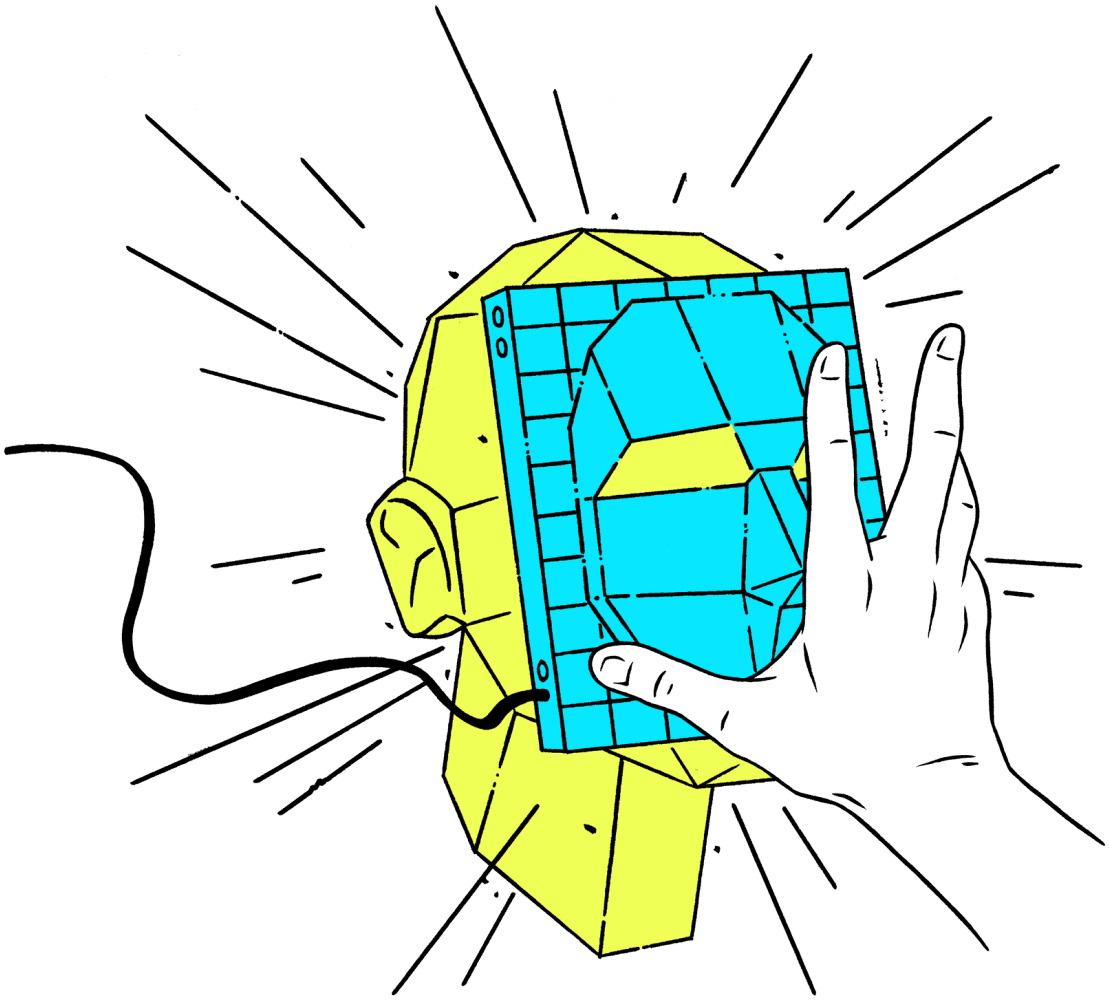

Wer sind wir? Warum künstliche Intelligenz immer ideologisch ist

Weil hinter jedem algorithmischen Modell bereits ein Modell der Welt steht, gibt es keine neutralen Daten. Das führt dazu, dass Ungerechtigkeit nicht nur wiederholt, sondern verstärkt wird. Die Verzerrungen zu beheben, wird schnell hochpolitisch.

Seit rund fünfzehn Jahren lautet die Grundidee von Big Tech: Die Welt besteht aus Daten, und Daten sind neutral. Hat man genug Daten, hat man auch ein Bild der Welt – und zwar derart, dass man sich sogar alle Theorien über diese Welt sparen kann. Kurz: Daten sollen Politik überflüssig machen.

