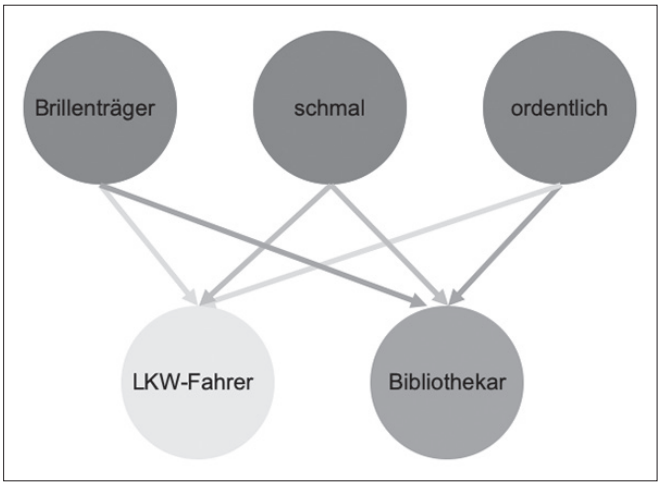
Abigail Morrison

Voreingenommenheit im Gehirn und im Algorithmus

Was ist überhaupt eine Voreingenommenheit? In der Alltagssprache bedeutet voreingenommen zu sein, dass man eine vorgefertigte Meinung hat, eine starke Präferenz für oder gegen etwas; sei es eine Idee, eine Ideologie, eine Person oder Personengruppe. Häufig führen solche Voreingenommenheiten, die bewusst oder unbewusst sein können, zu diskriminierendem Verhalten. Im kognitionspsychologischen Kontext hat der Begriff Voreingenommenheit eine etwas andere Bedeutung: nämlich unbewusste, systematische Fehler beim Wahrnehmen, Erinnern, Denken und Urteilen. Solche kognitiven Verzerrungen sind universal, unabhängig von Bildungsgrad oder Kultur; sie sind zudem äußerst hartnäckig und beeinflussen uns, auch wenn wir davon wissen und versuchen sie zu kompensieren.



Es gibt eine reiche Vielfalt an kognitiven Voreingenommenheiten.[1] Für den Vergleich zwischen den Fehlern menschlicher Gehirne und lernender Algorithmen ist das wichtigste Phänomen die Repräsentativitäts-Heuristik. Diese Heuristik beschreibt die Erkenntnis, dass Entscheidungen über die Wahrscheinlichkeit von Ereignissen dadurch beeinflusst werden, wie sehr sie einem Prototyp entsprechen. Die Repräsentativitäts-Heuristik ist die Ursache für sehr viele Denkfehler, zum Beispiel der Spielerfehlschluss, der Verknüpfungsfehlschluss und der Prävalenzfehler. Letzterer lässt sich durch ein einfaches Denkexperiment verdeutlichen:

Szenario

Andi ist ein schmaler Brillenträger mit ruhiger Stimme. Er hat eine Vorliebe für Ordnung und sein Lieblingskomponist ist Khachaturian.

Was ist wahrscheinlicher?

1. Andi ist Universitätsbibliothekar
2. Andi ist LKW-Fahrer

Unser instinktives Urteil ist, dass Andi sicherlich Bibliothekar ist – weil die Beschreibung doch viel besser zu unserer Vorstellung eines Bibliothekars passt. Auch wenn wir etwas von Statistik verstehen, und uns ein Bild machen, wie viele LKW-Fahrer es im Vergleich zu Universitätsbibliothekaren gibt, kostet es uns Mühe, das instinktive Urteil zu überwinden. Diese Wechselwirkung zwischen schnellem, instinktivem und langsamem, mühseligem Denken wurde gründlich und umfassend von Tversky und Kahneman untersucht (siehe Kahneman 2011 für eine zugängliche und unterhaltsame Zusammenfassung ihrer Recherchen).

