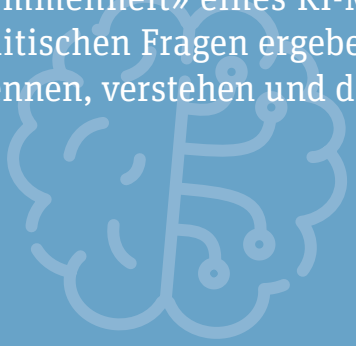
Ethik und Gesellschaft
Bias in KI: Gefahren erkennen, verstehen und beheben

Mit der Verbreitung von Anwendungen der künstlichen Intelligenz (KI) treten auch deren Schwächen ins Zentrum der öffentlichen Diskussion. Eine davon wird als «Bias» bezeichnet, das die «Voreingenommenheit» eines KI-Modells beschreibt. Neben ethischen und gesellschaftspolitischen Fragen ergeben sich daraus konkrete Gefahren, weshalb es zu erkennen, verstehen und die daraus resultierenden negativen Folgen zu beheben gilt.



Wenngleich bei weitem nicht so dynamisch wie in der freien Wirtschaft, ist der Einzug von KI weltweit auch im öffentlichen Sektor in vollem Gange. Beispiele dafür sind ein Pilotprojekt zur KI-gestützten Bearbeitung von Vorstössen im Kanton Zürich oder die Anwendung von Chatbots in deutschen Behörden. Die Algorithmen sollen helfen, wiederkehrende Tätigkeiten anhand klar definierter objektiver Kriterien automatisiert zu erledigen. Die Hoffnung ist, dass sich auf diese Weise potenzielle menschliche Schwächen wie Ermüdung oder diskriminierendes Verhalten ausschliessen lassen. Umso verwunderlicher scheinen deshalb die Berichte, laut derer KI-Anwendungen beispielsweise bei marginalisierten Personengruppen irreführende Ergebnisse produzieren oder sogar rassistische und sexistische Stereotypen aufweisen. Dieses Vorhandensein einer vermeintlichen Voreingenommenheit in Anwendungen künstlicher Intelligenz wird als Bias bezeichnet.

Wie entsteht Bias?

Bevor KI zum Einsatz kommen kann, werden Algorithmen auf grossen Datenmengen trainiert, die vorab von Menschen selektiert und ggf. annotiert wurden. Genau hier liegt zumeist der Ursprung von Bias. KI verfügt weder über Intelligenz noch über ein Verständnis ethischer Grundprinzipien. Entsprechend des Prinzips «Garbage In – Garbage Out» kann u. a. eine einseitige oder fehlerhafte Datenauswahl oder Annotation dazu führen, dass ein Algorithmus diese Einseitigkeit oder Verzerrung reproduziert. Wurde ein Modell beispielsweise mehrheitlich auf Daten von weissen männlichen Personen trainiert, führt dies oft dazu, dass das Modell unzuverlässige Ergebnisse bei Frauen oder Menschen anderer Ethnien liefert und dadurch für diese Personengruppen erhebliche Nachteile entstehen können. Der Bund hat diese Gefahren unlängst erkannt und definiert in seinen Leitlinien «Künstliche Intelligenz», dass bei der Entwicklung und dem Einsatz von KI die Menschenwürde und das Gemeinwohl an erster Stelle stehen sollen. Zudem müssten «Personen, Gruppen und Geschlechter» vor Diskriminierung und Stigmatisierung geschützt werden.

Umsichtiger Umgang mit Bias

Eine unerwünschte Benachteiligung oder Stigmatisierung kann daher schwerwiegende Folgen für die Betroffenen nach sich ziehen, wie Untersuchungen zum Einsatz von KI im öffentlichen und privaten Sektor bereits bezeugen. Idealerweise wird durch eine gezielte Diversifikation der Datengrundlage dem Entstehen von Bias vorgebeugt oder, sofern die entsprechenden Daten nicht vorliegen, auf die Grenzen der KI-Anwendung hingewiesen bzw. deren Einsatz auf bestimmte Anwendungsfälle begrenzt. Vor allem aber ist ein klares Verständnis über Bias unabdingbar, um eine korrekte Risikobeurteilung zu erreichen. Weil sich ein minimales Mass an Bias kaum vermeiden lässt, gilt es in erster Linie zu klären, inwiefern Bias Risiken mit sich bringt, die ein ressourcenintensives Eingreifen rechtfertigen würden. Um diese Frage zu beantworten, muss man sich auch die kulturellen und gesellschaftspolitischen Hintergründe bewusst machen, vor denen Vorurteile entstehen und wahrgenommen werden. So kann beispielsweise eine automatische Übersetzung des kontextlosen englischen Wortes «doctor» mit der männlichen Form «Arzt» (und nicht etwa «Ärztin») durchaus als Bias eingestuft werden. Allerdings resultiert hieraus kein direkter Schaden für die Nutzenden; beanstandet wird die Aufrechterhaltung gesellschaftlicher Stereotypen, die in einer egalitären Gesellschaft als problematisch angesehen werden kann. Es empfiehlt sich daher, von Fall zu Fall die Risiken von Bias einzuschätzen und in Zusammenarbeit mit Fachleuten mögliche sinnvolle Massnahmen zu eruieren.

Unsere Empfehlungen

1. Bias erkennen

Verwaltungen sollten ein Bewusstsein für die Existenz von Bias in KI-Anwendungen entwickeln und ihre Anwendungen stets auf den Prüfstand stellen. Vor allem in einer immer diverser werden Gesellschaft ist hierbei die Einbeziehung unterschiedlicher Personengruppen wichtig, um idealerweise der Entstehung von Bias vorzubeugen.

2. Bias verstehen

Sofern Bias erkannt wurde, bedarf es eines Risk-Assessments. Liegen erhebliche Gefahren für die Anbietenden oder Nutzenden vor, empfiehlt es sich, die Trainingsdaten auf Verzerrungen zu überprüfen.

3. Bias beheben

Sobald die Fehlerquelle ermittelt wurde, sollte die entsprechende KI-Anwendung mit neuen Daten verbessert werden. Wichtig ist dabei auch das kontinuierliche Feedback der Nutzenden (beispielsweise nach dem Prinzip «Human in the Loop»).

Mehr Informationen

Kontaktmöglichkeiten und weitere Informationen zu Bias in KI:
bfh.ch/ipst/public-sector-ai

Kontakt

Veton Matoshi
Wissenschaftlicher Mitarbeiter

veton.matoshi@bfh.ch
T +41 31 848 57 89

Prof. Dr. Marcel Gygli
Professur KI im öffentlichen Sektor

marcel.gygli@bfh.ch
T +41 31 848 64 90