www.jusletter-it.eu

Florent Thouvenin / Stephanie Volz / Soraya Weiner / Christoph Heitz

Diskriminierung beim Einsatz von Künstlicher Intelligenz (KI)

Technische Grundlagen für Rechtsanwendung und Rechtsentwicklung

Die zunehmende Verbreitung von Systemen der Künstlichen Intelligenz (KI) erhöht das Risiko der Diskriminierung von Menschen durch solche Systeme. Die Ursachen für Diskriminierung durch KI-Systeme sind vielfältig. Sie können insb. in der Definition von Modellstruktur und Trainingsverfahren und in nicht repräsentativen Trainingsdaten, aber auch in der Anwendung der Systeme liegen. Die Gefahr der Diskriminierung durch KI-Systeme lässt sich nur durch ein Zusammenwirken von rechtlichen und technischen Massnahmen verringern. Das erfordert eine vertiefte Zusammenarbeit von Informatiker:innen und Jurist:innen und ein gegenseitiges Verständnis der Begriffe und Konzepte. Der vorliegende Beitrag macht einen ersten Schritt, indem er Jurist:innen die technischen Grundlagen vermittelt, die für das Verständnis der Entstehung von Diskriminierung durch KI-Systeme erforderlich sind.

Beitragsart: Beiträge
Region: Schweiz, EU
Rechtsgebiete: Artificial Intelligence & Recht

Zitiervorschlag: Florent Thouvenin / Stephanie Volz / Soraya Weiner / Christoph Heitz, Diskriminierung beim Einsatz von Künstlicher Intelligenz (KI), in: Jusletter IT 4. Juli 2024

Inhaltsübersicht

I. Einleitung

[1] Die technische Entwicklung der letzten Jahre hat die Voraussetzungen für die Entwicklung und Anwendung von Systemen der sog. Künstlichen Intelligenz (KI-Systeme) geschaffen. Immer leistungsstärkere Computer, immer kostengünstigere Speichermöglichkeiten und die Verfügbarkeit enormer Datenmengen ermöglichen neue Analysen und automatisierte Entscheidun-

gen durch KI-Systeme in einer Vielzahl von Bereichen.[1] Diese Systeme sind längst Teil unserer Gesellschaft. Sie übernehmen Aufgaben und fällen Entscheidungen, die früher nur von Menschen erledigt bzw. getroffen werden konnten.[2] Mit dieser Entwicklung sind vielfältige Herausforderungen verbunden. Eine besonders wichtige ist die Gefahr, dass Menschen durch KI-Systeme diskriminiert werden.

[2] Dieser Beitrag richtet sich an Jurist:innen und andere Fachpersonen ohne Kenntnisse der Informatik. Er soll helfen, das Problem der Diskriminierung durch KI-Systeme besser zu verstehen. Zugleich bildet er die Grundlage für die Erarbeitung einer rechtlichen Lösung zur Erfassung von Diskriminierung durch KI-Systeme in der Schweiz.

II. Grundlagen

[3] Bereits bei der Definition von Diskriminierung können sich Schwierigkeiten ergeben, weil der Begriff in verschiedenen Bedeutungen gebraucht und in verschiedenen Disziplinen anders verstanden wird. Auch international bestehen grosse Unterschiede. In einem ersten Schritt sind deshalb die Begriffe zu klären.

A. Diskriminierung

1. Begriff

[4] Diskriminierung ist ein vielschichtiger Begriff, dem in verschiedenen wissenschaftlichen Disziplinen und Kontexten unterschiedliche Bedeutungen zukommen. Für den vorliegenden Beitrag ist entscheidend, wie der Begriff im Recht verstanden wird. Auch hier hat er verschiedene Bedeutungen.[3] Im Vordergrund steht der menschenrechtliche Gehalt,[4] der in der Bundesverfassung (Art. 8 Abs. 2 BV) zum Ausdruck kommt. Als Diskriminierung wird dort die Ungleichbehandlung von Personen in vergleichbaren Situationen verstanden, die an ein geschütztes Merkmal anknüpft und nicht qualifiziert gerechtfertigt ist.[5]

[5] Nach der Rechtsprechung des Bundesgerichts liegt eine Diskriminierung vor, «wenn eine Person ungleich behandelt wird allein aufgrund ihrer Zugehörigkeit zu einer bestimmten Gruppe, welche historisch oder in der gegenwärtigen sozialen Wirklichkeit tendenziell ausgegrenzt oder als minderwertig angesehen wird. Die Diskriminierung stellt eine qualifizierte Ungleichbe-

---

1 Mario Martini, Blackbox Algorithmus – Grundfragen einer Regulierung Künstlicher Intelligenz, Berlin 2019, S. 20; Carsten Orwat, Diskriminierungsrisiken durch Verwendung von Algorithmen, Antidiskriminierungsstelle des Bundes, Berlin 2019, S. 6 ff; ähnlich auch Brent Mittelstadt/Patrick Allo/Mariarosaria Taddeo/Sandra Wachter/Luciano Floridi, The ethics of algorithms: Mapping the debate, Big Data & Society 2016, S. 3.

2 Als Beispiele werden etwa Fliessbandarbeit, Kundendienst oder Hausarbeit genannt. Siehe dazu Yochanan E. Bigman et al., Algorithmic Discrimination Causes Less Moral Outrage Than Human Discrimination, Journal of Experimental Psychology 2022, S. 2 m.w.H.; Mittelstadt et al. (Fn. 1), S. 1, mit weiteren Beispielen.

3 Dazu Bernhard Waldmann, Das Diskriminierungsverbot von Art. 8 Abs. 2 BV als besonderer Gleichheitssatz, Bern 2003, S. 220.

4 Waldmann (Fn. 3), S. 220.

5 Rainer J. Schweizer/Kim Fankhauser, in Ehrenzeller/Egli/Hettich/Hongler/Schindler/Schmid/Schweizer (Hrsg.), Die schweizerische Bundesverfassung, St. Galler Kommentar, 4. Aufl., Zürich 2023, Art. 8 N 60; Bernhard Waldmann, in Waldmann/Belser/Epiney (Hrsg.), Bundesverfassung, Basler Kommentar, Basel 2015, Art. 8 N 58, N 87.

handlung von Personen in vergleichbaren Situationen dar, indem sie eine Benachteiligung von Menschen bewirkt, die als Herabwürdigung oder Ausgrenzung einzustufen ist, weil sie an Unterscheidungsmerkmalen anknüpft, die einen wesentlichen und nicht oder nur schwer aufgebbaren Bestandteil der Identität der betroffenen Personen ausmachen».[6]

[6] Die Definitionsmerkmale einer Diskriminierung sind damit (1) eine benachteiligende Ungleich- oder Gleichbehandlung in gleicher bzw. ungleicher Situation, die (2) an ein geschütztes Merkmal anknüpft und (3) das Fehlen einer qualifizierten Rechtfertigung.

a. Negative Ungleichbehandlung bzw. Gleichbehandlung

[7] Nicht jede Ungleichbehandlung ist eine Diskriminierung.[7] Erfasst werden Benachteiligungen, Herabwürdigungen, Stigmatisierungen und Ausgrenzungen in enger Verknüpfung mit der Menschenwürde.[8] Ungleichbehandlungen, die weder direkt noch indirekt auf geschützten Merkmalen basieren, werden nicht erfasst.[9] Um eine Diskriminierung zu bejahen, reicht es, dass Betroffene im Ergebnis diskriminiert werden. Eine Absicht ist nicht erforderlich.[10]

[8] Der im Recht verwendete Begriff der Diskriminierung hat stets einen negativen Gehalt.[11] Er ist abzugrenzen von der Diskriminierung im Sinn einer wertneutralen «Differenzierung»[12] oder «Ungleichbehandlung», wie er bspw. in der Informatik verstanden wird.[13]

b. Geschützte Merkmale

[9] Geschützte Merkmale sind persönliche Eigenschaften, die nicht oder nur schwer abgelegt werden können und die in der Vergangenheit Grund für Stigmatisierung und Herabwürdigung waren, bspw. Herkunft, Rasse, Geschlecht, Alter, Sprache, soziale Stellung, Lebensform, religiöse, weltanschauliche oder politische Überzeugung oder eine körperliche, geistige oder psychische Behinderung.[14] Die Bundesverfassung zählt einige geschützte Merkmale auf. Der Katalog ist aber

---

6 BGE 139 I 169, 174, E. 7.2.1; BSK BV-Waldmann (Fn. 5), Art. 8 N 60.

7 Tarkan Göksu, Rassendiskriminierung beim Vertragsabschluss als Persönlichkeitsverletzung, Freiburg 2003, Rz. 11, Rz. 15; SG BV-Schweizer/Fankhauser (Fn. 5), Art. 8 N 60.

8 Nadja Braun Binder et al., Künstliche Intelligenz: Handlungsbedarf im Schweizer Recht, in: Jusletter vom 28. Juni 2021, N 35; Giovanni Biaggini, in Biaggini (Hrsg.), Bundesverfassung der schweizerischen Eidgenossenschaft, Orell Füssli Kommentar, 2. Aufl., Zürich 2017, Art. 8 N 21; René Rhinow/Markus Schefer/Peter Uebersax, Schweizerisches Verfassungsrecht, 3. Aufl., Basel 2016, Rz. 1889; Jörg Paul Müller/Markus Schefer, Grundrechte in der Schweiz, Im Rahmen der Bundesverfassung, der EMRK und der UNO-Pakte, 4. Aufl., Bern 2008, S. 651, S. 684 ff.; SG BV-Schweizer/Fankhauser (Fn. 5), Art. 8 N 60 ff.; BSK BV-Waldmann (Fn. 5), Art. 8 N 47; Göksu (Fn. 9), Rz. 12.

9 Göksu (Fn. 7), N 11; Wildhaber Isabelle/Lohmann Melinda F./Kasper Gabriel, Diskriminierung durch Algorithmen – Überlegungen zum schweizerischen Recht am Beispiel prädiktiver Analytik am Arbeitsplatz, Zeitschrift für Schweizerisches Recht (ZSR), Band 138 (2019) I, Heft 5.

10 BGE 129 I 217, 227, E. 2.2.4; SG BV-Schweizer/Fankhauser (Fn. 5), Art. 8 N 61; BSK BV-Waldmann (Fn. 5), Art. 8 N 60; OFK BV-Biaggini (Fn. 8), Art. 8 N 18.

11 So etwa auch Orwat (Fn. 1), S. 24 Fn. 39.

12 Orwat (Fn. 1), S. 24 Fn. 39; siehe dazu die Diskussion in Susanne Beck et al., Künstliche Intelligenz und Diskriminierung, Whitepaper aus der Plattform Lernende Systeme, München 2019, S. 12.

13 Braun Binder et al. (Fn. 8), N 24.

14 Art. 8 Abs. 2 BV; BGE 147 I 73, 81, E. 6.1; BSK BV-Waldmann (Fn. 5), Art. 8 N 65 ff.

nicht abschliessend und die Auffassung, welche Merkmale geschützt sein sollen, kann sich über die Zeit verändern.[15]

c. Fehlende qualifizierte Rechtfertigung

[10] Eine Ungleichbehandlung in Anknüpfung an ein geschütztes Merkmal ist nicht per se unzulässig; sie kann gerechtfertigt sein, wenn sie sich sachlich begründen lässt und verhältnismässig ist.[16] Eine Ungleichbehandlung, die sich auf ein geschütztes Merkmal stützt, bedarf aber einer qualifizierten Rechtfertigung. Die Ungleichbehandlung muss dafür ein wichtiges und legitimes öffentliches Interesse verfolgen und zur Erreichung dieses Interesses geeignet und erforderlich sein, d.h. es darf kein anderes, milderes Mittel geben. Damit sie gerechtfertigt ist, muss die Ungleichbehandlung insgesamt verhältnismässig sein.[17] Dabei gelten für die verschiedenen geschützten Merkmale unterschiedlich strenge Anforderungen. Bei sozial zugeschriebenen Eigenschaften oder Rollenbildern (bspw. aufgrund der Herkunft) ist ein strenger Prüfungsmassstab anzuwenden und sachliche Gründe für eine Ungleichbehandlung werden kaum zu finden sein, während biologische Eigenschaften wie das Alter eine Ungleichbehandlung eher rechtfertigen können.[18]

2. Erscheinungsformen

[11] Diskriminierung kann in verschiedenen Formen auftreten. Eine direkte (auch: unmittelbare) Diskriminierung liegt vor, wenn direkt an ein geschütztes Merkmal angeknüpft wird. Von einer Diskriminierung durch Gleichbehandlung spricht man, wenn Fälle gleich behandelt werden, obwohl eine Differenzierung erforderlich wäre. Als indirekte Diskriminierung werden Konstellationen bezeichnet, bei denen Personen, die ein geschütztes Merkmal aufweisen, durch die tatsächlichen Auswirkungen einer an sich neutralen Regelung besonders stark benachteiligt werden. Bei algorithmischen Systemen besteht sodann die Gefahr von sog. Proxy-Diskriminierungen, die auftreten, wenn Systeme an Stellvertretern von geschützten Merkmalen anknüpfen.

a. Direkte Diskriminierung

[12] Eine direkte Diskriminierung liegt vor, wenn Menschen in vergleichbaren Situationen aufgrund eines geschützten Merkmals ungleich behandelt werden und diese Behandlung einen (oder mehrere) Menschen benachteiligt, ohne dass für die Ungleichbehandlung eine qualifizierte sach-

15 Früher galten auch Diskriminierungen aufgrund von Untertanenverhältnissen, der Geburt oder des Vorrechts des Ortes als unzulässig; siehe dazu SG BV-Schweizer/Fankhauser (Fn. 5), Art. 8 N 82; OFK BV-Biaggini (Fn. 8), Art. 8 N 24.

