KI demokratisieren Fairness und Transparenz lassen sich nicht durch Technik allein herstellen

Florian Eyert und Paola Lopez

Summary: The increasing use of Artificial Intelligence in the public sector raises the question how its negative effects can be addressed democratically. The algorithmic fairness metrics and technical instruments for creating transparency discussed in the context of current efforts for fairness and transparency in machine learning are not sufficient to do so. Rather, fairness must be thought of as a collective deliberation about justice, and transparency as a communicative prerequisite for this process.

Kurz gefasst: Der zunehmende Einsatz Künstlicher Intelligenz im staatlichen Bereich eröffnet die Frage, wie ihren negativen Effekten demokratisch begegnet werden kann. Die im Rahmen aktueller Bemühungen um Fairness und Transparenz im maschinellen Lernen diskutierten algorithmischen Fairness-Metriken und technischen Instrumente zur Herstellung von Transparenz genügen hierfür nicht. Vielmehr muss Fairness demokratisch als Aushandlungsprozess über Gerechtigkeit gedacht werden und Transparenz als kommunikative Voraussetzung dafür.

Künstliche Intelligenz, kurz KI, erfährt seit einigen Jahren einen enormen Zuwachs an Aufmerksamkeit in Wissenschaft, Medien, Kultur und Wirtschaft – sie wird oft als Schlüsseltechnologie des 21. Jahrhunderts bezeichnet. Auch im staatlichen Handeln nimmt KI, die heute meist auf dem datenbasierten Ansatz des maschinellen Lernens beruht, eine zunehmend bedeutende Rolle ein. Staatliche Akteure, so die häufige Erwartung, können Big Data nutzen, um optimal mit der Bevölkerung zu interagieren und auf diese Weise effizient ihre Aufgaben zu erfüllen. Doch die Effizienz hat einen Preis, den oft gerade strukturell benachteiligte Menschen zahlen. Ein zentrales Risiko beim Einsatz solcher Technologien stellen Verzerrungen oder „Biases" dar: Vorhandene soziale Ungleichheiten können sich in den algorithmischen Systemen niederschlagen und durch deren Rückwirkung auf die Gesellschaft automatisiert verstärkt werden (siehe hierzu den Beitrag von Paola Lopez in diesem Heft). Aufsehen erregte hier beispielsweise der von Julia Angwin und Koautor*innen untersuchte Fall der Software COMPAS, die in den USA zur gerichtlichen Beurteilung der Rückfallwahrscheinlichkeit von Straftäter*innen eingesetzt wird. Sie wiesen nach, dass Schwarze Straftäter*innen systematisch fälschlicherweise als gefährlicher eingeschätzt werden als andere – mit entsprechend strengeren Strafen. Verkompliziert wird dieses Fairness-Problem häufig durch die intransparente Architektur der Systeme, die es schwierig macht, den entstandenen Schaden überhaupt zu entdecken. Denn oft ist es durch die hohe technische Komplexität selbst Expert*innen nicht möglich, verlässliche Aussagen über die Entscheidungslogik der Systeme zu treffen. Darüber hinaus wird die Entwicklung von KI-Systemen im öffentlichen Bereich in den meisten Fällen privaten Unternehmen übertragen, ihre genaue Funktions- und Entscheidungsweise unterliegt also oft deren Geschäftsgeheimnis. Die diskriminierenden Effekte von KI-Anwendungen liegen so gewissermaßen innerhalb einer technischen „Black Box", notwendige Anpassungen können kaum eingeleitet werden.

Um den Herausforderungen von Bias und Intransparenz zu begegnen, gibt es in der Wissenschaft und in der Wirtschaft seit einigen Jahren Bestrebungen, Fairness und Transparenz im Machine Learning zu stärken. Zum einen wird versucht, Bias automatisiert zu erkennen und mit geeigneten Maßnahmen auszugleichen. Ein großer Teil davon sind in Programmiercode übersetzbare Fairness-Metriken, bei denen es zahlreiche unterschiedliche Optionen gibt, wie etwa die Arbeit von Sorelle Friedler und Koautoren zeigt. Ist ein KI-System fair, wenn die Ergebnisse unabhängig von Geschlecht, Alter oder sozioökonomischer Position sind? Oder dann, wenn es diskriminierungsrechtlich sensible Kategorien wie *race* gar nicht erst kennt? Oder aber dann, wenn nachteilige Entscheidungen mathematisch zufällig verteilt sind? Zum anderen wird daran gearbeitet, mithilfe technischer Verfahren die Transparenz zu steigern. Unter dem Stichwort *Explainable AI* wird etwa das Ziel verfolgt, sich selbsterklärende Machine-Learning-Systeme zu konstruieren. Die Frage dabei lautet: Nachdem Millionen von Daten verarbeitet und zu einer Entscheidung gebündelt werden – was waren die ausschlaggebenden Faktoren? Im Falle einer Bilderkennung können etwa bestimmte Bildregionen hervorgehoben werden, wie beispielsweise die Ohren in einem Tierfoto, die das Programm veranlassen, eine Katze zu identifizieren.

