HUMAN RESOURCES

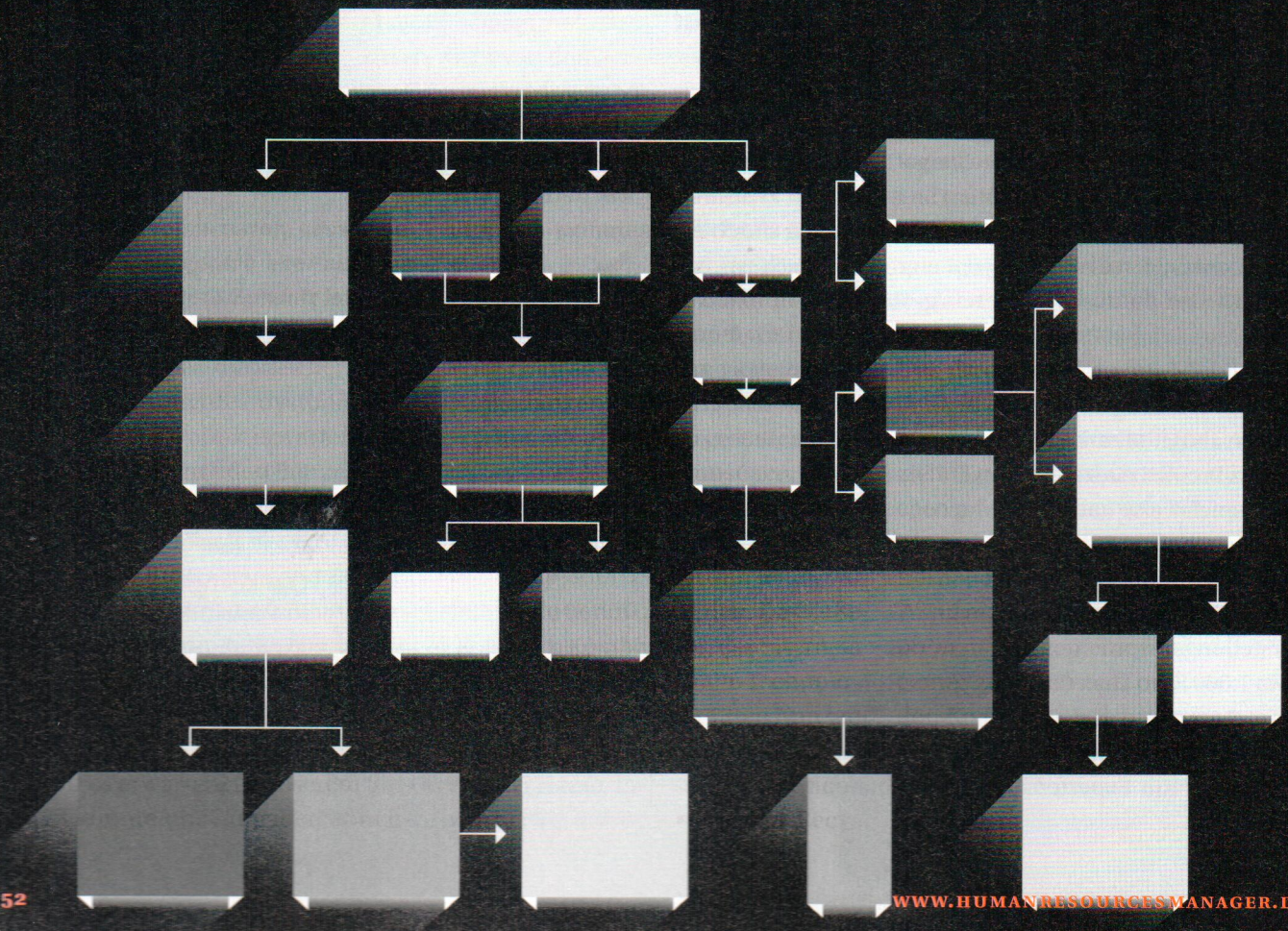
MANAGER

THEMA KÜNSTLICHE INTELLIGENZ

Entseelte Entscheidungen

Ein Interview von
Jeanne Wellnitz

Algorithmische Entscheidungssysteme sind lernende und damit auch verheißungsvolle Systeme: Sie sortieren Bewerbungen nach dem vermeintlich perfekten Kandidaten. Vielleicht können sie sogar die Leistung von Mitarbeitern einschätzen. Doch sind Maschinen tatsächlich objektiver als der Mensch? Vertragen wir automatisiertes Feedback besser? Und welche Gefahren lauern in diesen neuen Entwicklungen? Die Algorithmus-Designerin Katharina Zweig im Interview.



