GERT G. WAGNER

Künstliche Intelligenz verhindert Diskriminierung? Muss nicht – kann aber!

Gert G. Wagner ist Senior Research Fellow am DIW Berlin und Mitglied der Deutschen Akademie der Technikwissenschaften (acatech)
Der Kommentar gibt die Meinung des Autors wieder.

Algorithmen in Form mathematisch-statistischer Verfahren diskriminieren nicht, heißt es oft. Das stimmt aber nicht. Algorithmen sind weder per se neutral, noch ist es unmöglich, dass sogenannte selbstlernende Systeme diskriminieren, ganz im Gegenteil. Gerade bei diesen ist die Gefahr einer systematischen Diskriminierung durch die erlernten Modell groß, denn sie werden auf historischen Daten „trainiert", die sämtliche historischen und aktuellen Diskriminierungen unserer Gesellschaft nachbilden. Gleichwohl muss man Algorithmen nicht verteufeln, denn sie bieten auch die Möglichkeit, Diskriminierung offenzulegen und zu verhindern.

Früher kam man sexistischen oder anders diskriminierenden Entscheidern, die über ihre Motive schwiegen, nur schwer auf die Schliche. Wenn zum Beispiel ein Privatmann eine Wohnung vermietet, wird man ihm niemals Diskriminierung nachweisen können, wenn er nicht ausspricht, dass er etwa Menschen mit dunkler Hautfarbe oder Familien mit kleinen Kindern als Mieter nicht will. Würde dieser Vermieter jedoch einen Algorithmus einsetzen, etwa einen Fragebogen, der ihm hilft, BewerberInnen vorzusortieren – was Wohnungsbaugesellschaften übrigens machen –, hat dies einen entscheidenden Vorteil: Man kann sich diesen Algorithmus anschauen und testen.

Es wird zwar gerne behauptet, selbstlernende Algorithmen wären aufgrund ihrer Komplexität undurchschaubar, das ist aber nicht einmal die halbe Wahrheit. Selbst für neuronale Netze (das sind Computeralgorithmen, die als völlig undurchschaubar gelten) ist es möglich, herauszubekommen, welche „Nervenzellen" (Knoten) maßgeblich an einer automatischen Entscheidung beteiligt waren und was sie repräsentieren. Zudem – und das ist technisch immer möglich – kann man schlicht das Verhalten des Algorithmus prüfen, indem man den Entscheidungsprozess, der im Computer abläuft, gar nicht weiter betrachtet und lediglich Ergebnisse für eine Vielzahl von Beispieldaten bzw. Personen hinsichtlich zuvor definierter Kriterien von Fairness und Diskriminierung prüft. Das ist so, als würde die Stiftung Warentest eine Kaffeemaschine testen.

Die Stiftung Warentest braucht nicht den Bauplan zu kennen, um festzustellen, ob die Mühle gut mahlt und wie lange sie durchhält. Testen macht die interne Komplexität eines Algorithmus für die Qualitätskontrolle völlig irrelevant. Mit einem geeigneten „Testset" von Daten lässt sich also feststellen, ob ein Algorithmus diskriminiert.

Wobei freilich vorab sauber definiert werden muss, was wir unter Fairness überhaupt verstehen. Die Festlegung von Fairness-Kriterien ist weder trivial noch eindeutig. Beispiel Kreditvergabe: Ist die Kreditvergabe nur dann fair, wenn dieselben Anteile von Frauen und Männer, die einen Kredit wollen, auch bekommen? Dann werden Männer bevorzugt, wenn sie öfters als Frauen einen Kredit nicht zurückzahlen. Oder ist es fair, wenn *kreditwürdige* Frauen und Männer in gleichem Maße Kredite bekommen? Dann ist die Erfolgsquote von Frauen höher als die von Männern. Hört sich plausibel an. Aber man kann unter Fairness auch verstehen, dass tatsächlich kreditwürdige Frauen und Männer im gleichem Maße fälschlich als nicht-kreditwürdig eingestuft werden. Diesen Fehler der „zweiten Art" gibt es nämlich auch, und in diesem Beispiel würde er die Frauen stärker treffen. Dass alle drei Fairnesskriterien gleichzeitig erfüllt werden, ist theoretisch zwar möglich, aber in der Realität praktisch nie gegeben.

Nicht nur die KI-Forschung muss sich endlich damit befassen, was Fairness eigentlich in welcher Situation bedeutet. Und wir alle sollen uns anhand von Beispielfällen und einer systematischen Definition und Analyse von Fairness im Klaren darüber werden, was Algorithmen mit uns machen. Technisch ist das machbar – der Gesetzgeber muss es nur wollen und sicherstellen, dass es auch geschieht. Dann wird man bei Algorithmen gegebenfalls eingreifen können und sie „lernen" lassen, nicht zu diskriminieren.

Der Beitrag, der in längerer Form in der Süddeutschen Zeitung am 6. Mai 2019 erschienen ist, entstand in Zusammenarbeit mit der Informatikerin Meike Zehlike, Humboldt Universität Berlin und Max Planck Institut für Softwaresysteme in Saarbrücken.

DOI: https://doi.org/10.18723/diw_wb:2019-20-3

IMPRESSUM

DIW Berlin — Deutsches Institut für Wirtschaftsforschung e. V.
Mohrenstraße 58, 10117 Berlin
www.diw.de
Telefon: +49 30 897 89–0 Fax: –200
86. Jahrgang 15. Mai 2019

Vertrieb
DIW Berlin Leserservice, Postfach 74, 77649 Offenburg
leserservice@diw.de
Telefon: +49 1806 14 00 50 25 (20 Cent pro Anruf)

Gestaltung
Roman Wilhelm, DIW Berlin

Umschlagmotiv
© imageBROKER / Steffen Diemer

Satz
Satz-Rechen-Zentrum Hartmann + Heenemann GmbH & Co. KG, Berlin

Druck
USE gGmbH, Berlin

ISSN 0012-1304; ISSN 1860-8787 (online)

Abonnieren Sie auch unseren DIW- und/oder Wochenbericht-Newsletter unter **www.diw.de/newsletter**