Künstliche Intelligenz - Ethik und Recht

Hoeren / Pinelli

2022
ISBN 978-3-406-77388-4
C.H.BECK

umfassendes Spezialsortiment im Bereich Recht, Steuern und Wirtschaft mit rund 700.000 lieferbaren Fachbuchtiteln.

v. Lewinski/de Barros Fritz, Arbeitgeberhaftung nach dem AGG infolge des Einsatzes von Algorithmen bei Personalentscheidungen, NZA 2018, 620
v. Staudinger (Hrsg.), BGB, Band 2, 18. Aufl, Berlin 2018
Steege, Algorithmenbasierte Diskriminierung durch Einsatz von Künstlicher Intelligenz, MMR 2019, 715
Teubner, Digitale Rechtssubjekte? Zum privatrechtlichen Status autonomer Softwareagenten, AcP 218 (2018), 155
Verhoeven (Hrsg.), Digitalisierung im Recruiting – Wie sich Recruiting durch künstliche Intelligenz, Algorithmen und Bots verändert, Wiesbaden 2020
Wagner, Verantwortlichkeit im Zeichen digitaler Techniken, VersR 2020, 717
Wolf, Schuldnerhaftung bei Automatenversagen, JuS 1989, 899
Wischmeyer, Regulierung intelligenter Systeme, AöR 143 (2018), 1
Zech, Künstliche Intelligenz und Haftungsfragen, ZfPW 2019, 198

Kapitel 7. Rechtliche Bewertung möglicher Fairnessmaße

Kevekordes/Hauer/Amir Haeri

A. Einleitung

Die Gesellschaft als Ganzes hat sich dafür entschieden, verschiedene Formen der Diskriminierung aus dem Alltag zu verbannen. Diese Entscheidung spiegelt sich in diversen Gesetzen wider, wie zum Beispiel dem deutschen Grundgesetz oder dem Allgemeinen Gleichbehandlungsgesetz (AGG) auf deutscher Gesetzesebene, sowie der EU-Grundrechtecharta (GRCh) und der Datenschutzgrundverordnung (DS-GVO) auf europäischer Ebene.

Seit einigen Jahren ist der Einsatz von KI-basierten Empfehlungssystemen in vielen kritischen Anwendungsbereichen auf dem Vormarsch (zB gepanzerte Drohnen[1], Predictive Policing[2], uvm.[3]). Solche Systeme haben jedoch das Problem, dass sie mit Hilfe von Daten trainiert werden müssen, die eine historische Verzerrung enthalten können, was zu Diskriminierung führen kann.

Das Ziel für eine Gesellschaft muss es sein, eine faire algorithmische Entscheidungsfindung zu gestalten.[4] Der Begriff der Fairness ist allerdings vielschichtig. Philosophen debattieren seit Jahrtausenden die allgemeine Frage, was Fairness bedeutet (siehe zB das Prinzip der proportionalen Gleichheit von Aristoteles,[5] aber auch die Werke von Rawls[6] und Dworkin[7] in jüngerer Zeit). Maschinelle Lernmodelle bilden die Grundlage der KI-Entscheidungsfindung. Um Fairness bei KI-Entscheidungen

---

[1] *Altmann/Sauer* Survival 59 (2017), 117.

[2] *Shapiro* Nature 541 (2017), 458.

[3] *Faggella,* Everyday examples of artificial Intelligence and Machine Learning, *Emerj Artificial Intelligence Research*, 11.4.2020, https://emerj.com/ai-sector-overviews/everyday-examples-of-ai/ (geprüft am 16.4.2021).

[4] *Muller*, The Impact of Artificial Intelligence on Human Rights, Democracy and the Rule of Law., S. 18.

[5] *Aristoteles,* Nikomachische Ethik, 1131b, zitiert nach *Aristoteles/Gigon/Nickel*, Die Nikomachische Ethik, 2. Aufl. 2014, S. 199 f.

[6] *Rawls. John*, A Theory of Justice, 1971.