Dieses Phänomen entsteht nicht nur wegen gesellschaftlicher Vorurteile, sondern auch wegen der Hardware des Gehirns. Wahrscheinlichkeit zu berechnen, so wie es der Bayes'sche Satz verlangt, um über den Beruf von Andi zu entscheiden, ist keine natürliche Operation für Populationen von Nervenzellen, die über exzitatorische (erregende) oder inhibitorische (hemmende) Synapsen verbunden sind. Was das Gehirn allerdings sehr gut kann, ist Konzepte assoziieren (siehe Abbildung 1). Die stärkeren Verbindungen zwischen den Eigenschaften von Andi und ‚Bibliothekar' im Vergleich mit ‚LKW-Fahrer' führen dazu, dass ‚Bibliothekar' intensiver angeregt wird. Diese intensivere Aktivität wird vom Gehirn als höhere Wahrscheinlichkeit interpretiert, wobei die Häufigkeit der beiden Berufsgruppen nicht berücksichtigt wurde. Somit ist das Gehirn auch in Abwesenheit von Vorurteilen für solche Prävalenzfehler anfällig.

Abbildung 1: Assoziative Arbeitsweise des Gehirns. Die neuronalen Populationen, welche die Eigenschaften von Andi beschreiben, sind stärker mit Bibliothekar *als mit* LKW-Fahrer *assoziiert und regen deshalb ersteres stärker an.*

Es ist einfach zu sehen, wie die Repräsentativitäts-Heuristik unbewusst zu diskriminierenden menschlichen Urteilen führen kann. Wie kommt es zur Diskriminierung bei Algorithmen? Du et al. (2020) führen hierzu zwei Kategorien ein: 1) Diskriminierung bezüglich des *Ergebnisses* einer Klassifizierung, und 2) Diskriminierung bezüglich der *Qualität* der Klassifizierung. Die erste Kategorie kann weiter in 1a) Diskriminierung mittels *Input,* und 1b) Diskriminierung mittels *Repräsentation* unterteilt werden.

Diskriminierung mittels Input bedeutet, dass anhand der Trainingsdaten auf eine sensible Kategorie, beispielsweise eine gesetzlich geschützte, geschlossen wird. Obwohl solche Kategorien (wie zum Beispiel Geschlecht, ethnische Herkunft, usw.) in vielen Anwendungen nicht explizit angegeben werden, gibt es häufig korrelierte Eigenschaften (wie Postleitzahl, Name, Hobbys, Körpergröße, …). Diese sogenannten Proxy-Variablen sind besonders dann problematisch, wenn die Evaluationsfunktion durch menschliche Voreingenommenheit beeinflusst wurde. Ein mögliches Szenario ist ein KI-Werkzeug, um Lebensläufe zu evaluieren, bei dem die positiven Beispiele aus den Trainingsdaten die Kandidaten sind, die früher erfolgreich in der Firma aufgenommen wurden (siehe *https://tinyurl.com/r8sxx8xy*). Hatte die Firma historisch überwiegend Männer eingestellt, so lernt der Algorithmus, die Proxy-Variablen für Geschlecht zu identifizieren und in direkter Analogie zur Repräsentativitäts-Heuristik

die Lebensläufe männlicher Kandidaten an Hand dieser Variablen positiver zu evaluieren als die der Kandidatinnen. In den letzten Jahren konnten gehäuft Benachteiligungen von Minderheiten nachgewiesen werden, nicht nur bei der Rekrutierung, sondern auch in diversen und fundamentalen Lebensbereichen wie Gesundheit, Sicherheit, Zugang zu Krediten und der Behandlung im juristischen System. Diese Ergebnisse verdeutlichen, dass KI-Werkzeuge nicht dort naiv eingesetzt werden dürfen, wo historische Diskriminierungen in einer *Black Box* widergespiegelt und verstärkt werden können.

