Auf Fairness programmiert

In Zukunft werden Computer immer häufiger über Menschen entscheiden – sei es bei der Kreditvergabe oder bei der Bewertung von Bewerbern. Doch automatische Systeme, die dafür bereits eingesetzt werden, diskriminieren immer wieder einmal einzelne Personengruppen. **Niki Kilbertus** und **Bernhard Schölkopf**, Forscher am **Max-Planck-Institut für Intelligente Systeme** in Tübingen, wollen das ändern – mit fairen Algorithmen.

TEXT **TIM SCHRÖDER**

Auf den Fluren des Max-Planck-Instituts für Intelligente Systeme in Tübingen hängen viele Wandtafeln. Im Vorbeigehen notieren die Wissenschaftler darauf ihre Gedanken. Manche diskutieren hier auch neue Ideen miteinander, wenn sie sich auf den Gängen begegnen. „Das hilft sehr“, sagt Niki Kilbertus, „weil ich aktuell sehr viele Ideen entwickeln muss“; Konzepte, die weit über das hinausgehen, womit er sich bislang beschäftigt hat. Niki Kilbertus hat Mathematik und Physik studiert. Er weiß, wie man komplexe Sachverhalte formal korrekt löst und wie man Algorithmen programmiert. Doch mit seiner Promotion hat er den Pfad der formalen Sprache ein Stück weit verlassen. Denn er befasst sich mit einer Frage, die bereits seit einiger Zeit in der Öffentlichkeit heiß diskutiert wird: ob oder inwieweit Algorithmen fair sein können.

Ausgewogene Kalkulation: Algorithmen dürfen keine Personengruppen wie etwa Frauen oder Männer diskriminieren.

Foto: Alamy

Computer sind kühle Rechner, unbestechlich und irren sich nicht – könnte man meinen. Und doch kochte Ende 2018 eine Debatte darüber hoch, dass Computeralgorithmen Menschen diskriminieren. Es war bekannt geworden, dass ein großer Internethändler einen Computer bei Bewerbungen eine Vorauswahl treffen lassen wollte – und dass sich bereits in der Testphase gezeigt hatte, dass der Rechner Bewerbungen von Frauen öfter ablehnte als die von Männern. In den Medien gab es einen Aufschrei. Nicht zuletzt, weil Experten für die Zukunft erwarten, dass Computer mithilfe der riesigen Datenmengen, die heute verfügbar sind, immer häufiger über Menschen entscheiden werden. Es wäre skandalös, wenn sie dabei bestimmte Gruppen bevorzugten oder diskriminierten.

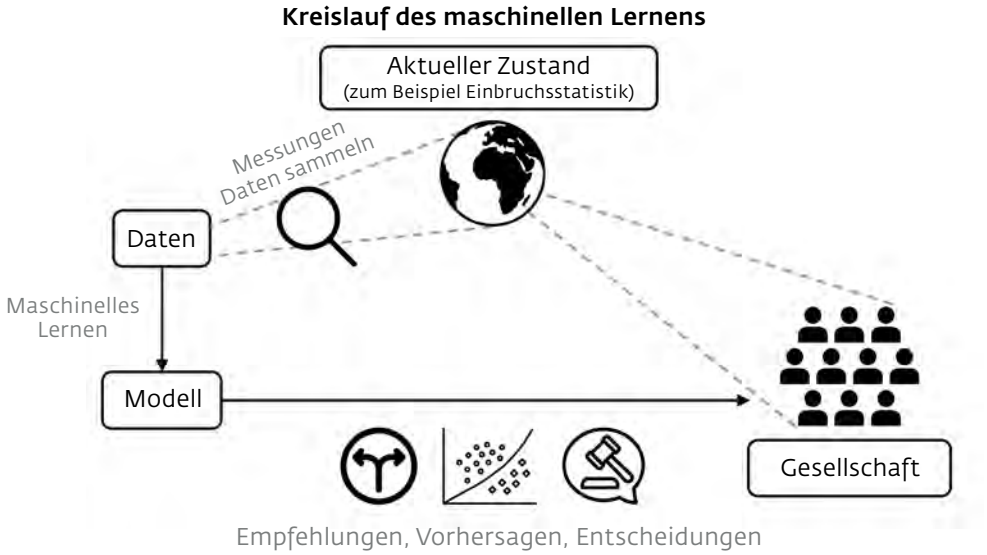
Angesichts solcher Szenarien haben sich weltweit Initiativen gegründet, die sich für Fairness in der künstlichen Intelligenz starkmachen – und nicht nur dafür. Zugleich fordern sie, dass die Unternehmen geradestehen für das, was ihre Algorithmen tun. Sie fordern Verantwortlichkeit, Accountability. Außerdem erwarten die Kritiker Transparenz, wie und warum die Rechenvorschriften eine bestimmte Entscheidung treffen. Die Rede ist von FAT: Fairness, Accountability, Transparency.