[7] *Dworkin* Philosophy & Public Affairs 10 (1981), 185 und *Dworkin* Philosophy & Public Affairs 10 (1981), 283.

zu gewährleisten, muss sie mathematisch definiert werden, da sie nur auf diese Weise in ein Lernmodell implementiert werden kann.[8]

Für viele mögliche Szenarien der KI-Entscheidungsfindung kann eine klare Definition dessen, was Fairness eigentlich bedeutet, nur durch die Betrachtung des Anwendungskontextes[9] ermittelt werden. Die Herausforderung liegt darin, dass es eine große Anzahl von Maßen zur Bestimmung einer Voreingenommenheit bzw. eines Bias (und damit sozusagen einer Unfairness) gibt, von denen sich einige gegenseitig ausschließen.[10] Um zu allgemeine Formulierungen zu vermeiden und mögliche Lösungen an einem konkreten Beispiel zu ermitteln, konzentriert sich dieser Aufsatz auf eine Ausarbeitung relevanter Fairnessmaße für Personalentscheidungen, gruppiert sie, betrachtet ihre technisch-mathematischen Vor- und Nachteile und bewertet – basierend auf deutschem und EU-Recht – die Rechtskonformität und Rechtsfolgen ihrer Anwendung.

B. KI in Personalentscheidungen

Die Idee, algorithmische Entscheidungsfindung als Grundlage für zukünftige Personalentscheidungen zu nutzen, zum Beispiel auf Basis einer Bewertung der Mitarbeiterleistung, verspricht potenziell bessere Ergebnisse und einen faireren und objektiveren Prozess.[11] Versprochen wird, dass maschinelles Lernen helfen kann, zB, indem es bestehende Vorurteile aufdeckt, welche dadurch gezielter angegangen werden können, oder indem es sogar frei von menschlichen Vorurteilen entscheidet.

Wenngleich Vorteile für den Einsatz von KI bei Personalentscheidungen bestehen mögen, birgt der Einsatz von KI gleichzeitig das erhebliche Risiko, bestehende Vorurteile und diskriminierende Effekte in der Gesellschaft zu verfestigen oder sogar zu verstärken.

Die Grundidee der Entscheidungsfindung durch maschinelles Lernen besteht darin, auf Basis vorhandener Datensätze, den sogenannten Trainingsdaten,[12] einen Satz von Entscheidungsregeln zu entwickeln, der als

[8] Das Thema wird in letzter Zeit so intensiv behandelt, dass sich eine gesamte Forschungsdomäne darum gebildet hat, die sich „Fairness, Accountability, and Transparency in Machine Learning“ nennt.

[9] *Güroğlu/van den Bos/Rombouts et al.* Soc Cogn Affect Neurosci 5 (2010), 414.

[10] *Zweig/Krafft* in *Kar/Thapa/Parycek* (Hrsg.), (Un)berechenbar? Algorithmen und Automatisierung in Staat und Gesellschaft, 2018, S. 204 (224).

[11] *Dattner/Chamorro-Premuzic/Buchband/Schettler*, The Legal and Ethical Implications of Using AI in Hiring, Harvard Business Review, 25.4.2019, https://hbr.org/2019/04/the-legal-and-ethical-implications-of-using-ai-in-hiring (zuletzt abgerufen am 16.4.2021).

[12] *Molnar*, Interpretable Machine Learning, S. 15.

sogenanntes statistisches Modell gespeichert wird. Daher neigt jedes statistische Modell dazu, strukturell genauso voreingenommen gegenüber bestimmten sozialen Gruppen wie der Trainingsdatensatz selbst zu sein.

Damit verbunden ist das Problem der mangelnden Transparenz im Entscheidungsprozess, das Auswirkung auf das Rechtssystem insgesamt hat. KI-basierte Black-Box-Modelle[13], insbesondere künstliche neuronale Netze, sind kaum erklärbar.[14] Die Ergebnisse, die die Modelle ausgeben, resultieren aus der Gewichtung von Tausenden von Verbindungen zwischen den Merkmalen des ursprünglichen Datensatzes. Die versteckten Schichten zwischen Eingabe- und Ausgabeschicht sind extrem schwer zu untersuchen. Technologien wie zB Shapley Values,[15] LIME[16], Anchors[17] oder Surrogate Models[18] scheinen vielversprechend, um die Black-Box nachvollziehbarer zu machen. Dennoch kann die immense Komplexität der Tausenden von Verbindungen zwischen den verborgenen Schichten, insbesondere künstlicher neuronaler Netze, bis heute nicht ausreichend erklärt werden. Dies stellt die übliche prozessorientierte Art der Feststellung von Diskriminierung in Frage, die sich auf den Entscheidungsprozess selbst statt auf dessen Ergebnis stützt.