Diese Ansätze sind in erster Linie technisch orientiert. Aus einer sozialwissenschaftlich informierten, interdisziplinären Perspektive wird aber deutlich, dass

sich demokratische Herausforderungen durch Bias und Intransparenz auf diese Weise nicht adäquat angehen lassen. Denn die technische Herangehensweise bringt grundlegende Begrenzungen mit sich. In einem ersten Schritt betrifft dies das Ziel der Fairness. Ein Blick in die Ideengeschichte des Begriffs zeigt, dass es sich hier keinesfalls um ein allgemein definierbares Ideal handelt. Während Teile der Philosophie und politischen Theorie an einer analytischen Präzisierung interessiert sind (wie im Fall von John Rawls und seiner formalisierten Fairness-Konstruktion), beschreibt die empirische Sozialwissenschaft, dass Fairness und dahinterstehende Vorstellungen von Gerechtigkeit grundlegend politische Phänomene sind, die in jeweiligen gesellschaftlichen Kontexten ausgehandelt werden und umkämpft sind. Auf die Frage, was fair und gerecht ist, gibt es also keine einfache Antwort: Wer welche sozialstaatlichen Ressourcen zugesprochen bekommt, wer wegen des Verdachts auf Steuerbetrug genauer unter die Lupe genommen wird oder wer wie lange Haftstrafen verbüßen soll – das sind hochgradig politische Angelegenheiten. KI kann vieles, doch das Politische aus solchen Fragen mittels mathematischer Raffinesse herauszurechnen, gehört nicht dazu. Den gesellschaftlichen Auswirkungen von KI-Anwendungen kann also nur eingeschränkt mit technischen Fairness-Maßen begegnet werden. Selbst wo dies möglich ist, muss die Auswahl des jeweils geeigneten Maßes außerhalb rein technischer Diskussionen stattfinden, ist doch jedes Maß mit spezifischen Wertvorstellungen verwoben. Wo staatliche Akteure Künstliche Intelligenz einsetzen, ist eine Auseinandersetzung über Gerechtigkeit und ihre Einschreibung in algorithmische Entscheidungssysteme unverzichtbarer Bestandteil demokratischer Legitimität: Was den Demos wesentlich betrifft, muss demokratisch ausgehandelt werden, statt es Programmierer*innen in Unternehmen zu überlassen.

Das führt direkt zur zweiten Herausforderung: der Frage, welche Art von Transparenz aus einer solchen Perspektive notwendig ist. Wenn KI-Systeme nur auf der Basis ausgehandelter Gerechtigkeitsnormen sinnvoll eingesetzt werden können, geraten die institutionellen und gesellschaftlichen Bedingungen dieser Aushandlungsprozesse in den Blick. Vor allem die Voraussetzungen für den Austausch von Argumenten hinsichtlich wünschenswerter oder zu vermeidender Effekte des staatlichen Einsatzes von Machine Learning gilt es sicherzustellen. Nur so kann die demokratische Übersetzung der Interessen Betroffener in die technische Gestaltung kollektiver Entscheidungsstrukturen gelingen. Zu diesen Voraussetzungen gehört aus unserer Sicht Transparenz in einem erweiterten Sinne, der einen angemessenen Wissensstand aller Betroffenen und einen institutionellen Rahmen für Kommunikation mit Systementwickler*innen und politisch Verantwortlichen einschließt. Wir verstehen Transparenz also nicht als technische Eigenschaft von Software, sondern als kommunikative Konstellation.

Das Herstellen von interpretierbarem Wissen über die internen Prozesse eines Machine-Learning-Systems mittels Explainable AI ist zwar eine wichtige Voraussetzung für die gesellschaftliche Einordnung Künstlicher Intelligenz. An Problemen wie proprietärem, also nur durch einzelne Unternehmen einsehbarem Code, verschlossenen Datenquellen und einem falsche Erwartungen weckenden Mediendiskurs ändert dieses Wissen aber kaum etwas. Ebensowenig hilft es Individuen ohne spezielle Bildung in Mathematik, Informatik oder Data Science dabei, sich souverän eine Meinung zu bilden, oder stellt es einen Dialog zwischen staatlichen Akteuren und der Bevölkerung her. Zudem zeigte jüngst die Forschung von Umang Bhatt und anderen, dass Explainable AI in der Praxis oft vor allem zum Finden von Fehlern im Code durch Programmierer*innen genutzt wird und nur in geringerem Maße den tatsächlich Betroffenen zugutekommt. Transparenz muss aus diesem Grund breiter gedacht werden. Es geht um mehr als technische Transparenz einzelner Entscheidungsprozesse für einzelne Expert*innen: Es muss eine grundlegende Transparenz des Gesamtsystems und seines Anwendungszusammenhangs gegenüber einer demokratischen Öffentlichkeit geben. Die Öffentlichkeit muss Kenntnisse darüber haben, was KI-Systeme können und auf welche Weise sie Ergebnisse und Entscheidungen generieren, welche Alternativen es gibt, wo KI-Systeme bereits verwendet werden und in welchen Kontexten ein Einsatz geplant ist – und zwar, bevor diese kostspielig entwickelt oder angekauft werden. Dieser Ansatz geht über die in der Debatte oft eingeforderte Steigerung digitaler Kompetenzen hinaus, wie sie etwa unter

