Aufruf: 28.02.2023).

AlgorithmWatch: Unding. O. D. Online unter: https://unding.de (letzter Aufruf: 20.02.2023).

AlgorithmWatch: Offener Brief: Jetzt algorithmenbasierte Diskriminierung anerkennen und Schutzlücken schließen! 2023. Online unter: https://algorithmwatch.org/de/offener-brief-diskriminierung-allgemeines-gleichbehandlungsgesetz/ (letzter Aufruf: 24.02.2023).

Antidiskriminierungsstelle des Bundes: Diskriminierungsformen. O. D. Online unter: https://www.antidiskriminierungsstelle.de/DE/ueber-diskriminierung/was-ist-diskri-

minierung/diskriminierungsformen/diskriminierungsformen-node.html (letzter Aufruf: 20.02.2023).

Antidiskriminierungsstelle des Bundes: AGG-Wegweiser. Erläuterungen und Beispiele zum Allgemeinen Gleichbehandlungsgesetz. 2022. Online unter: https://www.antidiskriminierungsstelle.de/SharedDocs/downloads/DE/publikationen/Wegweiser/agg_wegweiser_erlaeuterungen_beispiele.pdf?__blob=publicationFile&v=10 (letzter Aufruf: 20.02.2023).

Benjamin, Ruha/ McNealy, Jasmine: A New Jim Code? Berkman Klein Luncheon Series. 2019. Online unter: https://cyber.harvard.edu/events/new-jim-code (letzter Aufruf: 20.02.2023).

Berendt, Bettina: The AI Act Proposal: Towards the next transparency fallacy? 2022. In: Bundesministerium für Umwelt, Naturschutz, nukleare Sicherheit und Verbraucherschutz und Frauke Rostalski (Hg.): Künstliche Intelligenz. Wie gelingt eine vertrauenswürdige Verwendung in Deutschland und Europa?, S. 31–52. Online unter: https://www.mohrsiebeck.com/buch/kuenstliche-intelligenz-9783161612992 (letzter Aufruf: 20.02.2023).

Biddle, Sam: The Internet's New Favorite AI Proposes Torturing Iranians and Surveilling Mosques. In: The Intercept, 8. Dezember 2022. Online unter: https://theintercept.com/2022/12/08/openai-chatgpt-ai-bias-ethics/ (letzter Aufruf: 20.02.2023).

Birhane, Abeba/ Prabhu, Vinay Uday/ Kahembwe, Emmanuel: Multimodal datasets: misogyny, pornography, and malignant stereotypes. 2021. Online unter: https://arxiv.org/pdf/2110.01963.pdf (letzter Aufruf: 20.02.2023).

Birhane et. al.: Power to the People? Opportunities and Challenges for Participatory AI. 2022. In: EAAMO '22: Equity and Access in Algorithms, Mechanisms, and Optimization, S. 1–8. Online unter: https://arxiv.org/pdf/2209.07572v1.pdf (letzter Aufruf: 20.02.2023).

bitkom research: Die Menschen wollen KI – und haben auch Angst vor ihr. 2020. Online unter: https://www.bitkom-research.de/de/pressemitteilung/die-menschen-wollen-ki-und-haben-auch-angst-vor-ihr (letzter Aufruf: 20.02.2023).

Bundesministerium für Familie, Senioren, Frauen und Jugend: Dritter Gleichstellungsbericht der Bundesregierung. Digitalisierung geschlechtergerecht gestalten. 2021. Online unter: https://www.bmfsfj.de/resource/blob/184544/c0d592d2c37e7e2b5b4612379453e9f4/dritter-gleichstellungsbericht-bundestagsdrucksache-data.pdf (letzter Aufruf: 20.02.2023).

Buolamwini, Joy/ François, Camille/ Costanza-Chock, Sasha: Happy Hacker Summer Camp Season! A CRASH Project update, from the team at the Algorithmic Justice League. In: medium.com, 30. Juli 2021. Online unter: https://medium.com/@ajlunited/happy-hacker-summer-camp-season-e1f6fdaf7694 (letzter Aufruf: 28.02.2023).

