MEDIEN

NETZPOLITIK

# Algorithmen mit Vorurteilen

Joël Adami

**Über die Gefahr von vermeintlich neutralen Computern, die in wichtige Entscheidungsprozesse eingreifen.**

Algorithmen, die Dinge für uns sortieren, sind allgegenwärtig, zum Beispiel auf sozialen Netzwerken. Ob es wirklich sinnvoll ist, dass die Neujahrswünsche der Tante zum dritten Mal in der eigenen Facebook-Timeline auftauchen, weil sie so viele „Likes" haben, darüber machen sich wohl die wenigsten Gedanken. Als die automatische Gesichtserkennung von Google, die dazu gedacht ist, Fotos besser ordnen zu können, schwarze Menschen als „Gorillas" bezeichnete, gab es jedoch völlig zurecht einen Aufschrei. Und spätestens als die künstliche Intelligenz, die automatisch Bewerbungsunterlagen für Amazon durchsuchte, zum größten Teil weiße und männliche Bewerber*innen vorschlug, war vielen klar: So neutral, wie man vielleicht meinen könnte, sind Algorithmen oft gar nicht. Wie entstehen Vorurteile in Algorithmen und können sie wieder rausprogrammiert werden?

In vielen Bereichen sind Computer besser und schneller als Menschen. Vor allem dann, wenn es darum geht, große Datenmengen zu analysieren. Warum sollte man ihnen also nicht auch Aufgaben geben, die für Menschen sehr ermüdend sind - zum Beispiel die scheinbar immer gleichen Bewerbungsunterlagen vorzusortieren, so, dass nur die aussichtsreichsten Kandidat*innen übrig bleiben? Das Technologieunternehmen Amazon ging noch einen Schritt weiter und gab dem Algorithmus mit auf den Weg, möglichst jene Kandidat*innen zu bevorzugen, die jenen glichen, die in den letzten zehn Jahren bei der Firma angestellt wurden. Das Ergebnis war jedoch wenig revolutionär: Der Algorithmus reproduzierte die Vorurteile, die auch menschliche Sortierer*innen haben.

## Künstliche Vorurteile

Was ist eigentlich ein Algorithmus und wie kann ein Computer Vorurteile lernen? Der Begriff Algorithmus geht auf den Mathematiker Abu Dscha'far Muhammad ibn Musa al-Chwarizmi zurück. Als dessen 825 in Bagdad verfasstes mathematisches Lehrbuch im 12. Jahrhundert auf Latein übersetzt wurde, wurde aus al-Chwarizmi „Algorismi", was später zu Algorithmen wurde. Ein Algorithmus lässt sich mit einem Kochrezept vergleichen: Eine endliche Reihe von Anweisungen, die immer das gleiche Resultat ergeben.

Ein Schlagwort, das in den letzten Jahren immer häufiger in Verbindung mit künstlicher Intelligenz und Algorithmen zu hören ist, ist Machine Learning. Bei diesem „maschinellen Lernen" wird ein neuronales Netz simuliert, das wie das menschliche Gehirn funktioniert, nur meistens bedeutend kleiner ist. Dieses Netz verarbeitet einen Trainingsdatensatz, anhand dessen es Gesetzmäßigkeiten lernen soll. Um mit Machine Learning zum Beispiel Bilderkennung umsetzen zu können, muss das neuronale Netz sich also sehr viele Bilder und die dazugehörigen Beschreibungen „ansehen", bis es irgendwann erkennen kann, was darauf zu sehen ist. Wie gut es darin ist, liegt letzten Endes nicht nur an den Programmierkünsten seiner Macher*innen, sondern auch an der Qualität der Trainingsdaten. So kann es vorkommen, dass ein neuronales Netzwerk wie die Bilderkennung von Google beispielsweise denkt, die englische Queen trüge eine Badehaube – obwohl die Monarchin auf dem betreffenden Foto natürlich ihre Krone trägt. Oder der Porno-Erkennungsdienst des sozialen Netzwerkes tumblr, der Dünen für Nacktfotos hielt (woxx 1507). So lustig solche Beispiele im Einzelfall auch sein mögen, in ihrer Gesamtheit zeigen sie, dass kein System über sich selbst hinauswachsen kann - und bevor ein Algorithmus eine Entscheidungshilfe anbietet, fließen sehr viele menschliche Vorurteile in diesen Prozess mit ein.

