Frauke Kreuter, Christoph Kern, Patrick Oliver Schenk

Automatisierte Entscheidungen: Aspekte von Fairness, Datenqualität und Privacy

In:
Dössel, Olaf / Schäffter, Tobias / Rutert, Britta (Hrsg.): Künstliche Intelligenz in der Medizin.
ISBN: 978-3-949455-18-6
Berlin: Berlin-Brandenburgische Akademie der Wissenschaften, 2023
S. 98-111
(Denkanstöße aus der Akademie : eine Schriftenreihe der Berlin-Brandenburgischen Akademie der Wissenschaften ; 11)

Persistent Identifier: urn:nbn:de:kobv:b4-opus4-38102

AUTOMATISIERTE ENTSCHEIDUNGEN: ASPEKTE VON FAIRNESS, DATENQUALITÄT UND PRIVACY

Frauke Kreuter, Christoph Kern, Patrick Oliver Schenk

Ohne Zweifel sind die Möglichkeiten des Einsatzes von künstlicher Intelligenz (KI) in der Medizin vielversprechend. Dieser Band gibt Anregungen und lässt uns hoffnungsvoll in die Zukunft blicken. Damit sich diese Zukunft so realisiert, wie wir uns das erhoffen, ist es wichtig, einige Aspekte der KI-Anwendungen zu betrachten, die außerhalb der reinen Anwendung von Algorithmen liegen. Dazu gehören Fairness, Datenqualität und Privacy.

Ein Wort vorab zur Verwendung der Begrifflichkeiten: KI, Algorithmen und maschinelles Lernen haben für Statistiker und Datenwissenschaftler ganz spezifische Bedeutungen. Im breiten Sprachgebrauch werden sie allerdings oft in einem Atemzug genannt und austauschbar verwendet.[1] Das liegt an einer Gemeinsamkeit, die für diesen Beitrag von besonderer Bedeutung ist: Sie alle extrahieren Informationen und generieren Vorhersagen aus Daten. Die Aspekte, die dieser Beitrag behandelt, sind relevant für alle Situationen, in denen Entscheidungen mit Hilfe von Modellen getroffen werden, die auf Daten beruhen.

Im Vergleich zu klassischen statistischen Verfahren wie Regression haben KI-Verfahren eigene Stärken: Sie können unter potentiell sehr vielen Prädiktoren die relevantesten automatisch herausfiltern (z. B. in der Genomik sehr wichtig); auch sehr komplexe Strukturen wie nichtlineare Zusammenhänge oder Interaktionen können flexibel erkannt werden; die durchschnittliche Güte von Prognosen ist im Allgemeinen höher. Während die meisten heutigen KI-Anwendungen nur ein *allgemein* gut passendes Modell entwickeln, liegen Hoffnungen auch im vermehrten Einsatz von Verfahren zur Erkennung heterogener Effekte zwischen Gruppen oder Individuen (z. B. Präzisionsmedizin). Zu den Nachteilen von KI gehören insbesondere eine teils viel geringere Interpretierbarkeit, Verständlichkeit und Transparenz sowie ein oft viel höherer Bedarf an Daten (Beobachtungen) und Computer-Ressourcen. Maße dafür, wie unsicher die Ergebnisse von KI sind, werden noch entwickelt bzw. sind heute oft überoptimistisch. Auch wird die o. g. höhere

1 Was heute mit dem Oberbegriff KI bezeichnet wird, ist sehr oft ausschließlich maschinelles Lernen. Eine Einführung in letzteres bieten James G., Witten D., Hastie T., & Tibshirani R. 2021. An Introduction to Statistical Learning, Springer: New York, 2. Auflage, Buch und Kurs verfügbar via www.statlearning.com.

mittlere Prognosegüte typischerweise erkauft mit einem höheren, systematischen Bias – verzerrte Ergebnisse, die z. B., aber nicht immer, v. a. auf bestimmte Gruppen entfallen. Maße für die Datenabhängigkeit fehlen weitgehend.