Datenethikkommission der Bundesregierung: Gutachten der Datenethikkommission der Bundesregierung. 2019. Online unter: https://www.bmi.bund.de/SharedDocs/downloads/DE/publikationen/themen/it-digitalpolitik/gutachten-datenethikkommission.pdf?__blob=publicationFile&v=4 (letzter Aufruf: 20.02.2023).

Deutscher Bundestag: Kurzinformation. Bedeutung der Erwägungsgründe bei der Umsetzung von Richtlinien (am Beispiel der Richtlinie 2016/943). 2019. Online unter: https://www.bundestag.de/resource/blob/628192/23bfc89c0c4ffb15489850b0558ce23f/PE-6-020-19-pdf-data.pdf (letzter Aufruf: 20.02.2023).

Europäische Kommission: 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. 2021. Online unter: https://eur-lex.europa.eu/resource.html?uri=cellar:e0649735-a372-11eb-9585-01aa75ed71a1.0019.02/DOC_1&format=PDF (letzter Aufruf: 20.02.2023).

Europäische Kommission: Draft standardisation request to the European Standardisation Organisations in support of safe and trustworthy artificial intelligence. 2022. Online unter: https://ec.europa.eu/docsroom/documents/52376 (letzter Aufruf: 20.02.2023).

Europäische Kommission, Vertretung in Deutschland: Für vertrauenswürdige Künstliche Intelligenz: EU-Kommission legt weltweit ersten Rechtsrahmen vor. 2021. Online unter: https://germany.representation.ec.europa.eu/news/fur-vertrauenswurdige-kunstliche-intelligenz-eu-kommission-legt-weltweit-ersten-rechtsrahmen-vor-2021-04-21_de (letzter Aufruf: 20.02.2023).

Europäische Kommission, Vertretung in Deutschland: Künstliche Intelligenz (KI): Neue EU-Regeln zur Produkthaftung und harmonisierte Haftungsvorschriften. 2022. Online unter: https://germany.representation.ec.europa.eu/news/kunstliche-intelligenz-ki-neue-eu-regeln-zur-produkthaftung-und-harmonisierte-haftungsvorschriften-2022-09-28_de (letzter Aufruf: 24.02.2023).

Europarat: Ad hoc Committee on Artificial Intelligence (CAHAI). Feasibility Study. 2020. Online unter: https://rm.coe.int/cahai-23-2020-final-eng-feasibility-study-/1680a0c6da (letzter Aufruf: 20.02.2023).

European Union Agency for Fundamental Rights: Bias in Algortihms - Artificial Intelligence and Discrimination. 2022. Online unter: https://fra.europa.eu/sites/default/files/fra_uploads/fra-2022-bias-in-algorithms_en.pdf (letzter Aufruf: 20.02.2023).

Ferrer, Xavier et. al.: Bias and Discrimination in AI: a cross-disciplinary perspective. 2021. In: IEEE Technology und Society Magazine, Volume 40, Issue 2, S. 72-80. Online unter: https://arxiv.org/pdf/2008.07309.pdf (letzter Aufruf: 20.02.2023).

Gerdes, Anne: A participatory data-centric approach to AI Ethics by Design. 2022. In: Applied Artificial Intelligence Volume 36, Issue 1. Online unter: https://www.tandfonline.com/doi/full/10.1080/08839514.2021.2009222 (letzter Aufruf: 20.02.2023).

Google: Google Diversity Annual Report 2022. 2022. Online unter: https://static.googleusercontent.com/media/about.google/de//belonging/diversity-annual-report/2022/static/pdfs/google_2022_diversity_annual_report.pdf?cachebust=1093852 (letzter Aufruf: 28.02.2023).
Global Data: Meta: Workforce Diversity and Inclusion in 2022. 2022. Online unter: https://www.globaldata.com/data-insights/technology--media-and-telecom/meta-workforce-diversity-and-inclusion-2091197/ (letzter Aufruf: 20.02.2023).

Guijarro Santos, Victoria: It’s a match! Oder Rassismus? In: Digital Society Blog, Alexander von Humboldt Institut für Internet und Gesellschaft, 12. Juli 2021. Online unter: https://www.hiig.de/its-a-match-oder-rassismus/ (letzter Aufruf: 20.02.2023).