Betrachtet man als Beispiel für diese prozessorientierte Bewertung den Begriff „Gleichbehandlung" in der Richtlinie[19] 2006/54/EG[20], die sich mit der Umsetzung des Grundsatzes der Chancengleichheit und Gleichbehandlung von Männern und Frauen bei Personalentscheidungen befasst, so ist die prozessorientierte Bewertung bereits im Wort *Behandlung* versteckt. Das Augenmerk des Antidiskriminierungsrechts liegt nicht auf dem Ergebnis, sondern auf der *Behandlung* eines Individuums in einem bestimmten Fall, dh dem individuellen Prozess. Die Richtlinie definiert „unmittelbare Diskriminierung" als eine weniger günstige Behandlung einer Person im Vergleich zu einer anderen Person aufgrund des Geschlechts in einer vergleichbaren tatsächlichen, früheren oder hypotheti-

[13] Bei Black-Box-Modellen handelt es sich um KI-Modelle, deren innere Mechanik nicht einsehbar bzw. nicht nachvollziehbar ist.

[14] *Sokol/Flach* FAT* 2020 - Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency 2020, 56.

[15] Urspr. *Shapley* in Kuhn/Tucker (Hrsg.), Contributions to the Theory of Games, Volume II, 1953, S. 307; darauf aufbauend *Sundararajan/Najmi* PMLR 2020, 9269.

[16] *Ribeiro/S. Singh/Guestrin* KDD '16 2016, 1135.

[17] *Ribeiro/S. Singh/Guestrin* AAAI 32 (2018).

[18] *Craven/Shavlik* NIPS'95 1995, 24.

[19] Richtlinien bilden zusammen mit Verordnungen den wesentlichen Teil des sekundären EU-Rechts. Im Gegensatz zu Verordnungen sind Richtlinien nicht unmittelbar anwendbar, sondern ihre Bestimmungen müssen zunächst von den Mitgliedstaaten entsprechend ihrer eigenen Rechtssysteme umgesetzt werden.

[20] Richtlinie 2006/54/EG vom 5. Juli 2006 zur Verwirklichung des Grundsatzes der Chancengleichheit und Gleichbehandlung von Männern und Frauen in Arbeits- und Beschäftigungsfragen (Neufassung), ABl. EU 204/23 (zit. Gleichbehandlungsrichtlinie).

schen Situation.[21] Im Gegensatz dazu liegt eine „mittelbare Diskriminierung" vor, wenn dem Anschein nach neutrale Vorschriften, Kriterien oder Verfahren Personen aufgrund ihres Geschlechts in besonderer Weise benachteiligen, es sei denn, die betreffenden Vorschriften, Kriterien oder Verfahren sind durch ein rechtmäßiges Ziel sachlich gerechtfertigt und die Mittel sind zur Erreichung dieses Ziels angemessen und erforderlich.[22] Der Unterschied zwischen unmittelbarer und mittelbarer Diskriminierung liegt also in einer offen diskriminierenden *Behandlung* bzw. einem diskriminierenden Prozess (unmittelbare Diskriminierung) im Gegensatz zu einer scheinbar neutralen Bestimmung, einem Kriterium oder einer Praxis (mittelbare Diskriminierung). Das bedeutet, dass dasselbe Entscheidungsergebnis entweder auf einer unmittelbaren oder einer mittelbaren Diskriminierung beruhen kann; der Unterschied ergibt sich allein durch den Entscheidungsprozess, der dem Ergebnis zugrunde liegt. Die Idee, Diskriminierung aufgrund einer weniger günstigen Behandlung einer Person im Vergleich zu anderen zu definieren, findet sich im gesamten Antidiskriminierungsrecht der EU[23] und ist daher die Grundlage des Antidiskriminierungsrechts jedes Mitgliedstaates.

