INTERDISZIPLINÄRES FORSCHUNGSZENTRUM
für Technik, Arbeit und Kultur

IFZ - Electronic Working Papers

Faire KI - (wie) geht das?

Julian Anslinger

01 - 2021

FORSCHUNG VERÄNDERT

IFZ - Electronic Working Papers
ISSN: 2077-3102

Editiert von
IFZ - Interdisziplinäres Forschungszentrum für Technik, Arbeit und Kultur
Schlögelgasse 2
8010 Graz

Tel: +43/316/813909-0
E-Mail: office@ifz.at;
Web: https://www.ifz.at

Förderungen

Das Paper ist entstanden mit Informationen gewonnen im Projekt VEKIAA – Verantwortungsvolle Einbindung von KI-Assistenzsystemen am Arbeitsplatz, gefördert im Rahmen des Digitalisierungsfonds Arbeit 4.0 der Arbeiterkammer Wien und im Rahmen einer Auftragsarbeit der TU Graz.

Das Land Steiermark und die Stadt Graz unterstützen das IFZ mit einer Förderung.

Der Begriff Künstliche Intelligenz

Der Diskurs über faire KI ist geprägt von einer Vielzahl verschiedener Begriffe, Vorstellungen und Ansätze. Dies beginnt bereits bei dem Begriff „künstliche Intelligenz", einem Modewort, das viele unterschiedliche Methoden der Informatik unter sich vereint (Machine Learning, Deep Learning, induktive Programmierung, Bayessche Schätz-, Such- und Optimierungsmethoden usw.). Poole und Kolleg_innen umschreiben KI-Systeme als „Systeme, die ihre Umwelt erfassen und Handlungen setzen, die die Chancen maximieren, gesetzte Ziele zu erreichen." (Poole, Mackworth & Goebel, 1998). Der Begriff der künstlichen Intelligenz hat also wenig mit dem Intelligenzbegriff zu tun, den wir mit Menschen oder Tieren assoziieren. KI-Systeme werden für sehr spezifische Anwendungen geschaffen und funktionieren dann auch nur für diesen jeweiligen Zweck, eine breite „Artificial General Intelligence" oder „strong AI" bleibt weiterhin im Bereich der Utopien. Vor diesem Hintergrund plädieren manche Wissenschaftler*innen dazu, andere Begriffe wie „algorithmische Entscheidungssysteme" (Schröter, 2019) oder die genaue Methodenbezeichnung zu nutzen.

Was ist faire KI?

Auch im Hinblick auf faire KI ist der Diskurs weiterhin im vollen Gange. Oft synonym verwendet werden Begriffe wie ethische KI, vertrauenswürdige KI oder verantwortungsvolle KI. In einer qualitativen Inhaltsanalyse von 84 Ethik-Richtlinien fanden Anna Jobin und Kolleg_innen (2019) folgende häufig diskutierten Aspekte fairer KI: Gerechtigkeit, Transparenz, Schadlosigkeit, Verantwortung und Verantwortlichkeit, Datenschutz, Nutzen, Freiheit und Autonomie, Vertrauen, Nachhaltigkeit, Würde und Solidarität. Diese finden sich auch in ähnlicher Form in den Ethik-Leitlinien für vertrauenswürdige KI, der durch die Europäische Kommission eingesetzten Expert*innengruppe wieder, in der folgende Kernforderungen gestellt werden: Vorrang menschlichen Handelns und menschliche Aufsicht, technische Robustheit und Sicherheit, Datenschutz und Datenqualitätsmanagement, Transparenz, Vielfalt, Nichtdiskriminierung und Fairness, Gesellschaftliches und ökologisches Wohlergehen und Rechenschaftspflicht. Der Verordnungsvorschlag der Europäischen Kommission für KI-Systeme, welcher im April 2021 veröffentlicht wurde, und voraussichtlich frühestens im Jahr 2024 in adaptierter Form in Kraft treten wird, berücksichtigt viele dieser Aspekte in unterschiedlichem Ausmaß. In jedwedem Fall wird erwartet, dass die geplante Verordnung einen ähnlichen Einfluss auf die gesamte Welt haben wird, wie die Datenschutzgrundverordnung. Mindestens bis dahin gilt es, den Diskurs über faire KI voranzutreiben und über die vielen Möglichkeiten zu sprechen, in denen KI einen negativen Einfluss auf Gesellschaft und Umwelt haben kann. Zentrales Thema hierbei ist das Themengebiet der Biases.

Biases

Drei Arten von Biases scheinen eine besonders wesentliche Rolle zu spielen: Datenbasierte Biases, algorithmenbasierte Biases und kulturbasierte Biases.

