PRAXIS

Dem algorithmischen Bias auf der Spur

Doris Allhutter, *Institut für Technikfolgen-Abschätzung, Österreichische Akademie der Wissenschaften, Apostelgasse 23, 1030 Wien (dallhutt@oeaw.ac.at)*

Wer ist verantwortlich, wenn (halb-)automatisierte Systeme diskriminieren? Wie werden gesellschaftliche Stereotype und Diskriminierung durch *Machine Learning* und Künstliche Intelligenz (re-)produziert? Wo gibt es Interventionspunkte und welche Arten von Forschung und Regulierung sind erforderlich, um sicherzustellen, dass diese Interventionen effektiv sind? Zahlreiche Fälle von Diskriminierung durch Algorithmen aufgrund von Geschlecht, Ethnizität, Klasse und sexueller Orientierung haben in den letzten Jahren einen dringenden Handlungsbedarf deutlich gemacht. Unternehmen und Regierungsinstitutionen ziehen unter dem Motto von Effizienz und Kosteneinsparung zunehmend (halb-)automatisierte Entscheidungssysteme heran – oft sind sie ungetestet oder schlecht entwickelt, jedenfalls hinsichtlich der in sie eingeschriebenen strukturellen Ungleichheiten. Der Einsatz solcher Systeme in Bereichen wie Bildung, Beschäftigung, Gesundheit, Bankwesen, Polizeiarbeit und Terrorismusbekämpfung objektiviert damit diskriminierende Entscheidungen und läuft Gefahr, gesellschaftliche Ungleichheiten infrastrukturell zu verankern und langfristig zu verstärken.

In den letzten zehn Jahren hat sich rund um diese Problematik ein interdisziplinäres Forschungsfeld entwickelt, das sich erst unter der Bezeichnung *discrimination-aware data-mining* und aktuell unter *fairness, accountability and transparency in machine learning* (FAT) mit Bias in soziotechnischen Systemen auseinander setzt. Eine besondere Herausforderung dieser Forschung besteht in der Entwicklung geeigneter Methoden, um Bias aus Daten, Klassifizierungen und Modellen zu entfernen: Methoden des *Debiasing* sollen sowohl technischen als auch juristischen und emanzipatorischen Anforderungen genügen. Dieses wichtige Unterfangen basiert auf der Erkenntnis, dass informatische Systeme nicht objektiv sind. Gleichzeitig müssen aber auch hier eine Reihe an normativen Entscheidungen getroffen werden. *Debiasing* formalisiert im Grunde mathematisch, was als fair und frei von Diskriminierung verstanden wird. Wir können etwa fragen, wie soziale Kategorien wie Geschlecht, Ethnizität, Alter oder Klasse – z. B. bei der Vergabe von Krediten oder von staatlichen Fördermaßnahmen bei Arbeitslosigkeit – algorithmisch zueinander in Beziehung gesetzt werden und auf Basis welcher Annahmen dann getestet wird, ob ein System diskriminiert und für wen es fair ist.

Um diesen impliziten Normsetzungen auf die Spur zu kommen, beschäftigt sich ein am ITA durchgeführtes Projekt mit Praktiken des *Debiasing*. Konkret wird danach gefragt, welche impliziten Annahmen und Werte Forscher_innen der FAT-Community in ihren epistemischen Praktiken mobilisieren. Im Oktober 2018 fand dazu ein erster internationaler Workshop mit dem Titel „Debiasing and Discrimination-Awareness in Machine Learning: Exploring Implicit Assumptions and Context" statt. Im Rahmen des von Bettina Berendt (KU Leuven) und Doris Allhutter (ITA) durchgeführten Workshops nahm eine Gruppe von zehn Forscher_innen an einer kollektiven Dekonstruktion teil: Ziel dieses methodischen Herangehens ist es, konkrete Debiasing-Praktiken zu reflektieren und damit informatische Kernkonzepte und ihre impliziten Argumentationslogiken zu beleuchten. Problematisiert wurde beispielsweise die Fähigkeit eines Entscheidungsunterstützungssystems *entweder* möglichst korrekte *oder* faire Entscheidungen vorzuschlagen (der sogenannte *accuracy-fairness trade-off*). Entsprechende Reflexionen ermöglichen es der FAT-Community schrittweise zu erfassen, auf welch vielschichtige Weise ihre disziplinär geprägten Arbeitspraktiken mit gesellschaftlichen Machtverhältnissen verwoben sind. Dies kann nicht nur dazu beitragen, disziplinäre Konzepte und Methoden zu hinterfragen, sondern epistemische Praktiken zu transformieren.

Zum Weiterlesen

Allhutter, Doris (2019): Practices of debiasing in machine learning. An ethnographic study. Position Paper für den Workshop „Human-Centered Study of Data Science Work Practices" auf der CHI-Konferenz vom 4. bis 5. Mai 2019 in Glasgow.

Allhutter, Doris (2019): Of 'working ontologists' and 'high-quality human components'. The politics of semantic infrastructures. In: Janet Vertesi und David Ribes (Hg.): DigitalSTS. A field guide for science & technology studies. Princeton: Princeton University Press, S. 326–348.

Allhutter. Doris (in Vorbereitung): Automatisierte Diskriminierung. ITA-Dossier Nr. 43: https://www.oeaw.ac.at/ita/publikationen/ita-dossiers.

In dieser kostenpflichtigen Rubrik informieren NTA-Mitglieder über ihre Aktivitäten und unterstützen TATuP.
www.tatup.de/index.php/tatup/about/submissions