ARBEIT UND DIGITALISIERUNG 17.02.2021 | Eva Wolfangel

Mehr als ein Missverständnis

Künstliche Intelligenz schnappt rassistische und sexistische Vorurteile auf und zementiert sie so. Doch wollen Techkonzerne wirklich etwas ändern?

Sie sei gerade überraschend entlassen worden, teilte Timnit Gebru, eine der bekanntesten Ethik-Forscherinnen Googles, Anfang Dezember per Twitter mit. Binnen Stunden erregte der Vorfall große Aufmerksamkeit in der Tech-Community und erzeugte sehr viel Solidarität mit der afroamerikanischen Wissenschaftlerin. Für viele schien klar zu sein: Die Forscherin war mit ihrer Kritik zu unbequem geworden. Sowohl im Konzern selbst als auch durch die dort und anderswo entwickelte künstliche Intelligenz würden Minderheiten rassistisch diskriminiert, so ihr Vorwurf. Mehr als 2 500 Google-Mitarbeiter und mehr als 4 000 Unterstützerinnen aus der Wissenschaft unterschrieben einen Protestbrief.

Googles Chef für künstliche Intelligenz Jeff Dean sah sich angesichts des Protests genötigt, eine E-Mail zu veröffentlichen, die er an Gebrus Team geschickt hatte. Darin erklärt er, Kern des Streits sei ein Forschungsartikel gewesen, den Gebru mit anderen verfasst hatte. Er habe dessen Veröffentlichung untersagt, weil die aktuelle Forschung darin zu wenig berücksichtigt würde. Daraufhin habe Gebru gedroht zu kündigen – und Google habe dies noch am gleichen Tag akzeptiert. Gebru hingegen sagt, sie habe nicht gekündigt. Selbst wenn es sich um ein Missverständnis gehandelt hat: Es besteht der Verdacht, dass Google die Gelegenheit genutzt hat, eine unbequeme Mitarbeiterin loszuwerden.

Nun ist der Artikel veröffentlicht worden: Neben Gebrus Namen und dem externer Forscherinnen wie der Hauptautorin Emily Bender von der University of Washington tauchen keine beteiligten Google-Mitarbeiterinnen auf. Zwar sind sich Kritiker einig, dass der Artikel allein wohl kaum der Grund für eine Kündigung sein kann. Doch ein Blick darauf macht deutlich, was Google möglicherweise an Gebrus Position stört und dass der Konzern die Arbeit gegen rassistische und sexistische Stereotype im maschinellen Lernen einem teils absurden Wettkampf um die Größe von Künstliche-Intelligenz-Modellen unterordnet.

So übersetzt Google Translate beispielsweise „The doctor and the nurse“ immer als „Der Arzt und die Schwester“, obwohl es ebenso gut „Die Ärztin und der Pfleger“ heißen könnte.

Die Autorinnen argumentieren, dass der Trend zu immer größeren Modellen und immer mehr Trainingsdaten in der Computerlinguistik dazu führt, dass nicht nur massiv Ressourcen verbraucht werden (Strom für riesige Serverfarmen). Künstliche Intelligenz wird auf diese Weise auch immer schlechter kontrollierbar. Sie diskriminiert Minderheiten, ohne dass es den Entwicklern bewusst ist.

Dazu muss man wissen, wie künstliche Intelligenz (KI) im Bereich Sprache lernt. Vereinfacht gesagt bekommen die Systeme des maschinellen Lernens von Forscherinnen und Forschern zwei Zutaten vorgelegt: Input- sowie gewünschte Output-Daten. Beim maschinellen Übersetzen also beispielsweise große Mengen an Text, der in verschiedenen Sprachen vorliegt. Daraus lernen die Systeme dann selbst Zusammenhänge zwischen Sprachen. Für jede Form des maschinellen Sprachenlernens hat sich das Internet als gute Quelle erwiesen, schließlich ist es voller Sprachdaten.

Doch genau hier lauert auch eine Gefahr, betonen die Autorinnen: Große Datensätze, die auf Texten aus dem Internet basieren, würden „hegemoniale Standpunkte überrepräsentieren und Vorurteile codieren, die möglicherweise marginalisierte Bevölkerungsgruppen benachteiligen“. In der Tat hat die Fachrichtung des maschinellen Lernens seit vielen Jahren ein Problem der rassistischen und sexistischen Verzerrung. Das zeigt sich unter anderem in automatischen Übersetzungsprogrammen wie Google Translate, die beharrlich Rollenklischees verstärken, indem sie zum Beispiel Berufe aus Sprachen ohne grammatikalisches Geschlecht entsprechend übersetzen. So übersetzt Google Translate beispielsweise „The doctor and the nurse“ immer als „Der Arzt und die Schwester“, obwohl es ebenso gut „Die Ärztin und der Pfleger“ heißen könnte. Das passiert auch dann, wenn es sich – für Menschen – eindeutig erkennbar um eine Ärztin handelt, weil sie beispielsweise im Satz zuvor erwähnt wurde.

