Faire Algorithmen und die Fairness von Erklärungen

Informatische, rechtliche und ethische Perspektiven

The implications of conversing with intelligent machines in everyday life for people´s beliefs about algorithms, their communication behavior and their relationship building

Ein Projekt gefördert durch

VolkswagenStiftung

Impressum

Autoren:
André Artelt, Christian Geminn, Barbara Hammer, Arne Manzeschke, Lina Mavrina, Carina Weber

http://www.impact-projekt.de/

Kontakt:
Prof. Dr. Nicole Krämer

Telefon: +49 203 379 - 2482
E-Mail nicole.kraemer@uni-due.de

Universität Duisburg-Essen
Forsthausweg 2
47057 Duisburg

Details zur Publikation:
DOI: 10.17185/duepublico/76311
ISBN (Print): 978-3-940402-56-1

Veröffentlichende Institution: Universität Duisburg-Essen, Universitätsbibliothek, DuEPublico, Universitätsstraße 9-11, 45141 Essen, https://duepublico2.uni-due.de

1. Auflage, Juli 2022

Faire Algorithmen und die Fairness von Erklärungen: Informatische, rechtliche und ethische Perspektiven

Auf Künstliche Intelligenz (KI) gestützte Entscheidungssysteme haben einen immer größer werdenden Einfluss auf unser Leben zum Beispiel in Form von maschinellen Übersetzungsprogrammen[1], Werbe- und Produktempfehlungen im E-Commerce[2], Spam-Filtern[3], teilautonomen Fahrzeugen[4] und Sprachassistenten[5]. Obwohl KI-basierte Systeme zweifelsfrei zu beachtlichen Leistungen fähig sind und bereits nutzbringend in der Praxis eingesetzt werden, so sind ebenfalls Probleme sichtbar geworden. Einerseits finden sich Fehler in der Funktion (z.B. in der Bilderkennung[6]), andererseits ergeben sich sehr grundsätzliche Probleme in Form von unmoralischen Konsequenzen, die beispielsweise durch diskriminierende Ausgaben verursacht werden.[7] Eines der ersten (und bekanntesten) Systeme, in dem eine systematische Diskriminierung entdeckt worden ist, ist das COMPAS System (COMPAS – Correctional Offender Management Profiling for Alter-native Sanctions)[8]. Dieses System wurde in den USA zur Beurteilung der Rückfälligkeit von Strafgefangenen benutzt, es stellte sich jedoch heraus, dass People of Color (PoC) ein höheres Rückfallrisiko zugewiesen wurde als weißen Strafgefangenen – das heißt ein aus der Datenmenge als relevantes Muster identifiziertes Merkmal wurde von der KI positiv mit einem höheren Rückfallrisiko in Verbindung gebracht. Problematisch dabei ist, dass diese Korrelation von einem Menschen für die Entscheidung nicht als relevant angesehen würde, da sie nicht kausal ist. Es ist offensichtlich, dass eine solche Ausgabe der KI diskriminierend und inakzeptabel ist.

Als Konsequenz haben die Aspekte Erklärbarkeit und Fairness von KI-gestützten Entscheidungssystemen größere Beachtung erlangt.[9] Als ein Beispiel identifiziert die *IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems* fünf zentrale Prinzipien, die bei Design und Umsetzung intelligenter autonomer oder assistiver Systeme Beachtung finden

1 *Popel M./ Tomková M./ Tomek J./ Kaiser L.*, Transforming machine translation: a deep learning system reaches news translation quality comparable to human professionals, Nature communications 11.1, 2020, S. 1-15.

2 *Hessel T./ Pirker C./ Haltmeier M.,* Image-based fashion product recommendation with deep learning, International Conference on Machine Learning, Optimization, and Data Science. Springer, Cham, 2018.

3 *Tretyakov, K.*, Machine learning techniques in spam filtering, Data Mining Problem-oriented Seminar, MTAT. Vol. 3. No. 177. Citeseer, 2004.

4 *El Sallab A. / Abdou M./ Perot E./ Yogamani S.*, Deep reinforcement learning framework for autonomous driving, Electronic Imaging, AVM-023, 2017, S. 70-76.

5 *Tretyakov, K.*, Machine learning techniques in spam filtering, Data Mining Problem-oriented Seminar, MTAT. Vol. 3. No. 177. Citeseer, 2004.

6 *Papernot N./ McDaniel P./ Goodfellow I./ Jha S./ Celik Z. B./ Swami A.*, Practical black-box attacks against machine learning, Proceedings of the 2017 ACM on Asia conference on computer and communications security, 2017.

7 *Lo Piano, S.*, Ethical principles in machine learning and artificial intelligence: cases from the field and possible ways forward, Humanities and Social Sciences Communications, 2020, S. 1-7.

8 https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing 05, 2016.

9 *Guidotti R./ Monreale A./ Ruggieri S./ Turini F./ Giannotti F./ Pedreschi D*., A survey of methods for explaining black box models, ACM Computing Surveys, 2018, S. 1-42.
https://standards.ieee.org/wp-content/uploads/import/documents/other/ead_v2.pdf

sollten: Grundlegende Menschenrechte, Priorisierung des (menschlichen) Wohlbefindens, Pflicht zur Rechenschaft, Transparenz, Bewusstsein für die Möglichkeit des Technologiemissbrauchs durch Autonome/Intelligente Systeme.[10] Auch die Datenschutz-Grundverordnung (DSGVO) enthält hier relevante Vorgaben.

Der Aspekt der Fairness verlangt, dass ein KI-System keine diskriminierenden Ausgaben zeigt – siehe obiges Beispiel des COMPAS Systems. Der Aspekt der Fairness ist vor allem dann essenziell, wenn die vom KI-System vorgeschlagenen Empfehlungen (recommender system) oder automatisch exekutierten Entscheidungen (automatic decision making), Konsequenzen für die davon betroffenen Personen haben, die ihr Leben substantiell und längerfristig beeinflussen können. Zum Beispiel schlägt die EU-Kommission vor[11], KI-Systeme in unterschiedliche Gruppen gemäß deren Risiko/Konsequenzen für die Sicherheit und Bürgerrechte der Menschen einzuteilen, wobei unterschiedliche Einteilungen mit unterschiedlichen Anforderungen an die KI einhergehen. Es wurden verschiedene Methoden entwickelt, die in der Entwicklung von fairen und diskriminierungsfreien KI-Systemen zum Einsatz kommen. Das Problem hierbei liegt in der Unterdeterminierung des Begriffes „Fairness". Es hat sich gezeigt, dass es schwierig ist, mathematisch zu präzisieren, was Fairness in einem technischen System (z.B. einer KI) bedeutet und wie dieses algorithmisch umgesetzt werden kann – wir diskutieren diese Problematik ausführlicher in dem informatischen Teil dieses Policy Papers.

Erklärungen (durch KI-Systeme bzw. über sie) werden häufig verwendet, um Transparenz und Interpretierbarkeit zu realisieren – das heißt es wird eine Erklärung erstellt und die Grundlage von deren Berechnung in einer dem Menschen zugänglichen Form dargestellt. Es wurden viele unterschiedliche Erklärungsansätze entwickelt, die sich jeweils in der Zielgruppe als auch in dem Ziel der Erklärung unterscheiden. Es gibt beispielsweise Erklärungen, die versuchen, das KI-System als Ganzes zu erklären (globale Erklärung), aber auch Erklärungen, die lediglich einen sehr spezifischen und kleinen Teil adressieren (lokale Erklärung). Erklärungen stellen einen spezifischen Ansatz dar, um eine Interpretierbarkeit von KI-Modellen durch menschliche Betrachter*innen herzustellen.