Das gilt auch für das, was gerade so heftig diskutiert wird: Postkolonialismus, Genderforschung, Identitätspolitik. Doch in letzter Zeit wird deutlich, dass sich auch die Techbranche diesen Debatten nicht entziehen kann. Vor allem da nicht, wo der Datenglaube bisher die rasantesten Fortschritte gebracht hat: in der Entwicklung künstlicher Intelligenz – oder kurz: KI. Wo KI zu ethisch fragwürdigen Ergebnissen führt, steht nicht nur die angebliche Neutralität von Daten auf dem Prüfstand. Es zeigt sich auch, dass hier der Identitätsdiskurs die besten Argumente für seine Kritik am Status quo erhält.

Algorithmische Ungerechtigkeit

Was heute KI genannt wird, ist eigentlich ein Etikettenschwindel. Statt menschenähnlicher Intelligenz meint man damit eher automatisierte Systeme zur Entscheidung, Klassifikation oder Prognose. Meist arbeiten sie als kleine Helfer im Hintergrund und erstellen automatische Übersetzungen, schlagen uns Filme auf Netflix vor oder erkennen die Gesichter von Freunden, wenn wir Fotos auf Facebook hochladen. Oft basieren sie auf sogenannten «neuronalen Netzen» und sind, ganz rudimentär, den Synapsen und Neuronen des Gehirns nachempfunden.

Immer aber müssen sie mit Daten trainiert werden – je mehr, desto besser. Der Vorteil: Statt den KI-Systemen aufwendig Regeln über die Welt beizubringen, lernen sie anhand der Daten solche Regeln von ganz allein.

Doch gerade bei selbstlernenden Systemen ist schwer nachzuvollziehen, ob ihre Entscheidungen auch unseren Erwartungen entsprechen. Das Feld der KI-Ethik diskutiert diese Gefahr unter dem Begriff des *bias* – Verzerrungen, die bestehende Ungerechtigkeit wiederholen oder gar verstärken. So können KI-Systeme nicht nur den Rassismus oder den Sexismus widerspiegeln, den sie aus den Daten der Welt gelernt haben, sondern ihn auch noch potenzieren und so selbst aktiv Ungerechtigkeit hervorbringen.

Wie solche Ungerechtigkeit aussehen kann, zeigte 2018 eine Studie über Software zur Gesichtserkennung. Die untersuchten Systeme funktionierten bei Schwarzen deutlich schlechter als bei Weissen und bei schwarzen Frauen noch einmal schlechter als bei schwarzen Männern. Das Problem bestand darin, dass die Datengrundlage – eine grosse Menge Porträts, mit denen das KI-Modell trainiert wurde – nur wenige Fotos schwarzer Menschen enthielt.

Für viele KI-Ethikerinnen bestätigt dieser *data bias,* schlecht ausgewählte Daten, einen *human bias* – ein gesellschaftliches Vorurteil, das eine der Grundthesen identitätspolitischer Theorie ist: Weisse sind ein Standard, Nichtweisse dagegen eine Ausnahme. Derart trainiert, lautet das Argument, übernimmt das System eine Ungerechtigkeit, die es in der Welt bereits gibt. Es bestätigt in diesem Fall auch nebenbei die These von der sogenannten Intersektionalität von Diskriminierung: Schwarze Frauen werden hier

gleich doppelt benachteiligt. Der *bias* der Welt wird im *bias* der Daten wiederholt.

Mitautorin dieser Studie ist die KI-Ethikerin Timnit Gebru. Im Dezember machte sie Schlagzeilen, als Google sie überraschend entliess. Gebrus Forderung, *bias* ernster zu nehmen und die KI-Branche auch anhand identitätspolitischer Überlegungen ethischer zu gestalten, stiess auf Widerstand vor allen in der traditionellen, datengläubigen KI-Forschung. Denn für diese ist *data bias* kein ethisches, sondern nur ein technisches Problem – ein Fall des Prinzips *«garbage in, garbage out»:* Mit schlechten Daten kommt man zu schlechten Ergebnissen. Und nur weil in diesem Fall die Ergebnisse rassistisch sind, ist es die Welt noch lange nicht.

Für manche, wie den KI-Experten und Bestsellerautor Pedro Domingos, der Gebru auf Twitter heftig angriff, sind die KI-Ethiker in Wirklichkeit eine *«woke police»,* eine linke Moralpolizei, die durch antirassistische Ideologien in die Forschungsfreiheit eingreife.

Zum Interview

Die Wiener Wissenschaftlerin Doris Allhutter erforscht, wie Maschinen menschliche Stereotype erlernen. Und was man dagegen tun kann: «Plötzlich gehörten für Computer ‹Mexikaner› und ‹illegal› zusammen».

