Trau niemals einem Bild, das du nicht selbst bearbeitet hast. twitter/@osazuwa

«Plötzlich gehörten für Computer ‹Mexikaner› und ‹illegal› zusammen»

Die Wiener Wissenschaftlerin Doris Allhutter erforscht, wie Maschinen menschliche Stereotype erlernen. Und was man dagegen tun kann.

Ein Interview von Karin Cerny, 06.04.2021

Netflix weiss, welche Serie Sie als nächste schauen sollten. Basierend auf Ihrer bisherigen Auswahl und den Daten aller anderen Netflix-Nutzerinnen hat ein Computer einen Vorschlag berechnet, der Ihnen gefallen könnte.

Das Prinzip dahinter ist sogenanntes *machine learning,* ein Teilbereich der künstlichen Intelligenz. Es kommt bei Wetterprognosen zum Einsatz und bei selbstfahrenden Autos. Computer identifizieren mögliche Tumore auf Röntgenbildern oder geben Empfehlungen ab, die zunehmend unser Leben lenken – und darüber mitentscheiden, ob Sie etwa zu einem Bewerbungsgespräch eingeladen werden oder einen Kredit bekommen.

Zum Essay: Wer sind wir?

Algorithmen wiederholen Ungerechtigkeiten der realen Welt nicht nur – sie verstärken sie. Diese Verzerrungen zu beheben, wird schnell hochpolitisch: Warum künstliche Intelligenz immer ideologisch ist.

Im Grunde lernen Maschinen nicht völlig anders als wir Menschen: Aus Beispielen sammeln sie Erfahrung, die sie dann verallgemeinern. Allerdings lernen sie nicht einfach nur Beispiele auswendig, sondern erkennen Muster und Gesetzmässigkeiten. Umso wichtiger ist die Frage, woher die Maschinen ihr Wissen beziehen und welche gesellschaftlichen Stereotype sie dabei unhinterfragt bedienen.

Mit genau diesem Problem beschäftigt sich Doris Allhutter in ihrer Forschung am österreichischen Institut für Technikfolgen-Abschätzung in Wien. Sie ist Expertin für sogenannte *biases,* also Vorurteile, die Computersystemen eingeschrieben werden – und dazu führen, dass Personengruppen systematisch benachteiligt werden.

Frau Allhutter, Sie erforschen, wie Computer zu ihrem Alltagswissen kommen. Wie lernen Maschinen, was ein Mann ist und was eine Frau?
Eine grundlegende Annahme ist, dass sich die Geschlechter wesentlich voneinander unterscheiden. So definiert etwa Concept Net, eine Ressource, die Alltagswissen für *machine learning* aufbereitet, den Begriff «Mann» wie folgt: *«Man is a male person.»* Frauen werden hingegen über ihre reproduktiven Fähigkeiten definiert: *«Woman has a baby.»* Noch diskriminierender ist das Ergebnis, wenn es um zentrale Werte geht. *«Respect»* und *«honesty»* seien Männern wichtig. *«A woman wants to be loved and wants a man»* werden Frauen charakterisiert.

Woher kommen diese Klischees?
Wenn Systeme von Textdaten lernen, dann geht es meist über semantische Nähe, über Wörter, die häufig nah beieinander vorkommen. Zum Beispiel steht der Begriff «Programmierer» meist näher bei einem Männernamen. «Krankenpflege» näher bei Frauennamen. Man hinkt beim *machine learning* oft gesellschaftlichen Veränderungen hinterher, perpetuiert veraltete Rollenbilder.