16 SG BV-Schweizer/Fankhauser (Fn. 5), Art. 8 N 59; BSK BV-Waldmann (Fn. 5), Art. 8 N 56 ff., N 87; OFK BV-Biaggini (Fn. 8), Art. 8 N 26.

17 BSK BV-Waldmann (Fn. 5), Art. 8 N 87.

18 Für eine abgestufte Prüfung: Waldmann (Fn. 3), S. 592; BSK BV-Waldmann (Fn. 5), Art. 8 N 87; OFK BV-Biaggini (Fn. 8), Art. 8 N 26; Müller/Schefer (Fn. 8), S. 691 ff.; für zwei Prüfungsstandards: Ulrich Häfelin/Walter Haller/Helen Keller/Daniela Thurnherr, Schweizerisches Bundesstaatsrecht, 10. Aufl., Zürich/Basel/Genf 2020, N 776; Anne Peters, in Merten/Papier (Hrsg.), Handbuch der Grundrechte in Deutschland und Europa VII/2, Heidelberg 2007, § 211 N 57, N 59; Vincent Martenet, Géométrie de l'égalité, Zürich/Basel/Genf 2003, Rz. 885 ff.

liche Rechtfertigung vorliegt.[19] Eine direkte Diskriminierung ist bspw. gegeben, wenn ein KI-System bei einer Kreditvergabe einem französischsprechenden Menschen wegen seiner Sprache keinen Kredit gewährt.

b. Diskriminierung durch Gleichbehandlung

[13] Auch eine Gleichbehandlung kann eine Diskriminierung darstellen, wenn sie zu einer Benachteiligung von Trägern qualifizierter Merkmale führt und keine sachlichen Gründe für die Gleichbehandlung sprechen.[20] Die Diskriminierung durch Gleichbehandlung wird teilweise auch als Form der indirekten Diskriminierung bezeichnet.[21] Eine Diskriminierung durch Gleichbehandlung liegt bspw. vor, wenn Personen mit einer Lese- und Schreibschwäche dieselbe Zeit für das Lösen einer Prüfung erhalten wie Personen ohne eine solche Beeinträchtigung.

c. Indirekte Diskriminierung

[14] Die Rechtsordnung kennt zudem die sog. indirekte (auch: mittelbare oder faktische[22]) Diskriminierung.[23] Eine solche liegt vor, wenn eine an sich neutrale Regelung, also eine Regelung, die nicht an ein geschütztes Merkmal anknüpft, in ihren tatsächlichen Auswirkungen Angehörige einer (spezifisch) gegen Diskriminierung geschützten Gruppe besonders stark benachteiligt, ohne dass dies (qualifiziert) sachlich begründet wäre.[24] Die indirekte Benachteiligung muss allerdings von erheblicher Bedeutung sein, weil das Verbot der indirekten Diskriminierung nach der Rechtsprechung des Bundesgerichts nur die offensichtlichsten negativen Auswirkungen korrigieren soll.[25]

[15] Relevant ist die indirekte Diskriminierung in der Praxis vor allem in Fällen, in denen aufgrund einer neutralen Regelung wesentlich mehr Angehörige eines Geschlechts gegenüber denjenigen des anderen benachteiligt werden, ohne dass dies sachlich begründet wäre.[26] Im Zusammenhang mit der Frage, ob Frauen benachteiligt seien, weil «Frauenberufe» schlechter entlöhnt würden, hat das Bundesgericht festgehalten, dass eine Funktion in der Regel als typisch weiblich gilt, wenn der Frauenanteil wesentlich höher als 70% liege.[27] Eine rechtlich relevante indirekte Diskriminierung dürfte deshalb nur in wenigen Fällen gegeben sein.

---

19 SG BV-Schweizer/Fankhauser (Fn. 5), Art. 8 N 61; BSK BV-Waldmann (Fn. 5), Art. 8 N 62.

20 BSK BV-Waldmann (Fn. 5), Art. 8 N 64; das Bundesgericht hat die Diskriminierung durch Gleichbehandlung verschiedentlich zur indirekten Diskriminierung gezählt: BGE 139 I 169, 174 f., E. 7.2.2; BGE 138 I 205, 213 f., E. 5.5; BGE 135 I 49, 55, E. 4.3.

21 BGE 139 I 169, 174, E. 7.2.2; BGE 135 I 49, 55, E. 4.3; BSK BV-Waldmann (Fn. 5), Art. 8 N 64.

22 OFK BV-Biaggini (Fn. 8), Art. 8 N 20.

23 BSK BV-Waldmann (Fn. 5), Art. 8 N 60 ff.; OFK BV-Biaggini (Fn. 8), Art. 8 N 18; BGE 147 I 73, 81 f., E. 6.1.; BGE 126 II 377, 393 f., E. 6.c); auch zur indirekten Diskriminierung: BGE 141 I 241, 250 f., E. 4.3.2.

24 BGE 145 I 73, 85 f., E. 5.1; BGE 129 I 217, 224, E. 2.1; SG BV-Schweizer/Fankhauser (Fn. 5), Art. 8 N 62; BSK BV-Waldmann (Fn. 5), Art. 8 N 63; OFK BV-Biaggini (Fn. 8), Art. 8 N 20.

25 BGE 145 I 73, 86, E. 5.1; BGE 138 I 265, 267 f., E. 4.2.2; BSK BV-Waldmann (Fn. 5), Art. 8 N 63, N 89; OFK BV-Biaggini (Fn. 8), Art. 8 N 20.

26 BGE 124 II 409, 424 f, E. 7.

27 BGE 125 II 530, 532, E. 2.b); BGE 125 II 385, 387, E. 3.b). Die Rechtsprechung zur indirekten Diskriminierung basiert auf Art. 8 Abs. 3 BV, der eine faktische Gleichheit der Geschlechter fordert. Geht es um andere Merkmale, könnte der verlangte prozentuale Anteil noch höher ausfallen.

d. Proxy-Diskriminierung

[16] Im Zusammenhang mit algorithmischen Systemen wird oft der Begriff der Proxy-Diskriminierung verwendet. Dabei handelt es sich allerdings nicht um einen Rechtsbegriff. Bei der Proxy-Diskriminierung hängt eine Entscheidung von scheinbar neutralen Daten ab, im Ergebnis werden aber Personen mit geschützten Merkmalen diskriminiert,[28] weil ein Proxy (Stellvertreter) mit einem geschützten Merkmal mehr oder weniger stark korreliert.[29]

[17] Proxy-Diskriminierung tritt meist unabsichtlich auf, weil es viele Merkmale gibt, die mit geschützten Merkmalen korrelieren. So sind bspw. Grösse, Gewicht oder das Auftreten von bestimmten Krankheiten oft geschlechtsspezifisch (bspw. Brust- oder Prostatakrebs). *Machine learning*-Algorithmen sind anfällig, Entscheidungen zu generieren, die stark mit einem geschützten Merkmal korrelieren, auch wenn dieses selbst gar nicht verwendet wird. Dies ist allerdings in vielen Fällen nicht offensichtlich, auch nicht für die Entwickler:innnen der Systeme.[30] Wird bei einer Entscheidung absichtlich auf einen Proxy (oder mehrere Proxies) abgestellt, spricht man von *masking*. Dabei wird ein unverdächtiger Stellvertreter benutzt, um eine bewusste Diskriminierung zu verschleiern.[31] Historisch kommt diese Form der Proxy-Diskriminierung vom sog. *redlining*, bei dem in den USA bestimmte geografische Regionen von den Dienstleistungen eines Unternehmens ausgeschlossen wurden, weil dort überwiegend Afroamerikaner lebten. Da eine Unterscheidung aufgrund der «Rasse» nicht erlaubt war, wurde an die Region als Stellvertretermerkmal angeknüpft.[32]

[18] Das Problem der Proxy-Diskriminierung zeigt, dass sich Diskriminierung in der Regel nicht verhindern lässt, indem die geschützten Merkmale aus den Datensätzen entfernt werden.[33] Denn in den sehr grossen Datensätzen (Big Data), die für das Trainieren dieser Systeme verwendet werden, sind meist genügend andere Merkmale enthalten, die mit einem geschützten Merkmal korrelieren und die in der Summe zu einer (ungewollten) Rekonstruktion geschützter Merkmale führen. Die Erfahrung hat sogar gezeigt, dass geschützte Merkmale in algorithmischen Systemen explizit mitberücksichtigt werden müssen, um Diskriminierung überhaupt erkennen und nicht-diskriminierende Modelle entwickeln zu können.[34]

---

28 Gabriele Buchholtz/Martin Scheffel-Kain, Algorithmen und Proxy Discrimination in der Verwaltung: Vorschläge zur Wahrung digitaler Gleichheit, NVwZ 2022, S. 612.

29 Zur Problematik der Korrelationen siehe Martini (Fn. 1), S. 60; Mittelstadt et al. (Fn. 1), S. 5.

30 Siehe dazu hinten, II.B.4.

31 Janneke Gerards/Raphaële Xenidis, Algorithmic discrimination in Europe: Challenges and opportunities for gender equality and non-discrimination law, European Commission, European network of legal experts in gender equality and non-discrimination, 2020, S. 44 m.w.H.; Solon Barocas/Andrew D. Selbst, Big Data's Disparate Impact, California Law Review 2016, 692 f.

32 Anya E.R. Prince/Daniel Schwarcz, Proxy Discrimination in the Age of Artificial Intelligence and Big Data, Iowa Law Review, 2020, S. 1262, S. 1268 f.; Buchholtz/Scheffel-Kain (Fn. 28), S. 613. Redlining ist in den USA mittlerweile verboten; siehe dazu Barocas/Selbst (Fn. 31), S. 690 m.w.H.

33 Jon Kleinberg et al., Discrimination in the Age of Algorithms, Journal of Legal Analysis 2018, S. 137.

34 Indrė Žliobaitė/Bart Custers, Using sensitive personal data may be necessary for avoiding discrimination in data-driven decision models, Artificial Intelligence Law 2016, S. 185, S. 194 ff.; Hoffmann Hanna et al, Fairness by awareness? On the inclusion of protected features in algorithmic decisions, Computer Law & Security Review 2022/44, S. 1 – 2; 11; Kleinberg et al. (Fn. 33), S. 5, S. 34.

3. Bias und Fairness

[19] «Bias» und «Fairness» werden in der Informatik verwendet, um ethische Probleme im Zusammenhang mit der Funktionsweise und den Ergebnissen von Algorithmen zu beschreiben.

[20] Der Begriff «*Bias*» (Verzerrung) ist nicht mit dem Begriff «Diskriminierung» gleichzusetzen. Bias kann eine wertende Bedeutung wie Voreingenommenheit oder Vorurteil haben, aber auch wertneutral als Verzerrung oder Abweichung von einem Standard verstanden werden.[35] Der Begriff «*Algorithmic Bias*» ist nicht eindeutig definiert. Häufig nimmt er Bezug auf geschützte Gruppen und wird als systematische Benachteiligung gewisser Gruppen durch einen Algorithmus verstanden.[36]

[21] Im Zusammenhang mit Diskriminierung durch algorithmische Systeme ist oft von «*(Algorithmic) Fairness*» die Rede. Dieser Begriff wird vor allem in der Informatik verwendet; es handelt sich nicht um einen Rechtsbegriff. Aus Sicht des Rechts ist problematisch, dass es keine allgemein anerkannte und hinreichend präzise Definition von «Fairness» gibt. Viele Autor:innen[37] lehnen sich jedoch an das Problem der Diskriminierung an und verstehen fair als «diskriminierungsfrei», wobei der Begriff «Diskriminierung» durch Algorithmen nicht der rechtlichen Diskriminierung entspricht und sich nicht unbedingt auf Ungleichbehandlungen aufgrund von geschützten Merkmalen beschränkt.[38]

[22] Der Begriff der Fairness wird unterschiedlich formalisiert. Unterscheiden lassen sich insb. Gruppenfairness (*group fairness*) und individuelle Fairness (*individual fairness*).[39] Gruppenfairness stellt auf statistische Parität ab, d.h. die Ergebnisse des Algorithmus sollen «im Durchschnitt» für die verschiedenen Gruppen gleich sein.[40] Ein prominentes Beispiel für Gruppenfairness ist die sog. *demographic parity*. Dabei prüft man, ob die Wahrscheinlichkeit für eine bestimmte Entscheidung des Algorithmus für verschiedene Bevölkerungsgruppen gleich ist. So kann man bspw. die Akzeptanzquote von Bewerbenden für verschiedene Geschlechter prüfen – ist diese unterschiedlich, wäre das ein Verstoss gegen die Vorgabe der *demographic parity*. Die individuelle Fairness, die sich am rechtlichen Diskriminierungsbegriff orientiert, bedeutet, dass Personen mit ähnlichen Eigenschaften (mit Ausnahme der geschützten Merkmale) von algorithmischen Systemen dieselben Ergebnisse erhalten.[41] Konkret auf die Diskriminierung bezogen ist die sog. kausale Fairness (*causal fairness*).[42]

---

35 Xavier Ferrer et al., Bias and Discrimination in AI: A Cross-Disciplinary Perspective, IEEE Technology and Society Magazine 2020, S. 72; David Danks/Alex John London, Algorithmic Bias in Autonomous Systems, IJCAI 2017, S. 4691 f., S. 4696 legen dar, dass «Biases» sogar vorteilhaft sein können, um ein bestimmtes Ziel zu erreichen. Siehe dazu auch Batya Friedman/Helen Nissenbaum, Bias in Computer Systems, ACM Transactions on Information Systems 1996, S. 332, die sich ebenfalls zur Wertneutralität des Begriffs äussern, aber auf das negative Verständnis fokussieren.