Wenn es um automatisierte Entscheidungen und Algorithmen geht, ist Katharina Zweig ein häufig geladener Gast. Auf Konferenzen, Medienveranstaltungen oder in Diskussionsrunden beantwortet die Informatikprofessorin die brennende Frage: Brauchen wir eine Ethik für Algorithmen? Die 41-jährige Algorithmus-Designerin stellte vergangenes Jahr gemeinsam mit der Philosophin Lorena Jaume-Palasí, dem Datenjournalisten Lorenz Matzat und dem Reporter-ohne-Grenzen-Vorstandsmitglied Matthias Spielkamp die Initiative Algorithm Watch vor. Sie warnt davor, dass algorithmische Entscheidungssysteme, wenn sie nicht gut gemacht sind, diskriminierende Entscheidungen treffen können.

Frau Zweig, trifft eine Maschine bessere Entscheidungen als ein Personaler?

Katharina Zweig: Das hängt von der Situation ab. Maschinen können erst mal sehr viel mehr Daten verarbeiten, verrechnen sich nicht und man kann ihnen potenziell diskriminierende Eigenschaften vorenthalten. Jedoch kann es, erstens, natürlich sein, dass es gar keine 100-Prozent-Regel gibt, nach der man Menschen in gute und schlechte Mitarbeiter einteilen kann. Dann wird auch eine Maschine Fehler machen. Zweitens müsste man im Nachhinein überprüfen, ob sie tatsächlich weniger sowie weniger gravierende Fehler gemacht hat als der Mensch. Drittens lernen solche Systeme aus einer riesigen Menge von Daten – sind diese ungeeignet für die gewollten Entscheidungen, lernt die Maschine falsche Regeln.

Was könnte so eine falsche Regel sein?

In Deutschland wurde mehrfach nachgewiesen, dass Bewerber, deren Namen einen Migrationshintergrund suggerieren, seltener eingeladen werden als jene mit deutschem Namen – trotz identischer Lebensläufe. Wenn die Daten darüber, welche Bewerbung später erfolgreich war oder nicht, genau diesen Fehler enthalten, dann wird die Maschine das mitlernen. Und damit setzt sich die Diskriminierung in vermeintlich objektiven Berechnungen fort.

Die Maschine kann also gar nicht wirklich objektiv sein?

Es gibt die Hoffnung, dass man objektivere Entscheidungen erhält, wenn man ihr gewisse Daten vorenthält, die man einem Menschen gar nicht vorenthalten kann. Wie wollen Sie beispielsweise jeden Hinweis auf das mögliche Geschlecht aus einem Lebenslauf löschen – muss man dann „Fußballspielen" unter den genannten Hobbys streichen oder Mutterschutzangaben nicht machen dürfen? Dem Computer können wir diese Information leichter vorenthalten. Die Chancen einer automatischen Bewertung bestehen also darin, dass wir die Algorithmen dadurch tatsächlich objektiver machen können.

Und welche Fehlerquellen gibt es bei der Entwicklung eines Algorithmus?

Wir sprechen ja von algorithmischen Entscheidungssystemen; kurz ADM-Systeme (*Anm. d. Red.: Algorithmic Decision Making Systems*). Diese basieren auf einem klassischen Algorithmus, meistens aus dem Gebiet des maschinellen Lernens. Sie brauchen aber auch Daten – und oft werden die Resultate noch weiter verarbeitet und visualisiert. Manchmal lernen die Systeme auch kontinuierlich. In ADM-Systemen kann es an all diesen Stellen und weiteren Punkten Fehler geben.