Der toxische Feedback-Loop

Die Lösungsvorschläge von Domingo und Co. sind technisch. Sie lauten «Skalierung» oder «Kuratierung»: Entweder soll ein noch grösserer Satz an Daten solche Verzerrungen ausschleifen, oder man trainiert das Modell von Anfang an auf gleich viele weisse wie schwarze Gesichter.

Das Problem ist nur, dass diese lange bekannten *biases* nicht weniger werden: Ganze zwei Jahre nach dieser Studie zog die KI-Bildverbesserung «Pulse» 2020 Kritik auf sich. Sie kann verpixelte Fotos anhand gelernter Porträts wieder in klare Bilder hochrechnen. Als ein User auf Twitter aber einen verpixelten Obama in das System einspeiste, machte es aus ihm einen weissen Mann – den Standard. Wenn Problem und Lösung bekannt sind, warum hatte sich immer noch nichts verbessert?

Doch selbst bei einem perfekt paritätischen Datensatz bleiben andere Fragen. Was etwa ist mit der Entscheidung, Fotos in männlich und weiblich aufzuteilen? Dass non-binäre oder Transpersonen hier durchs Raster fallen, deutet bereits an, dass nicht allein die Qualität der Inputdaten das Problem ist. Es gibt schlicht keine neutralen Daten. Hinter allen KI-Modellen steht bereits ein Modell der Welt, lange bevor ihr Training überhaupt begonnen hat. Und das heisst: Künstliche Intelligenz ist immer ideologisch.

Solche Ideologien – von den ganz normalen Annahmen, die wir alle über die Welt machen, bis hin zu rassistischen oder sexistischen Denkmustern – sind vor allem dort gefährlich, wo KI als besonders neutral, besonders frei von Vorurteilen verkauft wird und Entscheidungen fällt, die reale Konsequenzen für Leben von Menschen haben. Traurige Berühmtheit erlangte «Compas», ein System, das unter anderem in Florida benutzt wird, um abzuschätzen, wie wahrscheinlich Angeklagte rückfällig werden. Die Ergebnisse haben Einfluss auf Entscheidungen über Haftstrafen und vorzeitige Entlassungen.

Die Rückfallwahrscheinlichkeit berechnet das System anhand von Trainingsdaten früherer Fälle, wobei neben Vorstrafen auch soziale und ökonomische Faktoren einfliessen. Doch auf diese Weise werden bereits Benachteiligte weiter benachteiligt. Wer etwa Sozialhilfe empfängt oder obdachlos ist, gilt bereits als *high risk*. Das Recherchenetzwerk Pro Publica zeigte, dass «Compas» so überproportional Schwarze benachteiligt: Das Modell, das aus den Fakten der Vergangenheit Normen über die Zukunft ableitet, hatte gelernt, Schwarze mit Verbrechen zu assoziieren.

Solche Unterdrückungseffekte kennt die Soziologie schon lange, aber in scheinbar neutralen Systemen sind sie schwer zu erkennen. «Die Frage ist», schreibt die KI-Forscherin Cathy O'Neil über Systeme wie «Compas», «ob wir menschliche Vorurteile eliminieren können oder sie nur hinter Technik verstecken.» Dass strukturell Benachteiligte unter solcher KI besonders leiden, nennt sie einen toxischen Feedback-Loop: Das Modell wiederholt nicht nur, sondern verstärkt den *bias* der Welt – es wird selbst zu einer Quelle von Ungerechtigkeit.

Plappernde, stochastische Papageien

Auch das Prinzip der Skalierung – mehr Daten bringen bessere Ergebnisse – scheint wenig zu nützen. Google feuerte Gebru aus seinem Ethik-Team für einen Artikel, in dem sie und ihre Mitautorinnen sogenannte grosse Sprachmodelle behandelten. Das sind KI-Systeme, die auf Unmengen von Text trainiert worden sind und so kohärente, korrekte Sätze produzieren können.

Auch hier hat sich bestätigt, dass mehr Daten bessere Performance bedeuten. So machte letztes Jahr das vom KI-Thinktank «Open AI» entwickelte Sprachmodell GPT-3 Furore, weil sein Output sich las, als sei er von einem Menschen geschrieben worden. Dabei folgt GPT-3 demselben Prinzip wie sein sehr viel weniger mächtiger Vorgänger, er ist aber mit zehnmal so vielen Trainingsdaten gefüttert worden. Switch-C, das neueste Modell von Google, ist noch einmal grösser.