36 Rachel K.E. Bellamy et al., AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias, 2018, S. 2; siehe auch Gerards/Xenidis (Fn. 31), S. 47.

37 Janine Strotherm et al., Fairness in KI-Systemen, 2023, S. 5 f.; Ninareh Mehrabi et al., A Survey on Bias and Fairness in Machine Leaning, 2021, S. 11 f.

38 Gerards/Xenidis (Fn. 31), S. 48.

39 Siehe insb. Mehrabi et al. (Fn. 37), S. 13.

40 Philipp Hacker, Teaching Fairness to Artificial Intelligence: Existing and Novel Strategies against Algorithmic Discrimination under EU Law, Common Market Law Review 2018, S. 1175; Ferrer et al. (Fn. 35), S. 73 f.

41 Hacker (Fn. 40), S. 1175 ; Janine Strotherm/Alissa Müller/Barbara Hammer/Benjamin Paasse, Fairness in KI-Systemen, abrufbar unter «arXiv:2307.08486v1», S. 1 ff., S. 12 ff.

42 Strotherm/Müller/Hammer/Paasse (Fn 41), 14 ff.; siehe dazu auch Drago Plecko/Elias Bareinboim, Causal Fairness Analysis, abrufbar unter «arXiv:2207.11385 », S. 1 ff.

[23] Was als «fair» gilt, hängt oft vom Kontext und vom kulturellen Umfeld ab.[43] In der Informatik wurde eine Reihe von Fairnessmetriken entwickelt, die auf verschiedenen normativen Annahmen darüber beruhen, was Fairness bedeutet.[44] Die Fairnessmetriken schliessen sich jedoch teilweise gegenseitig aus. Es kann mathematisch gezeigt werden, dass sich verschiedene Metriken für Gruppenfairness widersprechen.[45] Zudem kann eine Entscheidung, die auf Gruppenebene fair ist, für ein Individuum dennoch diskriminierend sein. Welche Fairnessmetriken in welchem Kontext eingesetzt werden können und sollen, ist deshalb offen.[46] Für den vorliegenden Zusammenhang ist wichtig, dass «Fairness» nicht mit Nicht-Diskriminierung im Rechtssinn gleichgesetzt werden kann. Namentlich ist nicht alles, was als «unfair» erscheinen mag, als Diskriminierung im rechtlichen Sinn zu qualifizieren, denn die Konzepte der «Fairness» gehen in aller Regel weiter als die Vermeidung von Diskriminierung.[47]

B. Technische Begriffe

1. Künstliche Intelligenz (KI)

[24] Die Definition des Begriffs «Künstliche Intelligenz» (KI) oder «*Artificial Intelligence (AI)*» war lange umstritten.[48] Auch in den Entwürfen für eine Verordnung des europäischen Parlaments und des Rates zur Festlegung harmonisierter Vorschriften für künstliche Intelligenz (KI VO) fanden sich verschiedene Definitionen. Die finale Fassung[49] definiert ein KI-System in Art. 3 (1) KI VO als «ein maschinengestütztes System, das für einen in unterschiedlichem Grade autonomen Betrieb ausgelegt ist und das nach seiner Betriebsaufnahme anpassungsfähig sein kann und das aus den erhaltenen Eingaben für explizite oder implizite Ziele ableitet, wie Ausgaben wie etwa Vorhersagen, Inhalte, Empfehlungen oder Entscheidungen erstellt werden, die physische oder virtuelle Umgebungen beeinflussen können». Diese Definition lehnt sich eng an die von der OECD verwendete Definition an.[50] Sie scheint sich nun nach längeren Diskussionen allgemein durchzusetzen.

[25] Ein Merkmal von «Künstlicher Intelligenz» ist eine mehr oder weniger ausgeprägte maschinelle Autonomie. Maschinelle Autonomie bedeutet einerseits, dass der Algorithmus zumindest bis zu einem gewissen Grad unabhängig von menschlicher Kontrolle und Einflussnahme agiert.[51] Andererseits werden dem Algorithmus – anders als bei der herkömmlichen Programmierung und

43 Bellamy et al. (Fn. 36), S. 1; Mehrabi et al. (Fn. 37), S. 11.

44 Eine Zusammenfassung der wichtigsten Fairnessmetriken findet sich bei Dana Pessach/Erez Shmueli, Algorithmic Fairness, 2020, S. 3 ff.

45 Solon Barocas/Moritz Hardt/Arvind Narayanan, Fairness and Machine Learning: Limitations and Opportunities, 2023, S. 64 ff.; Strotherm et al. (Fn. 37), S. 16 f.; Pessach/Shmueli (Fn. 44), 5 ff.

46 Corinna Hertweck/Christoph Heitz, A Systematic Approach to Group Fairness in Automated Decision Making, in 2021 8th Swiss Conference on Data Science, 2021, S. 1.

47 Bezogen auf das EU-Recht siehe Gerards/Xenidis (Fn. 34), S. 47; Pessach/Shmueli (Fn. 44), S. 7 ff. Zum Begriff der Diskriminierung im Schweizer Recht siehe oben, II.A.1.a).

48 Orwat (Fn. 1), S. 8 Fn. 10; Frederik J. Zuiderveen Borgesius, Strengthening Legal Protection against Discrimination by Algorithms and Artificial Intelligence, The International Journal of Human Rights 2020, S. 1573.

49 Verordnung des Europäischen Parlaments und des Rates zur Festlegung harmonisierter Vorschriften für Künstliche Intelligenz und zur Änderung bestimmter Rechtsakte der Union («KI VO»).

50 OECD, Recommendation of the Council on Artificial Intelligence, OECD/LEGAL/0449.

51 Erwägungsgrund 6 der KI VO.

Automatisierung – keine Regeln vorgegeben; vielmehr leitet der Algorithmus die Regeln in einem Lernprozess selbst ab. Eine besonders häufige Form ist das Lernen anhand von Trainingsdaten: Dabei lernt ein KI-System, optimale Ergebnisse zu erzielen (bspw. eine optimale Entscheidung zu treffen), indem es mit Daten aus der Vergangenheit trainiert wird. So kann man bspw. ein medizinisches Diagnosesystem mit Input-Daten von Patient:innen (bspw. Alter, Geschlecht, Vorerkrankungen, physiologische Marker) trainieren, bei denen das Vorhandensein einer bestimmten Krankheit (Output) erfasst wurde. Das System lernt dann, aus den Input-Daten den Output bestmöglich abzuleiten, d.h. eine Aussage darüber zu treffen, ob die Person diese Krankheit mit einer bestimmten Wahrscheinlichkeit hat.

[26] Die Art und Weise, wie ein KI-System von den Input-Daten zu einem Resultat («hat die Krankheit» vs. «hat die Krankheit nicht») gelangt, ist aber nicht durch einen Programmierer vorgegeben, sondern wird durch einen Lernalgorithmus anhand der Trainingsdaten – innerhalb gewisser Grenzen der zugrundeliegenden algorithmischen Struktur (bspw. eines neuronales Netzes einer bestimmten Topologie) – optimiert. Ein so trainiertes System wird dann verwendet, um für neue Fälle eine Vorhersage zu treffen.

[27] Für Menschen ist oft nicht zu erkennen, basierend auf welcher Regel ein bestimmtes Resultat generiert wurde. Ein trainiertes KI-System hat zwar intern einen Regelsatz kodiert. Dieser ist aber häufig durch tausende interne Parameter bestimmt, die sich meist weder direkt interpretieren noch durch einfache Entscheidungsregeln (wenn-dann) repräsentieren lassen. Entsprechend lässt sich oft nicht nachvollziehen, wie ein Resultat entstand. Die Frage nach den «Entscheidgründen» eines selbstlernenden Systems ist deshalb im Grunde unsinnig – denn die Antwort darauf ist immer: «Weil es mit den Trainingsdaten gut funktioniert hat.» Im Ergebnis sind die Entscheidungen deshalb für Aussenstehende in der Regel intransparent.[52] Erschwerend kommt hinzu, dass gewisse selbstlernende Systeme dynamisch sind und sich aufgrund neu hinzukommender Daten stets weiterentwickeln.

2. Algorithmen und algorithmische Systeme

a. Begriff

[28] Der Begriff «Algorithmus» wird vor allem in der Informatik verwendet. Als Algorithmus wird dort ein Verfahren bezeichnet, das dazu dient, eine spezifische Aufgabe zu lösen.[53] Ein Algorithmus besteht aus einer Folge von Anweisungen oder Handlungen. Algorithmen werden in der Regel als Programmcode in einer bestimmten Programmiersprache in einem Computersystem implementiert – der Algorithmus bezeichnet aber nicht den Programmcode, sondern die Folge von logischen Schritten, mit denen eine Aufgabe gelöst wird.

[29] Algorithmen sind formale Konstrukte, die es ermöglichen, aufgrund eines Inputs über einen vordefinierten Prozess zu einem Output (Resultat) zu gelangen.[54] Jeder Algorithmus hat einen bestimmten «Zweck» (*purpose*), verstanden als die zu lösende Aufgabe, und er beruht auf «Bedin-

52 Das rasch wachsende Gebiet der «Explainable AI» versucht, Methoden zu entwickeln, um für Menschen die Entscheidungsstruktur eines algorithmischen Systems erklärbar zu machen. Das ist eine anspruchsvolle Aufgabe, die heute bei weitem noch nicht gelöst ist. Siehe dazu etwa die Toolkits XAITK (https://xaitk.org/) und AI Fairness 360 (https://aif360.res.ibm.com/) (beide zuletzt aufgerufen am 3. Juni 2024).

53 Siehe dazu Thomas Ottmann/Peter Widmayer, Algorithmen und Datenstrukturen, 6. Aufl., Berlin 2017, S. 1.

54 Robin K. Hill, What an Algorithm Is, Philosophy & Technology 2016, S. 47.

gungen» (*provisions*), die erfüllt sein müssen, damit er die Aufgabe lösen kann. Die Aufgabe wird dabei durch die Erzeugung eines konkreten Outputs gelöst.[55]

[30] Als Beispiel mag eine Website dienen, die eine Liste von ausgewählten Hotels in Paris anzeigt, die direkt auf dieser Website gebucht werden können. Ein Algorithmus kann hier den Zweck haben, für jede Besucher:in der Website im Voraus die personalisierte Wahrscheinlichkeit zu bestimmen, ein ganz bestimmtes Hotel aus der angezeigten Liste zu buchen (*purpose*). Der Output des Algorithmus besteht in der Angabe dieser Wahrscheinlichkeit. Wenn der Algorithmus für die Berechnung dieser Wahrscheinlichkeit Alter und Geschlecht der Besucherin verwendet, ist die Verfügbarkeit der Werte für Alter und Geschlecht eine Bedingung (*provision*), damit der Algorithmus die Aufgabe lösen kann, eine personalisierte Wahrscheinlichkeit zu bestimmen.

[31] Wenn im öffentlichen Diskurs von Algorithmen die Rede ist, geht es normalerweise nicht um Algorithmen im Sinn von Verfahren, sondern um die konkrete Implementierung eines Verfahrens in einem bestimmten Programm, einer Software oder einem Informationssystem.[56] Eine solche Implementierung nennen wir algorithmisches System.[57]

[32] Algorithmische Systeme können die menschliche Analyse und Entscheidungsfindung inklusive daraus folgende Handlungen ergänzen oder ersetzen. Ein selbstfahrendes Auto oder ein Roboter sind höchst komplexe algorithmische Systeme, die zur Problemlösung (bspw. einen Passagier von A nach B zu bringen) ständig Daten verarbeiten und Aktionen ausführen. Für die von einer Diskriminierung Betroffenen spielt es in der Regel keine Rolle, ob das System nur Empfehlungen abgegeben hat, die von einem Menschen übernommen wurden, oder ob es sich um ein vollautomatisches System handelt, das ohne menschliche Einwirkung entscheidet.[58] Auch Computersysteme mit einfachsten Algorithmen, die nur Schritt-für-Schritt-Anleitungen ausführen, sollten deshalb von der Definition erfasst sein.[59]

b. Deterministische und nicht-deterministische Algorithmen

[33] Bei Algorithmen sind zwei Arten zu unterscheiden: Deterministische Algorithmen sind statisch und folgen einem eindeutig vorgegebenen Lösungsweg, der auch bei der Verarbeitung einer potenziell unbegrenzten Anzahl von Werten immer gleich bleibt.[60] Der gesamte Prozess ist eindeutig bestimmt und der gleiche Input liefert immer den gleichen Output.[61] Die Wirkungsweise ist damit jederzeit reproduzierbar.