Müssten dann nicht bei der Auswahl der Daten ein Soziologe, ein Psychologe und ein Moralphilosoph mit am Tisch sitzen?

Genau das ist das Problem. Die Implementierungsfälle, die wir begleiten konnten, zeigten: Bei der Entwicklung des ADM-Systems und der Datenauswahl sind oft keine der genannten Experten beteiligt. Auch die Schulung der Mitarbeiter fällt oft dürftig aus. Das ist einer der wichtigsten Gründe, warum wir Algorithm Watch gegründet haben. Wir sehen Chancen durch diese Systeme, aber eben auch viele Risiken, wenn sie schlecht gemacht sind.

Sie haben Algorithm Watch 2016 gemeinsam mit ihren Kollegen auf der Digitalkonferenz Re:publica vorgestellt. Dort sprachen Sie auch davon, dass Algorithmen entseelte Entscheidungen treffen. Was meinten Sie damit?

Wir müssen bei ADM-Systemen mehr festlegen, als wir es normalerweise bei Entscheidungen tun. Es gibt keinen Entscheidungsspielraum. Eine Freundin von mir lebt beispielsweise in einer großen Stadt und ihre Wohnung ist fünf Euro teurer, als sie für den Hartz-IV-Bezug sein dürfte. Ihr Jobvermittler nutzt seinen Ermessensspielraum und lässt sie und ihr Kind dennoch darin wohnen. Das kann ein Algorithmus prinzipiell nicht lernen, weil seine Entscheidung sonst wieder eine Regel wird. Eine Menge von festen Regeln ist aber kein Ermessensspielraum. Ich denke jedoch, dass wir genau diesen Spielraum manchmal brauchen, weil es schwierig ist, für selten auftretende Fälle sinnvolle Regeln aufzustellen, die auf Dauer gerecht sind.

Und das, obwohl es *lernende* Systeme sind?

Ja, solche Systeme lernen aus Daten nur die häufig auftretenden Strukturen, sie sind danach relativ fix. Sie lernen nur in dem Maß weiter, wie wir ihnen Futter bieten. Wenn ein Algorithmus eine Bewerbung aussortiert hat, obwohl der Kandidat unter anderen Gesichtspunkten geeignet gewesen wäre, können Sie dem Computer nicht sagen: „Das wäre eigentlich ein toller Kandidat gewesen, den du nach Hause geschickt hast. Merke dir das für das nächste Mal." Man sieht nur noch die

Talente, die vom Computer durchgelassen wurden.

In einem anderen Interview erwähnten Sie, dass eine Offenlegung der Algorithmen nicht ausreichen würde, um sie zu kontrollieren. Weshalb?

Die Personen, die das vorschlagen, stellen sich das zu leicht vor: Man schaut in den Code und es ist klar, was die Algorithmen eigentlich machen. Das ist nicht der Fall – es kostet viele Personenstunden, den Code eines anderen nachzuvollziehen. Diese Forderung reicht aber auch deswegen nicht aus, weil der Algorithmus nur ein Teil des Systems ist; man müsste auch alle anderen Aspekte analysieren. Zudem wäre der Zwang zur Offenlegung von Codes ein Innovationshemmnis. Das Offenlegen ist übrigens auch nicht die einzige Kontrollmöglichkeit.

Sondern?

Wenn ich zum Beispiel wissen will, ob ein ADM-System diskriminierend ist, schaue ich mir 200, 300 Entscheidungen an und prüfe, ob mehr Frauen oder mehr Männer abgewiesen wurden, ob Behinderte ausreichend berücksichtigt wurden oder ob ein Migrationshintergrund ein Hindernis war – in Abhängigkeit von den eingegangenen Bewerbungen natürlich. Dazu muss ich den Code nicht kennen. Ein ähnliches Projekt führen wir gerade mit Bezug auf Suchmaschinenergebnisse im Bundestagswahlkampf durch.

Wenn ein System in einer Unternehmensabteilung implementiert wird, wer kontrolliert es dort derart?