JEDE DISKRIMINIERUNG MUSS GENAU ANALYSIERT WERDEN

Niki Kilbertus arbeitet als Doktorand im Cambridge-Tübingen-Programm der Max-Planck-Gesellschaft und damit sowohl am Tübinger Max-Planck-Institut als auch am Pembroke College in Cambridge. Mit seinen Kollegen erforscht er den Part von FAT, der sich technisch umsetzen lässt: Die Wissenschaftler möchten Algorithmen Fairness beibringen. Dabei setzen sie auf maschinelles Lernen, sie befähigen den Computer also dazu, durch Lernen und Erfahrungen langsam besser zu werden.

Die Arbeit ist anspruchsvoll. Denn eine technische Lösung, mit der sich Fairness vom Fleck weg in allen Anwendungen realisieren lässt, gibt es bislang nicht. Und das wird sich vermutlich auch nicht ändern: „Eine Lösung für alle Fragestellungen kann es nicht geben. Jedes Problem, bei dem eine Diskriminierung in den Daten auftaucht, ist anders und muss erst einmal analysiert werden“, sagt Kilbertus. „Und dann kommt der zweite Schritt, der mindestens genauso aufwendig ist: Für das Problem aus der realen Welt müssen wir eine mathematische Beschreibung finden.“

Allerdings folgt der algorithmische Analyseprozess stets einem klaren Schema: Man sammelt Daten, füttert diese in ein Analyseprogramm ein, dessen Algorithmen dann eine Empfehlung wie



zum Beispiel „Bewerber ungeeignet“ ausspucken. Damit dieser automatisierte Prozess künftig stets fair abläuft, müsse man ihn auf der Datenseite und auf der Ausgabeseite verbessern. „Man muss zunächst untersuchen, welche Daten wie erhoben werden, und dann überprüfen, welche Entscheidungen der Computer am Ende ausgibt und warum“, erläutert Niki Kilbertus. Die Analysen berühren dabei stets auch den Datenschutz. „Immerhin dreht es sich in allen Fällen um sensible persönliche Daten. Wir müssen also auch Lösungen finden, mit denen man Daten analysieren kann, ohne dass die Daten offenliegen.“

In Tübingen arbeitet der junge Forscher in der Gruppe von Bernhard Schölkopf, Direktor am Max-Planck-Institut für Intelligente Systeme. Beide beschäftigen sich mit dem grundlegenden Thema der Kausalität im maschinellen Lernen – mit der Frage, inwieweit Computer sinnvolle Schlüsse zwischen verschiedenen Aspekten ziehen können. Und genau diese Frage ist auch beim Thema Fairness zentral.

Niki Kilbertus hat ein Beispiel parat: die Kreditvergabe und die Überprüfung, ob jemand kreditwürdig ist. Setzt man einfache Algorithmen ein, die simple kausale „Wenn-dann-Beziehungen“ durchführen, kann es kritisch werden. Wenn Bewohner eines bestimmten Stadtteils in der Vergangenheit ihre Kredite weniger oft begleichen konnten, könnte der Algorithmus beispielsweise den Wohnort als Indikator für künftige Bewerber verwenden. Er versteht dabei nicht, dass der Wohnort vermutlich keinen direkten kausalen Einfluss auf die Kreditwürdigkeit hat und dass es wohl eher andere relevante Faktoren gibt.