Sowohl Fairness als auch Erklärungen sind in der Informatik wie auch in anderen Disziplinen separat bereits ausführlich untersucht worden. Das Wechselspiel von Erklärungen und der Eigenschaft der Fairness von Modellen ist aktuell jedoch in wichtigen Teilen unklar. Manche Erklärungen, etwa solche, die auf die Nutzung eines sensiblen Attributs für eine Ausgabe hinweisen, können möglicherweise deutlich machen, dass in einer spezifischen Situation Fairness nicht gegeben ist. Umgekehrt wurde gezeigt, dass Erklärungen von Modellen, nicht nur die eigentlichen KI-Modelle, ebenfalls als unfair wahrgenommen werden können.[12] *Betrachten wir dazu folgendes Beispiel der Kreditvergabe. Wir haben ein System, welches Kreditanträge überprüft und anschließend eine Empfehlung ausgibt: Kredit kann genehmigt werden oder sollte abgelehnt werden. Insbesondere bei den abgelehnten Kreditanträgen gibt*

10 *The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems.* Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems, Version 2. IEEE, 2017. http://standards.ieee.org/develop/indconn/ec/autonomous_systems.html

11 https://ec.europa.eu/info/strategy/priorities-2019-2024/europe-fit-digital-age/excellence-trust-artificial-intelligence_de und https://eur-lex.europa.eu/legal-content/DE/TXT/?uri=CELEX%3A52021PC0206

12 *Artelt A./ Vaquet V./ Velioglu R./ Hinder F./ Brinkrolf J./ Schilling M./ Hammer B.,* Evaluating Robustness of Counterfactual Explanations, IEEE SSCI, 2021.

das System zusätzlich eine Erklärung an. Etwa eine sogenannte kontrafaktische Erklärung, welche beschreibt, was hätte anders sein müssen, damit die Empfehlung „Genehmigen" ausgegeben worden wäre: „Wenn das monatliche Gehalt 500€ höher wäre, hätte das System eine Genehmigung des Kredits empfohlen". Wenn sich jetzt die Erklärung für zwei Personen substantiell unterscheidet, obwohl die Personen sehr ähnlich sind, kann diese Erklärung als unfair empfunden werden. Verwendet man die gegebene kontrafaktische Erklärung als Handlungsempfehlung, dann könnten solche grundsätzlich schwerer zu erfüllenden Änderungen eine Person vor weiteren Versuchen abschrecken, während eine Person, die leichter zu erfüllende Handlungsempfehlungen erhält, diese in die Tat umsetzt. Es gilt also nicht nur zu beachten, dass die Ausgabe der KI selbst als fair wahrgenommen wird, sondern dass auch die Erklärungen selbst als faire Erklärungen für unterschiedliche Personen oder Situationen wahrgenommen werden.

In diesem Policy Paper betrachten und diskutieren wir die Kombination von Erklärung und Fairness – konkret betrachten wir die Fairness von Erklärungen. Insbesondere werden algorithmische, ethische und rechtliche Aspekte beleuchtet und diskutiert.

Fairness von Erklärungen – informatische Perspektive

Zuerst geben wir einen kurzen Überblick über die beiden Begriffe der Fairness und Erklärung von KI im Allgemeinen und betrachten dann ausführlich die Fairness von Erklärungen im Besonderen.

Fairness allgemein

Der Begriff der „Fairness" ist aus informatischer Sicht nicht präzise definiert. Jeder Mensch hat ein intuitives und gesellschaftlich geprägtes Verständnis von Fairness – siehe Absätze zur ethischen und rechtlichen Perspektive. Um die Fairness von einer KI beurteilen und algorithmisch realisieren zu können, muss der Begriff der Fairness mathematisch formalisiert werden. Es hat sich allerdings gezeigt, dass eine Formalisierung (d.h. Mathematisierung) des Begriffes Fairness nicht einfach und nicht nur auf eine Art möglich ist. Es gibt mittlerweile eine Vielzahl von unterschiedlichen Formalisierungen von Fairness[13], die einerseits alle intuitiv und sinnvoll erscheinen, andererseits aber trotzdem einige Situationen, die Menschen als unfair bezeichnen würden, nicht als angemessen abdecken. Außerdem hat sich herausgestellt, dass viele dieser Formalisierungen inkompatibel und widersprüchlich zueinander sind. Es ist unter Umständen unmöglich, mehrere dieser Formalisierungen gleichzeitig zu erfüllen und es besteht daher die Herausforderung, für spezifische KI-Komponenten und -Anwendungen jeweils geeignete sinnvolle Formalisierungen zu identifizieren.

Eine wichtige Unterscheidung bei der Formalisierung von Fairness sind Gruppenfairness[14] und individuelle Fairness[15]. In dem Ansatz der Gruppenfairness wird die Population in mehrere Gruppen – typischerweise sensitive Merkmale, welche keinen Einfluss auf die Ausgabe der KI haben sollen (z.B. Geschlecht, Hautfarbe und Herkunft) – aufgeteilt und es wird verlangt, dass die KI keine systematischen Unterschiede zwischen Mitgliedern dieser beiden Gruppen machen darf. Konkret bedeutet das, dass die KI Ausgaben generieren soll, die unabhängig von der Gruppenzugehörigkeit sind; es sollen also identische Ausgaben generiert werden, wenn sich zwei Personen ausschließlich durch die Gruppenzugehörigkeit unterscheiden. Da andere Merkmale (Wohnort, Beruf etc.) Informationen über das sensitive Merkmal beinhalten können und dieses damit implizit das Ergebnis beeinflusst auch wenn es selbst nicht explizit in der KI verwendet wird, liegt die Herausforderung in der Formalisierung und Realisierung von Unabhängigkeit: welche der nicht als sensitiv markierten Merkmale können in welcher Form für die Entscheidungsfindung berücksichtigt werden, auch wenn sie mit den sensitiven Merkmalen korrelieren? Mathematisch betrachtet gibt es unterschiedliche Möglichkeiten, Unabhängigkeiten relevanter Größen in diesem Kontext zu spezifizieren – dies führt wiederum zu der bereits erwähnten Vielzahl an unterschiedlichen möglichen Formalisierungen von Fairness. Diese widersprechen sich teilweise (beweisbar). Dies liegt unter anderem daran, dass diese Formalisierungen die Mengenverhältnisse von Vertretern der einzelnen Gruppen etwa unter den positiv bewerteten Empfehlungen oder auch den (in der Praxis vorwiegend unvermeidbaren) Fehlern des Algorithmus messen, statt

13 *Mehrabi N./ Morstatter F./ Saxena N./ Lerman K./ Galstyan A.*, A survey on bias and fairness in machine learning, arXiv preprint arXiv:1908.09635, 2019.

14 *Caton S./ Haas C.*, Fairness in machine learning: A survey, arXiv preprint arXiv:2010.04053, 2020.

15 *Dwork C./ Hardt M./ Pitassi T./ Reingold O./ Zemel R.*, Fairness through awareness, Proceedings of the 3rd innovations in theoretical computer science conference, 2012.

kausale Beziehungen zu modellieren. Das letztere ist in der Praxis oft nicht möglich, da die genauen kausalen Zusammenhänge unbekannt sind und durch die beobachteten Daten nicht vollständig charakterisiert werden.