Diskriminierung mittels Repräsentation bedeutet, dass die sensible Kategorie nicht im Input selber zu finden ist (entweder direkt oder als Proxy), sondern erst in tieferen Schichten eines Deep-Learning Netzwerks identifiziert werden kann. Dieses Problem tritt insbesondere bei Image-Daten auf; eine häufige Ursache dafür ist eine ungleichmäßige Repräsentation bei den Trainingsdaten, wenn zum Beispiel die Gruppe der positiven Beispiele mehr männliche als weibliche aufweist. Da viele Netzwerke durch Bilder aus dem Internet trainiert werden, ist das durchaus problematisch. Um sich selbst davon zu überzeugen, kann man „Consultant Doctor" und „Nurse" (die Begriffe auf Englisch für Facharzt/ärztin und Krankenpfleger:in sind geschlechtsneutral) in Google-Suchen angeben, und die Anzahl von Männern und Frauen zählen (zur Kontrolle die Information: in Deutschland sind 40 % der Ärzt:innen und 80 % der Krankenpfleger:innen weiblich). Natürlich geht es bei Diskriminierung nicht nur um das Geschlecht. Wenn man bei den obigen Suchergebnissen nach Repräsentationen von z. B. schwarzen Frauen, Frauen über 50, oder Menschen mit Behinderung sucht, würde man, allein nach den Bildern zu beurteilen, zu dem Schluss kommen, dass solche Menschen wohl keine Fachärzte sein können. Algorithmen, die mit solchen Datensätzen trainiert werden, werden zwangsläufig voreingenommen, da sie nicht über das strukturelle Hintergrundwissen verfügen, um Repräsentations-Missstände der Trainingsdaten zu kompensieren.

Auf *Diskriminierung bezüglich der Qualität der Klassifizierung* gehe ich nur kurz ein. Es heißt, dass ein bestimmtes Klassifizierungstool besser für eine demografische Gruppe funktioniert als für eine andere. Zum Beispiel können viele gängige Gesichtserkennungsanwendungen die Gesichter von Menschen mit heller Hautfarbe viel besser als die mit dunkler Hautfarbe erkennen und männliche Gesichter deutlich besser als weibliche. Der Unterschied im Identifikationserfolg zwischen weißen Männern und schwarzen Frauen kann bis zu 34 % betragen (Buolamwhini et al., 2018; Zou & Schiebinger, 2018). Dieses Problem ist normalerweise auf ungleiche Verteilungen in den *Trainingsdaten* zurückzuführen. Als Beispiel: im Imagenet Dataset 2012[2], welches als Basis für sehr viele Image-Klassifizierungs-Algorithmen dient, zeigen nur 41,6 % der Bilder von Menschen eine Frau; zudem gibt es praktisch keine Menschen über 60. In diesem Szenario passiert das *Parameterfitting* automatisch zugunsten der Mehrheitskategorie. Außerdem gibt es zwangsläufig mehr Variabilität bei den Minderheitskategorien, was die Performanz des Algorithmus in diesen Fällen reduziert.

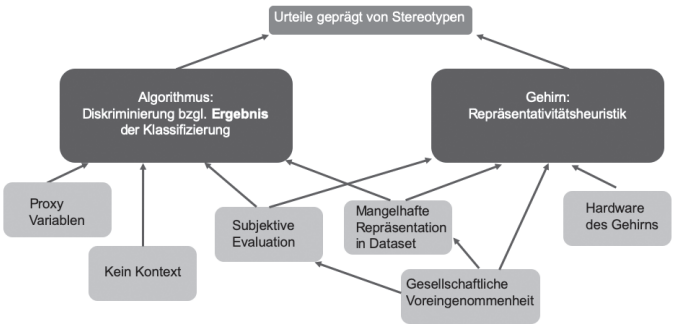


Abbildung 2: Die gemeinsamen und unterschiedlichen Ursachen diskriminierender Urteile beim Gehirn und beim Algorithmus

Insgesamt kann man argumentieren, dass Diskriminierung bezüglich des Ergebnisses einer Klassifizierung ein analoges Phänomen zur Repräsentativitäts-Heuristik ist. Sie haben ähnliche Effekte, nämlich Urteile, die von Stereotypen geprägt sind, und Gemeinsamkeiten bei ihren Ursachen, obgleich sie auch eigene Faktoren haben. Die Zusammenhänge sind (unvollständig) in Abbildung 2 illustriert.