Guijarro Santos, Victoria: A Crack in the Algorithm's Facade. A Fundamental Rights Perspective on "Efficiency" and "Neutrality" Narratives of Algorithms. 2022. In: Sven Quadflieg/ Klaus Neuburg/ Simon Nestler (Hg.), (Dis)Obedience in Digital Societies, S. 194-225. Online unter: https://www.transcript-open.de/doi/10.14361/9783839457634-009 (letzter Aufruf: 20.02.2023).

Hacker, Philipp: The European AI Liability Directives - Critique of a Half-Hearted Approach and Lessons for the Future. 2022. Online unter: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4279796 (letzter Aufruf: 24.02.2023).

Hardesty, Larry: Study finds gender and skin-type bias in commercial artificial-intelligence systems. In: MIT News, 11. Februar 2018. Online unter: https://news.mit.edu/2018/study-finds-gender-skin-type-bias-artificial-intelligence-systems-0212 (letzter Aufruf: 20.02.2023).

Heikkilä, Melissa: The viral AI avatar app Lensa undressed me—without my consent. In: MIT Technology Review, 12. Dezember 2022. Online unter: https://www.technologyreview.com/2022/12/12/1064751/the-viral-ai-avatar-app-lensa-undressed-me-without-my-consent (letzter Aufruf: 20.02.2023).

Heinrich Böll Stiftung Bremen: Interventionen// Glossar. Dominanzgesellschaft. 2020. Online unter: https://boell-bremen.de/de/2020/11/06/interventionen-glossar (letzter Aufruf: 20.02.2023).

Informations- und Dokumentationszentrum Antirassismusarbeit e.V.: Glossareintrag Diskriminierung. O. D. Online unter: https://www.idaev.de/recherchetools/glossar?tx_dpnglossary_glossary%5Baction%5D=list&tx_dpnglossary_glossary%5Bcontroller%5D=Term&tx_dpnglossary_glossary%5BcurrentCharacter%5D=D&cHash=c4fb7b9faf3e5d1c20c3bd2870ad4ec4 (letzter Aufruf: 20.02.2023).

Kalogeropoulos, Elena/ Lammers, Anne/ Müller-Brehm, Jaana/ Puntschuh, Michael: Wegweiser Digitale Debatten. Teil 2: Algorithmenvermittelte Diskriminierung. 2020. Hg. v. Innovationsbüro des Bundesministeriums für Familie, Senioren, Frauen und Jugend. Online unter: https://www.bmfsfj.de/resource/blob/186300/961021829a491933cf24e8f06ff8018f/wegweiser-digitale-debatten-teil-2-data.pdf (letzter Aufruf: 20.02.2023).

Kenway, Josh: Bug Bounties for Algorithmic Harms? Lessons from cybersecurity vulnerability disclosure for algorithmic harms discovery, disclosure, and redress. 2022. Hg. v. Algorithmic Justice League. Online unter: https://www.ajl.org/bugs

(letzter Aufruf: 08.02.2023).

Kettemann, Matthias C.: UNESCO-Empfehlung zur Ethik künstlicher Intelligenz. Bedingungen zur Implementierung in Deutschland. 2021. Hg. v. Deutsche UNESCO-Kommission. Online unter: https://www.unesco.de/sites/default/files/03-2022/DUK_Broschuere_KI-Empfehlung_DS_web_final.pdf (letzter Aufruf: 20.02.2023).

Koch, Bernard et. al.: Reduced, Reused and Recycled: The Life of a Dataset in Machine Learning Research. 2021. In: Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks, Volume 1. Online unter: https://arxiv.org/pdf/2112.01716.pdf (letzter Aufruf: 20.02.2023).

Kucera, Gregor: Das Bauchgefühl eint Mensch und Maschine. In: Wiener Zeitung, 29. Januar 2023. Online unter: https://www.wienerzeitung.at/nachrichten/wirtschaft/international/2176136-Das-Bauchgefuehl-eint-Mensch-und-Maschine.html (letzter Aufruf: 20.02.2023).

Lopez, Paola: Diskriminierung durch Data Bias. Künstliche Intelligenz kann soziale Ungleichheiten verstärken. 2021. In: WZB Mitteilungen, Heft 171, S. 26-28. Online unter: https://bibliothek.wzb.eu/artikel/2021/f-23704.pdf (letzter Aufruf: 20.02.2023).