Im Gegensatz zu Gruppenfairness steht bei der individuellen Fairness das Individuum und nicht eine bestimmte Gruppe im Fokus. Es muss insbesondere kein sensitives Attribut spezifiziert werden, sondern es sollen allgemein ähnliche Individuen ähnliche Resultate durch die KI hervorrufen. Das heißt eine Empfehlung zur Kreditwürdigkeit soll für Personen, die ähnliche Merkmale besitzen, gleich ausfallen. Hier besteht die Herausforderung darin, Ähnlichkeit von Individuen sinnvoll zu formalisieren. Ein Bezug zur Gruppenfairness ergibt sich, wenn die gewählte Ähnlichkeit sensitive Attribute entsprechend berücksichtigt, aber es sind hier auch andere Formalisierungen möglich. Auch bei der individuellen Fairness entsprechen unterschiedliche Formalisierungen unterschiedlichen Fairnesskriterien. Diese sind jedoch weniger problematisch als im Fall der Gruppenfairness, denn die genaue Formalisierung von individueller Fairness wird oft durch die Spezifizierung, welche Ähnlichkeiten keinen Einfluss auf die Ausgabe der KI haben sollen, für den jeweiligen Anwendungsbereich angepasst. Daher wird in diesem Bereich eine universelle, über verschiedene Domänen transferierende Formalisierung nicht erwartet.

In diesem Policy Paper befassen wir uns primär mit der individuellen Fairness, weil diese sehr allgemein angewandt werden kann und typischerweise von Benutzer*innen als intuitiv akzeptiert wird.

Erklärungen von KI-Systemen

Genau wie der Begriff der Fairness, so ist auch der Begriff „Erklärung“ oder allgemeiner „Interpretierbarkeit von KI-Modellen“ aus informatischer Sicht nicht wohl definiert.[16] Auch aus kognitionswissenschaftlicher Sicht ist nicht vollständig klar, was genau eine Erklärung eigentlich ist. Besonders prägnant drückt dies die Aussage „You know it when you see it“[17] aus (Deutsch: „Sie wissen es, wenn Sie es sehen“ – „es“ meint hier die Erklärung). Eine besondere Herausforderung in dem Kontext stellt die Tatsache dar, dass insbesondere Daten-getriebene KI-Modelle wie Modelle des maschinellen Lernens in Kontexten eingesetzt werden, in denen eine wissensbasierte oder exakte logische Modellierung nicht möglich ist. KI-Modelle sind daher nur bedingt und etwaig unter Verlusten der Genauigkeit durch symbolische oder logische Entitäten beschreibbar. Dies limitiert auch die unmittelbare Übertragbarkeit von etablierten Modellen der Erklärung wie dem deduktiv-nomologischen HO-Modell.[18] Entsprechend gibt es in der Informatik viele unterschiedliche Arten von Erklärungen von KI-Systemen.[19] Viele Erklärungsansätze versuchen die Wichtigkeit beziehungsweise den Beitrag einzelner Bestandteile der Eingabe für die Ausgabe der KI zu bestimmen und zu visualisieren. Ein sehr populärer Erklärungsansatz sind kontrafaktische

16 *Doshi-Velez F./ Kim B.*, Towards a rigorous science of interpretable machine learning, arXiv preprint arXiv:1702.08608, 2017.

17 *Offert, F.*, “I know it when I see it." Visualization and Intuitive Interpretability, arXiv preprint arXiv:1711.08042 2017.

18 *C. G. Hempel*: Aspects of Scientific Explanation and other Essays in the Philosophy of Science. New York/London 1965. Kap. 10: Studies in the Logic of Explanation (dt. Aspekte wissenschaftlicher Erklärung. Berlin/New York 1977)

19 *Guidotti R./ Monreale A./ Ruggieri S./ Turini F./ Giannotti F./ Pedreschi D*., A survey of methods for explaining black box models, ACM Computing Surveys, 2018, S. 1-42.

Erklärungen[20] – kontrafaktische Erklärungen selbst sind schon lange bekannt und haben ihre Ursprünge nicht in der Informatik, sondern in den Kognitionswissenschaften. Eine kontrafaktische Erklärung drückt aus, was bei der Eingabe (auf der Seite des Individuums) anders sein müsste, um eine bestimmte Ausgabe von der KI zu bekommen. In der Informatik wird diese Art von Erklärung dann auch als konkrete Handlungsempfehlung verstanden, die Nutzer*innen nahelegt, wie ein bestimmtes Ziel erreicht werden kann. Diese Handlungsempfehlung ist der Grund, warum kontrafaktische Erklärungen so populär sind. Da der Erklärung allerdings oft keine kausalen Modelle zugrunde liegen, ist nicht sichergestellt, dass diese Interpretation in der Praxis valide ist. Es gibt Evidenz dafür, dass von Menschen favorisierte und gebildete Erklärungen oft von kontrafaktischer Natur sind.[21]

Beispiel: *Stellen wir uns vor, eine Person hat sich bei einer Bank um einen Kredit beworben und dieser Kreditantrag wurde abgelehnt. Diese Person möchte jetzt wissen, warum der Kreditantrag abgelehnt worden ist. Eine mögliche kontrafaktische Erklärung könnte sein: Der Kredit wäre genehmigt worden, wenn die Person 500€ mehr im Monat verdienen würde.* Diese Art von Erklärung gibt sehr konkrete Handlungsempfehlungen, was diese Person tun müsste, damit der Kreditantrag beim nächsten Mal genehmigt werden würde.

Die Formalisierung und Berechnung von solchen kontrafaktischen Erklärungen von KI-Systemen sind einerseits teilweise schon sehr gut untersucht, andererseits gibt es in vielen Bereichen zum Beispiel der Fairness und Plausibilität noch erheblichen Forschungsbedarf.[22] Wegen ihrer Popularität und dem einfachen Verständnis beschränken wir uns in diesem Policy Paper auf die Betrachtung von kontrafaktischen Erklärungen für die Erklärung von KI-Systemen.

Fairness von Erklärungen

Nachdem wir Fairness und Erklärbarkeit von KI-Systemen getrennt voneinander betrachtet haben, werfen wir nun einen Blick auf die Kombination dieser beiden Aspekte. Konkret betrachten wir die individuelle Fairness von (kontrafaktischen) Erklärungen.[23] Während Fairness und Erklärungen als separate Aspekte in der Forschung schon seit einiger Zeit ausführlich untersucht werden – trotzdem gibt es, wie bereits erwähnt, in beiden Fällen noch einigen Forschungsbedarf –, ist die Fairness von Erklärungen ein deutlich weniger gut untersuchter Aspekt in der KI-Forschung.

Die individuelle Fairness von Erklärungen verlangt, dass ähnliche Individuen ähnliche Erklärungen bekommen. Insbesondere kontrafaktische Erklärungen können als unfair wahrgenommen werden, wenn ähnliche Individuen stark unterschiedliche Erklärungen bekommen, weil die konkreten Erklärungen in Form von Handlungsempfehlungen schwerer oder einfacher realisierbar sein könnten.[24]

20 *Wachter S./ Mittelstadt B./ Russell C.*, Counterfactual explanations without opening the black box: Automated decisions and the GDPR, Harv. JL & Tech. 31, 2017, S. 841-887.

21 *Miller, Tim.* "Explanation in artificial intelligence: Insights from the social sciences." Artificial intelligence 267 (2019): 1-38.

22 *Verma S./ Dickerson J./ Hines K.,* Counterfactual explanations for machine learning: A review, arXiv preprint arXiv:2010.10596, 2020.

23 *v. Kügelgen J./ Karimi A.-H./ Bhatt U./ Valera I./ Weller A./ Schölkopf B.*, On the fairness of causal algorithmic recours, arXiv preprint arXiv:2010.06529, 2020.

24 *Artelt A./ Vaquet V./ Velioglu R./ Hinder F./ Brinkrolf J./ Schilling M./ Hammer B.,* Evaluating Robustness of Counterfactual Explanations, IEEE SSCI, 2021.