[34] Nicht-deterministische Algorithmen enthalten dagegen Elemente, die nicht von vornherein bestimmt sind, bspw. Zufallskomponenten oder Trainingsdaten, die mit darüber entscheiden, wie das algorithmische System später entscheidet. Zu den nicht-deterministischen Algorithmen ge-

55 Hill (Fn. 54), S. 46 f.

56 Mittelstadt et al. (Fn. 1), S. 2.

57 Ähnlich auch Gutachten der Datenethikkommission der Bundesregierung, Berlin 2019, S. 34.

58 Zuiderveen Borgesius (Fn. 48), S. 1573.

59 Sinngemäss auch Orwat (Fn. 1), S. 3 ff. Siehe dazu auch Joshua A. Kroll et al., Accountable Algorithms, University of Pennsylvania Law Review 2017, S. 640 Fn. 14 mit verschiedenen Definitionen für «Algorithmus».

60 Tim W. Dornis, in Zentek/Gerstein (Hrsg.), Designgesetz mit Gemeinschaftsgeschmacksmusterrecht, Handkommentar, Baden-Baden 2022, Kap. 4 Künstliche Intelligenz und Design N 5; Martini (Fn. 1), S. 19.

61 Wolfgang Ertel, Grundkurs Künstliche Intelligenz, Eine praxisorientierte Einführung, 5. Aufl., Wiesbaden 2021, S. 22.

hören insb. «lernende» Algorithmen, die dynamisch sind und sich weiterentwickeln können.[62] Diese Algorithmen beschränken sich nicht auf die Befehle im Programmcode, sondern entwickeln sich durch den Einfluss von nicht von vornherein bestimmten Elementen weiter.[63] Da ein KI-System nach der Definition in der KI VO[64] anpassungsfähig mit einer gewissen Autonomie Schlüsse aus dem Input zieht, die nicht vollständig vorhergesagt werden können, beinhalten KI-Systeme nicht-deterministische Algorithmen.

3. Wichtigste Formen von Algorithmen

a. Regelbasierte Algorithmen und Expertensysteme

[35] Regelbasierte (*rule-* oder *knowledge-based*) Algorithmen folgen den ihnen vorgegebenen Regeln. Es handelt sich dabei um Wenn-Dann-Prozesse, die durch Entscheidungsbäume (*decision trees*) abgebildet werden können. Regelbasierte algorithmische Systeme können menschliche Entscheidungsfindungen abbilden.[65] Diese Systeme sind zwar nicht darauf ausgerichtet, selbst zu «lernen», sie sind aber nicht mit deterministischen Algorithmen gleichzusetzen, weil sie auch nicht-deterministische Elemente enthalten können, etwa wenn Expertensysteme Wahrscheinlichkeiten verwenden.

[36] Regelbasierte Algorithmen basieren auf vorhandenem explizitem Wissen (daraus leitet sich auch der Name «Expertensystem» ab) und implementieren dieses in Form eines Computercodes. Sie sind so gut wie das zugrundliegende Wissen.

b. Datenbasierte Algorithmen und neuronale Netze

[37] Im Gegensatz zu regelbasierten Systemen lernen datenbasierte Algorithmen aus Beispielen und funktionieren über das Erkennen von statistischen Zusammenhängen und Mustern in Daten. Das «Lernen aus Daten» reicht von relativ einfachen statistischen Modellen (bspw. Regressionsmodellen), die Zusammenhänge zwischen Variablen approximativ abbilden, bis zu komplexen neuronalen Netzen. In allen Fällen werden Zusammenhänge zwischen Variablen, die für eine Entscheidung nützlich sind, aus Daten abgeleitet. Der entscheidende Unterschied zu den Expertensystemen ist der Lernprozess. Datenbasierte Algorithmen können im Ergebnis nicht besser sein als die Qualität und Repräsentativität der zugrundeliegenden Trainingsdaten.

[38] Neuronale Netze sind computerinterne Darstellungen von Verknüpfungen zwischen Input-Variablen (bspw. verfügbare Daten) und Output-Variablen (bspw. eine Entscheidung). Sie sind dem Funktionsmechanismus des menschlichen Gehirns nachempfunden und versuchen, das Zusammenwirken von Neuronen im Gehirn zu replizieren.[66] Das «Lernen» eines neuronalen Netzes besteht in der Optimierung der Verknüpfungsstärke zwischen den Neuronen, die in verschiedenen Schichten (*layers*) organisiert sind. Diese Gewichtungen werden beim «Lernen» laufend an-

62 Martini (Fn. 1), S. 19 f., S. 43; Dornis (Fn. 60), N 6; Ertel (Fn. 61), S. 22.

63 Martini (Fn. 1), S. 19 ff., S. 43; Dornis (Fn. 60), N 6.

64 Siehe dazu vorne, II.B.1.

65 Orwat (Fn. 1), S. 4 f.; Gerards/Xenidis (Fn. 31), S. 32 mit dem Beispiel für einen Wenn-Dann-Prozess, wonach bei einer Geschwindigkeitsübertretung eine Busse ausgestellt wird.

66 Anne Lauscher/Sarah Legner, Künstliche Intelligenz und Diskriminierung, ZfDR 2022, S. 370; Johannes Scherk/Gerlinde Pöchhacker-Tröscher/Karina Wagner, Künstliche Intelligenz – Artificial Intelligence, 2017, S. 16 f.

gepasst, wodurch das neuronale Netz bessere Lösungen finden kann.[67] Damit können verschiedenste Zusammenhänge zwischen Input und Output abgebildet werden – und dies besser als mit klassischen statistischen Modellen. *Deep learning* bezeichnet ein Verfahren des maschinellen Lernens, bei dem neuronale Netze mit vielen Layers verwendet werden. Damit kann die Leistungsfähigkeit von neuronalen Netzen weiter gesteigert werden.[68]

4. Entwicklung von KI-Systemen

[39] Die Begriffe «Künstliche Intelligenz» und «*machine learning*» (Maschinelles Lernen) werden bisweilen synonym verwendet.[69] *Machine learning* bezeichnet aber eigentlich die technische Methode, um ein KI-System mittels Daten zu trainieren, also um das «Lernen» eines KI-Systems zu ermöglichen.[70] *Machine learning* ist derzeit der wichtigste Ansatz im Bereich der KI[71] und wird bspw. bei Gesichtserkennungssystemen, bei Sprachassistenten wie *Siri* und *Google Assistant* oder bei Übersetzungs- und Korrekturprogrammen sowie bei sog. LLMs (*large language models*) wie bspw. ChatGPT eingesetzt.[72]

[40] *Machine learning* basiert auf der Verarbeitung sehr grosser Mengen von Daten, namentlich Trainings-, Validierungs- und Testdaten. Die Trainings- und Validierungsdaten dienen dem Training des Algorithmus, die Testdaten der Evaluation und Überprüfung des trainierten Algorithmus. Die durch das Training erlernten Entscheidungsregeln von machine learning-Algorithmen sind für Menschen meist weder verständlich noch nachvollziehbar und damit intransparent, weshalb auf maschinellem Lernen beruhende algorithmische Systeme bisweilen als *black boxes* bezeichnet werden.[73] Das gilt selbst für relativ einfache Algorithmen, deren Wirkungsweise für Nicht-Experten oft nur schwer zu verstehen ist.

[41] Bei der Entwicklung von KI-Systemen, die auf *machine learning* beruhen, lassen sich drei Phasen unterscheiden: die Definition einer Modellstruktur und eines Trainingsverfahrens, das Trainieren und Validieren sowie das Testen.

a. Definition von Modellstruktur und Trainingsverfahren

[42] Der erste Schritt ist die Definition einer Modellstruktur für die Verknüpfung des Inputs mit dem Output und die Implementierung eines Trainingsverfahrens. Input und Output sind dabei durch eine komplexe, nicht-lineare mathematische Funktion verknüpft, welche die Struktur des Modells bildet. Die Parameter dieser Funktion (bei neuronalen Netzen können es hunderte, tausende oder mehr Parameter sein) werden durch das Training des Modells angepasst. Bei einem deep learning-Algorithmus besteht die Modellstruktur aus mehreren Schichten (*layers*), die aufgrund einer bestimmten Architektur miteinander verknüpft sind.

---

67 Martini (Fn. 1), S. 24 f.

68 Gerards/Xenidis (Fn. 31), S. 36 m.w.H.

69 Zuiderveen Borgesius (Fn. 48), S. 1574 m.w.H.

70 Orwat (Fn. 1), S. 8; Ertel (Fn. 61), S. 3, S. 201.

71 Harry Surden, Artificial Intelligence and Law: An Overview, Georgia State University Law Review 2019, S. 1315.

72 Zu Sprachassistenten siehe https://support.google.com/assistant/answer/11140942?hl=en und https://machinelearning.apple.com/research/siri-voices; zu ChatGPT siehe https://help.openai.com/en/articles/7842364-how-chatgpt-and-our-language-models-are-developed (alle zuletzt aufgerufen am 3. Juni 2024).

73 Mittelstadt et al. (Fn. 1); siehe auch Martini (Fn. 1), S. 28f.; Orwat (Fn. 1), S. 97.

[43] Das Trainingsverfahren bestimmt, wie die Modellparameter aufgrund der Trainingsdaten angepasst werden, sodass der Algorithmus eine optimale Performance erreicht. lm Rahmen der Programmierung des Trainingsverfahrens werden auch weitere Parameter (sog. Hyperparameter) festgelegt, bspw. Abbruchkriterien für den Trainingsprozess.

b. Trainieren und Validieren

[44] Beim Training lernt das System aus den Trainingsdaten, (eine) bestimmte (bei sog. *general purpose AI*-Modellen auch viele unbestimmte) Aufgabe(n) zu erfüllen. Ziel des Trainings ist es, die Beziehung zwischen Input (verfügbare Daten während des späteren Einsatzes) und Output (bpsw. gute Vorhersage, korrekte Empfehlung, richtige Entscheidung) möglichst optimal abzubilden.

[45] Das Resultat dieses Lernprozesses wird «Modell» genannt. Denn das Modell repräsentiert eine (approximative) Abbildung der Realität, wie sie sich in den Trainingsdaten in der Beziehung zwischen Input und Output zeigt. Grundlegend für alle Verfahren des maschinellen Lernens ist, dass die Systeme mit genügend Trainingsdaten «gefüttert» werden, sodass sie Korrelationen erkennen und berücksichtigen können.[74] Der «Lernprozess» ist beim maschinellen Lernen darauf beschränkt, Korrelationen zu entdecken und die Beziehung zwischen Input und Output möglichst gut abzubilden. Damit ist aber keine Aussage über eine Kausalität verbunden. Für menschliche Anwender besteht die Gefahr, dass auftretende Korrelationen mit Kausalitäten verwechselt werden.[75]

[46] Als einfaches Beispiel kann ein System dienen, das aufgrund der Informationen in E-Mails Spam-E-Mails erkennen soll. Input ist hier die E-Mail selbst (inklusive der Meta-Information, bspw. im Header), Output ist eine Variable Y, die im einfachsten Fall zwei Werte annehmen kann (binäre *Prediction*): Y=0 bedeutet: E-Mail ist kein Spam. Y=1 bedeutet: E-Mail ist Spam. Häufig setzt man auch probabilistische Modelle ein, bei denen der Output einen Score zwischen 0 und 1 liefert, wobei bspw. s=0.3 bedeutet, dass die E-Mail mit einer Wahrscheinlichkeit von 30% Spam ist.

[47] Das Modell, das nach dem Training vorliegt, gibt für jeden beliebigen Input einen Output aus. Ziel des Trainings ist die Entwicklung eines Modells, das möglichst in jedem Fall eine korrekte Aussage Y produziert. Das gelingt in der Regel nicht vollständig: Jedes trainierte Modell produziert gelegentlich auch falsche Aussagen. Je häufiger das Modell die richtige Aussage macht, desto besser ist es.

[48] Modelle können mit unterschiedlichen Trainingsformen trainiert werden, namentlich durch *supervised learning* (überwachtes Lernen), *semi-supervised learning* (teilweise überwachtes Lernen), *unsupervised learning* (nicht überwachtes Lernen) und *reinforcement learning* (bestärkendes oder verstärkendes Lernen)[76]. Bei allen diesen datenbasierten Lernmethoden besteht das Pro-

74 David Lehr/Paul Ohm, Playing with the Data: What Legal Scholars Should Learn About Machine Learning, UC Davis Law Review, S. 671 m.w.H. Siehe auch Orwat (Fn. 1), S. 8; Zuiderveen Borgesius (Fn. 48), S. 1574.

75 Siehe dazu hinten, III.B.

76 Alfred Früh/Dario Haux, Foundations of Artificial Intelligence and Machine Learning, 2022, S. 15 ff. mit weiteren Ausführungen.

blem, dass sie auf bestehenden Daten beruhen und die trainierten Modelle deshalb schnell veralten können, bspw. wenn sich die Umstände ändern.[77]

[49] Beim *supervised learning* lernt das Modell mit Daten, bei denen das Resultat (Output) bekannt ist: Für jeden Trainingsdatensatz sind sowohl die Input-Daten als auch der Output (oft «Label» genannt) gegeben. Die Labels wären zum Beispiel «Spam» und «kein Spam». Die Labels definieren die sog. *ground truth*; damit wird der Wert bezeichnet, den das Modell idealerweise reproduzieren soll. Dabei handelt es sich nicht zwingend um die objektive «Wahrheit»; vielmehr liegen den Labels oft menschliche Wertungen zugrunde.[78] Durch die gelabelten Trainingsdaten soll der Algorithmus Beziehungen, insb. Korrelationen und Muster, in den Daten erkennen und dadurch in die Lage versetzt werden, neue und ungelabelte Daten zu verarbeiten, d.h. für diese den richtigen Output zu berechnen. Im genannten Beispiel würde der Algorithmus zu erkennen versuchen, welche Attribute tendenziell eher in Spam-E-Mails vorkommen (bspw. unbekannter Absender oder der Begriff «Gewinn»). Diese Attribute werden durch das Modell in optimaler Weise so miteinander verrechnet, dass die Treffsicherheit maximiert wird.