Das ist besonders schwierig. Meistens gibt es eine grundlegende algorithmische Softwarelösung einer Firma, in welche die Daten der vergangenen zehn Jahre als Basis der Analyse eingegeben werden. Dann extrahiert der Algorithmus die Regeln, nach denen eine Bewerbung als geeignet befunden wird, oder die Werte, nach denen entschieden werden sollte, ob jemand befördert wird. Wir brauchen also nicht nur jemanden, der in dieses Softwareprodukt hineingucken kann, sondern es muss auch eine Begleitung geben, die den Lernvorgang betreut. Die Betriebsräte müssen darauf achten, dass das Training zur Nutzung des entstandenen ADM-Systems innerhalb der Firma korrekt läuft. Sie sollten außerdem die Datenauswahl überprüfen: Name und Geburtstort müssen beispielsweise nicht rein, Alter sicherlich schon. Bei der Bewertung von Leistung dürften zum Beispiel Krankentage nicht mit eingerechnet werden.

Katharina Zweig ist Forscherin und Universitätsprofessorin auf dem Gebiet der Algorithm Accountability. 2016 gründete sie zusammen mit Lorena Jaume-Palasí, Lorenz Matzat und Matthias Spielkamp die NGO Algorithm Watch.

Der Chatbot Tay von Microsoft hat vergangenes Jahr gezeigt, dass auch gut designte ADM-Systeme moralisch fragwürdige Ergebnisse aufweisen können: Er wurde von einer Gruppe mit rechtsradikalen Äußerungen gefüttert und hat dann selbst rechtsextreme Tweets abgesendet. Sollte ein Entwickler dem Algorithmus Moral implementieren, so dass er auf ihrer Grundlage entscheiden kann?

Das ist eine schwierige Frage. Ist der Designer eines solchen algorithmischen Systems für das Handeln seines Bots voll verantwortlich? Oder ist es nicht vielmehr die Gesellschaft, die den Bot in diese Richtung trainiert hat? Tay sollte eigentlich lernen, worüber Leute tweeten, und dann selbst Tweets absenden, die inhaltlich dazu passen. Dann haben sich Personen verabredet, ihn mit rechtsradikalen Informationen zu bombardieren, und dann – aus der Logik seiner Programmierung heraus folgerichtig – hat der Chatbot rechtsradikale Sprüche aus dem Netz gesucht und sie getweetet. In diesem speziellen Fall hätte man ihm vielleicht eine Liste von Wörtern mitgeben können, die garantiert tabu sind – aber das stellt auch immer eine vorweggenommene Zensur dar, die vermutlich genauso viel Unsinn erzeugt. Der Chatbot könnte dann ja beispielsweise auch keine Tweets absetzen, die Rechtsradika-

lismus verurteilen. Aus meiner Sicht wäre es besser, wenn wir dafür sorgen, dass sie „erzogen" werden können, also durch Korrekturen von außen ihr erlerntes Verhalten ändern.

So etwas könnte auch im HR-Kontext passieren.

Ja, natürlich. Nehmen wir an, es fängt beispielsweise ein neuer Manager an, der seine Mitarbeitergespräche führt und mit seinen Beurteilungen ein ADM System trainiert. Wäre er ein Mensch, der beispielsweise denkt, dass Frauen weniger leisten können als Männer, würde die Maschine seine Vorurteile lernen.

Was kann man dagegen tun?

Man könnte natürlich protokollieren, ob der Mensch, der die Maschine trainiert, solchen Vorurteilen unterliegt – und das wäre in diesem Ausmaß und dieser Granularität vermutlich zum ersten Mal in der Geschichte der Menschheit möglich. Und da wird es datenschutzrechtlich interessant: Darf man uns derart überwachen? Andere Fragen sind, ob die Maschine jegliche Art von Diskriminierung direkt vermeiden sollte – das könnte sie nämlich – oder ob die Firma, die die Software nutzt, dafür sorgen muss. Gibt es Berufe, wo es in Ordnung ist, zwischen Menschen beispielsweise nach Geschlecht oder Herkunft zu differenzieren? Diese Diskussion haben wir häufig vor deutschen Gerichten. Man sieht also, dass viele Probleme mit ADM-Systemen keine prinzipiell neuen sind. Es sind oft bekannte Probleme, die in neuem Gewand auf uns zukommen.