Um diesen Trend sorgen sich Gebru und ihre Mitautoren: Denn anders als von manchen erhofft, trägt die Grösse des Datensatzes keineswegs dazu bei, Diskriminierung zu verringern. Das Modell wurde recht wahllos auf Text aus dem Internet trainiert. Dass hier Sexismus und Rassismus nicht selten sind, ist kaum überraschend. Und so produziert auch GPT-3 Sätze, die von jüdischer Weltverschwörung bis zu Frauenhass reichen. Die KI ist eben keine wirkliche künstliche Intelligenz, sondern nur ein plappernder, «stochastischer Papagei», wie die Autorinnen ihren Artikel nennen.

Bias wird also mitskaliert. Dazu kommt, dass bei so grossen Datenmengen niemand mehr recht weiss, was eigentlich in ihnen steckt. Und weil zukünftige Systeme immer mehr Text für ihr Training benötigen, ist auch hier ein toxischer Feedback-Loop zu erwarten, wenn sie mit bereits von KIs generiertem Text gefüttert werden. Die Folge, so der Artikel, sei ein ethisches «Einrasten», das keine progressive Veränderung mehr zulasse, weil jede linguistische Neuerung immer nur einen verschwindend geringen Anteil an der Textmenge ausmacht. Das Modell ist strukturell konservativ.

Und der Skalierungsfetisch bringt noch andere Ungerechtigkeiten hervor. Das Training eines solchen Modells ist mit einem enormen Energieaufwand verbunden – das 55-Fache des durchschnittlichen Jahresverbrauchs einer einzelnen Person, wie die Autoren vorrechnen. Da die Performance der Modelle nicht linear anwachse, sondern für immer klei-

nere Zugewinne immer mehr Energie aufgewandt werden müsse, seien Sprachmodelle schon bald mittelbare Faktoren im Klimawandel. Und der treffe vor allem den globalen Süden: Wieder wird bestehende Ungerechtigkeit weiter verstärkt.

Eine Welt aus Wörtern

Was es einigen in der KI-Szene so schwer macht, solche ethischen Fragen zu diskutieren, ist nicht allein die Hoffnung, genug Daten würden auch die Welt abbilden. Vielmehr spielt derselbe Kulturkampf eine Rolle, der auch in Europa stattfindet und in dem Gendersternchen und Antidiskriminierung tiefe Gräben aufreissen, aber der seit Jahren eigentlich keine neuen Argumente mehr hervorbringt. Sich auf *biases* einzulassen, hiesse nämlich auch zuzugeben, dass es Punkte gibt, bei denen die *«woke police»* nicht ganz unrecht hat – und die sich nun sogar auch ganz konkret nachweisen lassen.

Das betrifft zuallererst die These, dass Ungerechtigkeiten miteinander verschränkt sind, dass sie Komplexe bilden, in denen am Ende diskriminierende Sprache auch etwas mit Klimagerechtigkeit zu tun hat. In solchen soziopolitischen Beziehungsgeflechten sind rein technische Lösungen nicht nur unzureichend, sondern würden mit einem Mal sehr viel mehr kosten als der Imagegewinn, den ein eigenes Ethik-Team mit sich brächte. Wenn Gebru fordert, bei jeder Investitionsentscheidung künftig immer die Folgen für die am meisten Marginalisierten einzubeziehen, ist klar, warum sie Google verlassen musste.

Aber auch andere identitätspolitische Grundannahmen scheinen sich zu bestätigen. Gerade bei Sprachmodellen wird das deutlich. In einem Kommentar auf «Heise» wies Autor Wolfgang Stieler darauf hin, dass Aktivistinnen für soziale Gerechtigkeit von einer konstruktivistischen Sicht auf die Welt ausgehen: Wörter bilden Wirklichkeit nicht bloss ab, sondern schaffen sie erst. So macht das generische Maskulinum andere Geschlechter unsichtbar, und das N-Wort führt historische Verletzung fort, selbst dann, wenn das in beiden Fällen nicht intendiert sein mag. Darüber wird im Moment heftig gestritten. Bei Sprachmodellen wie GPT-3 aber, schreibt Stieler, «trifft diese Theorie hundertprozentig zu. Sprache prägt ihr Verständnis von der Welt. Wörter und Sätze bilden für die KI die Basis für Zusammenhänge und Beziehungen zwischen den Dingen.»