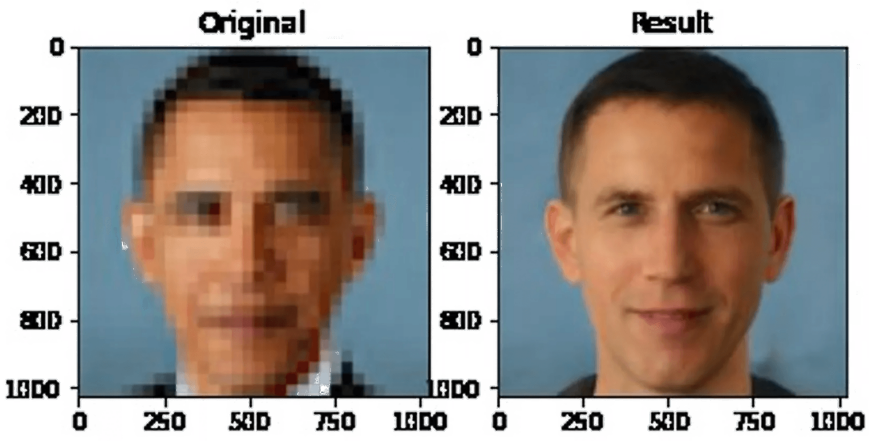
Woher beziehen die Maschinen ihre Informationen?
In früheren Expertensystemen schrieben Entwickler auf Basis dominanter gesellschaftlicher Annahmen fest, welches Wissen in Systeme gespeist wird. Damit gaben sie auch oft ihre männliche, weisse Mittelschicht-Sicht weiter. Inzwischen werden Informationen aus vielfältigen Quellen teils automatisch aus Big Data generiert. Das erscheint im Vergleich vielfältiger und demokratischer. Wobei auch da zu Beginn sogar Texte von Porno-Websites eingeflossen sind. Forscherinnen aus den USA wie Safiya Noble haben gezeigt, dass, wenn man den Begriff *«black girls»* in Google eingab, hauptsächlich sexualisierte oder pornografische Zuschreibungen auftauchten. Das wurde verbessert, trotzdem finden sich noch immer zahlreiche Stereotype.

Was muss man sich unter der Ressource Concept Net vorstellen?
Es gibt sogenannte Common-Sense-Ontologien, die aus Fakten über unsere Alltagswelt bestehen. Daten werden dahingehend aufbereitet, dass sie von Computern verstanden werden. Alltagssprache und -wissen wird so für Maschinen zugänglich gemacht. Schwierig wird es allerdings bei der Frage, was Fakten sind. Auf bestimmte Dinge können wir uns einigen. Zum Beispiel: Eine Zitrone ist sauer. Bei der Beschreibung von Männern und Frauen wird es komplizierter. Es gibt aber auch zahlreiche rassistische Beispiele, die man in dieser Quelle finden konnte.

Welche?
«Black man» wurde beschrieben durch folgende Sätze: «Ein schwarzer Mann greift in den Wäschetrockner.» Und: «ist mit dem Bösen verwandt» oder «wurde erschossen». Unter «schwarze Frau» findet sich in Concept Net

gar kein Eintrag. Das spiegelt wider, was oft in der Technik passiert, dass nämlich schwarze Frauen völlig unsichtbar sind.

Hat sich das inzwischen verändert?
Das gesellschaftliche Bewusstsein ist gewachsen. Man weiss inzwischen, dass es sogenannte *biases,* also Verzerrungen gibt, dass Daten bearbeitet und bereinigt werden müssen. Aber die Ergebnisse sind nach wie vor oft unbefriedigend. Die Beispiele, die ich vorhin genannt habe, stammten aus der Zeit vor dem sogenannten *debiasing,* also jenen Massnahmen, die darauf abzielen, Stereotype auszuscheiden. Danach stand bei *«black man»* nur mehr, das Konzept sei von «schwarz» und «Mann» abgeleitet, also keine absurde Tätigkeit, wie in den Wäschetrockner zu greifen. Aber noch immer, dass er in Beziehung zu «Dämon» und «böse» steht. «Weisser Mann» wurde hingegen beschrieben als «Kaukasier», «Europäer», «Good Old Boy» oder «White Trash». Mittlerweile wurden aber negative Assoziationen noch besser ausgefiltert.



Obama nach einer automatisierten Bildverbesserung. twitter/@Chicken3gg

Woher kommt die absurde Zuschreibung «Dämon»?
Aus dem englischen Wiktionary, einem Online-Wörterbuch, auf das beim *machine learning* zugegriffen wird. Beim «weissen Mann» ging es mehr um die Herkunft, und bei «schwarz» spielen figurative Bedeutungen aus alter Literatur mit. Aber auch semantische Nähe ist wichtig, wie ich eingangs ja schon beschrieben habe.

Aber ist das nicht fatal? Wenn «schwarzer Mann» im Internet oft neben «erschossen» steht, dann wird das einfach übernommen als Beschreibung?
Computer verstehen den Sinn ja nicht. Sie betrachten alles ohne Kontext. Sie wissen nicht, dass «erschossen» eine bestimmte Optik ergibt, dass Rassismus und Polizeigewalt mitspielen. Daher sind Kontextualisierung und *debiasing* wichtig. Das bleibt aber dennoch meist ungenügend.