KI-Forscher reagieren auf rassistische und sexistische Diskriminierung in Künstliche-Intelligenz-Systemen oft, indem sie die Menge der Trainingsdaten erhöhen.

KI-Forscher reagieren auf rassistische und sexistische Diskriminierung in Künstliche-Intelligenz-Systemen oft, indem sie die Menge der Trainingsdaten erhöhen und ebenso die der Parameter in der Hoffnung, dass die Ergebnisse repräsentativer werden. Parameter bezeichnen grob gesagt die Menge der Ausdifferenzierungsmöglichkeiten, die ein solches Netz hat. Manche vergleichen es mit der

Anzahl der Synapsen im menschlichen Gehirn und damit der Auswahl möglicher Verbindungen, was allerdings nur eine sehr grobe Annäherung ist.

Wie wenig allerdings riesige Mengen an Parametern gegen menschliche Vorurteile im maschinellen Lernen helfen, hat zuletzt das Beispiel des Sprach-KI-Modells GPT-3 von OpenAI gezeigt, das 175 Milliarden Parameter hat – und trotzdem rassistische und sexistische Verzerrungen aufwies, wie die OpenAI-Forscherinnen und - Forscher selbst schreiben.

Gebru und ihre Kolleginnen räumen ein, dass die riesigen Modelle zwar erfolgreich seien, wenn es um spezifische Anwendungsfälle ginge. Die gesellschaftlichen Nachteile überwiegen aber aus ihrer Sicht, da sie Sprache zwar erfolgreich imitieren, letztlich aber nicht verstehen. Da sie aber in der Lage sind, kongruente Texte zu produzieren, kommen sie der Neigung des menschlichen Gehirns entgegen, Sinn in Sprache zu finden. Aus ethischer Perspektive sei es wichtig, „den Nutzen ebenso wie die Risiken der Nachahmung von Menschen zu untersuchen“. Gebru und ihre Kolleginnen schlagen vor, kleinere Modelle zu verwenden und ebenso kleinere Datensätze, um Zeit und Kapazitäten zu haben, diese zu kuratieren.

Es sei keine Frage, die sich allein technisch lösen lasse, auch nicht mit kleineren Modellen und kuratierten Datensätzen: „Diese Vorurteile liegen in unserer Kultur, das sind wir.“

Der Ansatz von Bender, Gebru und Kolleginnen gehe allerdings nicht weit genug, sagt Joanna Bryson, Technikethikerin an der Hertie School of Governance. „Viele denken, man muss nur die richtigen Trainingsdaten wählen, um Vorurteile aus dem maschinellen Lernen zu bekommen.“ Aber es sei keine Frage, die sich allein technisch lösen lasse, auch nicht mit kleineren Modellen und kuratierten Datensätzen: „Diese Vorurteile liegen in unserer Kultur, das sind wir.“ Bryson hatte 2017 mit Kolleginnen unter anderem aus der Hirnforschung gezeigt, dass in unserer Art zu denken ähnliche Verzerrungen vorliegen.

Natürlich verwundert es nicht, dass die Vorurteile, die in den Trainingsdaten stecken, von den lernenden Maschinen übernommen werden. Über viele Jahre haben Forscherinnen versucht, diesen Bias zu eliminieren, indem sie beispielsweise bestimmte Teile oder Informationen aus den Daten entfernten – doch die Systeme des maschinellen Lernens sind so gut darin, Muster in Daten zu finden, dass sie die rassistischen oder sexistischen Zusammenhänge oft selbst wieder

rekonstruierten. Das zeigte sich am inzwischen berühmt gewordenen Beispiel der Software Compas, die in den USA Richter beraten sollte bei der Entscheidung, ob sie einen Gefangenen vorzeitig entlassen wollen. Das System benachteiligte systematisch Afroamerikaner – auch wenn die Information über die Hautfarbe gar nicht vorlag. So hatten Rechercheure von Pro Publica zeigen können, dass die Gerichtssoftware die Hautfarbe eines Straffälligen aus Drittvariablen, etwa dem Wohnort und dem Namen, errechnet hatte.