Die prozessbezogene Bewertung wird noch deutlicher, wenn man sich die Definition der mittelbaren Diskriminierung im Detail ansieht.[24] Betrachtet man zB einen Arbeitsplatz, der ein Mindestmaß an körperlicher Kraft erfordert, ist es nur logisch, dass statistisch gesehen, aufgrund biologischer Gegebenheiten, mehr Männer als Frauen beschäftigt werden. Im Falle eines Rechtsstreits müsste ein Richter beurteilen, ob die formulierten Anforderungen angemessen sind und die daraus resultierende Verzerrung in diesem speziellen Fall akzeptabel ist.

Diese Denkweise eignet sich für Entscheider, die niedergeschriebene Praktiken und Kriterien befolgen, um im Einzelfall eine möglichst unvoreingenommene Entscheidung treffen zu können. Sie entspricht der klassischen Methode, eine abstrakte Regel auf einen Einzelfall anzuwenden.

So gut diese Denkweise für Menschen funktionieren mag, so schlecht oder zumindest nicht geeignet ist sie für Machine Learning-Modelle. Die beschriebene, prozessbasierte Denkweise setzt voraus, dass der tatsächli-

[21] Art. 2 Abs. 1 lit. a Gleichbehandlungsrichtlinie.
[22] Art. 2 Abs. 1 lit. b Gleichbehandlungsrichtlinie.
[23] Art. 2 Gleichbehandlungsrichtlinie; Art. 2 Richtlinie 2000/43/EG vom 29. Juni 2000 zur Anwendung des Gleichbehandlungsgrundsatzes, ohne Unterscheidung der Rasse oder der ethnischen Herkunft, ABl. EU 180/22; Art. 2 Richtlinie 2000/78/EG vom 27. November 2000 zur Festlegung eines allgemeinen Rahmens für die Verwirklichung der Gleichbehandlung in Beschäftigung und Beruf, ABl. EU 303/16; Art. 2 Richtlinie 2004/113/EG vom 13. Dezember 2004 zur Verwirklichung des Grundsatzes der Gleichbehandlung von Männern und Frauen beim Zugang zu und bei der Versorgung mit Gütern und Dienstleistungen; ABl. EU 373/37.
[24] Vgl. Art. 2 Abs. 1 lit. b Gleichbehandlungsrichtlinie.

che Entscheidungsprozess – auch der Maschine – transparent ist oder zumindest nachgebildet werden kann. Um zu verstehen, ob eine neutrale Bestimmung zu einer mittelbaren Diskriminierung geführt hat, muss sie in dieser prozessbasierten Sichtweise transparent gemacht werden.

C. Mögliche Fairnessmaße

Für reine Blackbox-Modelle, insbesondere für künstliche neuronale Netze mit Deep Learning, lässt sich die erforderliche Transparenz nur schwerlich herstellen. Dies ist einer der Gründe, warum es grundsätzlich gemäß Art. 22 Abs. 1 DS-GVO verboten ist, dass KI vollautomatische Einstellungsentscheidungen trifft. Der Einzelne darf nicht einem technischen und intransparenten Verfahren ausgeliefert sein, ohne die zugrundeliegenden Annahmen und Bewertungskriterien nachvollziehen und ggf. seine Rechte und Interessen geltend machen zu können.[25]

Wenn der Entscheidungsprozess nicht transparent und erklärbar gemacht werden kann, fällt entweder die Idee einer prozessorientierten Bewertung in sich zusammen oder die neue Technologie kann nicht in solchen Entscheidungsprozessen eingesetzt werden, bei denen ein Einblick in die Entscheidungslogik maßgeblich ist.