Bedeutet das, dass sich aktuelle Debatten etwa zu «Black Lives Matter» und #MeToo auch negativ auf Computer auswirken können? Weil zum Beispiel Frauen noch mehr mit Gewalt, Hass und Vergewaltigung assoziiert werden?
Öffentliche Debatten haben oft eigenartige Auswirkungen. Dazu wird in den USA gerade viel geforscht. Zu Beginn von Donald Trumps Amtszeit wollte er ja eine Mauer zwischen den USA und Mexiko bauen. Der Diskurs über illegale Einwanderer war so stark, dass «illegal» und «Mexikaner» für Computer zusammengehörten.

Seit wann gibt es *debiasing* überhaupt?
In Europa hat das unter dem Titel *discrimination-aware data mining* um 2008 begonnen. Mittlerweile nennt sich das Forschungsfeld *fairness in machine learning*. Von der Mainstream-Computerwissenschaft wurde es lange als Humbug abgetan. Erst als mit sehr plakativen Beispielen, die durch die Medien gingen, die Aufmerksamkeit wuchs, hat sich eine internationale Community gebildet.

Grosse Unternehmen waren also zu Beginn zurückhaltend?
Sobald ich zugebe, dass meine Systeme Verzerrungen haben, lassen sie sich schlechter verkaufen. Computer basieren auf dem Mythos, dass sie objektiv und neutral sind. Aber mittlerweile gibt es auch bei Unternehmen wie Google Forschungsgruppen, die sich für Ethik und Diversität einsetzen. Die grossen Konzerne heften sich gern Diversität auf die Fahnen. Sobald es zu tief gehender Kritik kommt, werden wieder die alten Machtverhältnisse hergestellt. Man möchte sich schmücken, aber nicht grundsätzlich infrage stellen. Im Vorjahr hat Google seine renommierte schwarze Künstliche-Intelligenz-Ethik-Forscherin Timnit Gebru entlassen, was für viel Aufsehen sorgte. Auslöser des Konflikts war die Veröffentlichung eines Forschungspapiers, in dem die Wissenschaftlerin kritisierte, dass Techfirmen zu wenig tun, um Geschlechterstereotype und beleidigende Sprache in Systemen der künstlichen Intelligenz zu verhindern. Im Februar traf ihre Kollegin Margaret Mitchell dasselbe Schicksal. Sie hatte ihre internen Nachrichten durchsucht, um Beispiele zu finden, die belegen, dass Gebru im Betrieb diskriminiert wurde.

Wie funktioniert *debiasing* genau?
Auf Plattformen wie «Amazon Mechanical Turk» oder «Crowd Flower» erledigen Datenarbeiter sogenannte *microtasks*. Das ist mittlerweile eine grosse Industrie, die vorwiegend an Menschen aus dem Globalen Süden ausgelagert ist. Um ein Beispiel zu nennen: Eine *microtask*-Aufgabe wäre die Frage: «Ist die Erderwärmung ein Teil der Domäne Klimawandel?» Sie lässt sich mit «Ja» oder «Nein» beantworten. Oder es gibt Begriffspaare, und die Arbeiter sollten entscheiden, ob es sich um Stereotype oder eine adäquate Unterscheidung zwischen den Geschlechtern handelt.

Aber diese Arbeiter können doch auch Vorurteile haben, die sie an Computer weitergeben?
Genau, es kann auch hier unterschiedliche Ansichten je nach kulturellem Kontext geben. Crowdarbeiter gibt es in den USA, in Europa überwiegend in ärmeren Ländern, und auch in Indien. Viele von ihnen sind gut ausgebildet, kommen aber aus einer prekären Schicht, machen den Job, um sich über Wasser zu halten. Oft werden auch Mütter angesprochen, die sich etwas dazuverdienen wollen. Es sind unzählige Daten, die zu Geld gemacht werden sollen. Aber es kostet auch viel, um sie für Computer lesbar zu machen. Deshalb nimmt man billige Arbeitskräfte.

Ist es nicht sexistisch, wenn Männer und Frauen nur durch Unterschiede definiert werden?
Durchaus. Es gab die Idee, dass sich Geschlechterstereotype leicht beseitigen lassen. Es muss nur die Hierarchie zwischen zwei geschlechtsspezifischen Begriffen entfernt werden. Was dabei aber nicht hinterfragt wird, ist die Geschlechterbinarität selbst. Männer und Frauen müssen sich für Computer grundsätzlich unterscheiden. Es wird immer nach einer Differenz gesucht. Und das zu einer Zeit, in der Gendernormen zunehmend infrage gestellt werden. Noch komplexer gestaltet sich die Aufgabe, Rassismen zu entfernen. Man versucht dafür mit *sentiment analysis* positive und negative Assoziationen und Vorurteile zu erkennen.