Ein Beispiel, das dies sehr gut illustriert, stammt aus Österreich. Das Arbeitsmarktservice, das dortige Arbeitsamt, versucht, die Chancen von Arbeitslosen auf dem Jobmarkt mittels eines Algorithmus zu bestimmen. Die Arbeitssuchenden werden in drei Kategorien eingeteilt: „sehr gute", „mittlere" und „geringe Integrationschancen". Bisher wird das Programm getestet, ab 2020 soll es dann auch Konsequenzen haben. Absurderweise könnte diese Einteilung dazu führen, dass die Menschen, die in die unterste Kategorie fallen, gar keine oder wenig Unterstützung bei der Arbeitssuche erhalten. Die Logik des Systems: In Menschen, die kaum Chancen bei der Arbeitssuche haben, sollen möglichst wenig Ressourcen gesteckt werden.

Eine Person, die solche Beispiele kritisch untersucht, ist Sabrina Burtscher. Die Informatikstudentin aus Wien hat im Rahmen einer Lehrveranstaltung namens „Critical algorithm studies" an der Technischen Universität Wien ihr Interesse für das Thema entdeckt: „Ich habe aber auch bei meinem Engagement bei der Studierendenvertretung gemerkt, dass manche Systeme, zum Beispiel Aufnahmeverfahren, unfair oder benachteiligend sein können." Zu dem AMS-System hat sie mehrere Vorträge gehalten, zum Beispiel bei der österreichischen „PrivacyWeek" letzten Herbst oder im Rahmen des jährlichen Hacker*innenkongresses des Chaos Computer Clubs in Deutschland im Dezember.

„In den AMS-Algorithmus sind die klassischen Diskriminierungsachsen wie Geschlecht, Herkunft, Alter und

Bunte Lichtinstallationen sind ein Teil der Hacker*innenkonferenz 35C3 gewesen – Vorträge über die Gefahren von Algorithmic Bias ein anderer.

FOTO: YVES SORGE.

Bildung geflossen. Es ist ein Glücksfall, dass der Algorithmus veröffentlicht wurde, denn so kann man sich anschauen, wie diese systematische Diskriminierung entsteht", sagt Burtscher im Gespräch mit der woxx. Frauen haben im AMS-Algorithmus schlechtere Karten als Männer, auch wenn alle anderen Faktoren gleich sind. Sie haben also höhere Chancen, in eine Kategorie zu fallen, in der sie als schlecht vermittelbar gelten – mit allen Konsequenzen, die dies haben kann.

## Diskriminierung in Algorithmen gegossen

Das Phänomen hat einen Namen: „Algorithmic Bias", also Algorithmische Verzerrung. „Mit Algorithmic Bias wird die systematische Benachteiligung von spezifischen Gruppen durch einen Algorithmus bezeichnet", erklärt Burtscher. „Üblicherweise basiert diese Benachteiligung darauf, wie der Algorithmus trainiert worden ist oder wie er Dinge gewichtet. Der Algorithmus lernt dann halt aus unterschiedlichen Gründen eine systematische Benachteiligung." Oft sind es also die Trainingsdaten, mit denen Machine-Learning-Programme lernen sollen, in denen sich schon Ungleichheiten abbilden. Bei den schwarzen Menschen, die von Google als Gorillas gekennzeichnet worden sind, liegt die Erklärung nahe: In den Fotos, die zum Training der Gesichtserkennung eingesetzt wurden, waren vor allem weiße Menschen zu sehen. Nicht unbedingt ein Wunder in einer Industrie, die weiß und männlich dominiert ist.