Zuzugeben, dass nun technisch zumindest plausibel wird, was identitätspolitische Theorie und antirassistische Aktivisten immer behauptet haben, heisst für viele, über ihren eigenen Schatten zu springen. Und neben bloss weltanschaulichem Abscheu bereiten auch die praktischen Lösungen Bauchschmerzen. Denn wenn alle KI ideologisch ist, kann der Ausgleich eines *bias* nur durch einen wiederum ideologischen Eingriff geschehen. Wie aber sähen «besser kuratierte Daten» genau aus? Welche Texte würden aussortiert, welche stärker gewichtet? Ist das nicht die Umerziehung, die von konservativer Seite so gern an die Wand gemalt wird? Und würde ein Sprachmodell, das keine sexistische und rassistische Sprache kennt, nicht eine ganz andere Welt widerspiegeln als die unsere?

Datenpolitik

Hier wird es für viele heikel: Wenn KI aus Daten der Vergangenheit Handlungsanweisungen für die Zukunft schliesst, dann muss man, will man *biases* beseitigen, sie mit einem Idealbild der Wirklichkeit füttern. Kuratierung von Daten heisst dann zu formulieren, in welcher Welt wir le-

ben möchten. Und das ist, so weit haben kritische Stimmen wie Domingos recht, eine immens politische Frage.

Unrecht haben sie, wenn sie meinen, es gebe auch neutrale Kuratierung, oder wenn sie gar sagen, dass nichts zu tun unpolitisch wäre. Denn dass KI-Modelle nicht neutral sind, ist kaum zu bestreiten: Der private Thinktank Open AI, der auch GPT-3 entwickelt hat, zeigte in einer Analyse, dass sein jüngstes Modell ein eigenes Neuron ausgebildet hatte, das gleichermassen auf «Islam» und auf «Terrorismus» anspringt.

KI-Systeme codieren gesellschaftliche Visionen. In welcher Welt wir leben wollen, sollte aber in öffentlichen, demokratischen Prozessen ausgehandelt, nicht von privatwirtschaftlichen Unternehmen festgelegt werden. Das leisten auch interne Ethik-Abteilungen nicht, die bei Marktzwängen immer nachgeben müssen. Öffentliche Proteste bringen ebenfalls wenig, wenn es keine Mittel gibt, im Zweifelsfall gegen solche Produkte wirksam Einspruch zu erheben oder sie zu verbieten.

Statt unverbindlicher Ethikrichtlinien braucht es eine Gesetzgebung, wie sie die EU angekündigt, aber noch nicht umgesetzt hat. Dazu gehört als letztes, extremes Mittel auch die Vergemeinschaftung: Sollten KI-Modelle zu einem übermächtigen Ort von Gesellschaftsentwürfen werden, spricht nichts dagegen, sie tatsächlich öffentlicher Kontrolle zu unterstellen, indem man sie in die öffentliche Hand überführt.

Diese Optionen auszuhandeln, ist aber zuallererst Aufgabe der Zivilgesellschaft und der öffentlichen Debatte. Einen wichtigen Beitrag leisten NGOs wie «Our Data Bodies», die die Öffentlichkeit auch gegen Lobbystimmen aufklären. In Europa ist «Algorithm Watch» Vorreiter: Neben ihrem Grundsatzreport «Automating Society» (für Deutschland und die Schweiz) über den zunehmenden Einfluss automatisierter Entscheidungssysteme hat die Organisation auch die Website Unding.de ins Leben gerufen: Sie hilft Betroffenen, die Opfer solcher Systeme wurden, bei den Unternehmen gezielt Einspruch zu erheben, solange es keine klaren, rechtlich gesicherten Beschwerdewege gibt.

In diesem Kontext könnten auch die festgefahrenen Feuilletondebatten über identitätspolitische Theorie und ihre Gegner wieder konkreter werden, statt schon tausendfach gehörte Argumente aufzuwärmen. Und wenig ist konkreter als Folgen, die sich ergeben, werden die Ideologien beider Seiten tatsächlich technisch implementiert. Denn Ideologien sind es allemal. Sie durchziehen Daten und technische Systeme – Neutralität gibt es hier nicht.

Weicht man dieser Debatte aus, mag man sich weiter genüsslich über die politisch korrekte Sprachpolizei echauffieren. Aber die Entscheidung über das Bild der Welt, das uns KI-Systeme dann zeigen, fällt anderswo – bei Google, in Mountain View.

In einer früheren Version waren Links fehlerhaft gesetzt, wir entschuldigen uns für diesen Fehler.

Zum Autor

Hannes Bajohr studierte Philosophie, deutsche Literatur und neuere und neueste Geschichte. Er ist Junior Fellow am Collegium Helveticum, dem Institute for Advanced Studies der ETH Zürich, der Universität Zürich und der Zürcher Hochschule der Künste.