Siehe folgendes Beispiel bezüglich einer Kreditvergabe aus der Einleitung. *Bei den abgelehnten Kreditanträgen gibt das System zusätzlich eine kontrafaktische Erklärung aus, die angibt, was hätte anders sein müssen, damit die Empfehlung „Genehmigen" ausgegeben worden wäre:„Wenn das monatliche Gehalt 500€ höher wäre, hätte das System eine Genehmigung des Kredits empfohlen". Wenn jetzt beispielsweise Frauen grundsätzlich schwieriger zu erfüllende Änderungen wie zusätzliche Sicherheiten durch Wertanlagen, welche nicht einfach beschafft werden können und entweder vorhanden sind oder nicht, vorgeschlagen werden als Männern, kann man die Erklärungen als unfair wahrnehmen.* Dabei sind in der Praxis solche unfairen Erklärungen für eine einzelne Person in der Regel nicht einfach erkennbar, da einzelne Personen sich meist nur mit einer sehr kleinen Anzahl anderer Personen vergleichen können. Nichtsdestotrotz können solche unfairen Erklärungen substantielle Konsequenzen für die einzelne Person haben. Daher ist es wichtig, die Fairness von algorithmischen Entscheidungen und auch von sie begleitenden Erklärungen zu fordern. Auch hier ist die Herausforderung wieder die genaue Formalisierung, was Ähnlichkeit von Individuen und von Erklärungen bedeutet. An dieser Stelle sei bemerkt, dass auch andere Arten von Erklärungen unfair sein können.[25] Wir beschränken uns hier jedoch auf kontrafaktische Erklärungen, weil Unfairness bei diesen offensichtlicher und einfacher zu verstehen ist.

Es gibt in der Literatur dargestellte Erkenntnisse, die zeigen, dass die Standardansätze und Formalisierungen für das Berechnen von kontrafaktischer Erklärung anfällig für individuelle Unfairness sind. Konkret steigt zudem etwa die zu erwartende Unfairness, je mehr Attribute/Merkmale die Eingabe hat.[26] Das bedeutet, dass es ohne weitere Maßnahmen sehr wahrscheinlich ist, dass eine generierte kontrafaktische Erklärung auf die eine oder andere Art unfair ist, wobei wir uns auf den Begriff der individuelle Fairness beziehen.

Maßnahmen, die die individuelle Fairness von kontrafaktischen Erklärungen garantieren, sind noch in der Entwicklung und zurzeit Gegenstand aktiver Forschungsarbeiten. Ein vielversprechender Ansatz ist der Versuch, individuelle Fairness durch zusätzliche Plausibilitätsbedingungen zu erreichen.[27] Der Begriff der Plausibilität referiert dabei auf die zusätzliche Bedingung, dass eine kontrafaktische Erklärung in der Praxis ausführbare Handlungen vorschlägt. Die Mathematisierung und algorithmische Realisierung der Anforderung der Plausibilität von kontrafaktischen Erklärungen ist oftmals noch ein Problem, denn viele Methoden für das Berechnen von kontrafaktischen Erklärungen liefern aktuell nicht zwingend eine plausible und durchführbare Erklärung – zum Beispiel könnten Teile der Handlungsempfehlung für das betreffende Individuum unmöglich zu realisieren sein.[28] In der aktuellen Forschung findet sich eine Anzahl an Methoden, die Aspekte der Plausibilität

25 *Alvarez-Melis D./ Jaakkola T. S.,* Towards robust interpretability with self-explaining neural networks." arXiv preprint arXiv:1806.07538, 2018; *Anders C. J./ Pasliev P./ Dombrowski A.-K./ Müller K.-R./ Kessel P.,* Fairwashing explanations with off-manifold detergent, International Conference on Machine Learning. PMLR, 2020.

26 *Artelt A./ Vaquet V./ Velioglu R./ Hinder F./ Brinkrolf J./ Schilling M./ Hammer B*., Evaluating Robustness of Counterfactual Explanations, IEEE SSCI, 2021.

27 *Artelt A./ Vaquet V./ Velioglu R./ Hinder F./ Brinkrolf J./ Schilling M./ Hammer B.,* Evaluating Robustness of Counterfactual Explanations, IEEE SSCI, 2021.

28 *Artelt A./ Hammer B.,* Convex density constraints for computing plausible counterfactual explanations, International Conference on Artificial Neural Networks. Springer, Cham, 2020.

realisieren.[29] Erste Ergebnisse deuten darauf hin, dass plausible kontrafaktische Erklärungen, in denen explizit etwa durch die Garantie einer gewissen (auf den beobachteten Daten basierenden) Wahrscheinlichkeit auf Plausibilität der Handlungsempfehlung geachtet worden ist, weniger anfällig für Unfairness sind als kontrafaktische Erklärungen, die ohne solche zusätzlichen Bedingungen generiert wurden.[30] Es ist noch offen, inwieweit dies vom spezifischen KI-System abhängt. Darüber hinaus gehend handelt es sich derzeit um empirische Beobachtungen und keine formalen Beweise. Zudem sind andere Ansätze denkbar, die individuelle Fairness von KI-Systemen und deren Erklärungen ermöglichen.

29 *Van Looveren A./ Klaise J.,* Interpretable counterfactual explanations guided by prototypes, arXiv preprint arXiv:1907.02584, 2019; *Poyiadzi, Rafael, et al.* "FACE: Feasible and actionable counterfactual explanations." Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. 2020.

30 *Artelt A./ Vaquet V./ Velioglu R./ Hinder F./ Brinkrolf J./ Schilling M./ Hammer B.,* Evaluating Robustness of Counterfactual Explanations." IEEE SSCI, 2021.

Fairness im Datenschutzrecht

Fairness im alten und im neuen Datenschutzrecht

Nach Art. 5 Abs. 1 lit. A der Datenschutz-Grundverordnung (DSGVO) ist Fairness einer der Grundsätze für die Verarbeitung personenbezogener Daten, wenngleich die deutsche Sprachfassung den englischen Begriff der Fairness durch das Begriffspaar „Treu und Glauben“[31] ersetzt.[32] Damit knüpft die Verordnung an Art. 6 Abs. 1 lit. A der Vorgängerrichtlinie 95/46/EG an. Anders als die Datenschutz-Grundverordnung verknüpfte die Richtlinie 95/46/EG Fairness (bzw. auch hier in der deutschen Fassung der Richtlinie „Treu und Glauben“) ganz explizit mit Transparenz und der Bereitstellung von Informationen an die von einer Verarbeitung ihrer personenbezogenen Daten betroffenen Personen. So forderten Art. 10 lit. C und Art. 11 lit. C der Richtlinie, der betroffenen Person weitere Informationen bereitzustellen (etwa zu den Empfängern der Daten), sofern diese Informationen „unter Berücksichtigung der spezifischen Umstände, unter denen die Daten erhoben werden, notwendig sind, um gegenüber der betroffenen Person eine Verarbeitung nach Treu und Glauben zu gewährleisten“ (englisch: „to guarantee fair processing“). Auch Art. 8 Abs. 2 Satz 1 der Charta der Grundrechte der Europäischen Union bestimmt, dass personenbezogene Daten nur nach Treu und Glauben (bzw. „fairly“) verarbeitet werden dürfen.