Wo wollen wir also ausschließlich menschliche Entscheidungen haben?

Ich denke, dass Entscheidungen grundsätzlich davon profitieren, wenn sie auf der Basis von dafür erhobenen Daten getroffen werden. Diese Datenanalyse ist also zu trennen von der Nutzung eines ADM-Systems, das Menschen direkt kategorisiert oder bewertet. Ansonsten gibt es drei Punkte, die darüber entscheiden sollten, ob man ein ADM-System einsetzen will:

Erstens: Tritt der zu bewertende Fall zu selten auf, wird ein ADM-System nicht genügend Datenpunkte haben, um sinnvolle Regeln ableiten zu können. Irgendwelche Regeln wird es aber aufstellen, die dann objektiv scheinen.

Zweitens: Wenn der Fall zwar oft auftritt, die wesentlichen Daten aber schwer in guter Qualität zu erhalten sind, sollte ein ADM-System nicht verwendet werden. Beispiele dafür sind Leistungsbewertungen von Mitarbeiterinnen und Mitarbeitern aus Sensordaten: Denkt jemand, der auf einen Bildschirm starrt, nach oder ist er abgelenkt? Solange das nicht gut unterschieden werden kann, sollten darauf keine Entscheidungen aufgebaut werden.

Drittens: Wenn es aller Erfahrung nach keine einfachen Regeln gibt, mit denen eine Person zu bewerten ist, sondern viele miteinander im Konflikt stehende Eigenschaften zu bedenken sind, verschaffen ADM-Systeme nur scheinbar eine objektivere Bewertung.

Aber wie gesagt: Auch bei menschlichen Entscheidern könnte es sich lohnen, besser festzuhalten, wie gut ihre Entscheidungen letztlich waren und ihnen eine deutlich bessere Datengrundlage für ihre Entscheidungen liefern.

„Es hängt im Grunde nicht so sehr davon ab, ob wir von Menschen oder Maschinen beurteilt werden, sondern ob wir wissen, wie die Entscheidungen gefällt werden."

Ist denn die Maschine überhaupt kostengünstiger als der Mensch?

Wir wissen nicht, ob automatisierte Systeme tatsächlich kostengünstigere und effizientere Entscheidungen treffen als die Menschen, die dafür ausgebildet sind. Solche Systeme sind in der Anschaffung und im Training erst mal teuer. Wenn wir durch ihre Entscheidungen gute Bewerber nach Hause schicken, weil sie ungewöhnliche Lebensläufe haben, verursacht das Kosten, die hingegen schwierig zu entdecken sind. So ist es auch, wenn Mitarbeiter demotiviert werden, weil sie zwar tatsächlich weniger Leistung erbracht haben, dies aber an einer persönlichen Notsituation lag, die die Maschine nicht erkennt. Nur ein Mensch kann sehen, dass dieser Mitarbeiter trotzdem hart gearbeitet hat, wenn er denn da war, aber weniger vor Ort sein konnte, weil er beispielsweise seine kranke Mutter pflegen musste. Genau an diesen Stellen werden Maschinen, die nach starren, einsichtslosen Kriterien vorgehen, Fehler machen, die Kosten nach sich ziehen. Daher ist es notwendig, dass wir eine Zeit lang beide Systeme parallel laufen lassen und dann entscheiden, ob das KI-System wirklich besser ist.

Glauben Sie, dass wir Feedback von Maschinen besser aushalten als von Menschen?

Es gibt einen Teil dieser lernenden Algorithmen, die ihre Entscheidungen nicht erklären können. Sie können uns keinen Einblick in das „Warum" gewähren. Das werden wir, denke ich, nicht akzeptieren. Ich kann mir vorstellen, dass ein Leistungssystem mit transparenten Kriterien für uns leichter anzunehmen ist. Es hängt im Grunde nicht so sehr davon ab, ob wir von Menschen oder Maschinen beurteilt werden, sondern ob wir wissen, wie die Entscheidungen gefällt werden – ob wir dem System letztendlich vertrauen. •