[50] Beim *unsupervised learning* entwickelt der Algorithmus gestützt auf nicht *gelabelte* Daten selbständig ein Modell, unter Ausnutzung von Strukturen in den Daten und Ähnlichkeiten zwischen Datensätzen. Wegen des Fehlens einer *ground truth* ist die Nachvollziehbarkeit der Resultate geringer als beim *supervised learning*.[79] *Unsupervised learning* eignet sich insb., um aus grossen Datenmengen bislang unbekannte Erkenntnisse zu gewinnen. Da dieses Lernverfahren keine gelabelten Daten benötigt, eignet es sich gut für einen kontinuierlichen Einsatz, weil jeder neue Datensatz als Lerndatensatz verwendet werden kann. Dieser Ansatz ist damit flexibler und dynamischer.[80]

[51] Beim *semi-supervised learning* werden sowohl *gelabelte* als auch *ungelabelte* Daten ins System eingegeben, was Kosten und Zeit spart. Zudem kann das Zugeben von ungelabelten Daten zur Optimierung des Modells beitragen.[81]

[52] *Reinforcement learning* bezeichnet ein Lernverfahren, bei dem ein algorithmisches System mittels Feedback zu immer besseren Lösungen gelangt. Der Algorithmus erhält ein Ziel und lernt durch *trial-and-error*.[82] Innerhalb einer definierten Umgebung – bspw. in einem Unternehmen – trifft das System laufend Entscheidungen – bspw. die Auswahl geeigneter Investitionsobjekte oder die Einstellung von Bewerbern. Trifft das System erfolgreiche Entscheidungen, erhält es positives Feedback, bspw. von menschlichen Begutachtern, bei nicht erfolgreichen negatives. Aufgrund dieser Information adaptiert das System sein Verhalten, um so die Qualität der Entscheidung

77 Gerards/Xenidis (Fn. 31), S. 40.

78 Jakob Feldkamp/Quirin Kappler/Maximilian Poretschkin/Anna Schmitz/Erik Weiss, Rechtliche Fairnessanforderungen an KI-Systeme und ihre technische Evaluation – Eine Analyse anhand ausgewählter Kreditscoring-Systeme unter besonderer Berücksichtigung der zukünftigen europäischen KI-Verordnung, ZfDR 2024, S. 60 ff., S. 104; Meike Zehlike/Alex Loosle/Hakan Jonsson/Emil Wiedemann/Philipp Hacker, Beyond Incompatibility: Trade-offs between Mutually Exclusive Fairness Criteria in Machine Learning and Law, abrufbar unter «arXiv:2212.00469v3», S. 5, 26; ähnlich Zuiderveen Borgesius (Fn 49), S. 16.

79 Gerards/Xenidis (Fn. 31), S. 35; Früh/Haux (Fn. 76), S. 11.

80 Gerards/Xenidis (Fn. 31), S. 38 m.w.H.

81 Früh/Haux (Fn. 31), S. 16 m.w.H.

82 Johannes Kevekordes, in Hoeren/Sieber/Holznagel (Hrsg.), Handbuch Multimedia-Recht, 60. Aufl., München 2024, Teil 29.1 KI im Überblick N 21.

schrittweise zu verbessern. Welche Daten als Trainingsdaten genutzt werden und wie das auf die Entscheidungen folgende Feedback erfolgt, spielt eine bedeutende Rolle.[83]

[53] Teil des Trainings ist das *Validieren* zur Prüfung der Qualität des Modells. Für das Validieren können separate Datensätze verwendet werden, die Validierungsdaten können aber auch Teil der Trainingsdaten sein. Mit dem Validieren wird die Leistung des Modells während des Trainings überprüft, um eine Überanpassung (*overfitting*) zu verhindern[84]. Eine Überanpassung liegt vor, wenn das Modell zwar für die Trainingsdaten gute Ergebnisse liefert, aber nicht für neue Daten, wenn das Modell also nicht generalisieren kann.

c. Testen

[54] Nachdem das Modell trainiert wurde, wird es mit Testdaten überprüft. Als Testdaten werden Daten verwendet, die dieselbe Struktur wie die Trainingsdaten aufweisen, aber nicht für das Trainieren oder Validieren benutzt wurden. Typischerweise werden dabei Input-Daten verwendet, bei denen der richtige Output bekannt ist. Bei Eingabe der Input-Daten generiert das Modell aufgrund der im Training etablierten Beziehung zwischen Input und Output für jeden Input einen Output. Durch den Vergleich dieses Outputs mit dem richtigen Output kann die Performance das trainierten Modells geprüft werden.

[55] Wenn die Ergebnisse der Tests zufriedenstellend sind, kann der Algorithmus eingesetzt werden. In vielen Fällen entwickelt sich der Algorithmus danach nicht mehr weiter, er lernt also nicht aus neuen Input-Daten. In diesen Fällen wird ein einmal trainiertes Modell so, wie es ist, in der Praxis eingesetzt, und nur gelegentlich einem Update unterzogen. In anderen Fällen werden *machine learning*-Algorithmen so verwendet, dass sie während des Einsatzes dazulernen und sich kontinuierlich weiterentwickeln.[85]

5. Anwendung von KI-Systemen

[56] Das fertig entwickelte Modell wird in der Regel verwendet, um Entscheidungen zu fällen oder Prognosen zu erstellen. In dieser Phase geht es nicht mehr um die Entwicklung, sondern um die Anwendung des Systems. Die Phasen der Entwicklung und Anwendung können sich allerdings überlappen, wenn ein KI-System laufend oder regelmässig neu trainiert wird, bspw. unter Verwendung weiterer (allenfalls bei der Anwendung des Systems entstandener) Daten.

[57] Beim fertig entwickelten Modell ist zwischen Modellqualität und Diskriminierungsfreiheit als unterschiedliche Eigenschaften zu unterscheiden.[86] Es ist durchaus möglich, dass ein qualitativ optimales Modell zu Diskriminierung führt, bspw. weil es mit Daten trainiert wurde, die eine

---

83 Früh/Haux (Fn. 76), S. 18; diese Trainingsmethode wurde teilweise auch bei ChatGPT angewendet; siehe dazu https://www.technologyreview.com/2023/03/03/1069311/inside-story-oral-history-how-chatgpt-built-openai/; https://help.openai.com/en/articles/6783457-what-is-chatgpt (beide zuletzt aufgerufen am 13. Mai 2024).

84 Siehe Art. 3 Bst. 30 KIVO.

85 So etwa beim sog. *online learning* im Gegensatz zum *batch learning*; siehe dazu Steven C.H. Hoi et al., Online learning: A Comprehensive Survey, SMU Technical Report 2018, S. 2; Eva Julia Lohse, Regulierung von Künstlicher Intelligenz, in Chibanguza/Kuss/Steege (Hrsg.), Künstliche Intelligenz, Recht und Praxis automatisierter und autonomer Systeme, München 2021, N 9.

86 Orwat (Fn. 1), S. 101.

diskriminierende Situation spiegeln. Umgekehrt kann ein Modell von geringer Qualität diskriminierungsfrei sein.

III. Problemstellung

A. Gefahr und Nutzen von KI

[58] KI-Systeme sind zwar anfällig für diskriminierende Entscheide, sie können aber auch dazu beitragen, Diskriminierung zu verringern oder zu verhindern.[87] Zum einen können sie helfen, von Menschen verursachte Diskriminierung zu entdecken,[88] vor allem aber ist es ungleich einfacher, «Biases» in algorithmischen Systemen zu bereinigen, als die Voreingenommenheit und diskriminierenden Neigungen von Menschen zu überwinden.[89]

B. Korrelationen

[59] Algorithmische Systeme lernen durch das Erkennen von Mustern; sie sind darauf trainiert, in Daten Korrelationen zu erkennen.[90] Als Korrelation wird die Beziehung zwischen zwei oder mehreren Variablen bezeichnet.[91] Es handelt sich dabei um ein statistisches Mass, das ausdrückt, wie stark die Beziehung zwischen diesen Variablen ist. Datengetriebene Algorithmen sind in der Lage, Korrelationen in grösserem Mass zu erkennen, als dies bislang möglich war.[92] Auf maschinellem Lernen beruhende Algorithmen können aber keine Kausalitäten erkennen.[93]

[60] Das juristische Denken ist hingegen von der Suche nach Kausalitäten geprägt, verstanden als direkter Zusammenhang von Ursache und Wirkung. Wer bspw. von einer Person Schadenersatz verlangt, muss nachweisen können, dass das Verhalten der schädigenden Person für den entstandenen Schaden kausal war.

[61] Algorithmische Systeme und juristisches Denken unterscheiden sich damit grundlegend. Aus diesen konzeptionellen Unterschieden können sich (Verständnis-)Probleme ergeben, so etwa, wenn Nutzer:innen von algorithmischen Systemen Korrelationen mit Kausalitäten gleichsetzen.[94] Denn das Bestehen einer Korrelation zwischen zwei Variablen sagt nichts über den Grund

87 Martini (Fn. 1), S. 334; Ferrer et al. (Fn. 35), S. 70.

88 Bigman et al. (Fn. 2), S. 19, m.w.H; Ferrer et al. (Fn. 35), S. 79.

89 Bigman et al. (Fn. 2), S. 20.

90 Barocas/Hardt/Narayanan (Fn. 45), S. 214.

91 Duden, «Korrelation»; siehe dazu auch Kasper Gabriel, People Analytics in privatrechtlichen Arbeitsverhältnissen, Baden Baden, 2021, S. 32 m.w.H.

92 Lauscher/Legner (Fn. 66), S. 369; ähnlich auch Raphaële Xenidis/Linda Senden, EU non-discrimination law in the era of artificial intelligence: Mapping the challenges of algorithmic discrimination, in Bernitz et al. (Hrsg.), General Principles of EU Law and the EU Digital Order, S. 153 f.

93 Die Analyse von kausalen Zusammenhängen zwischen beobachteten Variablen ist Gegenstand der sog. *Causal Inference*, ein im Kontext der Datenanalyse relativ neues Forschungsgebiet. Siehe dazu Pearl (2009) für eine Einführung aus statistischer Sicht (abrufbar unter https://www.amazon.de/Causality-Judea-Pearl/dp/052189560X). Die Konzepte sind noch nicht in die heutige Praxis des *machine learning* eingeflossen.

94 Barocas/Hardt/Narayanan (Fn. 45), S. 13; Mittelstadt et al. (Fn. 1), S. 4 f. m.w.H.; spezifisch zu «cum-hoc-ergo-propter-hoc Fehlschlüssen» Martini (Fn. 1), S. 60 m.w.H. Am Gericht der EU (EuG) prüft der Gemeinschaftsrichter etwa, «ob die vorhandenen Daten kausale Aussagen (z.B. aufgrund wissenschaftlicher Experimente) erlauben oder ob blosse Korrelationen (aufgrund einer Analyse von Beobachtungsdaten) feststellbar sind»; siehe dazu Tilmann Altwicker, Statistikbasierte Argumentation in der EU-Rechtsprechung, EuZ 2021, S. 144.

dieser Korrelation und nichts über das Vorliegen einer kausalen Beziehung aus. Auch eine hohe Korrelation zwischen zwei Variablen bedeutet nicht, dass zwischen diesen Variablen eine Kausalität besteht und sie sagt nichts über die Richtung der Kausalität aus.[95] Dass zwei Variablen korrelieren, ist oft die Folge einer dritten Variable, die kausal beide Variablen beeinflusst. So besteht bspw. eine Korrelation zwischen dem Verkauf von Fondue und Skiunfällen, der vermehrte Verkauf von Fondue ist aber weder kausal für die grössere Zahl von Skiunfällen noch sind die Skiunfälle kausal für den Verkauf von Fondue. Kausal ist vielmehr eine dritte Variable, nämlich die Jahreszeit (Winter). Eine Korrelation kann zwar ein Indiz für eine mögliche Kausalität sein, sie ist aber kein Nachweis.

[62] Die Bedeutung der Unterscheidung wird bei der rechtlichen Beurteilung von Diskriminierungen durch algorithmische Systeme deutlich: Eine direkte Diskriminierung durch ein solches System liegt vor, wenn ein geschütztes Merkmal kausal für eine benachteiligende Ungleichbehandlung ist, wenn das algorithmische System bei einer Entscheidung also auf das geschützte Merkmal abstellt. Oft ist aber schwierig zu erkennen, auf welchen Merkmalen die Entscheidung eines algorithmischen Systems beruht. Trifft das System für eine geschützte Gruppe nachteilige Entscheidungen, liegt immerhin ein Indiz für eine Diskriminierung vor, das Anlass zu einer vertieften Prüfung geben sollte. Ergibt die Prüfung, dass das geschützte Merkmal tatsächlich für die Entscheidung kausal ist, liegt eine direkte Diskriminierung vor. Andernfalls ist zu prüfen, ob eine andere Variable (oder mehrere andere Variablen) kausal für die Benachteiligung ist. Korreliert diese Variable (oder die mehreren Variablen) mit einem geschützten Merkmal, kann eine Proxy-Diskriminierung vorliegen, wenn eine hohe Übereinstimmung mit einem geschützten Merkmal besteht.[96] Das kann bspw. der Fall sein, wenn ein algorithmisches System zur Berechnung von Kreditscores auf Postleitzahlen abstellt und in Wohnkreisen mit Postleitzahlen, die zu einem tieferen Kreditscore führen, der Anteil an Menschen ausländischer Herkunft deutlich höher ist als in den anderen Wohnkreisen.