dem Stichwort *digital literacy* oder *algorithmic literacy* befürwortet wird. Diese Konzepte bleiben bei einer Bringschuld individueller Bürger*innen stehen, die eine solche Kompetenz erwerben müssen. Transparenz als kommunikative Konstellation bedeutet auch, dass Orte geschaffen werden, an denen Kommunikation sowohl über Fakten als auch über Wertentscheidungen in KI-Systemen stattfinden kann und an denen Entwickler*innen und Verantwortliche der Systeme kommunikativ erreichbar sind. Da die Auswirkungen von KI-Systemen im staatlichen Handeln potenziell jede*n treffen kann, verstehen wir auch hier die Herstellung der Voraussetzungen für die gesellschaftliche Verhandlung um öffentliche KI-Systeme nicht als individuelle, sondern als öffentliche Aufgabe.

Folgt man diesem Argumentationsgang, dann muss Fairness demokratisch als Aushandlungsprozess über Gerechtigkeit verstanden und Transparenz inklusiv als kommunikative Voraussetzung für diese Aushandlungen gedacht werden. Welche konkreten Anforderungen ergeben sich daraus? Wichtig ist es zunächst, sich nicht dem Narrativ der unaufhaltsamen Digitalisierung zu ergeben, der man sich eben früher oder später fügen muss. Vielmehr geht es darum, als Gesellschaft zu entscheiden: Möchte man ein KI-System in einem bestimmten Kontext anwenden, und wenn ja: wie? Expert*innen und politische Entscheidungsträger*innen stehen hierbei in der Verantwortung, Verhandelbarkeit und umfassende Transparenz aktiv herzustellen. Dies wäre als integraler Bestandteil eines jeden von öffentlicher Hand geplanten KI-Projekts zu betrachten. Dazu müssen neue Formate und Orte des Austauschs konzipiert werden. Ein allererster Schritt hierfür können öffentlich einsehbare Register verwendeter KI-Systeme sein, wie sie die Städte Helsinki und Amsterdam im vergangenen Jahr eingeführt haben. Idealerweise würden dort auch erst geplante Projekte aufgeführt und durch verständliche und auch Begrenzungen offenlegende Informationskampagnen begleitet werden. Darauf aufbauend wären Diskussionsformate, Bürger*innendialoge, aktive Kontakte zu entsprechenden NGOs, öffentliche Foren und andere Beteiligungsformen sinnvoll, um Räume der Aushandlung herzustellen. Auch Offenheit gegenüber zivilgesellschaftlicher Kritik muss aktiv institutionalisiert werden. Schließlich gehört zu einer tatsächlichen demokratischen Aushandlung auch die Möglichkeit, KI-basierte Systeme in gewissen Kontexten abzuschaffen oder gar nicht erst einzusetzen, wenn sie zu viele Risiken bergen oder bestehenden Gerechtigkeitsvorstellungen entgegenstehen. Auch diese Option muss Teil des aktuellen gesellschaftlichen Experiments mit den Möglichkeiten und Grenzen sein, Computer in gesellschaftliche Entscheidungssysteme einzubinden.

Literatur

Angwin, Julia/Larson, Jeff/Mattu, Surya/Kirchner, Lauren: Machine Bias. In: ProPublica, 2016, 23. Mai. Online: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing (Stand 11.02.2021).

Bhatt, Umang/Xiang, Alice/Sharma, Shubham/Weller, Adrian/Taly, Ankur/Jia, Yunhan/Ghosh, Joydeep/Puri, Ruchir/Moura, José M. F./Eckersley, Peter: „Explainable Machine Learning in Deployment". In: FAT '20 Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 2020, S. 648–657. DOI: 10.1145/3351095.3375624.*

Friedler, Sorelle A./Scheidegger, Carlos/Venkatasubramanian, Suresh/Choudhary, Sonam/Hamilton, Evan P./Roth, Derek: „A Comparative Study of Fairness-Enhancing Interventions in Machine Learning". In: FAT '19: Proceedings of the Conference on Fairness, Accountability, and Transparency, 2019, S. 329–338. DOI: 10.1145/3287560.3287589.*

Florian Eyert ist Doktorand am Weizenbaum-Institut für die vernetzte Gesellschaft in der Forschungsgruppe 18 „Quantifizierung und gesellschaftliche Regulierung". In seiner Dissertation befasst er sich mit der Digitalisierung von Governance-Prozessen und der politischen Dimension von Computermodellen und Künstlicher Intelligenz. *[Foto: Thu-Ha Nguyen]*

florian.eyert@wzb.eu

Paola Lopez ist Gastwissenschaftlerin in der Forschungsgruppe Politik der Digitalisierung und am Weizenbaum-Institut für die vernetzte Gesellschaft in der Forschungsgruppe 18 „Quantifizierung und gesellschaftliche Regulierung". Vgl. auch ihren diesem Text vorangehenden Artikel in diesem Heft.

paola.lopez@wzb.eu