Eine Erläuterung zu Datenverarbeitung nach Treu und Glauben beziehungsweise fairer Daten-verarbeitung erhält die Datenschutz-Grundverordnung nicht (mehr). Erwägungsgrund 38 der Richtlinie 95/46/EG enthielt noch folgende Erklärung: „Datenverarbeitung nach Treu und Glauben setzt voraus, dass die betroffenen Personen in der Lage sind, das Vorhandensein einer Verarbeitung zu erfahren und ordnungsgemäß und umfassend über die Bedingungen der Erhebung informiert zu werden, wenn Daten bei ihnen erhoben werden.“ Damit scheint faire Datenverarbeitung vor allem transparente Datenverarbeitung zu bedeuten. Anders als die Datenschutz-Grundverordnung kannte die Richtlinie jedoch keinen expliziten Grundsatz der Transparenz, der hier vielmehr als Voraussetzung für die Verarbeitung nach Treu und Glauben verstanden wurde.[33] In der Datenschutz-Grundverordnung wird Transparenz als eigenständiger Grundsatz neben Treu und Glauben beziehungsweise Fairness genannt. Damit ist unklar, welche Rolle der Grundsatz von Treu und Glauben in der Datenschutz-Grundverordnung spielt und in welcher Beziehung er zum Grundsatz der Transparenz steht.[34] Darauf, dass Transparenz auch weiterhin mit Fairness verwoben ist, weist die gemeinsame Nennung der Grundsätze von Rechtmäßigkeit, Verarbeitung nach Treu und Glauben und

31 Treu und Glauben ist ein aus dem Zivilrecht bekannter Grundsatz, der zu redlichem Verhalten verpflichtet.

32 Zur Herstellung sprachlicher Konsistenz und zur Vermeidung einer Verwechslung mit dem zivilrechtlichen Rechtsgrundsatz von Treu und Glauben wird vorgeschlagen, die Wendung „Verarbeitung nach Treu und Glauben“ in Art. 5 Abs. 1 lit. a DSGVO der deutschen Sprachfassung der Datenschutz-Grundverordnung durch „Fairness“ zu ersetzen. S. *Roßnagel A./ Geminn C.,* Datenschutz-Grundverordnung verbessern, Baden-Baden, 2020, S. 101, 116.

33 Artikel-29-Datenschutzgruppe, Leitlinien in Bezug auf die Einwilligung gemäß Verordnung 2016/679, WP259 rev.01, 10. April 2018, Rn. 3.

34 Es wird deshalb vorgeschlagen, den Grundsatz über eine Änderung von Erwägungsgrund 39 DSGVO zu präzisieren und klar von den Grundsätzen von der Rechtmäßigkeit und der Transparenz abzugrenzen. S. *Roßnagel A./ Geminn C.,* Datenschutz-Grundverordnung verbessern, Baden-Baden, 2020, S. 101.

Transparenz in Art. 5 Abs. 1 lit. A DSGVO hin sowie die Wendung „faire und transparente Verarbeitung“[35].[36]

Letztlich könnte und sollte der Grundsatz der Verarbeitung nach Treu und Glauben in der Datenschutz-Grundverordnung die Rolle einer Auffangklausel spielen, die ungerechte Praxisergebnisse verhindern soll – zum Beispiel wenn eine Verarbeitung zwar formell und materiell rechtmäßig erfolgt, dabei aber in unbilliger Weise die Macht des Datenverarbeiters zum Nachteil der betroffenen Person ausgenutzt wurde.[37] Damit bleibt jedoch weiter offen, welche konkreten Folgen Fairness gerade bezogen auf Erklärungen von KI-gestützten Entscheidungssystemen gegenüber den Nutzenden haben kann. Der Europäische Datenschutzausschuss hat zum Grundsatz von Treu und Glauben festgestellt, er umfasse „unter anderem die Anerkennung der vernünftigen Erwartungen der betroffenen Personen, die Beachtung etwaiger nachteiliger Folgen, die die Verarbeitung für sie haben kann, und die Berücksichtigung der Beziehung und der potenziellen Auswirkungen eines Ungleichgewichts zwischen ihnen und dem Verantwortlichen“.[38] Dies kann auf die Feststellung verkürzt werden, dass solche Verhaltensweisen als unfair zu werten sind, „die Vertrauen missbrauchen“.[39] Zentral ist auch die Feststellung, dass der Grundsatz von Treu und Glauben auch „nicht dadurch negiert oder auf irgendeine Weise abgeschwächt“ wird, dass eine Einwilligung eingeholt wird.[40]

Pflichten des Verantwortlichen und Rechte der betroffenen Person

Das Datenschutzrecht fordert von dem für die Verarbeitung personenbezogener Daten Verantwortlichen eine Reihe von transparenzbezogenen Maßnahmen und hier insbesondere die Bereitstellung von Informationen. Umgekehrt steht der betroffenen Person unter anderem ein Anrecht zu, zu erfahren „nach welcher Logik die automatisierte Verarbeitung personenbezogener Daten erfolgt und welche Folgen eine solche Verarbeitung haben kann, zumindest in Fällen, in denen die Verarbeitung auf Profiling beruht“.[41] Der betroffenen Person müssen nach Art. 13 Abs. 2 lit. F und 14 Abs. 2 lit. G DSGVO im Falle einer automatisierten Entscheidungsfindung „aussagekräftige Informationen über die involvierte Logik sowie die Tragweite und die angestrebten Auswirkungen einer derartigen Verarbeitung für die betroffene Person“ bereitgestellt werden. Dies wird jedoch auf Situationen beschränkt, in denen eine automatisierte Entscheidung der betroffenen Person „gegenüber rechtliche Wirkung entfaltet oder sie in ähnlicher Weise erheblich beeinträchtigt“.[42] Wegen

35 Erwägungsgründe 39 Satz 4, 60 Satz 1 und 2, 71 Satz 6 sowie Art. 13 Abs. 2, 14 Abs. 2, 40 Abs. 2 lit. a DSGVO.

36 S. zur Verknüpfung *Malgieri G.*, The concept of Fairness in the GDPR, in: Proceedings of FAT* '20, January 27–30, 2020. ACM, New York, NY, USA, S. 154 (155 f.); Nach Artikel-29-Datenschutzgruppe, Leitlinien in Bezug auf die Einwilligung gemäß Verordnung 2016/679, WP259 rev.01, 10. April 2018, Rn. 2 sind die Grundsätze untrennbar verbunden.

37 S. *Roßnagel A./ Geminn C.,* Datenschutz-Grundverordnung verbessern, Baden-Baden, 2020, S. 46 f.; *Roßnagel A.*, in: *Simitis/Hornung/Spiecker gen. Döhmann*, Datenschutzrecht, 2019, Art. 5 DSGVO, Rn. 47.

38 Leitlinien 2/2019 für die Verarbeitung personenbezogener Daten gemäß Artikel 6 Absatz 1 Buchstabe b DSGVO im Zusammenhang mit der Erbringung von Online-Diensten für betroffene Personen, Version 2.0, 8. Oktober 2019, Rn. 12; Artikel-29-Datenschutzgruppe, Leitlinien für Transparenz gemäß der Verordnung 2016/679, WP2060 rev.01, 11. April 2018, Rn. 30 spricht von „berechtigten Erwartungen der betroffenen Person“.

39 *Roßnagel A.,* in: Simitis/Hornung/Spiecker gen. Döhmann, Datenschutzrecht, 2019, Art. 5 DSGVO, Rn. 47.

40 Artikel-29-Datenschutzgruppe, Leitlinien in Bezug auf die Einwilligung gemäß Verordnung 2016/679, WP259 rev.01, 10. April 2018, Rn. 1.

41 Erwägungsgrund 63 Satz 3 DSGVO.

42 Art. 22 Abs. 1 DSGVO.

dieser starken Verengung wird eine Ausweitung gefordert, die auch Situationen erfasst, in denen beispielsweise eine automatisierte Entscheidung erfolgt, einer Person einen höheren Kaufpreis abzuverlangen.[43] Es sollen also auch Situationen erfasst werden, in denen die Wirkung der automatisierten Entscheidung geringer ist als eine rechtliche Wirkung.