Wie klassische Statistik erkennt und repliziert KI Muster in Daten, und zwar unabhängig davon ob diese Muster nun auf Datenprobleme oder auf „echte", uns interessierende Zusammenhänge zurückgehen. Wohnen den Daten, anhand derer ein Modell trainiert wird, Probleme inne, ist nicht zu erwarten, dass KI diese von selbst erkennt und löst.[2]

Im ersten der drei Teile dieses Beitrags stellen wir einige KI-Anwendungen vor. Gemeinsam ist den Anwendungen, dass sie Daten aus unterschiedlichen Quellen verwenden und Vorhersagen für unterschiedlichste Gruppen treffen. Gemeinsam ist ihnen auch, dass sie eine Balance zwischen der Qualität der Vorhersage insgesamt und der Vorhersagegüte für einzelne Gruppen finden müssten. Werden die gruppenspezifischen Vorhersagen nicht berücksichtigt, so kommt es leicht zu unintendierten Ungleichbehandlungen einzelner Gruppen. Mittlerweile steht eine Reihe von Metriken zur Verfügung, um Ergebnisse von KI-Anwendungen auf dieses Problem hin zu prüfen. Ob die Algorithmen sich fair verhalten, ist hierbei eine zentrale Frage.

Die mögliche Ungleichbehandlung kann, muss aber nicht unbedingt ein Problem sein. Denn aus soziologischer Sicht kann eine ungleiche Behandlung verschiedener Bevölkerungsgruppen durchaus intendiert sein. Dies werden wir im zweiten Teil diskutieren. Das Problem der Fairness wird sich nicht rein technisch lösen lassen, sondern bedarf einer gesellschaftlichen Diskussion über allgemeine Gerechtigkeitsprinzipien. Aber auch wenn sich eine Gesellschaft darauf einigt, welche Gerechtigkeitsprinzipien angewendet werden sollen, können die KI-Anwendungen nur so gut sein, wie die Daten, die in sie hineinfließen.

Ein Problem, das sich bei der Sammlung oder Generierung von Daten immer stellt, ist das systematische Fehlen von Informationen. Dies kann verschiedenste Ursachen haben; eine, an der wir etwas ändern können, ist die Frage des Datenzugangs. Wir sind in Deutschland aus gutem Grund darauf bedacht, den Datenschutz hochzuhalten. Ohne Zweifel ist die Selbstbestimmung darüber, welche

2 Ein einfaches Mehr der gleichen Daten löst das Problem nicht. Fehlt eine gesellschaftliche Gruppe völlig oder hat eine Variable einen systematischen Messfehler, ändert sich allein durch eine höhere Beobachtungszahl nichts an der fehlenden Repräsentanz oder dem Messfehler.

Informationen von einem und über einen verwendet werden können, ein hohes Gut. Es ist jedoch für die oder den Einzelne/-n eine schwer zu bewältigende Aufgabe, in jeder Situation zu entscheiden, ob eigene Daten freigegeben werden sollten oder nicht. Die daraus resultierenden Konsequenzen zu überblicken ist nicht nur schwierig, sondern oft unmöglich. Wir regen im dritten Teil deshalb an, einen normativen Blick auf diese Frage zu werfen und zu überlegen, in welchem Kontext die Nutzung welcher Daten legitim ist. Anstatt die Datennutzung über Datentypen zu regeln, könnte es sich lohnen, mehr darauf zu achten, mit welchem Ergebnis Daten genutzt werden. Technisch stehen bereits viele Lösungen zur Verfügung, die es ermöglichen würden, Daten für die Hebung des Allgemeinwohls besser zu nutzen. Dieser Aufgabe sollten wir uns stellen.

AUTOMATISIERTE ENTSCHEIDUNGEN IM ÖFFENTLICHEN SEKTOR

Häufig steht das Ergebnis eines Algorithmus nicht für sich, sondern ist nur der erste Schritt: als Basis von Entscheidungen (zweiter Schritt). Überall dort, wo knappe Ressourcen eingesetzt werden, besteht das Bedürfnis, den Einsatz dieser Ressourcen so zu optimieren, dass das Ergebnis möglichst effizient ist. Datenbasierte Vorhersagen über z. B. den Bedarf einer Zuteilung oder den möglichen Ausgang einer Handlung können beim Einsatz knapper Ressourcen helfen. Sind x Einheiten einer Ressource auf mehr als x Personen zu verteilen, kann ein Algorithmus z. B. diejenigen x Individuen mit dem größten Bedarf ermitteln. Koppelt man die Entscheidung dann fest an das Ergebnis des Algorithmus (die x Personen mit dem größten Bedarf werden bedacht, alle anderen nicht), d. h. automatisiert man die Entscheidung, spricht man von Automated Decision Making (ADM).[3] Dies ist vor allem dann attraktiv, wenn sich Ereignisse stabil aus vorhandenen Daten vorhersagen lassen, also wenn (und besser: nur, dann wenn) sich immer gleiche Muster in Daten finden lassen, die verlässlich auf ein Problem oder ein Ereignis hinweisen. Ein Vorteil der Automatisierung von Entscheidungen liegt in der Beschleunigung von Prozessen. Ebenso erhofft man, dass durch die feste Anbindung an das Ergebnis des Algorithmus nur noch relevante Größen einen Einfluss auf die Entscheidung haben und diese objektiver als durch Menschen gefällte Entscheidungen sind.