Market, Karla/ Ahouzi, Afrae/ Debus, Pascal: Fairness in Regression - Analysing a Job Candidates Ranking System. 2022. In: Daniel Demmler/ Daniel Krupka/ Hannes Federrath (Hg.): INFORMATIK 2022, Gesellschaft für Informatik, S. 1275-1285. Online unter: https://dl.gi.de/bitstream/handle/20.500.12116/39483/trustai_04.pdf?sequence=1&isAllowed=y (letzter Aufruf: 20.02.2023).

Masakhane. O.D. Online unter: https://www.masakhane.io (letzter Aufruf: 20.02.2023).

Miceli, Milagros et al.: Documenting Data Production Processes: A Participatory Approach for Data Work. 2022. In: Proceedings of the ACM on Human-Computer Interaction, Volume 6, Issue CSCW2, Art. 510. Online unter: https://arxiv.org/pdf/2207.04958.pdf (letzter Aufruf: 20.02.2023).

Miceli, Milagros/ Posada, Julian: The Data-Production Dispositif. 2022. In: Proceedings of the ACM on Human-Computer Interaction, Volume 6, Issue CSCW2, Art. 460. Online unter: https://arxiv.org/pdf/2205.11963.pdf (letzter Aufruf: 24.02.2023).

Michot, Sarah et. al.: Algorithmenbasierte Diskriminierung. Warum Antidiskriminierungsgesetze jetzt angepasst werden müssen Policy Brief #5 des Digital Autonomy Hubs. 2022. Online unter: https://digitalautonomy.net/fileadmin/PR/Digitalautonomy/PDF/DAH_Draft_Policy_Brief__5.pdf (letzter Aufruf: 24.02.2023).

Müller, Eduard: KI-Zertifizierung. Normungsroadmap 2.0: „Neue Möglichkeiten nutzen, ohne Risiken zu übersehen". In: Tagesspiegel Background, Digitalisierung & KI, 9. Dezember 2022. Online unter: https://background.tagesspiegel.de/digitalisierung/normungsroadmap-2-0-neue-moeglichkeiten-nutzen-ohne-risiken-zu-uebersehen (letzter Aufruf: 20.02.2023).

Muller, Catelijne: It's Human Rights Day: How AI impacts virtually all human rights. 2020. In: allai.nl, o. D. Online unter: https://allai.nl/its-human-rights-day-how-ai-impacts-virtually-all-human-rights/ (letzter Aufruf: 20.02.2023).
NDR Talk Show: KI-Expertin Kenza Ait Si Abbou Lyadini. 2022. Hg. v. ARD. Online unter: https://www.youtube.com/watch?v=QJqFQoK4MT8 (letzter Aufruf: 20.02.2023).

Nekoto et al.: Participatory Research for Low-resourced Machine Translation: A Case Study in African Languages. 2020. In: Findings of the Association for Computational Linguistics: EMNLP 2020, S. 2144-2160. Online unter: https://arxiv.org/abs/2010.02353 (letzter Aufruf: 20.02.2023).

Neue deutsche Medienmacher*innen e.V.: NdM-Glossar. Schwarz. o. D. Online unter: https://glossar.neuemedienmacher.de/glossar/prefix:s/ (letzter Aufruf: 20.02.2023).

Nonnecke, Brandie/ Dawson, Philip: Human Rights Impact Assessments for AI: Analysis and Recommentations. 2022. Online unter: https://www.accessnow.org/cms/assets/uploads/2022/11/Access-Now-Version-Human-Rights-Implications-of-Algorithmic-Impact-Assessments_-Priority-Recommendations-to-Guide-Effective-Development-and-Use.pdf (letzter Aufruf: 20.02.2023).

Orwat, Carsten: Diskriminierungsrisiken durch Verwendung von Algorithmen. 2019. Hg. v. Antidiskriminierungsstelle des Bundes. Online unter: https://www.antidiskriminierungsstelle.de/SharedDocs/downloads/DE/publikationen/Expertisen/studie_diskriminierungsrisiken_durch_verwendung_von_algorithmen.pdf?__blob=publicationFile&v=3 (letzter Aufruf: 20.02.2023).