Es sind aber nicht nur der Algorithmus oder die Trainingsdaten wichtig, sondern auch die Darstellung der Ergebnisse kann entscheidend sein. „Oft treffen solche Algorithmen ja keine endgültigen Entscheidungen, sondern zeigen eine Vorsortierung an und geben einen Überblick. Wie diese Auflistung oder Sortierung dargestellt wird und welche Faktoren angezeigt werden, ist kritisch dafür, wie letzten Endes die Qualität der Entscheidungen ist", gibt Burtscher zu bedenken.

Dieses Phänomen ist bereits seit Längerem bekannt. Eine Studie ergab 1982, dass die allermeisten Flugbuchungen, die von Angestellten in Reisebüros oder an Flughäfen über das Buchungssystem Sabre getätigt wurden, auf der ersten Bildschirmseite erschienen, über die Hälfte waren sogar das allererste Ergebnis. Das lag unter anderem daran, dass die Bildschirme damals klein und das Nutzer*inneninterface umständlich war. American Airlines, die am meisten Anteile an dem System hielten, ließen den Algorithmus so manipulieren, dass ihre Flüge stets an erster Stelle erschienen – selbst dann, wenn die Verbindung weder die schnellste noch die billigste war. Ein heute bekanntes Phänomen sind die Google-Suchergebnisse – kaum jemand klickt sich hier auf die zweite Seite durch.

Dies lässt sich auch in Studien nachweisen. Bettina Berendt und Sören Preibusch von der Universität Leuven haben Proband*innen gebeten, in die Rolle von Bankangestellten zu schlüpfen, die über die Vergabe eines Darlehens entscheiden. Dabei wurden Faktoren wie Herkunft oder Geschlecht auf drei unterschiedliche Weisen dargestellt. Das Ergebnis: Bei jener Variante, bei der das System solche geschützten Merkmale unterschlug, entschieden die Proband*innen gemäß der Vorgaben und hielten alle Regeln bei der Kreditvergabe ein. Die Darstellungsform kann also auch dabei helfen, Diskriminierungen zu vermindern.

## Untransparente Blackboxes

Ist es möglich, den Bias, die Vorurteile wieder aus Algorithmen herauszubekommen, oder welche zu erstellen, die erst gar keine enthalten? „Es hängt immer an der Fragestellung und an der Art von Daten, die genutzt werden. Grundsätzlich ist es ein Problem, dass Machine-Learning-Systeme oft untransparente Blackboxes sind, in die nicht mal jene reinschauen können, die die Systeme trainiert haben", so Burtscher. Transparenz ist sowieso oft alleine deswegen nicht gegeben, weil die algorithmischen Systeme dem Geschäftsgeheimnis unterliegen und somit nicht offengelegt werden. Auch wenn man einem Machine-Learning-System verbietet, das Attribut Geschlecht zu verwenden, kann es trotzdem vorkommen, dass es einen Weg findet, Männer zu bevorzugen – zum Beispiel über Hobbys oder bestimmte Universitäten, die in Bewerbungsunterlagen vorkommen.

Sabrina Burtscher plädiert dafür, sich des Problems bewusst zu werden: „Es ist problematisch, wenn Menschen glauben, es könnte einen neutralen Algorithmus geben. An jeder Stelle im Entwicklungszyklus einer Software kann Bias mit einfließen, also muss man auch an jedem Schritt ansetzen, diese Vorurteile zu bekämpfen." Die Lösung dieses Problems liegt aber vielleicht gar nicht darin, möglichst neutrale Algorithmen zu schaffen, sondern das Ziel genauer zu definieren: „Wenn Amazon beispielsweise für mehr Diversität am Arbeitsplatz sorgen möchte, könnten sie den Algorithmus so trainieren, dass zum Beispiel mehr schwarze Frauen zu Bewerbungsgesprächen eingeladen werden."

---

Mehr über Sabrina Burtschers Arbeiten und Vorträge findet sich unter https://pascoda.fairydust.space/ und auf twitter unter @pascoda.