3 Die Übergänge dazu, dass die Verknüpfung nicht eins-zu-eins ist, sondern noch Entscheidungsspielraum besteht (Semi-automated Decision Making), ist durchaus fließend. Die klassische Situation hingegen kennzeichnet, dass die Resultate eines Algorithmus nur eine Information unter vielen für (einen) menschliche(n) Entscheider darstellen.

ADM im Gesundheitswesen

Wie sehen automatisierte Entscheidungen im Gesundheitswesen aus? In den USA nutzen Krankenversicherungen Algorithmen, um das zukünftige Krankheitsrisiko und damit den Bedarf für einen besonders hohen Versorgungsaufwand abzuschätzen. Diejenigen mit einem hohen Risiko werden dann mit besonderen Vorsorgeprogrammen unterstützt.[4] In Israel nutzen Ärzt:innen Vorhersagemodelle, um zu entscheiden, welche Patienten welche personalisierten Behandlungen bekommen sollen.[5] In Chicago werden Verfahren des maschinellen Lernens genutzt um vorherzusagen, welche HIV-Positiven wahrscheinlich aus der kontinuierlichen Behandlung und Betreuung herausfallen und damit sich selbst und andere gefährden könnten.[6] Im US-Bundesstaat Kansas wird versucht, mit Hilfe von KI die Spirale von unbehandelten psychiatrischen Erkrankungen und Verhaftungen zu durchbrechen. Personen, die laut Modell eine hohe Wahrscheinlichkeit haben, straffällig zu werden, werden je nach vorhandenen Ressourcen priorisiert behandelt.[7] Hinter diesen Einsätzen der Algorithmen verbirgt sich die Hoffnung, die begrenzten Vorhersagefähigkeiten von Menschen mit statistischen, algorithmischen Vorhersagetechniken zu verbessern oder gar zu ersetzen.[8]

Was kann da schiefgehen?

Der Einsatz von Algorithmen, der im amerikanischen Justizsystem eingeführt wurde, um Entscheidungen effizienter zu gestalten und der Subjektivität von Richtern vorzubeugen, hat jedoch zu Ernüchterung geführt. In Untersuchungen der gemeinnützigen Nachrichtenredaktion ProPublica, zeigte sich, dass nur 20 Prozent der Personen, für die Gewaltverbrechen vorhergesagt worden waren, diese auch tatsächlich begingen. Bei der Vorhersage, wer wieder straffällig werden

4 Obermeyer Z, Powers B, Vogeli C, & Mullainathan S. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. Science 366(6464): 447–453.

5 Dagan N, Cohen-Stavi CJ, Avgil Tsadok M. et al. 2019. Translating clinical trial results into personalized recommendations by considering multiple outcomes and subjective views.npj Digital Medicine 2(81).

6 Ramachandran A, Kumar A, Koenig H. et al. 2020. Predictive Analytics for Retention in Care in an Urban HIV Clinic. Scientific Reports 10(6421).

7 Rodolfa KT, Lamba H & Ghani R. 2021. Empirical observation of negligible fairness–accuracy tradeoffs in machine learning for public policy. Nature Machine Intelligence 3: 896–904.

8 Engelmann J & Puntschuh M. 2020. KI im Behördeneinsatz: Erfahrungen und Empfehlungen. Fraunhofer-Institut für Offene Kommunikationssysteme, Berlin. http://publica.fraunhofer.de/eprints/urn_nbn_de_0011-n-6350714.pdf

würde, machte der Algorithmus bei schwarzen und weißen Angeklagten in etwa gleichem Maße, aber auf sehr unterschiedliche Weise Fehler. Der Algorithmus stufte schwarze Angeklagte fast doppelt so häufig fälschlicherweise als rückfällig ein wie weiße Angeklagte (falsch-positiv), während weiße Angeklagte wesentlich häufiger fälschlicherweise als nicht-rückfällig eingestuft wurden (falsch-negativ).[9]

Seit ProPublica auf diese gruppenspezifischen Vorhersageunterschiede aufmerksam gemacht hat, überschlägt sich die KI-Forschung mit der Entwicklung und dem Einsatz von Maßzahlen bzw. Metriken zur Bestimmung von ‚Fairness'. Zugleich schreiben Regularien vor, dass in Antidiskriminierungsgesetzen definierte, geschützte Merkmale (wie Geschlecht, Alter, Herkunft) nicht zum Training von KI-Algorithmen verwendet und Entscheidungen nicht allein auf Algorithmen basieren dürfen.