Apropos Vorurteile. Bei Bildern wurde festgestellt, dass der Computer, wenn er das Gesicht eines Mannes mit dem Körper ergänzen soll, diesem einen Anzug anzieht. Bei einer Frau eher einen Bikini, selbst bei Politikerinnen. Wie funktioniert diese Bilderkennung?
Computer können auf das Bild selbst nicht zugreifen. Um es zu interpretieren, versuchen sie, es mit Sprache zu verbinden. Nehmen wir ein Beispiel: Ein Pferd steht vor einem Gebäude. Der Computer weiss vom Lernen aus Textdateien, dass ein Pferd wahrscheinlicher vor einem Stall steht als vor einer Kirche.

Bilderkennung ist also noch komplexer als Textlernen?
Wir helfen ja täglich mit. Immer wenn man anklicken soll, ob Autos oder Zebrastreifen auf Bildern zu sehen sind, dann trainieren wir Computer. Wir sind unbezahlte Arbeitskräfte, ohne es zu wissen. Überall werden Daten gesammelt: Die Barbie-Puppe nimmt Sprache im Kinderzimmer auf, Smartphones können mithören, Gesichtserkennungssysteme in Geschäften erkennen, wohin wir blicken.

Big Brother ist längst Gegenwart. Sollen wir deswegen Angst haben?
Viele denken sich: Ich habe ja nichts zu verbergen. Ich kann etwas gratis nutzen, ist mir doch egal, wenn meine Daten gesammelt werden und ich dann Werbung zugespielt bekomme. Der breiten Masse ist nicht klar, dass es gar nicht um individuelle Überwachung geht, sondern darum, dass gesellschaftliche Muster abgeleitet werden. Es geht vor allem darum, möglichst viele Daten zu sammeln und dann die Systeme damit zu trainieren.

Was muss sich Ihrer Meinung nach verändern?
Die Geschlechterforschung betont seit dreissig Jahren, dass Computersysteme gesellschaftliche Ungleichheiten reproduzieren. Langsam wächst das kritische Bewusstsein, verstärkt durch das Interesse an künstlicher Intelligenz. Mit rein technischen Lösungen kommt man aber nicht weiter. Es ist ein gesellschaftliches Problem. Systeme könnten besser gemacht werden, indem sie den sozialen Kontext besser verstehen würden. Dann müsste man mitdenken, dass es Diskriminierung und Rassismus gibt. Und wenn jemand erschossen wurde, dass dies eine bestimmte diskriminierende Praxis miteinschliesst.

Gerechtigkeit müsste eine Kategorie beim Programmieren werden?
Ja, die Computerwissenschaften müssen anfangen, politischer zu denken. Rassismus und Sexismus sollten als strukturelle Probleme betrachtet werden und nicht als etwas, was einzelnen Personen zufällig passiert. Unser Fairnessbegriff geht von der Frage aus, ob eine Entscheidung, die ein System trifft, für eine individuelle Person fair ist oder nicht. Es geht nicht darum, ob sie auch gesellschaftlich gerecht ist. In interdisziplinärer Zusammenarbeit kann hingegen genau beleuchtet werden, wie soziale Kategorien wie Geschlecht, Ethnizität, Alter oder Klasse algorithmisch zueinander in Beziehung gesetzt werden. Der Hausverstand der Computer muss gerechter werden. Aber natürlich ist das ein Projekt, das nie abgeschlossen sein wird.

In einer früheren Version haben wir geschrieben, dass Krebszellen auf Röntgenbildern identifiziert werden können. Das ist falsch, wir haben dies korrigiert – und bedanken uns herzlich für den Hinweis aus der Leserschaft.

Zur Autorin

Karin Cerny lebt in Wien. Sie schreibt regelmässig über Theater, Literatur und Kulturpolitik im Wochenmagazin «Profil» sowie Reise- und Modegeschichten für «Rondo», die Beilage der Tageszeitung «Der Standard». Für die Republik schrieb sie zuletzt über die Bedeutung von Künstlerinnen für den Surrealismus sowie über historische Stoffe in TV-Serien.