Perrigo, Billy: Exclusive: OpenAI Used Kenyan Workers on Less Than $2 Per Hour to Make ChatGPT Less Toxic. In: Time.com, 18. Januar 2023. Online unter: https://time.com/6247678/openai-chatgpt-kenya-workers/ (letzter Aufruf: 24.02.2023).

Poretschkin, Maximilian et. al.: Leitfaden zur Gestaltung vertrauenswürdiger Künstlicher Intelligenz. KI-Prüfkatalog. 2021. Hg. v. Fraunhofer-Institut für Intelligente Analyse- und Informationssysteme IAIS. Online unter: https://www.iais.fraunhofer.de/content/dam/iais/fb/Kuenstliche_intelligenz/ki-pruefkatalog/202107_KI-Pruefkatalog.pdf (letzter Aufruf: 20.02.2023).

Puchelt, Jonas: Wer haftet eigentlich, wenn KI eines Tages den Menschen ersetzt? In: it-daily.net, 30. Januar 2023. Online unter: https://www.it-daily.net/it-management/digitalisierung/wer-haftet-eigentlich-wenn-ki-eines-tages-den-menschen-ersetzt (letzter Aufruf: 20.02.2023).

Rameil, Lukas: Hat Künstliche Intelligenz ein Bewusstsein? In: Südkurier, 9. Januar 2023. Online unter: https://www.suedkurier.de/ueberregional/rundblick/ki-technologie-hat-kuenstliche-intelligenz-ein-bewusstsein;art1373253,11425179 (letzter Aufruf: 20.02.2023).

Rentsch, Susanne: „Computer sagt nein“ - Gesellschaftliche Teilhabe und strukturelle Diskriminierung im Zeitalter Künstlicher Intelligenz. 2023. In: In: Andreas Wagener/ Carsten Stark (Hg.): Die Digitalisierung des Politischen. Theoretische und praktische Herausforderungen für die Demokratie, S. 23-44. Online unter: https://link.springer.com/chapter/10.1007/978-3-658-38268-1_2 (letzter Aufruf: 20.02.2023).

Schwartmann, Rolf: Wenn Maschinen die Macht übernehmen. In: Frankfurter Allgemeine Zeitung, 25. Januar 2023. Online unter: https://www.faz.net/einspruch/chatgpt-wenn-maschinen-die-macht-uebernehmen-18629187.html (letzter Aufruf: 20.02.2023).

Sperber, Sonja et. al.: Gender Data Gap and its impact on management science – Reflections from a European perspective. 2022. In: European Management Journal, Volume 41, Issue 1, S. 2-8. Online unter: https://reader.elsevier.com/reader/sd/pii/S0263237322001554?token=A7D7348A4E152B27E128ED6A55155D3646EAA579099C61DBD96022D8B164AAC3CB3F71EB261D6578064D705C836D3E77&originRegion=eu-west-1&originCreation=20230116103159 (letzter Aufruf: 20.02.2023).

The Alan Turing Institute: AI Standards Hub. O. D. Online unter: https://aistandardshub.org/ai-standards-search/ (letzter Aufruf: 20.02.2023).

Veale, Michael/ Zuiderveen Borgesius, Frederik: Demystifying the Draft EU Artificial Intelligence Act – Analysing the good, the bad, and the unclear elements of the proposed approach. 2021. In: Computer Law Review International, Band 22, Heft 4, S. 97-112. Online unter: https://www.degruyter.com/document/doi/10.9785/cri-2021-220402/html?lang=de (letzter Aufruf: 20.02.2023).

Wachter, Sandra/ Mittelstadt, Brent/ Russel, Chris: Bias Preservation in Machine Learning: The Legality of Fairness Metrics Under EU Non-Discrimination Law. 2021. In: West Virginia Law Review, Volume 123, Issue 3, S. 735-790. Online unter: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3792772 (letzter Aufruf: 20.02.2023).

Wahlster, Wolfgang/ Winterhalter, Christoph (Hg.): Deutsche Normierungsroadmap Künstliche Intelligenz. Ausgabe 2. 2022. Online unter: https://www.din.de/resource/blob/891106/57b7d46a1d2514a183a6ad2de89782ab/deutsche-normungsroadmap-kuenstliche-intelligenz-ausgabe-2--data.pdf (letzter Aufruf: 20.02.2023).