1. Datenbasierte Biases: Ein zentraler datenbasierter Bias ist der sogenannte Sample-Bias, der entsteht, wenn die einem Algorithmus zugrundeliegenden Daten nicht die

(relevante) Wirklichkeit widerspiegeln. Bekannt geworden sind beispielsweise Bilderkennungsalgorithmen, die Probleme damit hatten, dunkelhäutige Frauen korrekt als solche zu identifizieren, während die Gesichter von Männern oder weißen Menschen häufiger richtig klassifiziert wurden (z. B. Buolamwini & Gebru, 2018; Zou & Schiebinger, 2018). Ein anderer Bilderkennungsalgorithmus wiederrum identifizierte Männer, die in Küchen abgebildet wurden, fälschlicherweise als Frauen (Zhao et al., 2017). In beiden Fällen ist die schlechte Datenlage schuld. Im ersten Beispiel enthielt der zum Trainieren verwendete Datensatz hauptsächlich Fotos von weißen Personen und Männern; im zweiten Beispiel waren im zugrundeliegenden Datensatz kochende Männer weniger repräsentiert als kochende Frauen.

2. Algorithmenbasierte Biases: Algorithmenbasierte Biases können u.a. dann auftreten, wenn Nutzer*innen nicht ausreichend in die Erstellung der Software miteingebunden werden. Beispielsweise ist es relevant darauf zu achten, wie man die Ergebnisse von KI-Algorithmen so darstellt, dass diese nutzbringend und richtig interpretiert werden können. Forschungsarbeiten konnten demonstrieren, dass Visualisierungen leichter und schneller wahrgenommen werden können als Datentabellen. Es hat sich jedoch gezeigt, dass Visualisierungen verhindern können, dass sich Menschen tiefergehend mit dem Betrachtungsgegenstand auseinandersetzen. In einer Studie von der Universität Michigan und Microsoft zeigte sich zum Beispiel, dass selbst Datenwissenschaftler*innen Visualisierungen von KI-Ergebnissen zur Interpretation heranzogen, auch wenn diese für sie nicht vollständig nachvollziehbar waren (Kaur, 2020). Darüber hinaus erhöhten die teilweise nicht nachvollziehbaren Visualisierungen trotzdem das Vertrauen der Expert*innen in den Machine-Learning Algorithmus. Dass Menschen von Computern dargestellte Ergebnisse fälschlicherweise als besonders objektiv und korrekt wahrnehmen, ist ein Phänomen, das als Automation Bias beschrieben wird (Mehrabi, 2021). Zusammenfassend muss auch bei der Programmierung von Algorithmen genauestens berücksichtigt und bestenfalls untersucht werden, wie sich die Algorithmen und deren Ergebnisdarstellung auswirken könnten.

3. Kulturbasierte Biases: Man stelle sich einen KI-Algorithmus vor, dessen Daten so ausgewählt sind, dass sie die Wirklichkeit perfekt repräsentieren und der Algorithmus darüber hinaus ohne jegliche Fehlentscheidungen programmiert wurde. Eine perfekte nicht-gebiaste KI, oder? Nein, nicht zwangsläufig. In einigen Fällen von gebiasten KI-Systemen, kann es passieren, dass der Bias nicht aus Daten oder Algorithmen resultiert, sondern in der Gesellschaft besteht und in Daten oder Algorithmus reproduziert wurde.
   Ein Beispiel hierfür ist ein von Amazon eingesetzter KI-Algorithmus, der von 2014 bis 2015 genutzt wurde, um die Lebensläufe von Bewerber*innen zu bewerten (Dastin, 2018). Zum Trainieren des Algorithmus wurden historische aufgezeichnete Lebenslaufdaten und deren Erfolgsraten früherer Bewerber*innen genutzt. Um diskriminierende Effekte von vorneherein auszuschließen, wurde der Algorithmus so eingestellt, dass Geschlecht als Feature nicht berücksichtigt werden sollte. Das Machine-Learn-

ing-Programm fand jedoch einen indirekten und unvorhergesehenen Weg, männliche und weibliche Bewerber*innen zu unterscheiden. Und zwar stellte es fest, dass erfolgreiche Bewerbungen für Jobs im Technik-Bereich zumeist von Personen eingereicht wurden, die keine mit Weiblichkeit verknüpften Attribute aufwiesen. So führte beispielsweise die Erwähnung des Wortes Frau (wie z. B. in Frauensportverein) automatisch zu einer schlechteren Bewertung der jeweiligen Lebensläufe. Die KI setzte somit ein in der westlichen Gesellschaft verankertes Phänomen, dem Umstand, dass vergleichsweise wenige Frauen in Technikberufen tätig sind, fort. Bewerbende Frauen bekamen somit eventuell gar nicht die Chance, ihre Technikkompetenz unter Beweis zu stellen und wurden von vorneherein ausgeschlossen. In anderen Worten: Sie wurden diskriminiert.