[63] Die Ungleichbehandlung durch ein algorithmisches System lässt sich rechtfertigen, wenn ein sachlicher Grund für die Ungleichbehandlung vorliegt. Das wäre bspw. der Fall, wenn ein algorithmisches System für ältere Personen höhere Krankenkassenprämien berechnet, weil die höheren Prämien aufgrund der höheren Gesundheitskosten bei älteren Menschen gerechtfertigt sind. Die Ungleichbehandlung muss aber verhältnismässig sein, was hier davon abhängen wird, wie viel höher die Prämien für die älteren Personen sind.

C. Skalierung

[64] Algorithmische Systeme implementieren Entscheidungsverfahren, die beliebig oft angewandt werden können. Wenn ein System einmal entwickelt ist, sind die Kosten für jede Entscheidung minimal. Algorithmische Systeme werden deshalb oft für Anwendungsfälle entwickelt, bei denen eine sehr grosse Zahl von Entscheidungen zu treffen ist.[97] Ein diskriminierendes Entscheidungsverhalten eines algorithmischen Systems skaliert deshalb viel mehr und hat damit häufig

95 Gabriel (Fn 91), S. 32; Gerards/Xenidis (Fn. 31), S. 44.

96 Siehe dazu vorne, II.A.2.d).

97 Aziz Z. Huq, A Right to a Human Decision, Virginia Law Review 2020, S. 613; Gerards/Xenidis (Fn. 31), S. 46; Gutachten der Datenethikkommission der Bundesregierung (Fn. 57), S. 167.

viel weitreichendere Auswirkungen als das diskriminierende Verhalten eines (einzelnen) Menschen.[98]

[65] Die Skalierung kann allerdings auch positive Effekte haben, indem sie Diskriminierung sichtbar und damit korrigierbar macht.[99] Wenn einzelne Personen – bspw. bei der Steuerveranlagung – regelmässig diskriminierende Entscheide treffen, wird dies kaum auffallen. Ein algorithmisches System kann dagegen prinzipiell mit unendlich vielen Testfällen überprüft werden. Diskriminierendes Verhalten lässt sich so deutlich besser entdecken und statistisch nachweisen. Diese Erkenntnisse können die Grundlage für Verbesserungen bilden, die es erlauben, nicht-diskriminierende Systeme zu entwickeln. Anders als bei Menschen kann diskriminierendes Verhalten bei algorithmischen Systemen nicht nur gemessen, sondern auch nicht-diskriminierendes Verhalten weitgehend garantiert werden; Voraussetzung ist allerdings, dass die für das Testen verwendeten Datensätze für den späteren Einsatz repräsentativ sind. Auch bei diesem Einsatz lässt sich Diskriminierung an den Ergebnissen in der Regel besser erkennen als bei Menschen, weil das System eine grössere Anzahl von Entscheidungen trifft, als wenn ein Mensch entscheidet.

D. Rückkoppelungseffekte (*feedback loops*)

[66] Bestehende Voreingenommenheiten und Diskriminierungen können in algorithmischen Systemen fortgesetzt werden und sich noch verstärken.[100] Dabei gibt es verschiedene Ausprägungen von Rückkoppelungseffekten (*feedback loops*).

[67] Bei *machine learning-Algorithmen* ergibt sich ein *feedback loop*, wenn der Output eines Modells als Trainingsdaten für ebendieses oder für ein anderes Modell verwendet wird. Ein Empfehlungsalgorithmus, der bei einem Streamingdienst Filme vorschlägt, beeinflusst, welche Filme die Nutzenden ansehen werden. Die häufiger angeschauten Filme werden dann wiederum anderen Nutzenden vorgeschlagen.[101] Eine solche Verstärkung kann auch zu ungewollten, diskriminierenden Resultaten führen.[102] Wenn ein algorithmisches System männliche Bewerber für eine bestimmte Arbeitsstelle als geeigneter qualifiziert und sie deshalb eher vorschlägt, werden mehr männliche Bewerber angestellt. Dies verstärkt wiederum die Annahme, dass männliche Bewerber für die Stelle besser geeignet sind, weil sie öfter angestellt wurden.[103]

[68] *Feedback loops* entstehen oft ungewollt. Ein besonders problematisches Beispiel ist die sog. vorhersagende Polizeiarbeit (*predictive policing*), bei der Straftaten präventiv verhindert werden sollen, indem der Einsatz der Polizei an Orten konzentriert wird, für die ein hohes Risiko

98 Gerards/Xenidis (Fn. 31), S. 46.

99 Gutachten der Datenethikkommission der Bundesregierung (Fn. 57), S. 167.

100 Paul F. Langer/Jan C. Weyerer, Diskriminierungen und Verzerrungen durch Künstliche Intelligenz, Entstehung und Wirkung im gesellschaftlichen Kontext, in Oswald/Borucki (Hrsg.), Demokratietheorie im Zeitalter der Frühdigitalisierung, Wiesbaden 2020, S. 223 f.; Orwat (Fn. 1), S. 83, S. 89 f.; Barocas/Hardt/Narayanan (Fn. 45), S. 13 ff.

101 https://developers.google.com/machine-learning/glossary?hl=en#feedback-loop (zuletzt aufgerufen am 3. Juni 2024).

102 Barocas/Hardt/Narayanan (Fn. 45), S. 14 f.

103 So ähnlich z.B. das Bewerbungstool von Amazon, das gemäss Amazon aber nie im Einsatz war https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G (zuletzt aufgerufen am 3. Juni 2024).

für Straftaten vorhergesagt wird.[104] Wegen des verstärkten Einsatzes der Polizei wird dann auch mehr kriminelles Verhalten entdeckt, womit sich die Vorhersage zu bestätigen scheint. Tatsächlich handelt es sich allerdings um eine sich selbst erfüllende Vorhersage (*self-fulfilling prediction)*.[105] Werden Daten dieser zusätzlich entdeckten kriminellen Aktivitäten gesammelt und als Trainingsdaten für das System verwendet, ohne den verstärkten Einsatz der Polizei zu berücksichtigen, wird die Vorhersage weiter verfestigt.

[69] Diese Beispiele zeigen, dass sich Ungleichbehandlungen durch *feedback loops* mit der Zeit verstärken können.

E. Zielkonflikte von Datenschutz und Genauigkeit

[70] Diskriminierung durch algorithmische Systeme ist für die Entwickler:innen der Systeme oft nur schwer zu erkennen, insb. bei der sog. Proxy-Diskriminierung. Dieses Problem wird durch die Vorgaben des Datenschutzrechts noch verstärkt: Daten, die sich auf geschützte Merkmale beziehen, sind oft als besonders schützenswerte Personendaten zu qualifizieren, an deren Bearbeitung das Datenschutzrecht erhöhte Anforderungen stellt. Als besonders schützenswerte Personendaten gelten bspw. Daten über die Religion, die sexuelle Orientierung, die ethnische Herkunft und die Rasse. Das schweizerische Datenschutzrecht kennt den Grundsatz der Verhältnismässigkeit (Art. 4 Abs. 2 DSG), der unter anderem verlangt, dass so wenig Daten wie möglich gesammelt und bearbeitet werden.[106] Das europäische Datenschutzrecht sieht denselben Grundsatz ausdrücklich als Grundsatz der Datenminimierung vor (Art. 5 Abs. 1 Bst. c DSGVO). Bei der Bearbeitung von besonders schützenswerten Personendaten ist besondere Zurückhaltung geboten. Nach der DSGVO ist die Bearbeitung dieser Daten gar grundsätzlich verboten, wobei eine Reihe weitreichender Ausnahmen besteht (Art. 9 DSGVO).

[71] Aufgrund der Vorgaben des Datenschutzrechts dürfen besonders schützenswerte Personendaten oft weder gesammelt noch in einem System erfasst werden, weil deren Bearbeitung für das mit dem System verfolgte Ziel nicht notwendig ist. Wenn diese Daten im System aber fehlen, lassen sich systematische Zusammenhänge zwischen den geschützten Merkmalen und den Entscheidungen eines algorithmischen Systems nicht erkennen, so dass Diskriminierung nicht entdeckt und nachgewiesen werden kann. So ist es u.a. aus datenschutzrechtlicher Sicht in der Regel nicht zulässig, Bewerbende nach Religion oder sexueller Orientierung zu fragen und diese Merkmale im Dossier zu erfassen. Wenn es aber Daten gibt, die mit der Religion oder der sexuellen Orientierung korrelieren und die Bearbeitung dieser Daten durch das algorithmische System zu einer Proxy-Diskriminierung führt, lässt sich diese nicht erkennen und nachweisen. Die EU versucht, dieses Dilemma mit einer erleichterten Bearbeitung von besonders schützenswerten Daten im Zusammenhang mit Hochrisiko-KI-Systemen zu lösen.[107] In der Schweiz gibt es bisher

---

104 Stephan Richter/Sonja Kind, Predictive Policing, Büro für Technikfolgen-Abschätzung beim deutschen Bundestag, 2016, S. 4 m.w.H.; Kristian Lum/William Isaac, To predict and serve?, 2016, S. 16 m.w.H.; siehe dazu auch Frederik J. Zuiderveen Borgesius, Discrimination, Artificial Intelligence, and Algorithmic Decision-Making, 2018, S. 12.

105 Barocas/Hardt/Narayanan (Fn. 45), S. 14.

106 Bruno Baeriswyl, in Baeriswyl/Pärli/Blonski (Hrsg.), Datenschutzgesetz, Stämpflis Handkommentar, 2. Aufl., Bern 2023, Art. 6 N 21.

107 Art. 10 Abs. 5 KI VO enthält eine Ausnahme vom Verbot der Bearbeitung von besonderen Kategorien personenbezogener Daten gemäss Art. 9 DSGVO. Diese entsprechen weitgehend den besonders schützenswerten Personenda-

keine entsprechende Regelung. Das Problem kann (und sollte) aber im Rahmen der Auslegung und Anwendung des Grundsatzes der Verhältnismässigkeit berücksichtigt werden.

[72] Versucht man, Proxy-Diskriminierungen zu verhindern, indem möglichst wenige Merkmale von Personen in einem algorithmischen System erfasst werden, besteht die Gefahr, dass der Algorithmus weniger präzise Resultate generiert. Auch das Bestreben, Algorithmen diskriminierungsfrei zu gestalten, indem die Nicht-Diskriminierung als zusätzliches Ziel des Algorithmus eingeführt wird, führt in der Regel zu qualitativ weniger guten Resultaten.[108]

IV. Ursachen für algorithmische Diskriminierung

[73] Risikofaktoren für Diskriminierung bestehen auf verschiedenen Stufen der Entwicklung und Anwendung algorithmischer Systeme.[109] Die Identifikation von Diskriminierungsursachen ist bei komplexen algorithmischen Systemen selbst bei aufwändigen Tests äusserst schwierig und manchmal sogar unmöglich.[110]

A. Entwicklung

1. Definition von Modellstruktur und Trainingsverfahren

[74] Die Ursache für Diskriminierung kann schon in der Definition der Modellstruktur und in der Festlegung des Trainingsverfahrens liegen, weil in diese Design-Entscheide der Entwickler:innen persönliche Einschätzungen und Wertungen einfliessen.[111]

[75] Auch die Kombination verschiedener Teile eines algorithmischen Systems kann zu diskriminierenden Ergebnissen führen. So wird bei der Entwicklung oft auf bestehende Komponenten anderer Systeme, bspw. auf Teile von Programmcode (*snippets*), Datenbanken und bereits trainierte Modelle zurückgegriffen, die wiederum von anderen Entwickler:innen und aus einem anderen Kontext stammen und deren Entscheidungen enthalten.[112]

[76] Das Aufdecken der zu Diskriminierung führenden Design-Entscheidungen der Entwickler:innen ist besonders schwierig, weil algorithmische Systeme oft von grossen Teams über län-

ten nach Schweizer Recht. Gemäss Art. 10 Abs. 5 KI VO dürfen Anbieter von Hochrisiko-KI-Systemen solche Daten bearbeiten, wenn es unbedingt notwendig ist, um Verzerrungen im System zu erkennen und zu korrigieren. Dabei müssen jedoch angemessene Vorkehrungen für den Schutz der Grundrechte und Grundfreiheiten der betroffenen Personen getroffen werden. Dazu zählen (vorrangig) die Anonymisierung der Daten oder, wenn durch diese der Zweck der Datenverarbeitung beeinträchtigt würde, die Pseudonymisierung oder Verschlüsselung.

108 Siehe dazu vorne, III.E.

109 Siehe dazu Harini Suresh/John Guttag, A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle, 2021, S. 1 ff.; Gerards/Xenidis (Fn. 31), S. 37.