Weinberg, Lindsey: Rethinking Fairness: An Interdisciplinary Survey of Critiques of Hegemonic ML Fairness Approaches. 2022. In: Journal of Artificial Intelligence Research, Volume 74, S. 75-109. Online unter: https://www.jair.org/index.php/jair/article/view/13196/26797 (letzter Aufruf: 20.02.2023).

Werder, Karl/ Ramesh, Balasubramaniam/ Zhang, Rongen: Establishing Data Provenance for Responsible Artificial Intelligence Systems. 2022. In: ACM Transactions on Management Information Systems, Volume 13, Issue 2, Art. 22. Online unter: https://dl.acm.org/doi/abs/10.1145/3503488 (letzter Aufruf: 24.02.2023).

Wolfangel, Eva: Ausgebeutet, um die KI zu zähmen. In: Zeit Online, 20. Januar 2023. Online unter: https://www.zeit.de/zustimmung?url=https%3A%2F%2Fwww.zeit.de%2Fdigital%2F2023-01%2Fchatgpt-ki-training-arbeitsbedingungen-kenia (letzter Aufruf: 20.02.2023).

Zhang, Daniel et al.: Artificial Intelligence Index 2022 Annual Report. 2022. Hg. v. AI Index Steering Committee, Stanford Institute for Human-Centered AI. Online unter: https://aiindex.stanford.edu/wp-content/uploads/2022/03/2022-AI-Index-Report_Master.pdf (letzter Aufruf: 20.02.2023).

IMPRESSSUM

Herausgeber und inhaltlich Verantwortlicher i. S. d. § 55 Abs. 2 RStV:
Philipp Otto

iRights.Lab GmbH
Oranienstr. 185
D-10999 Berlin
Telefon: +49 (0)30 40 36 77 230
Fax: +49 (0)30 40 36 77 260
E-Mail: zvki@irights-lab.de

Geschäftsführer*in: Philipp Otto, Dr. Wiebke Glässer
Registergericht: Amtsgericht Berlin-Charlottenburg
Registernummer: HRB 185640 B
Finanzamt für Körperschaften II
USt-IdNr.: DE311181302

Projektleitung: Philipp Otto

Autor*innen: Dr. Gergana Baeva, Franziska Busse, Jaana Müller-Brehm, Tom Völkel

Chefredaktion: Jaana Müller-Brehm

Redaktion: Merlin Münch, Philipp Otto, Verena Till

Inhaltliche Mitarbeit: Michael Puntschuh, Paul Ritzka, Tom Völkel

Gestaltung und Illustration: Marta Ricci Design mit Shutterstock AI Generator

Lektorat: text|struktur

Alle Bilder dieser Ausgabe wurden mit einem KI-Bildgenerator erstellt. Dabei haben wir auf eine kostenpflichtige Anwendung zurückgegriffen, deren Anbieter angibt, die Trainingsdaten zu prüfen und ihre Verwendung zu entlohnen.
Es wurden zahlreiche Prompts ausprobiert, um zu diesen Ergebnissen zu kommen. Wenn Sie mehr über das Vorgehen und die genutzte Anwendung wissen möchten, kontaktieren Sie uns: zvki@irights-lab.de.

Die Online-Version von *Missing Link* und weitere Informationen zum Projekt *ZVKI* finden Sie unter: www.zvki.de.

Projektpartner*innen *ZVKI*: *iRights.Lab, Fraunhofer AISEC, Fraunhofer IAIS, Freie Universität Berlin*

Gefördert durch: *Bundesministerium für Umwelt, Naturschutz, nukleare Sicherheit und Verbraucherschutz (BMUV)*

Das Projekt *ZVKI* wird vom unabhängigen Think Tank *iRights.Lab* verantwortet und durchgeführt. Das *iRights.Lab* entwickelt Strategien und praktische Lösungen, um die Veränderungen in der digitalen Welt vorteilhaft zu gestalten. Wir unterstützen öffentliche Einrichtungen, Stiftungen, Unternehmen, Wissenschaft und Politik dabei, die Herausforderungen der Digitalisierung zu meistern und die vielschichtigen Potenziale effektiv und positiv zu nutzen.

