Diskriminierung durch Data Bias Künstliche Intelligenz kann soziale Ungleichheiten verstärken

Paola Lopez

Summary: Artificial intelligence that makes or guides decisions based on large amounts of data can create discriminatory effects and reinforce existing inequalities – especially when it is used in social contexts such as the labor market, distribution of social welfare or credit matters. Three types of bias can be differentiated: purely technical bias, socio-technical bias and societal bias.

Kurz gefasst: Künstliche Intelligenz, die auf der Grundlage großer Datenmengen Entscheidungen trifft oder anleitet, kann diskriminierende Effekte erzeugen und vorhandene Ungleichheiten automatisiert verstärken – vor allem dann, wenn sie in gesellschaftlichen Kontexten wie dem Arbeitsmarkt, sozialstaatlichen Verteilungen oder Kreditfragen zur Anwendung kommt. Drei Arten von Bias können unterschieden werden: rein technischer, soziotechnischer oder aber gesellschaftlicher Bias.

Immer häufiger wird auf Künstliche Intelligenz zurückgegriffen, wenn es um Entscheidungen geht, die Menschen und ihre Gestaltungsspielräume betreffen. Algorithmische Systeme werden eingesetzt, um die Verteilung sozialstaatlicher oder medizinischer Ressourcen zu unterstützen. In den USA wird in strafrechtlichen Verfahren bei der Bestimmung von Strafmaß und Bewährungsauflagen häufig eine algorithmische Risikoprognose bezüglich der Gefährlichkeit von Menschen befragt. Große Unternehmen verwenden bei Neu-Einstellungen algorithmische Hiring-Systeme, die Bewerbungen automatisiert vorauswählen, Finanzdienstleister prüfen die Kreditwürdigkeit. Denn hochkomplexe mathematische Methoden in Verbindung mit riesigen Datenmengen stellen eine Art digitalen Erfahrungsschatz bereit, den man sich zunutze machen kann. Außerdem sollen Entscheidungsprozesse durch Algorithmen effizienter und objektiver werden, weil subjektive menschliche Befindlichkeiten eingehegt werden, so die Argumente.

Doch von akademischer, politischer, aktivistischer und journalistischer Seite regt sich Kritik: Gerade algorithmische Systeme könnten diskriminierende Effekte nach sich ziehen und damit bestehende soziale Ungleichheiten verstärken – automatisiert und oft unbemerkt. Denn in vielen Entscheidungssituationen besteht ein Machtgefälle. Werden diese Entscheidungen oder Prozesse, und sei es teilweise, auf ein algorithmisches System ausgelagert, so können gravierende Probleme auftreten. Entscheidet ein Mensch fehlerhaft oder diskriminierend in Bezug auf andere Menschen, so ist davon nur eine begrenzte Menge von Menschen betroffen. In einem demokratischen Rechtsstaat gibt es in vielen Bereichen – zumindest theoretisch – Mittel, um diskriminierende Entscheidungen zu beanstanden. Doch im Versprechen von massenhaften, automatisierten Abläufen steckt die Möglichkeit, Schaden in großem Umfang anzurichten: Ein System, das automatisiert diskriminiert, trifft kontinuierlich und effizient große Mengen von Entscheidungen. Kleine Übel wirken sich also unverhältnismäßig stark aus – gerade wegen der datenbasierten Effizienz. Menschen in marginalisierten gesellschaftlichen Positionen sind davon besonders betroffen, weil sie sich am wenigsten wehren können.

„Data bias" lautet ein zentrales Schlagwort, wenn es in politischen, medialen und akademischen Diskursen um Diskriminierung durch Künstliche Intelligenz geht. Warum sind Daten überhaupt so wichtig? Den meisten aktuell gebräuchlichen algorithmischen Systemen liegt ein datenbasiertes Paradigma zugrunde: Bei ihrer Entwicklung werden große Mengen an Daten statistisch analysiert. Die in den Daten gefundenen Muster werden schließlich auf neuen Input übertragen. Für die automatisierte Objekterkennung auf Bildern zum Beispiel werden bei der Entwicklung des Systems große Mengen an Beispielbildern eingespeist, sodass das Computerprogramm durch die schiere Menge an Bilddaten mit Bezeichnung „Blume" zu erkennen „lernt", wie eine Blume aussieht: welche Formen Blumen haben können, welche Farben und Farbkontraste und so weiter. Von diesem „Lernprozess" stammt auch die Bezeichnung des „Machine Learning". Daten bilden also die mathematische Grundsubstanz solcher Systeme und stecken damit auch ihre Grenzen ab. Denn ein algorithmisches System kann höchstens so viel wissen wie die zugrunde liegenden Daten: Werden einem Programm zur Blumenerkennung ausschließlich Bilddaten von Sonnenblumen und Gänseblümchen eingespeist, so wird es eine Rose nicht als Blume erkennen können. In gesell-