110 Mittelstadt et al. (Fn. 1), S. 2; Zuiderveen Borgesius (Fn. 104), S. 35 m.w.H. Zu den Schwierigkeiten siehe auch Sandra Wachter/Brent Mittelstadt/Chris Russell, Why fairness cannot be automated: Bridging the gap between EU non-discrimination law and AI, Computer Law & Security Review 2021, S. 19 ff.

111 Gerards/Xenidis (Fn. 31), S. 41; Langer/Weyerer (Fn. 100), S. 224 ff. Dagegen könnte helfen, dass Entwicklungsteams möglichst divers zusammengesetzt werden.

112 Siehe dazu etwa European Union Agency for Fundamental Rights, Bias in Algorithms, Artificial Intelligence and Discrimination, Wien 2022, S. 95 ff.; siehe dazu auch das Beispiel in Felix Neutatz/Ziawasch Abedjan, What is «Good» Training Data? – Data Quality Dimensions that Matter for Machine Learning, in Bundesministerium für Umwelt, Naturschutz, nukleare Sicherheit und Verbraucherschutz und Frauke Rostalski (Hrsg.), Künstliche Intelligenz, Tübingen 2022, S. 8.

gere Zeit entwickelt werden, was es praktisch unmöglich macht, den Entwicklungsprozess und die im System enthaltenen Wertungen, Biases und Abhängigkeiten ganzheitlich zu verstehen.[113]

2. Trainieren

[77] *Machine learning*-Systeme suchen Muster in existierenden Daten, die sie für Vorhersagen für neue Daten verwenden[114]. Sind die Daten, die für das Training eines *machine learning*-Modells verwendet werden, «verzerrt» bzw. *biased*, wird auch die Vorhersage für die neuen Daten *biased* sein. Dieser Mechanismus, der bisweilen als *garbage in*, *garbage out* bezeichnet wird, kann zu diskriminierenden Vorhersagen führen.[115]

[78] Datensätze können aus verschiedenen Gründen verzerrt sein. Zum einen sind Daten stets ein Abbild der Vergangenheit und repräsentieren damit früher bestehende Zustände und Probleme. Bestehende Stereotypen und Diskriminierungen fliessen so in die algorithmischen Systeme ein (*historical bias*).[116] *Historical biases* können bspw. bei Übersetzungstools vorkommen, indem geschlechtsneutrale Bezeichnungen im Englischen, bspw. *teacher* oder *doctor*, in der deutschen Übersetzung zur männlichen Form führen, hier also zu «Lehrer» oder «Arzt».[117]

[79] Damit KI-Systeme für alle Personengruppen korrekte Aussagen treffen können, sollten diese Gruppen in den Trainingsdaten im richtigen Verhältnis repräsentiert sein.[118] Oft sind aber nicht für alle Personengruppen gleich viele Daten verfügbar; namentlich sind Minderheiten in Datensätzen oft untervertreten.[119] Ist eine Personengruppe in einem Datensatz unterrepräsentiert, wird das mit diesen Daten trainierte Modell für diese Gruppe in der Regel eine schlechtere Vorhersage liefern oder mehr Fehlentscheide produzieren (*representation bias*). Eine solche Verzerrung kann auftreten, wenn Daten eine Situation spiegeln, in der eine Ungleichbehandlung bestand, etwa bei Daten über historische Beschäftigungsverhältnisse, in denen Frauen unterrepräsentiert sind, weil ihnen der Zugang zu bestimmten Berufen verwehrt war.[120] Verzerrungen können sich aber auch bei Daten ergeben, die aus der Nutzung von Diensten oder Produkten stammen, die von gewissen Personengruppen weniger oder gar nicht genutzt werden. So sind bspw. Datensätze, die aus *Social Media* stammen, besonders unausgewogen, weil gewisse demographische Gruppen, bspw. ältere Menschen, Personen aus bestimmten geographischen Regionen oder Personen mit limitiertem Internetzugang zu wenig vertreten sind.[121] Armut, die geographische Lage oder eine bestimmte Lebensform sind weitere Faktoren, die dazu führen können, dass von bestimmten Personen

113 Mittelstadt et al. (Fn. 1), S. 6 f.

114 Siehe dazu vorne, II.B.4.

115 Philipp Müller-Peltzer/Valentin Tanczik, Künstliche Intelligenz und Daten, Data-Governance nach der geplanten KI-Verordnung, RDi 2023, S. 455; Mittelstadt et al. (Fn. 1), S. 4 f.; Orwat (Fn. 1), S. 79; Zuiderveen Borgesius (Fn. 48), S. 1574; Suresh/Guttag (Fn. 109), S. 3 ff.

116 Lorenzo Belenguer, AI bias: exploring discriminatory algorithmic decision-making models and the application of possible machine-centric solutions adapted from the pharmaceutical industry, AI and Ethics, 2022, S. 3 f.; Orwat (Fn. 1), S. 80 f.

117 Für ein ähnliches Beispiel zu Neopronomen siehe Lauscher/Legner (Fn. 66), S. 371.

118 Orwat (Fn. 1), S. 79 f.; Hacker (Fn. 40), S. 1147; Suresh/Guttag (Fn. 109), S. 3 ff.

119 Kevekordes, Handbuch Multimedia-Recht (Fn. 82), Teil 29.1 KI im Überblick N 38; Suresh/Guttag (Fn. 109), S. 4.

120 Orwat (Fn. 1), S. 80 m.w.H.

121 Orwat (Fn. 1), S. 81; Barocas/Selbst (Fn. 31), S. 684 f.

weniger Daten vorliegen.[122] Auch bei Systemen der sog. generativen KI sind die verwendeten Trainingsdaten regelmässig nicht repräsentativ. Wenn untervertretene Personen auf Grundlage von Daten über die Mehrheit beurteilt werden, kann dies dazu führen, dass bestehende Vorteile strukturell bevorteilter Mehrheiten verstärkt werden, weil die Minderheiten schlechter beurteilt werden.[123] Eine Unterrepräsentation in den Daten führte bspw. dazu, dass ein algorithmisches Bilderkennungssystem Menschen mit schwarzer Hautfarbe nicht korrekt erkennen konnte.[124]

[80] Ein anderer Grund für Bias in den Trainingsdaten liegt darin, dass die Qualität der Daten für verschiedene Gruppen unterschiedlich sein kann (*measurement bias*).[125] Gerade für Minderheiten ist die Qualität der erhobenen Daten oft systematisch schlechter als für die Mehrheit. So sind bspw. im Gesundheitsbereich Daten von sozial und finanziell schlechter gestellten Personen oft lückenhaft, weil sie schlechteren Zugang zur medizinischen Versorgung haben.[126]

[81] Ein Risiko für Diskriminierung besteht auch bei der Aufbereitung des Trainingsdatensatzes, etwa beim Labeln von Trainingsdaten. Die Vergabe der Labels kann bspw. einen *Gender-Bias* aufweisen, etwa wenn Frauen für einen Beruf systematisch als weniger geeignet eingeschätzt werden als sie es tatsächlich sind. Werden solche Labels für das Trainieren eines algorithmischen Systems verwendet, der beruflichen Erfolg vorhersagen soll, wird das System Frauen diskriminieren.

3. Validieren und Testen

[82] Auch das Validieren und Testen eines trainierten Modells kann zu Diskriminierung führen, wenn die Validierungs- und Testdaten nicht ausgewogen sind oder die Qualität des Modells mit standardisierten Benchmark-Datensätzen überprüft wird, die für die geplante Verwendung des KI-Systems nicht repräsentativ sind (*evaluation bias*).[127]

B. Anwendung

[83] Die Ursachen für Diskriminierung können auch in der Anwendung algorithmischer Systeme liegen. Denn Personen, welche diese Systeme verwenden, müssen deren Ergebnisse in der Regel interpretieren. Der dabei bestehende Spielraum kann bewusst oder unbewusst in diskriminierender Weise genutzt werden, etwa wenn die zuständige Person das Ergebnis so interpretiert, dass Angehörige einer Gruppe mit geschützten Merkmalen benachteiligt werden (*interpretation bias*).[128]

[84] Zu Diskriminierung kann die Anwendung algorithmischer Systeme auch führen, wenn ein sog. *aggregation bias* besteht. Dabei handelt es sich um eine systematische Verzerrung, die entsteht, wenn ein Algorithmus auf alle Personen einer Population angewendet wird, obwohl er für

122 Jonas Lerman, Big Data and Its Exclusions, Stanford Law Review Online 2013, S. 57; Barocas/Selbst (Fn. 31), S. 684 f.

123 Kevekordes, Handbuch Multimedia-Recht (Fn. 82), Teil 29.1 KI im Überblick N 38.

124 https://incidentdatabase.ai/cite/49/ und https://www.theguardian.com/technology/2016/sep/08/artificial-intelligence-beauty-contest-doesnt-like-black-people (beide zuletzt aufgerufen am 3. Juni 2024).

125 Daniel Wolff, KI-Biases im Gesundheitswesen – Teil 1, DuD 2022, S. 737; Suresh/Guttag (Fn. 109), S. 5.

126 Wolff (Fn. 125), S. 737 m.w.H.

127 Suresh/Guttag (Fn. 115), S. 6.

128 Mittelstadt et al. (Fn. 1), S. 8 m.w.H.; Ferrer et al. (Fn. 35), S. 72.

manche Gruppen schlecht oder gar nicht funktioniert. Ein *aggregation bias* kann sich ergeben, wenn ein Prognosealgorithmus für die Mehrheit der Bevölkerung optimiert wurde, weil nur Daten über die Mehrheit für das Training verwendet wurden. Ein so trainierter Algorithmus kann für Vorhersagen über Minderheiten ungeeignet sein, was zu Diskriminierung führen kann. Dasselbe gilt, wenn ein für die Allgemeinheit entwickeltes System auf Gruppen oder Fälle angewendet wird, für die eine gesonderte Betrachtung angezeigt wäre.[129]

[85] Zu Diskriminierung kann es zudem kommen, wenn algorithmische Systeme in Situationen verwendet werden, für die sie nicht vorgesehen waren. Das wird als *transfer context bias* bezeichnet.[130] Wenn ein algorithmisches System für die Beurteilung einer bestimmten Personengruppe entwickelt wurde, kann es bei der Anwendung auf eine andere Personengruppe zu ungenauen oder falschen Ergebnissen führen.

[86] Ein weiteres Problem ist die sog. Automatisierungsverzerrung (*automation bias*). Gemeint ist damit ein (zu) hohes Vertrauen in algorithmische Systeme, die von den Anwendern bisweilen für unfehlbar gehalten werden.[131] So wird algorithmischen Systemen unterstellt, dass sie wegen der fehlenden Emotionalität weniger diskriminierend seien als Menschen, die entscheiden. Dass algorithmische Systeme emotionslos sind, bedeutet aber nicht, dass sie wertfrei bzw. «unbiased» sind.[132] Es ist allerdings unklar, ob und in welchen Konstellationen die Gefahr eines *automation bias* besteht und wie gross sie tatsächlich ist. Relevant ist diese Verzerrung möglicherweise vor allem, weil Menschen nur zurückhaltend von den Empfehlungen algorithmischer Systeme abweichen, v.a. wenn sie die Abweichung gegenüber Vorgesetzten begründen müssen und die selbst gefällte Entscheidung falsch sein könnte (sog. *Default-Effekt*).[133]

V. Technische (und andere) Lösungsansätze

A. Sensibilisierung und Aufklärung

[87] Diskriminierungen durch algorithmische Systeme lassen sich nur verhindern, wenn sich alle Beteiligten, von den Entwickler:innen bis zu den Anwender:innen, der Gefahr bewusst sind. Alle Beteiligten müssen deshalb für die Problematik sensibilisiert und über die Lösungsansätze aufgeklärt werden. Erforderlich ist dafür ein Grundverständnis der technischen Funktionsweise algorithmischer Systeme und der rechtlichen Regelung von Diskriminierung. Die Sensibilisierung und Aufklärung sollte innerhalb der Organisation erfolgen, welche die algorithmischen Systeme entwickelt und/oder anwendet, bspw. im Rahmen interner Schulungen.

129 Orwat (Fn. 1), S. 86; Suresh/Guttag (Fn. 109), S. 5.

130 Danks/London (Fn. 35), S. 4691 ff.; Ferrer et al. (Fn. 35), S. 72.

131 Orwat (Fn. 1), S. 22 Fn. 37; Gerards/Xenidis (Fn. 31), S. 42 m.w.H.

132 Martini (Fn. 1), S. 335; Mittelstadt et al. (Fn. 1), S. 1, S. 7 m.w.H.; Bigman et al. (Fn. 2), S. 19.

133 Gerards/Xenidis (Fn. 31), S. 42; Gutachten der Datenethikkommission der Bundesregierung (Fn. 57), S. 213.

B. Anforderungen an Trainingsdaten

[88] Da Diskriminierung durch algorithmische Systeme oft durch die Verwendung problematischer Trainingsdaten verursacht wird,[134] sind qualitativ hochwertige Trainingsdaten essenziell. Zur Vermeidung von Diskriminierung ist entscheidend, dass die Trainingsdaten mit Blick auf die vorgesehene Anwendung der algorithmischen Systeme repräsentativ sind.[135] Repräsentative Datensätze können durch das Hinzufügen fehlender Daten oder durch die Gewichtung von Daten erreicht werden.[136] Zu diesem Zweck kann ein Datensatz auch mit synthetischen Daten ergänzt werden.