DER ALGORITHMUS MÜSSTE SEINE ANNAHMEN ÜBERPRÜFEN

Oft ist es so, dass das System mit historischen Daten über Kreditrückzahlungen trainiert wird und dann aus den persönlichen Daten wie etwa dem Wohnort neuer Bewerber deren Kreditwürdigkeit berechnet: Wie wahrscheinlich ist es, dass der Bewerber seinen Kredit zurückzahlt? Da aber zum Beispiel auch an einem aus Sicht des Algorithmus schlecht beleumundeten Ort eine

Oben Mit der Welt rückgekoppelt: Ein Algorithmus des maschinellen Lernens löst eine Frage, etwa in welchen Stadtteilen mit vielen Einbrüchen zu rechnen ist. Dafür wird er mit Daten zum aktuellen Zustand gefüttert, zum Beispiel mit Daten zur wirtschaftlichen und sozialen Lage sowie zur Kriminalität. Daraus entwickelt der Algorithmus ein Modell, anhand dessen er Vorhersagen, Entscheidungen oder Empfehlungen ausgibt. Wenn höhere Polizeipräsenz in einem Viertel dann Verbrechen verhindert, verändert das die Gesellschaft und so den Zustand der Welt. Daran muss der Algorithmus das Modell mit neuen Daten anpassen.

Rechte Seite Freiraum für Ideen: Mateo Rojas-Carulla, Niki Kilbertus und Nadine Rüegg (von links) tauschen sich über unterschiedliche Aspekte der künstlichen Intelligenz aus.

Grafik: MPI für intelligente Systeme

Foto: P. Junker/MPI für intelligente Systeme

kreditwürdige Person wohnen kann, kann der Computer manche Personen unbeabsichtigt diskriminieren.

„Im Grunde müsste der Algorithmus so ausgestattet sein, dass er seine Annahmen regelmäßig überprüft", sagt Niki Kilbertus. Etwa, indem er nach bestimmten Kriterien doch hin und wieder einer Person einen Kredit gibt, die zunächst als nicht kreditwürdig eingestuft ist. Wirtschaftswissenschaftler nennen dieses Vorgehen auch *explore versus exploit*. *Explore* bedeutet, neue Lösungen zu testen, zu explorieren. Mit *exploit* ist hingegen gemeint, einen bestehenden Ansatz, so gut es geht, auszunutzen, um zum Beispiel den Aufwand für eine neue Entwicklung zu vermeiden. Im Fall der Kreditvergabe würde das System beim *Explore*-Ansatz gelegentlich Kredite entgegen den ursprünglichen Regeln vergeben. Zahlt die Person den Kredit wider Erwarten doch zurück, muss der Algorithmus angepasst und verbessert werden.