schaftlichen Kontexten geht es freilich nicht um das Erkennen von Blumen, sondern um automatisierte Gesichtserkennung und Identitätsfeststellung in Strafprozessen oder um Lebenslaufdaten in Kontexten des Recruitings oder um prognostizierte finanzielle Bonität oder Kreditwürdigkeit mittels bestimmter personenbezogener Daten. Nun wurde schon viel über Daten geschrieben und geforscht: Die Science and Technology Studies bearbeiten seit Jahrzehnten, also schon vor dem Aufkommen aktueller Digitalisierungsdebatten, verschiedene Problemfelder in diesem Zusammenhang. Alleine die Frage, was in Daten gemessen wird und was demgegenüber unsichtbar bleibt, ist Stoff vieler Diskussionen. Bei der Kategorienbildung und der Frage, was überhaupt quantifizierbar ist, kommt es zwangsweise zu Vereinfachungen. Daten und Zahlen lassen eine heterogene, komplexe Welt homogen und damit kontrollierbar erscheinen.

„Data bias", also die Verzerrung, die sich aus der intensiven Produktion und Verarbeitung von Daten ergibt, kann unterschiedlichen Mustern folgen. Um algorithmische Systeme analytisch einschätzen und politisch kritisieren zu können, ist es wichtig, klare Begriffe in Bezug auf dieses Phänomen zu haben. Ich habe eine Typologie entwickelt, die drei Arten von Datenbias unterscheidet. Ein zentrales Unterscheidungsmerkmal ist dabei, ob und inwiefern sich die digitalen Daten von dem realen Phänomen unterscheiden und wenn ja, ob die Abweichung auf gesellschaftliche Ungleichheiten zurückgeht. Diese Typologie setzt auf einem bereits 1996 erschienenen Modell von Batya Friedman und Helen Nissenbaum auf, das sich auf ganz allgemeine Computersysteme bezieht.

Zunächst gibt es den rein technischen Bias, den ich recht weitläufig definiere und der jede Art von technischer oder konzeptueller Fehlmessung und Fehlkonzeption umfasst. Hier liegt eine Diskrepanz vor zwischen dem, was man in Daten abbilden oder messen möchte, und dem, was abgebildet oder gemessen wird. Diese Abweichung ist jedoch nicht in strukturellen, gesellschaftlichen Ungleichheiten begründet. Beispiele wären hier die Falschübertragung der Postleitzahl in einer Datenbank oder technische Fehlschlüsse bei der Vergabe von Sozialleistungen. Algorithmische Systeme, deren Datengrundlage einem rein technischen Bias unterliegen, können komplett falsche und unsinnige Ergebnisse erzeugen. Die Leidtragenden sind jene Menschen, die sich ohnehin in einer prekären Lage befinden. Wird eine Sozialleistung oder eine Versicherungsleistung oder ein Kredit fälschlicherweise nicht gewährt, so ist die Berichtigung von algorithmischen Fehlern stets mit vielen Ressourcen, Geduld und entsprechendem Know-how verbunden, wie Virginia Eubanks in ihrem Buch „Automating Inequality" ausführlich beschreibt. Eine Fehldarstellung in Daten kann – selbst wenn sie nicht in strukturell ungleichheitsbezogener Weise entsteht – strukturell unterschiedliche Auswirkungen je nach gesellschaftlicher Position der Individuen haben.

Die zweite Art des Bias in Daten bezeichne ich als soziotechnischen Bias. Dieser beschreibt eine systematische Abweichung der Datengrundlage von dem Phänomen, das dargestellt werden soll. Diese Abweichung kann jedoch nicht durch einen vermeintlich neutralen technischen Fehler erklärt werden, sondern ist vielmehr in der soziotechnischen Dimension zu suchen. Bestimmte Gruppen, die strukturell benachteiligt werden, sind in den entsprechenden Daten unsichtbar, übermäßig sichtbar oder verzerrt abgebildet. Ein viel diskutiertes Beispiel hierfür ist eine weitverbreitete Gesichtserkennungssoftware. Die Informatikerin Joy Buolamwini zeigte, dass dieses System einem „racial bias" unterliegt: Es erkennt hellere Gesichter systematisch besser und Gesichter mit dunkler Hautfarbe schlecht bis gar nicht. In einem eindrucksvollen TED-Talk demonstriert sie, wie ihr eigenes Gesicht durch das Programm zunächst gar nicht – und dann doch als Gesicht erkannt wird, sobald sie sich eine weiße Plastikmaske aufsetzt. Das ist eine direkte Übersetzung der Tatsache, dass dieses Programm gelernt hat, dass (nur) helle Gesichter Gesichter sind – bei der Erstellung des Systems wurde schlichtweg eine zu geringe Vielfalt von Bilddaten von Gesichtern eingespeist. Die Entwickler*innen des Systems und die für die Datengrundlage Verantwortlichen machten Schwarze Menschen digital unsichtbar.