Fairness im Kontext von automatisierten Entscheidungen

Fairness im Kontext von automatisierten Entscheidungen adressiert Erwägungsgrund 71 Satz 6 DSGVO, wo es heißt, um „der betroffenen Person gegenüber eine faire und transparente Verarbeitung zu gewährleisten, sollte der für die Verarbeitung Verantwortliche geeignete mathematische oder statistische Verfahren für das Profiling verwenden, technische und organisatorische Maßnahmen treffen, mit denen in geeigneter Weise insbesondere sichergestellt wird, dass Faktoren, die zu unrichtigen personenbezogenen Daten führen, korrigiert werden und das Risiko von Fehlern minimiert wird“. Darüber hinaus sollt der Verantwortliche „personenbezogene Daten in einer Weise sichern, dass den potenziellen Bedrohungen für die Interessen und Rechte der betroffenen Person Rechnung getragen wird und unter anderem verhindern, dass es gegenüber natürlichen Personen aufgrund von Rasse, ethnischer Herkunft, politischer Meinung, Religion oder Weltanschauung, Gewerkschaftszugehörigkeit, genetischer Anlagen oder Gesundheits-zustand sowie sexueller Orientierung zu diskriminierenden Wirkungen oder zu einer Verarbeitung kommt, die eine solche Wirkung hat“.[44] Die Artikel-29-Datenschutzgruppe führte in der Folge am Beispiel einer Bonitätsprüfung aus, der Verantwortliche müsse in der Lage sein, der betroffenen Person seine Einschätzung zur Bonität „und die zugrunde liegenden Überlegungen zu erklären. Der Verantwortliche erläutert, dass er dadurch faire und verantwortungsvolle Entscheidungen bezüglich der Kreditvergabe treffen kann“.[45] Für ein „unfaires Profiling“ unter Verstoß gegen den Grundsatz von Treu und Glauben wird folgendes Beispiel genannt: „Ein Datenbroker verkauft Verbraucherprofile an Finanzunternehmen, ohne dass die Verbraucher einwilligen oder wissen, welche Daten davon betroffen sind. Bei diesen Profilen werden die Verbraucher in Kategorien eingeteilt (wie zum Beispiel ‚lebt auf dem Land und kommt knapp über die Runden‘, ‚ums Überleben kämpfender Angehöriger einer ethnischen Minderheit in zweitklassiger Stadt‘, ‚Schwerer Start‘, ‚jung und alleinerziehend‘) oder mit Punkten ‚bewertet‘, wobei deren finanzielle Schwäche im Mittelpunkt steht. Die Finanzunternehmen bieten diesen Verbrauchern dann Überbrückungskredite und andere eher ungewöhnliche Finanzdienstleistungen an (kostspielige Kredite und andere finanziell riskante Produkte).“[46]

Folgerungen

43 *Roßnagel A./ Geminn C.,* Datenschutz-Grundverordnung verbessern, Baden-Baden, 2020, S. 82 ff.

44 S. auch Artikel-29-Datenschutzgruppe, Leitlinien zu automatisierten Entscheidungen im Einzelfall einschließlich Profiling für die Zwecke der Verordnung 2016/679, WP251 rev.01, 6. Februar 2018, wo jedoch von „fairem, nichtdiskriminierendem und sachlich richtigem Profiling“ gesprochen wird (ebd., S. 16), wodurch wiederum eine Trennung von Fairness und Nichtdiskriminierung impliziert wird. An anderer Stelle werden Fairness und Nichtdiskriminierung jedoch gleichgestellt (S. 11).

45 S. auch Artikel-29-Datenschutzgruppe, Leitlinien zu automatisierten Entscheidungen im Einzelfall einschließlich Profiling für die Zwecke der Verordnung 2016/679, WP251 rev.01, 6. Februar 2018, S. 28.

46 Artikel-29-Datenschutzgruppe, Leitlinien zu automatisierten Entscheidungen im Einzelfall einschließlich Profiling für die Zwecke der Verordnung 2016/679, WP251 rev.01, 6. Februar 2018, S. 11. Die Gruppe übernimmt das Beispiel dabei aus einem Bericht des U.S. Senats.

Die Pflicht zur Transparenz der Datenverarbeitung bezieht sich nach der Artikel-29-Datenschutzgruppe auf „die Unterrichtung der betroffenen Personen im Zusammenhang mit der nach Treu und Glauben erfolgenden Verarbeitung“ als einer von drei Kernbereichen der Transparenz.[47] Quintessenz der Bedeutung des Grundsatzes von Treu und Glauben (und damit für eine faire Bereitstellung von Informationen) soll sein, dass die betroffene Person „nicht später von der Art und Weise überrascht werden sollte, in der ihre personenbezogenen Daten verwendet worden sind“.[48] Die betroffene Person soll insbesondere „bei komplexen, technischen oder unerwarteten Verarbeitungsvorgängen neben der Bereitstellung der nach den Artikeln 13 und 14 vorgeschriebenen Informationen […] gesondert und eindeutig formuliert“ über die wichtigsten Folgen der Verarbeitung aufgeklärt werden.[49] Ferner darf durch den Sprachgebrauch bei den bereitgestellten Informationen – beispielsweise durch die Verwendung von Modalverben wie „könnte“ – der Grundsatz von Treu und Glauben nicht untergraben werden.[50] Auch eine rechtzeitige Information der betroffenen Person soll sich neben der Transparenzpflicht auch auf den Grundsatz von Treu und Glauben stützen; zudem ist die betroffene Person aktiv über Änderungen und Aktualisierungen und auch über deren Auswirkungen zu informieren.[51] Die bereitgestellten Informationen sollten auch „Angaben über die Verarbeitung, welche sich am stärksten auf die betroffene Person auswirkt, und die Verarbeitungsvorgänge, mit denen Letztere ggfs. Nicht gerechnet hat, enthalten“.[52]

In der Gesamtschau nimmt Fairness bezogen auf Erklärungen aus datenschutzrechtlicher Sicht vor allem die Rolle einer Untermauerung bestehender Transparenzpflichten ein. Nach Art. 12 Abs. 1 Satz 1 DSGVO sind Informationen und Mitteilungen an die betroffene Person, „die sich auf die Verarbeitung beziehen, in präziser, transparenter, verständlicher und leicht zugänglicher Form in einer klaren und einfachen Sprache zu übermitteln; dies gilt insbesondere für Informationen, die sich speziell an Kinder richten“. Darüber hinaus besteht jedoch gerade bei automatisierten Entscheidungen auch ein Bezug zum Thema Diskriminierung, dessen Konturen allerdings höchst unscharf bleiben. Klar ist, dass Fairness beziehungsweise dem datenschutzrechtlichen Grundsatz von Treu und Glauben ein eigenständiger Gehalt zukommen muss, der vom Verordnungsgeber auch explizit gewollt ist. Solange aber Gerichte und hier insbesondere der Europäische Gerichtshof Fairness im Sinne der Datenschutz-Grundverordnung nicht als entscheidenden Faktor für die Rechtswidrigkeit einer Datenverarbeitung mobilisieren und dem Grundsatz so Konturen verleihen, bleiben seine praktischen Auswirkungen gering.

47 Artikel-29-Datenschutzgruppe, Leitlinien für Transparenz gemäß der Verordnung 2016/679, WP2060 rev.01, 11. April 2018, Rn. 2.