[89] Das Generieren qualitativ hochwertiger Datensätze ist allerdings nicht immer einfach. Oft ist gar nicht erkennbar, dass Daten nicht repräsentativ sind, bspw. weil sie pseudonymisiert oder aggregiert sind oder weil gewisse, für die Prüfung der Repräsentativität notwendige Merkmale und Attribute im Datensatz fehlen.[137] Fraglich ist auch, wie Repräsentativität erreicht werden kann, wenn für gewisse Bevölkerungsgruppen keine oder zu wenig Daten verfügbar sind. Hinzu kommt, dass das Beschaffen von Daten oft mit hohen Kosten verbunden ist. Gerade Startups können sich solche Datensätze oft nicht leisten. Das fördert die Tendenz, sich mit einem funktionierenden System zufrieden zu geben und Diskriminierungen in Kauf zu nehmen. Aus ökonomischen Gründen bestehen jedenfalls kaum Anreize, die Trainingsdaten möglichst diskriminierungsfrei zu gestalten.[138]

[90] Wegen ihrer zentralen Bedeutung enthält die KI VO eine Reihe von Vorgaben zur Qualität von Trainings-, Validierungs- und Testdaten, die bei Hochrisiko-KI-Systemen verwendet werden. Mit Blick auf die Vermeidung von Diskriminierung schreibt die KI VO namentlich vor, dass die Daten auf mögliche Verzerrungen (Biases) zu untersuchen sind, welche die Gesundheit und Sicherheit von Personen beeinträchtigen, sich negativ auf die Grundrechte auswirken oder zu einer nach dem Unionsrecht verbotenen Diskriminierung führen können, insb. wenn die Outputs die Inputs künftiger Verwendungen beeinflussen (Art. 10 Abs. 2 Bst. f KI VO). Zudem sind geeignete Massnahmen zur Erkennung, Verhinderung und Abschwächung möglicher Biases zu ergreifen (Art. 10 Abs. 2 Bst. f KI VO) und relevante Datenlücken zu beheben (Art. 10 Abs. 2 Bst. g KI VO). Überdies müssen Trainings-, Validierungs- und Testdatensätze sowie Labels relevant und hinreichend repräsentativ sein, angemessen auf Fehler überprüft werden und im Hinblick auf den beabsichtigten Zweck so vollständig wie möglich sein. Ausserdem müssen die Datensätze die geeigneten statistischen Merkmale aufweisen, gegebenenfalls auch bezüglich der Personen oder Personengruppen, für die das Hochrisiko-KI-System bestimmungsgemäss verwendet werden soll (Art. 10 Abs. 3 KI VO).

[91] Die Bestimmungen in der KI VO hinterlassen allerdings viele Fragen. Schon die Begriffe Verzerrung bzw. «Bias» sind unklar, zumal eine Definition fehlt.[139] Zudem ist ungewiss, wann

134 Siehe dazu vorne, III. A.

135 Zum Problem des *representation bias* siehe vorne, IV.A.2.

136 Patricia Shaw, in Hervey/Lavy (Hrsg.), The Law of Artificial Intelligence, London 2021, N 3-027.

137 Shaw, in Hervey/Lavy (Fn. 136), N 3-032.

138 Dazu auch Hacker (Fn. 40), S. 1150.

139 Müller-Peltzer/Tanczik (Fn. 115), S. 457.

statistische Merkmale für einen konkreten Kontext eines KI-Systems geeignet sind.[140] Vor allem aber erfassen die Vorgaben zur Datenqualität nur eine (wenn auch wichtige) Ursache von Diskriminierung durch algorithmische Systeme. Denn richtige und repräsentative Trainingsdaten sind zwar meist eine notwendige, aber keine hinreichende Bedingung, dass mit solchen Daten trainierte Systeme diskriminierungsfrei sind.[141]

C. Non-Discrimination by Design

[92] Im Datenschutzrecht gilt das Prinzip *Privacy by Design* (Art. 7 DSG), welches die für eine Datenbearbeitung Verantwortlichen verpflichtet, das Einhalten der Vorgaben des Datenschutzrechts beim Design der Bearbeitungsprozesse durch technische und organisatorische Massnahmen sicherzustellen.[142] Ein solcher Ansatz könnte auch beim Problem der Diskriminierung durch algorithmische Systeme verfolgt werden. Ein Prinzip der *Non-Discrimination by Design* würde verlangen, dass algorithmische Systeme von Anfang an diskriminierungsfrei gestaltet werden. Der Grundsatz würde nicht nur die Entwicklung der Systeme erfassen, sondern auch ein laufendes Testen und Überwachen bei der Anwendung erfordern.

[93] Mit einem *by Design*-Ansatz wird eigentlich nur eine Selbstverständlichkeit zum Ausdruck gebracht, nämlich: dass die Einhaltung der Rechtsordnung nicht erst sichergestellt werden kann, wenn bereits ein Regelverstoss vorliegt, weshalb die Ausgestaltung von Prozessen oder Systemen von Anfang an darauf ausgerichtet sein muss, die Vorgaben des Rechts einzuhalten. Da sich Diskriminierungen durch algorithmische Systeme nur durch eine entsprechende Ausgestaltung der Systeme verhindern oder zumindest verringern lassen, kann es dennoch hilfreich sein, dies durch einen ausdrücklichen *Non-Discrimination by Design*-Ansatz zu verdeutlichen. Bei algorithmischen Systemen, die sich weiterentwickeln, bleibt dieser Ansatz nicht auf das ursprüngliche Design des Systems beschränkt. Vielmehr muss über den gesamten Lebenszyklus immer wieder geprüft werden, ob die Anwendung eines Systems aufgrund der Weiterentwicklung nicht doch zu Diskriminierung führt. Trifft dies zu, muss das System verbessert werden.

[94] Für die Umsetzung eines *Non-Discrimination by Design*-Ansatzes muss zum einen sichergestellt werden, dass die verwendeten Daten keinen Bias enthalten, der während des Trainings des algorithmischen Systems internalisiert werden kann.[143] Das reicht aber nicht, um Diskriminierung zu verhindern: Einen Algorithmus so zu optimieren, dass er eine bestimmte Aufgabe bestmöglich erfüllt, bedeutet keineswegs, dass Diskriminierung ausgeschlossen wird. Im Gegenteil kann gerade das Streben nach einer «optimalen» Lösung zu einem diskriminierenden System führen.[144] Wenn Diskriminierungsfreiheit nicht explizit als zusätzliches Ziel definiert wird, stellt sie sich höchstens zufällig ein. Fehlt eine solche Vorgabe, wird ein optimierter Algorithmus oft diskriminierende Entscheidungen fällen.

---

140 Martin Ebers/Veronika R.S. Hoch/Frank Rosenkranz/Hannah Ruschemeier/Björn Steinrötter, Der Entwurf für eine EU-KI-Verordnung: Richtige Richtung mit Optimierungsbedarf, Eine kritische Bewertung durch Mitglieder der Robotics & AI Law Society, Rdi 2021, S. 534; Müller-Peltzer/Tanczik (Fn. 115), S. 455.

141 Siehe dazu vorne, IV.

142 SHK DSG-Baeriswyl (Fn. 106), Art. 7 N 6.

143 Alexander Tischbirek, Artificial Intelligence and Discrimination: Discriminating Against Discriminatory Systems, in Wischmeyer/Rademacher (Hrsg.), Regulating Artificial Intelligence, Cham 2020, S. 106; siehe dazu vorne, V.B.

144 Siehe dazu vorne, II.B.4.

[95] Der hier vorgeschlagene Ansatz würde erfordern, dass die Diskriminierungsfreiheit explizit als Bedingung im Algorithmus festgelegt wird, auch wenn das die Genauigkeit der Ergebnisse reduzieren kann.[145] Das bedeutet einerseits, dass KI-Systeme bei Entscheidungen oder Empfehlungen in der Regel weder auf geschützte Merkmale noch auf Proxies für solche Merkmale abstellen sollten. In gewissen Fällen ist das aber entweder nicht möglich oder nicht sinnvoll. Der Verzicht auf die Berücksichtigung geschützter Merkmale kann allenfalls sogar dazu führen, dass erforderliche Unterscheidungen nicht vorgenommen werden, sodass es zu einer Diskriminierung aufgrund von Gleichbehandlung[146] kommt. In diesen Konstellationen muss deshalb sichergestellt werden, dass die Möglichkeit der qualifizierten Rechtfertigung einer Ungleichbehandlung[147] im Algorithmus explizit berücksichtigt wird. Erste Ansätze dazu sind in der Literatur bereits vorhanden.[148] Bei dieser Frage, die sich an der Schnittstelle von Rechtswissenschaft und Informatik bewegt, besteht allerdings Bedarf nach vertiefter, interdisziplinärer Forschung.

VI. Fazit

[96] Die zunehmende Verbreitung von algorithmischen Systemen erhöht die Gefahr der Diskriminierung durch solche Systeme. Das ist aus zwei Gründen besonders problematisch: Zum einen werden algorithmische Systeme meist auf eine Vielzahl von Fällen angewendet, weshalb Diskriminierung durch solche Systeme regelmässig viele Menschen trifft (Skalierung). Zum andern sind Entscheidungen von algorithmischen Systemen regelmässig nur beschränkt erklärbar und nachvollziehbar, was den Nachweis der Diskriminierung erschwert.

[97] Die Verwendung von algorithmischen Systemen schafft erhebliche Risiken für Diskriminierung, bietet aber auch Chancen, um Diskriminierung zu verringern oder zu verhindern. Zum einen können algorithmische Systeme helfen, von Menschen verursachte Diskriminierung zu entdecken. Zum andern ist es ungleich einfacher, «Biases» in solchen Systemen zu bereinigen, als die Voreingenommenheit und diskriminierenden Neigungen von Menschen zu überwinden.

[98] Das Risiko der Diskriminierung durch algorithmische Systeme lässt sich nur durch ein Zusammenwirken von rechtlichen und technischen Massnahmen verringern. Das erfordert eine vertiefte Zusammenarbeit von Jurist:innen und Informatiker:innen, die auf einem gegenseitigen Verständnis der Regeln, Mechanismen und Herausforderungen beruht. Dieser Beitrag macht einen ersten Schritt, indem er versucht, interessierten Jurist:innen die technischen Grundlagen zu vermitteln, die sie benötigen, um die Ursachen für Diskriminierung durch algorithmische Systeme und die Mittel zur Bekämpfung zu verstehen. Die Darstellung der rechtlichen Grundlagen sollte Informatiker:innen helfen, den rechtlich relevanten Begriff der Diskriminierung besser zu verstehen und Diskriminierungen im Rechtssinn von weiteren Konstellationen abzugrenzen, die in der Informatik als Diskriminierung verstanden werden, rechtlich aber zulässig sind. Dazu gehört insb. die in der Informatik bisher zu wenig beachtete Möglichkeit der Rechtfertigung von Ungleichbehandlungen.

145 Siehe dazu vorne, III.E.

146 Siehe dazu vorne, II.A.2.b).

147 Siehe dazu vorne, II.A.1.b).

148 Michele Loi/Anders Herlitz/Hoda Heidari, Fair equality of chances for prediction-based decisions, Economics and Philosophy, Published online 2023: 1-24.

[99] Ein Zusammenwirken von Recht und Technik ist auch bei den Massnahmen erforderlich, die getroffen werden müssen, um Diskriminierung durch algorithmische Systeme zu vermeiden. Dieser Ansatz kann als *Non-Discrimination by design* bezeichnet werden. Ausgangspunkt ist dabei die Frage, welche Ungleichbehandlungen rechtlich als Diskriminierung zu qualifizieren sind und wie Diskriminierung durch algorithmische Systeme entsteht. Auf dieser Grundlage können technische, rechtliche und weitere Lösungen entwickelt werden, die dazu beitragen, das Auftreten rechtlich relevanter Diskriminierung zu verhindern. Im Vordergrund stehen dabei die Aufklärung und Sensibiliserung der Entwickler:innen und Anwender:innen von algorithmischen Systemen, Anforderungen an die Qualität der Trainingsdaten, die explizite Vorgabe des Ziels der Nicht-Diskriminierung an algorithmische Systeme sowie das laufende Testen und Überwachen der Systeme bei der Anwendung. Das kohärente Zusammenwirken dieser Ansätze sollte ermöglichen, Diskriminierungen durch algorithmische Systeme massgeblich zu verringern.

Florent Thouvenin, Prof. Dr., Lehrstuhl für Informations- und Kommunikationsrecht, Vorsitzender des Lenkungsausschusses des Center for Information Technology, Society, and Law (ITSL), Universität Zürich.

Stephanie Volz, Dr. iur, Wissenschaftliche Geschäftsführerin, Center for Information Technology, Society, and Law (ITSL), an der Universität Zürich, Rechtsanwältin.

Soraya Weiner, wissenschaftliche Mitarbeiterin von 2022–2023 am Center for Information Technology, Society, and Law (ITSL). Derzeit arbeitet sie als Juristin im Bereich Datenschutz und doktoriert an der Universität Zürich. Dieser Beitrag gibt die persönliche Meinung der Autorin wieder und entspricht nicht notwendigerweise der Position ihrer Arbeitgeberin.

Christoph Heitz, Prof. Dr., Mitglied Institutsleitung des Instituts für Datenanalyse und Prozessdesign, School of Engineering, Zürcher Hochschule für Angewandte Wissenschaften.