Ein weiteres Beispiel für die Fortschreibung von kulturell bestehenden Ungleichheiten unterschiedlicher Gruppen ist ein vom Arbeitsmarktservice (AMS) programmierter Algorithmus. Der vom AMS als „Arbeitsmarktchancen-Assistenz-System“ (AMAS) bezeichneter Algorithmus soll zur Bewertung von Jobchancen dienen und die Einschätzung der Förderwürdigkeit von Arbeitssuchenden durch Arbeitsvermittler*innen unterstützen. Personen mit, durch den Algorithmus berechneten, hohen und niedrigen Jobchancen sollen hierbei weniger gefördert werden als Personen mit mittleren Jobchancen. Der Algorithmus basiert zwar nicht auf Verfahren, die es zuließen, ihn als künstliche Intelligenz zu bezeichnen, er eignet sich jedoch als anschauliches Beispiel für kulturbasierte Biases von Algorithmen, auch wegen seines potentiell immensen Einflusses. Eine von der AK OÖ beauftragte und durch das Institut für Technikfolgenabschätzung durchgeführte Studie hat das AMAS untersucht (Allhutter, 2020). Die Studie konnte zeigen, dass das AMAS Personen aufgrund von Personenmerkmalen wie Geschlecht, Alter, Migrationshintergrund und Behinderung unterschiedlich in ihren Jobchancen bewertet. Die Bewertung basiert auf historisch aufgezeichneten Daten, die zeigen, dass Personen mit bestimmten Merkmalen oder Merkmalskombinationen eher einen Job gefunden haben als andere. Beispielsweise schätzt das AMAS die Jobchancen von Frauen mit dem Merkmal Kinderbetreuungspflichten schlechter ein als diejenigen von Männern mit Kinderbetreuungspflichten. Dieser Unterschied in den Jobchancen mag sich aufgrund verschiedener Gegebenheiten in der Gesellschaft (z. B. aufgrund von diskriminierenden Einstellungspraktiken; s. bspw. Hipp, 2019) tatsächlich so zeigen, sollte sich aber eigentlich nicht negativ auf die Förderungsempfehlungen von Arbeitssuchenden auswirken. Eine ähnliche Herabstufung im AMAS-Ratingsystem findet sich für Frauen mit Migrationshintergrund. Statt diese Personengruppen besonders zu fördern, um sie besser auf dem Arbeitsmarkt zu integrieren und eine langfristige Gleichstellung zu erreichen, entscheidet sich das AMS für die Diskriminierung dieser Gruppen und eine Fortschreibung von gesellschaftlich bestehenden Ungleichheiten.

Biases verhindern

Wie können Biases und weitere negative Auswirkungen von KI-Systemen in Unternehmen nun verhindert werden? Hier eine kurze und naturgemäß unvollständige Übersicht wichtiger Maßnahmen.

- Die Entwicklung eines KI-Systems sollte nie abgeschlossen sein. Eventuelle Biases und weitere Auswirkungen müssen von Beginn an mitbedacht und im Laufe des Prozesses berücksichtigt werden. Das beginnt mit dem Unternehmensinteresse und der Frage danach, welche KI-Systeme und Daten tatsächlich benötigt werden; geht über den Prozess der Technologieentwicklung (d. h. Entwicklung, Training und Testung des Algorithmus, der Implementierung) bis hin zur Implementierung und einem langfristigen Monitoring.

- Dem Training zugrundeliegende Daten sollten möglichst die relevante Wirklichkeit widerspiegeln. Ändern sich die Umstände oder wird die Relevanz von bisher unberücksichtigten Variablen deutlich, sollte ein Algorithmus neu trainiert werden.

- Besteht die Gefahr des Fortschreibens gesellschaftlich bestehender Ungleichheiten, sollten algorithmische Korrekturmaßnahmen getroffenen werden. Beispielsweise könnten Daten mit unterschiedlichen Gewichtungen versehen werden, um bestehende Biases in den Trainingsdaten zu korrigieren.

- KI-Systeme sollten möglichst transparent gestaltet sein um eventuelle Biases bereits während des Betriebes erkennen oder im Nachhinein nachvollziehen zu können. Das Annotieren der Trainingsdaten sollten einem genau festgelegten Standard folgen und wenn möglich, parallel durch mehrere Menschen durchgeführt werden, sodass die Übereinstimmung der annotierenden Personen berechenbar ist.