48 Laut Artikel-29-Datenschutzgruppe, Leitlinien für Transparenz gemäß der Verordnung 2016/679, WP2060 rev.01, 11. April 2018, Rn. 10 als wichtigem Aspekt des Grundsatzes von Treu und Glauben.

49 Laut Artikel-29-Datenschutzgruppe, Leitlinien für Transparenz gemäß der Verordnung 2016/679, WP2060 rev.01, 11. April 2018, Rn. 10.

50 Laut Artikel-29-Datenschutzgruppe, Leitlinien für Transparenz gemäß der Verordnung 2016/679, WP2060 rev.01, 11. April 2018, Rn. 13.

51 Laut Artikel-29-Datenschutzgruppe, Leitlinien für Transparenz gemäß der Verordnung 2016/679, WP2060 rev.01, 11. April 2018, Rn. 27, 29, 31, 48.

52 Laut Artikel-29-Datenschutzgruppe, Leitlinien für Transparenz gemäß der Verordnung 2016/679, WP2060 rev.01, 11. April 2018, Rn. 36.

Fairness aus ethischer Perspektive

Der Begriff der Fairness

> Did you ensure an adequate working definition of "fairness" that you apply in designing AI systems? Is your definition commonly used? Did you consider other definitions before choosing this one? Did you ensure a quantitative analysis or metrics to measure and test the applied definition of fairness?[53]

An dieser Handlungsempfehlung für Entwickler*innen von Systemen künstlicher Intelligenz der High-Level Expert Group on AI (AI HLEG) der europäischen Kommission lässt sich bereits ablesen, dass es durchaus denkbar ist, bei der Arbeit mit KI unterschiedlichste Definitionen des Fairnessbegriffs zu implementieren.
Unterschiedliche Disziplinen und unterschiedliche Subsysteme haben auch unterschiedliche Konzepte von Fairness – konkret wird es an dieser Stelle selten. Fairness erscheint als normativer Begriff, der eine Situation in ihrer Ganzheit beschreibt; so kann der Ablauf eines Fußballspiels beispielsweise abschließend als fair bezeichnet werden, wobei nicht jeder Spielzug einzeln betrachtet wird. Fairness scheint intuitiv; selbst in einschlägigen Lexika finden sich keine Begriffsdefinitionen, sondern meist der Verweis auf „Gerechtigkeit". Diese ist eng mit der Fairness verknüpft, lässt sich ein gesellschaftlicher Zustand schließlich dann als gerecht bezeichnen, wenn er fair ist, was bedeutet, dass ihm alle Beteiligten als gleichberechtigte, vernünftige und freie Personen zustimmen können.[54] Wie bereits in der Literatur beschrieben[55], gibt es auch bei der Entwicklung und Programmierung von Algorithmen keine einheitliche Auffassung darüber, was solche als „fair" klassifiziert und wie sich sicherstellen lässt, dass diese zu „fairen" Ergebnissen gelangen oder gar „faire" Erklärungen für die getroffenen Entscheidungen liefern. Der Fairnessbegriff wird hier uneigentlich gebraucht, da ein Algorithmus für sich nicht fair sein oder fair handeln kann.

„Menschen-wie-du-Berechnungen" – fair?

Individuelle Fairness durch plausible kontrafaktische Erklärungen zu erreichen und dabei Handlungsempfehlungen zu geben, die umsetzbar sind – dies ist das Ziel von fairen Erklärungen der Informatik.[56] Doch Maschinen sind von sich aus nicht in der Lage, Gerechtigkeitsfragen zu berücksichtigen, sie können lediglich Zusammenhänge zwischen bereits vorhandenen Daten feststellen. Die Aufgabe der Reflexion liegt beim Menschen als moralischem Subjekt, nur dieser kann Daten, Entscheidungen und Erklärungen auswerten und mit Gründen beurteilen, was gerecht und „fair" ist.[57] Geschieht dies jedoch nicht, so können Algorithmen ähnlich selbsterfüllender Prophezeiungen Ungleichheiten

53 Vgl. *High-Level Expert Group on AI (AI HLEG),* Ethics Guidelines for Trustworthy AI, Brüssel, 2019.
54 Vgl. *Dimitriou M./Schweiger G.,* Fairness und Fairplay. Eine interdisziplinäre Annährung, in: *Dimitriou/Schweiger (Hrsg.),* Fairness und Fairplay. Interdisziplinäre Perspektiven, Wiesbaden, 2015, S. 15.
55 Vgl. *André Artelt, Barbara Hammer:* Efficient computation of counterfactual explanations and counterfactual metrics of prototype-based classifiers. Neurocomputing 470: 304-317 (2022)
56 Vgl. *André Artelt, Barbara Hammer*: Convex Density Constraints for Computing Plausible Counterfactual Explanations. ICANN (1) 2020: 353-365
57 Vgl. *Martini M.,* Blackbox Algorithmus. Grundfragen einer Regulierung Künstlicher Intelligenz, Berlin, 2019, S. 49.

reproduzieren und so diskriminieren, ohne, dass dies von Entwickelnden, Forschenden oder Anwendenden beabsichtigt war.[58] So entsteht mitunter ein Vergrößerungsglas für Vorurteile, Stigmata und Diskriminierung.

Wie „fair" ist es außerdem, wie beispielsweise O'Neil es nennt, solche „Menschen-wie-du"-Berechnungen[59] in Form einer kontrafaktischen Erklärung auszugeben? So begnügt man sich damit, zu fragen, wie sich ähnliche Menschen in der Vergangenheit verhielten beziehungsweise welche Ergebnisse und Entscheidungen ihre Daten hervorgerufen haben. Nach O'Neil müsste an dieser Stelle idealerweise gefragt werden, wie sich eben dieses Individuum, für das die Entscheidung vom Algorithmus getroffen wird, in der Vergangenheit verhalten hat.[60] Eine weitere Frage muss an dieser Stelle lauten: Wann wollen wir Unterschiede, wann wollen wir Ähnlichkeiten in Daten und Attributen erkennen und wann und wie wollen wir Menschen in Gruppen einteilen? Sensitive Attribute spielen unter Umständen eine wichtige Rolle bei der Entscheidungsfindung des Algorithmus, wie bei der Mikrokreditvergabe in Indien, die hauptsächlich an Frauen erfolgt, da diese „zuverlässiger" seien.[61]

Die substanzielle und prozedurale Dimension der Fairness

Entscheidungen und die damit zusammenhängenden Erklärungen Künstlicher Intelligenzen werden intuitiv oftmals als fairer wahrgenommen, da Algorithmen gemeinhin als objektiver gelten – dieser Objektivitätsbias muss immer wieder aufs Neue beleuchtet und darf keinesfalls außer Acht gelassen werden. Algorithmen sind niemals neutral, sie reflektieren die Wertvorstellungen ihrer Programmierer*innen und Auftraggeber*innen. Nachbessern kann man an dieser Stelle nur schwer, vielmehr muss die Rechtslage angepasst werden. Beispielsweise sollte es für die Betroffenen möglich sein, eine „automatische" Diskriminierung durch einen Algorithmus feststellen zu können und sollte eine solche vorliegen, entsprechend dagegen vorgehen zu können.[62] Hieran lassen sich die unterschiedlichen Dimensionen der Fairness im Zusammenhang mit automatisierten Entscheidungen ablesen: die substanzielle sowie die prozedurale Dimension der Fairness. Erstere bezieht sich auf die gleiche und gerechte Verteilung, hier soll sichergestellt sein, dass Individuen und Gruppen frei von unfairen Bias, Diskriminierung und Stigmatisierung behandelt werden. Die prozedurale Dimension der Fairness beinhaltet die Möglichkeit, die getroffenen Entscheidungen und ausgegebenen Erklärungen eines KI-Systems anzufechten und wirksame Rechtsmittel gegen diese einlegen zu können. Dazu muss die für die Entscheidung verantwortliche Instanz identifizierbar sowie die Entscheidungsprozesse der KI erklärbar und einsehbar sein. Wie auch in der Informatik sowie im Recht sind also somit Transparenz, Erklärbarkeit und Verantwortlichkeit die Schlüsselbegriffe.[63] Transparenz allein reicht nicht aus. Selbst wenn die Entscheidungen und Erklärungen nachvollziehbar und durchschaubar sind, so ist das noch kein Garant für Fairness. Doch gehört es in jedem Fall zu

58 Vgl. *Martini M.,* Blackbox Algorithmus. Grundfragen einer Regulierung Künstlicher Intelligenz, Berlin, 2019, S. 54.

59 *O´Neil C.*, Angriff der Algorithmen, München, 2017, S. 198.