Die dritte Art des Bias nenne ich gesellschaftlichen Bias. Hier findet sich in den Daten keine Falschabbildung der Realität, sondern eine korrekte Abbildung von

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". Am Institut für Rechtsphilosophie der Universität Wien arbeitet die Mathematikerin an einer interdisziplinären Dissertation über datenbasierte algorithmische Systeme und Künstliche Intelligenz im Kontext des staatlichen Handelns. *(Foto: privat)*

paola.lopez@wzb.eu

strukturellen Ungleichheiten, die in der Gesellschaft vorherrschen. Es wird der aggregierte gesellschaftliche Bias abgebildet, betont und verstärkt. Ein Beispiel ist etwa ein automatisiertes Hiring-System, das von Amazon entwickelt und mittlerweile wieder abgeschafft wurde: Es unterlag einem gravierenden „gender bias" und schätzte Frauen systematisch als weniger geeignet und weniger fähig ein als Männer. Dieser Bias erklärt sich mit einem Blick auf die zugrunde liegenden Daten: Das System verwendete Daten aus dem Amazon-Unternehmen und prognostizierte anhand der in der Vergangenheit erfolgreichen Mitarbeiter*innen, welche Bewerber*innen am geeignetsten sind. Ganz ähnlich funktioniert der Algorithmus des Arbeitsmarktservice (AMS) in Österreich, der sich derzeit in einer rechtlichen Aushandlungssituation befindet. Dabei handelt es sich um ein algorithmisches Prognosesystem, das eingesetzt werden soll, um die Segmentierung von Erwerbsarbeitslosen in drei Gruppen mit unterschiedlichem Zugang zu arbeitsmarktpolitischen Förderressourcen anzuleiten. Dieses bewertet die Chancen von Individuen am Arbeitsmarkt entlang verschiedener Daten und kommt zu dem Ergebnis, dass ein weiblicher Geschlechtseintrag negative Auswirkungen auf die Chancen hat, genauso wie ein Alterseintrag über 30, noch gravierender über 50, eine Nicht-EU-Staatsangehörigkeit, gesundheitliche Beeinträchtigungen oder Betreuungspflichten – diese jedoch nur bei Frauen. Auch hier wird der strukturelle Bias der Gesellschaft, in diesem Fall des österreichischen Arbeitsmarkts, sichtbar. Die Datengrundlage, anhand der das System „gelernt" hat, Prognosen anzustellen, besteht nämlich aus vergangenen Arbeitsmarktdaten. Die negative Auswirkung eines weiblichen Geschlechtseintrags auf die Chancen am Arbeitsmarkt zeigt also, dass in der Datengrundlage des Systems Frauen systematisch langsamer und weniger nachhaltig in den Arbeitsmarkt integriert wurden. Was zunächst nicht mehr ist als die Abbildung einer Realität, wird zum Problem, wenn darüber Ressourcen an Individuen verteilt werden. Dabei kann es zu Diskriminierung kommen, die ungleiche Realität wird fortgeschrieben.

Alle drei Arten von Bias können erheblichen Schaden anrichten, vor allem wenn die entsprechenden Systeme eingesetzt werden, um Entscheidungen über Menschen zu treffen. Falsche algorithmische Ergebnisse oder ein schlichtes „Computer says no" aufgrund von rein technischem Bias können nur mit Geduld und Ressourcen richtiggestellt werden, die manche Menschen weniger haben als andere. Soziotechnischer Bias kann repariert werden, wenn die entsprechende Datengrundlage erweitert wird. Dass das jedoch ein mehr als ambivalentes Unterfangen sein kann, zeigt etwa folgendes Beispiel: So soll ein Subunternehmer von Google zur Verbesserung der Datengrundlage von Gesichtserkennungssoftware gezielt Schwarze Wohnungslose, die eher nicht widersprechen würden, anvisiert haben, um mittels Foto ihre Gesichtsdaten zu sammeln. Gesellschaftlicher Bias schließlich, der in einem datenbasierten algorithmischen System sichtbar wird, kann nur durch breite gesellschaftliche Transformationsprozesse eingehegt werden. Diese Prozesse können angestoßen werden, wenn die algorithmischen Systeme kritisch analysiert und als Diagnosewerkzeuge zum Erkennen vorhandener Ungleichheiten genutzt werden.

Literatur

Buolamwini, Joy: How I'm fighting Bias in Algorithms. 2016. Online: https://www.ted.com/talks/joy_buolamwini_how_i_m_fighting_bias_in_algorithms. (Stand 09.02.2021).

Eubanks, Virginia: Automating Inequality: How High–tech Tools Profile, Police, and Punish the Poor. New York: St. Martin's Press 2017.

Friedman, Batya/Nissenbaum, Helen: „Bias in Computer Systems". In: ACM Transactions on Information Systems, 1996, Jg. 14, H. 3, S. 330–347.

Lopez, Paola: Reinforcing Intersectional Inequality via the AMS Algorithm in Austria. In: Proceedings of the STS Conference Graz 2019, S. 280–309.