Sorelle Friedler, eine US-Informatikerin, die sich mit Ethik beschäftigt, betont: „Man kann nicht alle Attribute löschen, die möglicherweise zu Diskriminierung führen könnten, denn dann müsste man fast alle Daten löschen." Zumal uns Menschen vieles davon gar nicht bewusst ist.

Vorurteile, Verzerrungen und Ungerechtigkeiten erkennt erst, wer den Status quo an einem gesellschaftlichen Ideal misst. Was fehlt, sind also Formeln für Gerechtigkeit. Nur: Wie sollen die aussehen?

Vorurteile, Verzerrungen und Ungerechtigkeiten erkennt erst, wer den Status quo an einem gesellschaftlichen Ideal misst. Das müssten Menschen den Maschinen zuerst einmal vermitteln – in eindeutigen mathematischen Formeln. Was fehlt, sind also Formeln für Gerechtigkeit. Nur: Wie sollen die aussehen? Nehmen wir als Beispiel ein relativ simples, nämlich jenes, dass Männer und Frauen auf dem Arbeitsmarkt gleiche Chancen haben sollen. Klingt ziemlich unstrittig. Doch woran soll ein Algorithmus „gleiche Chancen" erkennen? Ist jede Gruppe ungleich, in der Frauen und Männer nicht fifty-fifty sind?

Noch einen Schritt weitergedacht: Wie illusorisch erscheint es, soziale Verzerrungen algorithmisch ausgleichen zu sollen? Sollte zum Beispiel eine Software, die Bewerber automatisch für ein Vorstellungsgespräch auswählt, darauf achten, dass Männer und Frauen in der Auswahl genau gleich verteilt sind? Oder sollte sie sich an der Geschlechterverteilung aller Bewerbungen orientieren? Oder das Geschlecht gänzlich ausblenden? Und welche Faktoren müssten für eine gerechte Verteilung ebenfalls bedacht werden? Qualifikation? Hautfarbe? Alter? – So schnell zerrinnt das vermeintlich simple Beispiel in einer langen Liste von Fragen.

Bryson hat einen anderen Vorschlag: „Wir müssen unsere Kultur ändern. Allerdings ist perfekte Fairness nicht möglich, da alle menschlichen Erfahrungen von der Vielfalt der Lebenswelten ausgehen und es keine Möglichkeit gibt, alle Unterschiede auszugleichen."

Immer wieder wurde gemutmaßt, dass Google letztlich nur das erforschen lässt, was marktwirtschaftlich gesehen Vorteile bringt. Die aktuellen Vorfälle bestätigen diesen Verdacht.

Unsere Kultur zu ändern, das ist ein langer, wenn nicht gar unmöglicher Weg. Das andere Extrem ist allerdings das, was gerade in manchen der Ethikabteilungen der großen Konzerne wie Google geschieht. Immer wieder wurde gemutmaßt, dass Google letztlich nur das erforschen lässt, was marktwirtschaftlich gesehen Vorteile bringt. Die aktuellen Vorfälle bestätigen diesen Verdacht.

Nur wenige Wochen, nachdem Jeff Dean untersagt hatte, den kritischen Artikel über große Sprachmodelle unter dem Namen von Google-Mitarbeitern zu publizieren, veröffentlichte Google Deep Mind am 11. Januar einen Rekord: ein riesiges Sprachmodell mit 1,6 Billionen Parametern, das anhand eines 750 Gigabyte großen Datensatzes mit Texten aus Wikipedia, Reddit und anderen Internetseiten trainiert wurde – „zwei Größenordnungen größer als Wikipedia“, wie Google-KI-Forscher betonen. Da wäre ein Artikel, der genau diesen Größenwahn kritisiert, wohl zur Unzeit gekommen. Von möglichen Nachteilen ist in der Veröffentlichung keine Rede, ebenso wenig vom Problem künstlicher Intelligenz mit gesellschaftlichen Stereotypen aus Internetartikeln.

Von einer fairen Zukunft mit künstlicher Intelligenz sind wir weit entfernt, solange in der KI-Forschung vor allem „höher, weiter, schneller“ zählt. Ethik darf nicht lediglich ein Lippenbekenntnis sein, das nicht mehr zählt, sobald es nicht mehr ins Marketing passt.



Eva Wolfangel

Eva Wolfangel ist Kulturwissenschaftlerin und Journalistin. Schwerpunkt ihrer Reportagen, die unter anderem in der ZEIT, dem Spiegel und dem GEO-Magazin erschienen sind, sind Zukunftstechnologien wie künstliche Intelligenz und virtuelle Realität.