Im Allgemeinen lassen sich viele bekannte Probleme rund um die algorithmische Entscheidung grob in zwei Kategorien einteilen: 1) Probleme im Zusammenhang mit dem Design und dem Einsatz von algorithmischen Entscheidungsmodellen und 2) Probleme im Zusammenhang mit den Trainingsdaten. Beide Arten von Problemen können zur Verstärkung oder Integration von Vorurteilen beim maschinellen Lernen führen. Während sich die erste Art von Problemen jedoch um etablierte Entwicklungsprozesse dreht und durch Experimente und bessere Erfahrungen gelöst werden muss, kann die zweite Art mathematisch adressiert und somit optimiert werden. Was beide Arten von Problemen jedoch auslassen, ist die Frage, wie man Bias bzw. Voreingenommenheit bei Machine Learning-Modellen eigentlich definiert. Wenn die algorithmenbasierte Entscheidungsfindung auf breiter Basis in die Praxis umgesetzt werden soll, ist es notwendig, sein Augenmerk verstärkt auf die Definition und Bewertung von algorithmischem Bias zu legen.[26] Die wesentliche Frage besteht darin, welche Verteilung von Ergebnissen als diskriminierend angesehen werden kann. Bei dem Versuch, genau diese Frage zu beantworten, entsteht die Idee, Fairnessmaße zu definieren und zu verwenden. Da über

[25] Siehe *Scholz* in Simitis-DSGVO, Art. 22 DSGVO, Rn. 3.
[26] *Guggenberger* MMR 2019, 777 (778).

20 Fairnessmaße[27] existieren und die Diskussion darüber, welche Fairnessmaße unter welchen Umständen verwendet werden sollten, noch in den Kinderschuhen steckt[28], muss die logische Folgefrage dann lauten, ob und wie verschiedene Fairnessmaße in die bestehende Rechtsordnung integriert werden können.

Die technische Komponente der meisten algorithmischen Entscheidungssysteme ist ein Klassifikationsmodell: Angenommen, es würden sich 40 Personen auf eine offene Stelle bewerben, 10 Frauen und 30 Männer. Ein Klassifikationsmodell soll entscheiden, welche der 40 Personen zu einem Vorstellungsgespräch eingeladen werden und welche nicht. Die „Eingabe" dieses Modells, also die Grundlage, auf der Entscheidungen getroffen werden sollen, ist die in Zahlen umgewandelte (man spricht hierbei von einer Operationalisierung Sammlung als relevant eingestufter Merkmale jedes Bewerbers). Die „Ausgabe" des Modells ist eine Empfehlung, ob eine der Eingabe entsprechende Person eingeladen werden soll, oder nicht. Die Klassifikation kommt auf Grundlage einer statistischen Auswertung historischer Daten zu Stande, man bezeichnet diese historischen Daten als „Ground Truth". Da es nur die beiden Klassen „wird zum Vorstellungsgespräch eingeladen" und „wird nicht zum Vorstellungsgespräch eingeladen" gibt, liegt hier eine binäre Klassifizierung vor.

Fairnessmaße berücksichtigen nun im Kontext der binären Klassifizierung eine der Klassen als das gewünschte Ergebnis für eine Person (genannt die positive Klasse, im Beispiel also für ein Vorstellungsgespräch berücksichtigt zu werden) und die andere Klasse als das unerwünschte Ergebnis für eine Person (genannt die negative Klasse, im Beispiel also nicht für ein Vorstellungsgespräch berücksichtigt zu werden).

Betrachtet man nun ein Klassifikationsproblem, setzt sich dessen Eingabe zusammen aus den sensitiven Attributen $A_1, \ldots, A_m$ (zB Geschlecht oder religiöse Überzeugung), gegen die eine Diskriminierung nach der Antidiskriminierungsgesetzgebung rechtswidrig ist, und aus den übrigen Attributen $X_1, \ldots, X_n$. Außerdem sei Y die Zielausgabe (je nach Problemstellung entweder die Ausgabe, die der Ground Truth entspricht oder die ideal gewünschte Ausgabe) und R die Systemausgabe, also die Ausgabe des Klassifikationsmodells. Um das Problem zu vereinfachen, soll sich auf eine binäre Klassifikation $R = \{0,1\}$ und die binären Merkmale $X = \{0,1\}$ konzentriert werden (zB hat Berufserfahrung oder nicht). Ohne, dass die Ergebnisse der Untersuchung ihre Allgemeingültigkeit verlieren, soll angenommen werden, dass es nur ein sensitives Attribut A gibt, wel-

[27] *Verma* FairWare '18 2018, 1 (2).

[28] *Hutchinson/Mitchell* FAT*'19 2019, 49; *Friedler/Scheidegger/Venkatasubramanian et al.* FAT*'19 2019, 329.