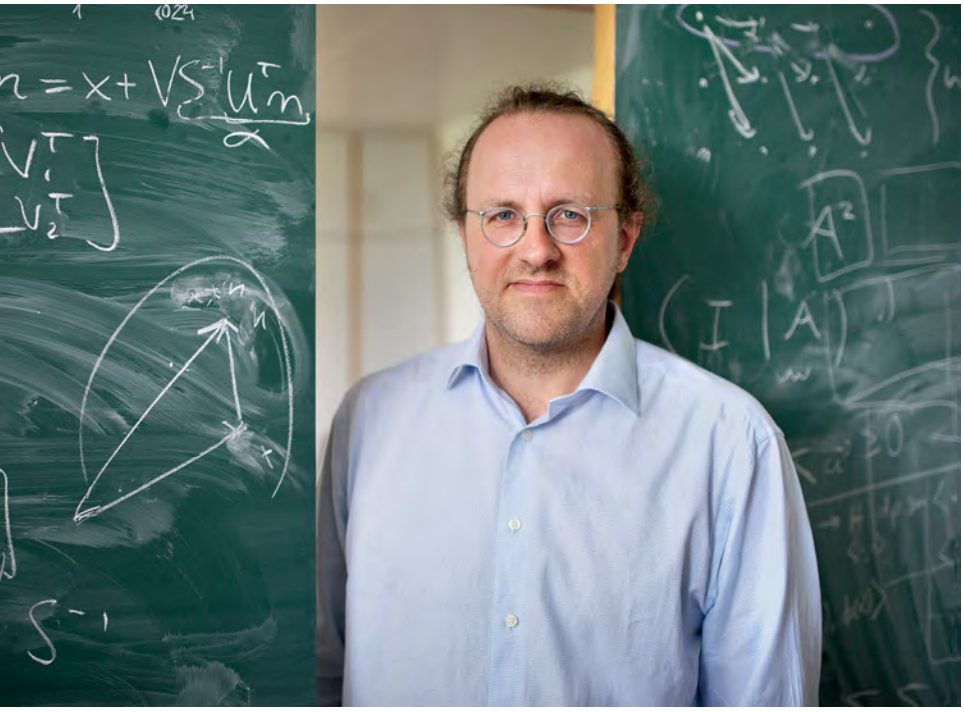
Das Beispiel zeigt, dass die bislang genutzten Daten in bestimmten Fällen für eine faire Entscheidungsfindung nicht ausreichen. Niki Kilbertus: „In Sachen Fairness besteht die Herausforderung darin, Daten so gut wie möglich in einen realistischen, kausalen Zusammenhang zu stellen." So lasse sich vermeiden, dass ein Algorithmus die Daten nur als eine große Ansammlung von Zahlen betrachtet, zwischen denen er dann munter wilde Korrelationen herstellt, die mit kausalen Zusammenhängen aber oft nichts zu tun haben dürften.

ES KOMMT AUF DIE WAHREN KAUSALEN ZUSAMMENHÄNGE AN

Niki Kilbertus veranschaulicht das mit einem abwegigen Beispiel. So könnte ein Algorithmus erkennen, dass Menschen, die Kredite mit höherer Wahrscheinlichkeit zurückzahlen, eher ordentlich sind und zu Hause Filzgleiter unter ihre Stühle kleben. Da eine Rechenvorschrift per se keine Zusammenhänge versteht, könnte sie daraus schließen, dass es sinnvoll ist, neuen Kreditnehmern eine Packung Filzgleiter zu schenken, um ihre Kreditwürdigkeit zu erhöhen. „Nur wenn wir die wahren kausalen Zusammenhänge kennen, können wir sinnvolle Fragen stellen wie: Hätte die Person zurückgezahlt, wenn sie anderswo leben würde?", sagt Niki Kilbertus. Falsche Kausalitäten etwa zwischen dem Besitz von Filzgleitern und der Zahlungstreue ließen sich hingegen ausschließen.

Wie wichtig es ist, Kausalitäten richtig zu erkennen, zeigt auch das Beispiel des Simpson-Paradoxons. Dieses wurde nach dem britischen Statistiker Edward H. Simpson benannt, der in den 1950er-Jahren gezeigt hat, dass sich durch eine bestimmte Kombination von Daten aus verschiedenen Gruppen, etwa Frauen und Männern, scheinbar paradoxe Situationen ergeben. Das klassische Beispiel für ein Simpson-Paradoxon stammt aus dem Jahr 1973. Damals gab es große Aufregung, weil an der US-amerikanischen University of California in Berkeley von den Frauen, die sich beworben hatten, ein geringerer Anteil zum Studium zugelassen worden war als von den männlichen Bewerbern.