Diese Erkenntnis heißt weder, dass wir diskriminierende Ergebnisse einfach hinnehmen, noch, dass wir auf den Einsatz von KI-Algorithmen verzichten müssen. In der Tat gibt es eine Reihe von technischen Ansätzen, um die Voreingenommenheit eines Algorithmus zu reduzieren. Ich werde hier ein paar davon kurz schildern:

- Erstens, Trainingsdaten können angepasst werden, um die Repräsentation unterschiedlicher Gruppen auszugleichen. Yang et al. (2020) haben dies mit großer Wirkung bei dem ImageNet-DataSet durchgeführt.

- Zweitens, um Proxy-Variablen zu vermeiden, kann man das Netzwerk analysieren, um die Variablen der Input-Daten zu identifizieren, die besonders wichtig für die Klassifizierung

Abigail Morrison

Prof. Dr. **Abigail Morrison** ist Forscherin in Computational Neuroscience am Forschungszentrum Jülich und an der RWTH Aachen. Ihr wissenschaftlicher Fokus ist Lernen und Informationsverarbeitung in puls-gekoppelten neuronalen Netzwerken, mit weiteren Interessen an den Schnittstellen zwischen Neurowissenschaften und künstlicher Intelligenz und am Einsatz von Hochleistungsrechnern in der Neurowissenschaft.

sind. Im nächsten Schritt kann man eine Korrelationsanalyse durchführen, um mögliche Zusammenhänge zwischen den so gefundenen Variablen und Mitgliedschaft in einer gesetzlich geschützten Kategorie zu identifizieren. Besteht so eine Korrelation, so hat man eine Proxy-Variable gefunden, die aus den Daten ausgeschlossen werden kann.

- Drittens, um Diskriminierung mittels Repräsentation zu kontern, steht die *Concept-Activation-Vector*-Technik zur Verfügung (Kim et al., 2018). Hier untersucht man, ob das Netzwerk indirekt geschützte Kategorien mitlernt, und verwendet diese Information als Trainingssignal, um diesem Lernen entgegenzuwirken.

Selbstverständlich verlangen diese Ansätze mehr Reflektion, mehr Arbeit und mehr Zeit als einfach mal schnell ein Netzwerk zu trainieren. Dennoch haben Algorithmen den großen Vorteil, im Gegensatz zum menschlichen Gehirn, dass sie technische Systeme sind, die beliebig auseinandergenommen und inspiziert werden können, so dass wir wirksame technische Lösungen erarbeiten können. Daher besteht die Hoffnung, dass der richtige, reflektierte Einsatz von KI Diskriminierung in der Gesellschaft reduziert – allerdings nur, wenn wir auf jeder Ebene darauf bestehen, dass KI-Werkzeuge diesbezüglich kritisch geprüft und unabhängig validiert werden.

Referenzen

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

Du M, Yang F, Zou N, Hu X (2020) Fairness in Deep Learning: A Computational Perspective. IEEE Intelligent Systems.