60 Vgl. *O´Neil C.*, Angriff der Algorithmen, München, 2017, S. 199.

61 Vgl. ZF hilft e.V. 2021, Internetquelle (https://www.zf-hilft.de/site/zfhilft/de/help_projects/small_loans_india/small_loans_india.html, 23.11.2021).

62 Vgl. *Stalder F.,* Algorithmen, die wir brauchen. Überlegungen zu neuen technopolitischen Bedingungen der Kooperation und des Kollektiven, in: *Otto/Gräf (Hrsg.):* 3TH1CS. Die Ethik der digitalen Zeit, Bonn, 2018, S. 51.

63 Vgl. *High-Level Expert Group on AI (AI HLEG),* Ethics Guidelines for Trustworthy AI, Brüssel, 2019.

Fairness aus ethischer Perspektive

einem transparenten Umgang mit Algorithmen, technische Aufklärung zu leisten in dem Sinne, dass Anwender*innen und Nutzer*innen niemals technischer Neutralität, sondern vielmehr menschlichen Interessen gegenüberstehen. Zusätzlich gibt es bei selbstlernenden Algorithmen einen undefinierten Bereich, in dem diese selbst Akzentuierungen und Modifikationen vornehmen, die unter Umständen nicht die Interessen des Auftraggebenden betreffen. Dies liegt in der Logik eines selbst lernenden Algorithmus: Findet er gewisse Bestärkung und Bestätigung, so entwickelt er sich im Rahmen einer gewissen Eigenständigkeit weiter, die mit der Datenlage und Kommunikationsumgebung einhergeht. Diese „Eigenständigkeit" der Technik lässt es zu einer ethischen Herausforderung werden, eine Fairnessbewertung diesen Algorithmen gegenüber überhaupt erst anzusetzen, da hier keine autonomen Subjekte für sich genommen handeln. Zwar können Algorithmen selektieren, entscheiden und auch in gewissem Sinne manipulieren, doch können sie darüber nicht mit sich selbst und anderen, moralischen Subjekten in einen Dialog treten und somit auch keine moralischen Einsichten gewinnen. Von einer Moral der Algorithmen kann man also an dieser Stelle nicht sprechen.[64]

Tests und Prozeduren, die den Algorithmus explizit auf diskriminierendes Verhalten abklopfen, sind ebenfalls eine wichtige Handlungsempfehlung für die Arbeit mit „fairen" Algorithmen. Auch diese sind Teil der prozeduralen Dimension der Fairness. Schließlich müssen auch die eingelesenen Daten immer wieder einer genauen Prüfung unterzogen werden. Die Entscheidungen und Erklärungen können schließlich nur so fair sein, wie es auch die Daten selbst sind, die eingespeist wurden. Hier sind gezielt hohe Qualitätssicherungsmaßnahmen zu fordern – die eingespeisten Daten müssen zureichend und qualitativ hochwertig sein. Die moralischen Vorstellungen der Programmierer*innen und Auftraggeber*innen müssen explizit und der ethischen Reflexion zugänglich gemacht werden. Lediglich moralische Vorgaben zu fordern, greift an dieser Stelle nicht: Es sollte nichts verdeckt in den Algorithmus eingeschrieben sein, denn so entzieht es sich der Verhandelbarkeit. Dies bedeutet auch, dass in vielen Fällen Gerechtigkeit über Profit gestellt werden muss.[65]

64 Vgl. *Reichmann W.*, Die Banalität des Algorithmus, In: *Krotz/Karmasin/Rath (Hrsg.):* Maschinenethik. Normative Grenzen autonomer Systeme, Wiesbaden 2019, S. 147.

65 Vgl. *O´Neil C.*, Angriff der Algorithmen, München, 2017, S. 277.

Gesamtfazit

Erklärbarkeit und Fairness sind essenzielle Aspekte für den zunehmenden Einsatz von KI-gestützten Entscheidungssystemen in der Praxis. Sowohl der Bereich der Erklärbarkeit als auch der Bereich der Fairness sind bereits in der Vergangenheit ausführlich, wenn auch noch nicht vollständig, untersucht worden. Deutlich weniger gut beleuchtet wurde bisher die Fairness von Erklärungen selbst.

In diesem Policy Paper haben wir primär die individuelle Fairness von kontrafaktischen Erklärungen betrachtet und argumentiert, dass auch Erklärungen selbst im uneigentlichen Wortsinn „unfair" sein können und so nicht nur die Fairness der KI selbst, sondern auch die der Erklärungen einen wichtigen Aspekt darstellt. Konkret bedeutet der Begriff eines fairen Algorithmus, dass der Algorithmus für ähnliche Nutzer*innen ähnliche Ausgaben produziert. Fairness von Erklärungen hingegen verlangt, dass die Gründe/Erklärungen für ähnliche Nutzer*innen ebenfalls ähnlich sind – das heißt wenn der Algorithmus eine ähnliche Ausgabe liefert, dann soll auch der Grund beziehungsweise die Erklärung für eben dieses Verhalten ähnlich sein. Es ist wichtig hervorzuheben, dass aus Fairness vom Algorithmus nicht zwingend folgt, dass die Erklärungen ebenfalls fair sind. Aus diesem Grund ist es von großer Wichtigkeit, dass neben dem bereits sehr verbreiteten Aspekt der algorithmischen Fairness auch die Fairness von Erklärungen berücksichtigt wird, um eine ganzheitliche Abdeckung des Fairness-Aspektes garantieren zu können.

Erklärbarkeit ist der Schlüsselbegriff, wenn algorithmenbasierte Entscheidungssysteme „fair" werden sollen. Dies beginnt bereits bei entsprechend geprüften, qualitativ hochwertigen Daten, führt über die ethische Reflexion der Auftraggeber*innen, Programmierer*innen schließlich auch zur Verhandelbarkeit der Ergebnisse und Erklärungen betroffener Anwender*innen.

Es ist von enormer Wichtigkeit, sich im Hinblick auf die Datenverarbeitung und die Arbeit mit Algorithmen einen Fairnessbegriff zu erarbeiten, der interdisziplinär und übergreifend geltend gemacht werden kann und so auch einen entscheidenden Faktor für rechtliche Grundsätze darstellen kann. Fairness sollte nicht nur eine Untermauerung der Transparenzpflicht darstellen, vielmehr sollte ihr ein eigenständiger Gehalt zukommen.

IMPACT

Ein gemeinsames Projekt von

UNIVERSITÄT DUISBURG ESSEN

UNIVERSITÄT BIELEFELD

UNIKASSEL VERSITÄT

Evangelische Hochschule Nürnberg

Gefördert durch

VolkswagenStiftung

DOI: 10.17185/duepublico/76311
URN: urn:nbn:de:hbz:465-20220722-080422-6