Doch lag hier kein Fall von Diskriminierung vor. Eine Analyse der Daten ergab später, dass sich Frauen eher für die stark überlaufenen geistes- und gesellschaftswissenschaftlichen Fächer ange-



Lehre für Maschinen: Bernhard Schölkopf bringt Algorithmen bei, in Daten die wahren kausalen Zusammenhänge zu erkennen. Das trägt auch dazu bei, dass Computer faire Entscheidungen treffen.

meldet hatten – und deshalb insgesamt häufiger abgelehnt wurden als Männer, die sich für die weniger nachgefragten Fächer wie Chemie oder Ingenieurswissenschaften beworben hatten. Das Paradoxon bestand darin, dass der prozentuale Anteil der zugelassenen Frauen innerhalb der meisten Studiengänge sogar höher war als der der angenommenen männlichen Bewerber; über alle Studiengänge gemittelt, überwog aber etwas der prozentuale Anteil der männlichen Bewerber, die angenommen worden waren. „Erst wenn wir die richtigen Daten haben und wissen, für welches Department sich die Frauen beworben haben, können wir den kausalen Zusammenhang richtig verstehen: Das Geschlecht beeinflusst die Fächerwahl. Dann können wir diese Situation richtig interpretieren“, sagt Kilbertus.

Bernhard Schölkopf geht bei der Frage nach den richtigen Kausalitäten sogar noch weiter. Für ihn stellt sich auch die Frage, wo Fairness beginnt. Etwa bei der Frage, ob in den USA Afroamerikaner bei einer Bewerbung die gleichen Chancen haben wie ihre weißen Mitbewerber? „Oder muss man schon früher anfangen, etwa bei dem Gedanken, dass schwarze Kinder nicht dieselben Bildungschancen haben wie weiße – was ihre ganze Vita und später ihre Chancen, eine Arbeit zu finden, beeinflusst?“ Schölkopf fragt, ob man nicht auch solche Aspekte einfließen lassen müsste, um eine wirklich faire Rechenvorschrift zu kreieren – einen Algorithmus mit mehreren Ebenen der Fairness sozusagen.

AUCH MENSCHEN HANDELN NICHT IMMER FAIR

Allerdings mahnt er auch, die Kirche im Dorf zu lassen. „Menschen entscheiden täglich über andere Menschen. Warum sich ein Mensch so oder so entscheidet, ist oftmals völlig intransparent. Von einem Algorithmus aber verlangen wir, dass dieser immer hundertprozentig richtig und fair entscheiden muss.“ Letztlich führe das Studium der Fairness bei Maschinen auch zu der Einsicht, dass Menschen ebenfalls nicht immer fair handeln – einfach deshalb, weil sie beispielsweise aufgrund mangelnder Information oder Erfahrung Fehler machen. Genau wie Maschinen. Ein Hautarzt beispielsweise, der in seiner Karriere nur hellhäutige Menschen auf Hautkrebs untersucht hat, könnte bei dunkelhäutigen eher falsch diagnostizieren und möglicherweise einen Tumor übersehen. Dunkelhäutige Menschen würden bei ihm also nicht mit derselben Qualität behandelt werden wie hellhäutige. Die vermeintliche Unfairness kann aber damit erklärt werden, dass dem Arzt die Erfahrung mit bestimmten Patientengruppen fehlt.

Ob ein Algorithmus in Zukunft fair entscheidet, hängt also auch davon ab, wie Unternehmen oder Menschen allgemein Fairness definieren oder empfinden. „Über allem steht natürlich, dass ein Algorithmus stets richtig entscheiden soll“, sagt Niki Kilbertus. „Doch was richtig ist, das muss in jedem Falle erst einmal geklärt werden.“

Dabei müsse man noch etwas bedenken: Entscheidungen von Computern können einen aktiven Einfluss auf die Welt haben. Ein Beispiel sind moderne Programme, die aus Einbrüchen und Diebstählen in einer Stadt ermitteln, in welchen Vierteln am ehesten mit weiteren Einbrüchen zu rechnen ist. Die Aussage des Algorithmus wird dazu führen, dass dort mehr Polizei präsent ist, und möglicherweise auch dazu, dass dort mehr Straftaten aufgedeckt werden. Doch ist denkbar, dass dort in Wahrheit nicht mehr Straftaten begangen worden sind, sondern dass mehr aufgedeckt wurden, weil die Polizei dort besonders häufig auf Streife war.