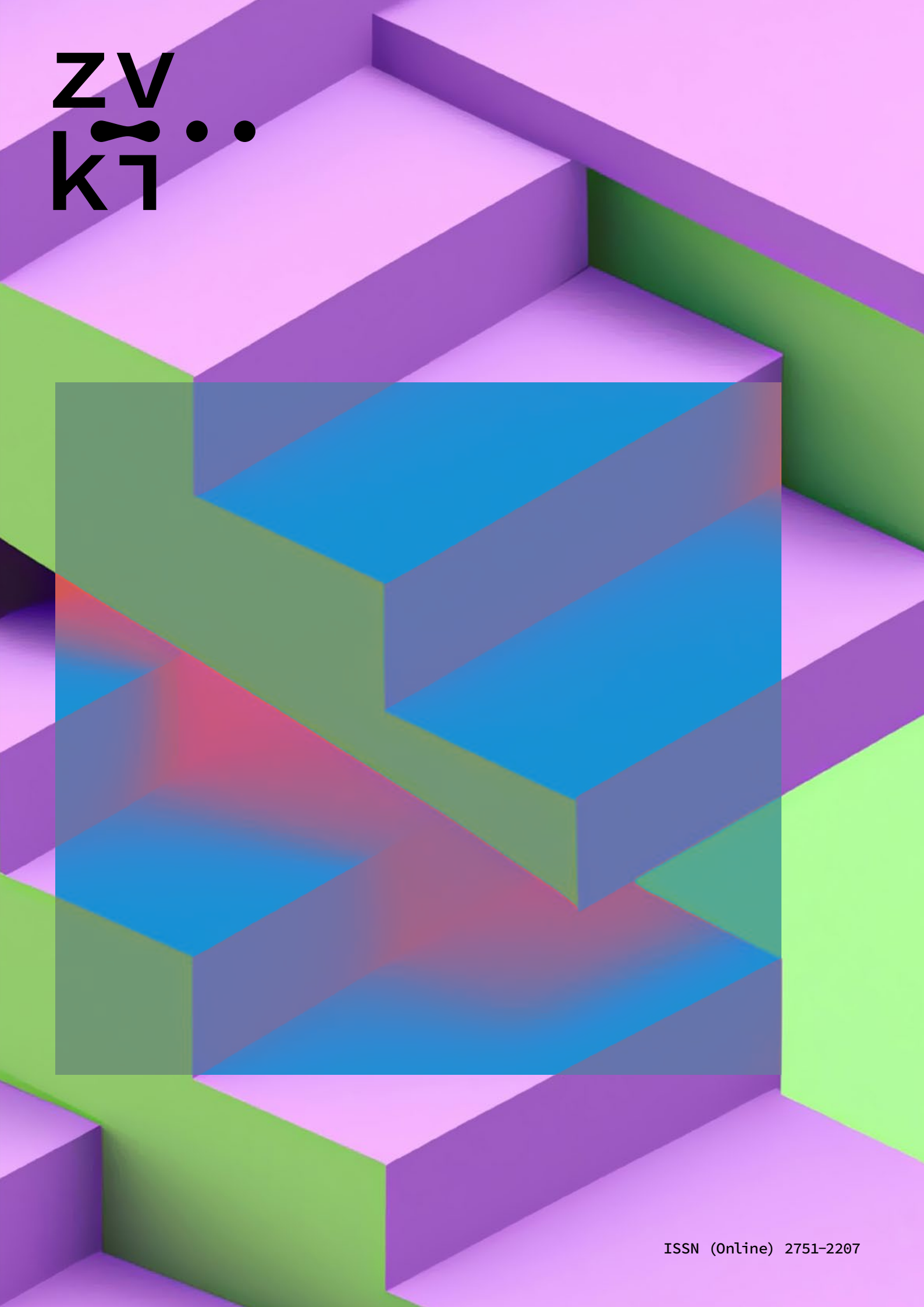
Weitere Informationen über das *iRights.Lab* finden Sie unter www.irights-lab.de.

Gefördert durch:

Bundesministerium
für Umwelt, Naturschutz, nukleare Sicherheit
und Verbraucherschutz

aufgrund eines Beschlusses
des Deutschen Bundestages

iRights.Lab
Think Tank für die digitale Welt



ISSN (Online) 2751-2207