- Ethische sowie gender- und diversitätsbezogene Fragestellungen sollten bereits während der KI-Programmierung berücksichtigt werden. Gute Tipps für eine Vorgehensweise in verschiedenen wissenschaftlichen Feldern finden sich zum Beispiel auf: https://genderedinnovations.stanford.edu

- „I-Methodology" (d. h. die Einschreibung von fast ausschließlich eigenen Erfahrungen durch Technik-Entwickler*innen) sollte verhindert werden. Hierfür lohnt sich, neben User-centered-Design, eine inter- und transdisziplinäre Arbeitsweise in möglichst gemischten Teams. Diversität, Inter- und Transdisziplinarität erhöht die Robustheit von Technologien und wissenschaftlichen Erkenntnissen.

- Die Bedürfnisse und Charakteristiken aller potentieller Betroffenen sollten stets im Auge behalten werden. Hierfür lohnt es sich, Betroffene gleich von Anfang an einen gemeinsamen Tisch zu holen und zum Beispiel in Workshops Interessen und Bedürfnisse abzuwägen. Dieser Prozess ist wunderbar im WeBuildAi-Framework von Min Kyung Lee (2019) beschrieben.

- In Betrieben sollten Arbeitnehmer*innenvertretungen stets im direkten Dialog mit der Unternehmens-IT stehen, um den Bedürfnissen der Arbeitnehmenden deutliches Gehör zu verschaffen. Ein mögliches Vorgehen haben wir in der Publikation How to make it fAIr – Methoden partizipativer Technikgestaltung für das Anwendungsfeld der Künstlichen Intelligenz des Projekts dAIalog.at beschrieben. Darüber hinaus wird im Projekt VEKIAA (Verantwortungsvolle Einbindung von KI-Assistenzsystemen am Arbeitsplatz) ein Handbuch veröffentlicht (Sommer 2022), welches noch einmal tiefergehend auf diesen Prozess eingehen wird.

Literatur

Doris Allhutter et al. (2020). Der AMS-Algorithmus. Eine Soziotechnische Analyse des Arbeitsmarktchancen-Assistenz-Systems (AMAS). url: http://e-pub.oeaw.ac.at/0xc1aa5576_0x003bfdf3.pdf.

Joy Buolamwini & Timnit Gebru (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In: Conference on fairness, accountability and transparency. PMLR. S. 77-91.

Jeffrey Dastin (2018). Amazon scraps secret AI recruiting tool that showed bias against women. url: https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/a...;

Lena Hipp (2019). Do Hiring Practices Penalize Women and Benefit Men for Having Children? Experimental Evidence from Germany. In: European Sociological Review 36.2 (Nov. 2019), S. 25-{264. issn: 0266-7215. doi: 10.1093/esr/jcz056. eprint: https://academic.oup.com/esr/article-pdf/36/2/250/33018077/jcz056.pdf. url: https://doi.org/10.1093/esr/jcz056.

Anna Jobin, Marcello Ienca, & Effy Vayena (2019). The global landscape of AI ethics guidelines. In: Nature Machine Intelligence 1, Nr. 9: 389–99, https://doi.org/10.1038/s42256-019-0088-2.

Harmanpreet Kaur et al. (2020). Interpreting Interpretability: Understanding Data Scientists' Use of Interpretability Tools for Machine Learning. In: CHI Conference on Human Factors in Computing Systems Proceedings, 1–14, https://doi.org/10.1145/3313831.3376219.

Min Kyung Lee et al. (2019). ACM Hum.-Comput. Interact. 3, CSCW, Article 181 (November 2019), 35 pages. DOI:https://doi.org/10.1145/3359283

Ninareh Mehrabi et al. (2021). A survey on bias and fairness in machine learning. In: ACM Computing Surveys (CSUR) 54.6, S. 1-35.

Ulf Mellström (2009). The intersection of gender, race and cultural boundaries, or why is computer science in Malaysia dominated by women? In: Social studies of science 39.6, S. 885-907.

David Poole & Alan Mackworth & Randy Goebel. (1998). Computational Intelligence: A Logical Approach.

Welf Schröter (2019). Der mitbestimmte Algorithmus. Hg.: Welf, Schröter. Talheimer Verlag.

Anita Thaler et al. (2021). How to make it fAIr – Methoden partizipativer Technikgestaltung für das Anwendungsfeld der Künstlichen Intelligenz. https://daialog.at/wp-content/uploads/2021/09/dAIalog.at_-Methoden-part...;

Jieyu Zhao et al. (2017). Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. doi: 10.18653/v1/D17-1323.

James Zou & Londa Schiebinger (2018). AI can be sexist and racist | it's time to make it fair. In: Nature 559.7714, S. 324-326.