„Die Ergebnisse eines Algorithmus können also falsche Schlüsse nach sich ziehen – wir sprechen hier von einem Feedback-Loop, bei dem der Algorithmus das reale Leben beeinflusst.“ Ein ähnliches Beispiel ist die Stauwarnung von Navigationsdiensten. Ist es auf einer Straße voll, empfiehlt der Service, auf

Foto: Wolfram Scheible

andere Strecken auszuweichen – auf denen sich dann kurze Zeit später der Verkehr staut. Niki Kilbertus mahnt aus diesem Grund, nicht nur die Fairness im Blick zu behalten, sondern auch die Feedbackeffekte.

Er hat in seiner Promotion bereits einige Diskriminierungsfälle analysiert und erste Versuche gemacht, das jeweilige Problem mathematisch zu beschreiben. „Die Programmierarbeit braucht noch einmal mindestens so viel Zeit wie die Analyse jedes Falls." Doch noch ist für ihn und seine Kollegen die Analyse die Hauptarbeit. „Jeder Algorithmus arbeitet mit bestimmten Kriterien, aus denen eine Aussage abgeleitet wird. Wir versuchen jetzt herauszufinden, wo die Kriterien brechen, wo die Schwachstelle ist."

Eine interessante Frage sei etwa, warum der Algorithmus des bereits erwähnten Internethändlers im vergangenen Jahr scheinbar Bewerberinnen benachteiligt hat. Zwar sei nicht bekannt, wie der Algorithmus im Detail arbeitete, Kilbertus hat aber eine Vermutung. „Man darf davon ausgehen, dass versucht wurde, die Bewerbungen anonymisiert zu analysieren." Dem System waren also wohl weder Geschlecht noch Name bekannt. Dennoch hat es nicht funktioniert. „Andere Studien zeigen, dass Frauen in Bewerbungen oftmals soziales Engagement oder entsprechende Tätigkeiten angeben, Männer hingegen signalisieren eher Dominanz und wirken kompetitiver", erklärt Kilbertus. „Das können dann genau die Fähigkeiten sein, die das Unternehmen sucht." Bernhard Schölkopf ergänzt: „Es ist falsch anzunehmen, dass wir grundsätzlich Fairness erreichen, indem wir dem Algorithmus bestimmte Angaben wie das Geschlecht vorenthalten." Diese *fairness by unawareness*, Fairness durch Unwissenheit, klappe längst nicht immer.

Wie auch immer sich Algorithmen Fairness beibringen lässt, im Vergleich zu den rein mathematischen Formalismen, mit denen sich Niki Kilbertus früher beschäftigte, handelt es sich eher um ein weiches Kriterium. Auch wenn er die Möglichkeit schätzte, Aussagen klar beweisen oder falsifizieren zu können, empfindet er den neuen Aspekt als bereichernd. „Es ist sehr interessant, an einem solchen gesellschaftlich relevanten Thema zu arbeiten", sagt er. Dass für seine Forschungsarbeit mathematisches und informatisches Wissen nicht ausreichen, wurde ihm ziemlich schnell klar. Deshalb hat er sich auch in Sozialwissenschaften und in juristischen Fragestellungen weitergebildet. Und das ist hoffentlich genug Rüstzeug, um wirklich faire Algorithmen zu finden. ◂

AUF DEN PUNKT GEBRACHT

- **Algorithmen entscheiden immer häufiger über Menschen, dabei kommt es immer wieder zu Diskriminierungen, nicht zuletzt, weil in Daten falsche kausale Zusammenhänge wie etwa zwischen dem Wohnort und der Kreditwürdigkeit einer Person hergestellt werden.**
- **Um Algorithmen Fairness beizubringen, analysieren die Forscher des Tübinger Max-Planck-Instituts für Intelligente Systeme in jedem einzelnen Fall, welche Daten die Rechenvorschrift verwendet und wie sie daraus zu Entscheidungen kommt.**
- **Anhand ihrer Erkenntnisse zu den tatsächlichen kausalen Zusammenhängen, die die Antwort auf eine Frage wie etwa „Welche Person wird einen Kredit zurückzahlen?" liefert, formulieren die Forscher für das jeweilige Problem eine mathematische Beschreibung.**