Kahneman D (2011) Thinking, Fast and Slow. Macmillan. ISBN 978-1-4299-6935-2

Kim B, Wattenberg M, Gilmer J, Cai C, Wexler J, Viegas F, Sayres R (2018) Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). International Conference on Machine Learning (ICML)

Yang K, Qinami K, Fei-Fei L, Deng J, Russakovsky O (2020) Towards Fairer Datasets: Filtering and Balancing the Distribution of the People Subtree in the ImageNet Hierarchy. Conference on Fairness, Accountabiility and Transparency (FAT*)

Zou J, Schiebinger L (2018) AI can be sexist and racist—it's time to make it fair. Nature 559 (pp. 324-326)

Anmerkungen

1 *Siehe https://de.wikipedia.org/wiki/Liste_kognitiver_Verzerrungen für eine ausführliche Liste.*

2 *https://www.image-net.org/index.php*

Silja Samerski

Kontrollierte Selbstbestimmung

Wie Überwachung im Gesundheitswesen unter die Haut geht

Die Digitalisierung des Gesundheitswesens verbessert die Versorgung und verhilft Patient:innen zu mehr Selbstbestimmung, so versprechen IT-Unternehmen, Krankenkassen und Expertengremien. Die elektronische Patientenakte (ePA) soll beispielsweise nicht nur Ärzt:innen, sondern auch den Patient:innen selbst einen besseren Überblick u.a. über Befunde, Arztbriefe und Abrechnungsdaten ermöglichen. Digitale Angebote wie Gesundheitsportale, virtuelle Gesundheitsassistenten und Symptom-Checker sollen zudem die Gesundheitskompetenz der Bevölkerung und damit auch die Patientensouveränität erhöhen (Sachverständigenrat Gesundheit 2021; Techniker-Krankenkasse). Wird die lang geforderte Patienten-Selbstbestimmung im Gesundheitswesen mithilfe der Digitalisierung endlich Wirklichkeit?

erschienen in der FIfF-Kommunikation, herausgegeben von FIfF e.V. - ISSN 0938-3476 www.fiff.de

Es gibt gute Gründe, diesem Vers[illegible]nd dort alle medizinischen Daten zu sein. Bereits heute zeichnet sic[illegible]nten hätten dann nur noch das rene Datensouveränität von Patie[illegible]und Informationen zu „verschatten" – in Estland beispielsweise ist ein solches Opt-out-System sundheitswesen kein großes Gewicht hat. Angesichts des Datenhungers von Forschung und Industrie haben zahlreiche realisiert. Auch die Corona-Krise zeigt, wie groß die Bereitschaft Gremien und Experten das Recht auf informationelle Selbstbestimmung bereits für veraltet erklärt und neue Datenschutzkonzepte vorgelegt, die Big Data nicht behindern (Deutscher Ethikrat 2017, SVR Gesundheit 2021). Unter dem Banner *Sharing is caring*, oder noch drastischer: *Datenschutz tötet* arbeiten Wissenschaftler:innen und Digitalunternehmen an der Aufweichung von Datenschutz und Datensouveränität. Bisher ist vorgesehen, dass Patient:innen selbst darüber entscheiden können, welche Daten sie in der ePA speichern lassen und mit welchen Ärzt:innen sie diese teilen möchten. Werden die begehrten Gesundheitsdaten jedoch erst in einer Datenbank gespeichert, wird es schwierig sein, den gläsernen Patienten langfristig zu verhindern. Der Sachverständigenrat (SVR)-Gesundheit schlägt in seinem aktuellen Gutachten (2021) bereits vor, die ePA allen Menist, Datenschutz und Persönlichkeitsrechte angesichts von Gesundheitsbedrohungen und Gesundheitsverheißungen hintanzustellen. Binnen kürzester Zeit ist es normal geworden, sich im Alltag mit sensiblen Gesundheitsdaten wie dem Immunitätsstatus auszuweisen, um Zugang zum Restaurant oder Theater zu erhalten. Die Infrastruktur und die Gewöhnung, die in Krisenzeiten geschaffen wurden, sollen nach der Krise nicht abgebaut, sondern ausgebaut werden: Der französische Rüstungskonzern *Thales* hat beispielsweise vor, die Impfpässe zu einer digitalen ID weiterzuentwickeln: „So-called digital *vaccination passports* will play a key role in enabling citizens to access all manner of services and will act as a precursor to the rollout of mobile digital ID" so ist auf der Webseite von Thales zu lesen (Theyras 2021).