Weder Regularien noch Maßzahlen haben bisher zum gewünschten Erfolg geführt.

- Selbst wenn geschützte Merkmale nicht für das Modelltraining verwendet werden, können die Vorhersagen unterschiedlich gut für geschützte Gruppen sein.[10] Wenn geschützte Merkmale komplett aus den Daten entfernt werden, ist es umgekehrt sehr schwer, überhaupt zu evaluieren, ob eine bestimmte Gruppe systematisch benachteiligt wird.

- Selbst wenn menschliche Entscheider den Vorschlag des Algorithmus lediglich als eine Informationsquelle hinzuziehen sollen, werden sie doch stark davon beeinflusst und neigen dazu, Vorgeschlagenes zu bestätigen.[11]

- Selbst wenn Metriken zur Überprüfung der ‚Fairness' von Algorithmen zur Verfügung stehen, leiden sie an dem inhärenten Problem, dass sie Zielgrößen formulieren, die von Algorithmen nicht alle gleichzeitig eingehalten werden können.[12]

9 Angwin J, Larson J, Mattu S, Kirchner L. (2016). Machine Bias. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

10 Pope DG, and Sydnor JR. 2011. Implementing Anti-discrimination Policies in Statistical Profiling Models. American Economic Journal: Economic Policy 3(3): 206–231.

11 Goddard K, Roudsari A, & Wyatt JC. 2012. Automation bias: a systematic review of frequency, effect mediators, and mitigators. Journal of the American Medical Informatics Association 19(1): 121–127.

12 Berk R, Heidari H, Jabbari S, Kearns M, & Roth A. 2021. Fairness in Criminal Justice Risk Assessments: The State of the Art. Sociological Methods & Research 50(1): 3–44.

Der letzte Punkt ergibt sich daraus, dass wir zum einen in der Praxis in den allermeisten Situationen unterschiedliche Basisraten für verschiedene gesellschaftliche Gruppen vorfinden. Zum Beispiel geht man davon aus, dass Frauen eher zu Depressionen neigen oder dass junge Männer für die überwiegende Mehrheit der Gewaltverbrechen verantwortlich sind. Zum anderen sind Modelle nie perfekt und die Modellgüte unterscheidet sich oft zwischen den Gruppen. Derartige Unterschiede wirken sich auf die statistisch erreichbare Fairness aus, d. h. auf die Frage, welche Metriken gleichzeitig erfüllt werden können. Abzuwägen gilt z. B., ob man einen Anstieg von falsch-negativen Vorhersagen in Kauf nimmt, um mehr Fairness zwischen den Gruppen herzustellen.[13]

Auch wenn ethische Prinzipien in der Medizin eine optimale Behandlung aller gebieten, gibt es auch hier zahlreiche Situationen, in denen aufgrund bestehender Ressourcenknappheit (feste Gesamtbudgets, fixe Kapazitäten oder Wartezeiten) Entscheidungen über Behandlungsprioritäten oder die Zuteilung bestimmter Vorsorgemaßnahmen getroffen werden müssen. Derzeit werden in Deutschland Leitlinien der Fachgesellschaften in diesen Entscheidungssituationen verwendet, individuelle Interpretationen und Abwägungen sind in der Regel aber unvermeidbar. Wenn KI systematisch eingesetzt werden soll, wird eine Diskussion über Leitlinien und Interpretationen umso wichtiger. Denkbar ist, dass beim Einsatz von KI bestimmte systematische Verzerrungen in ganz anderer Weise skalieren,[14] individuelle Biases eines einzelnen Entscheiders jedoch überschrieben werden. Das heißt, dass anstatt eine Reihe Entscheider mit unterschiedlichen Biases zu haben, wird beim Einsatz von KI für alle Entscheidungen der Bias der Mehrheit wirksam. Im oben erwähnten Gerichtskontext kann man sich das wie folgt vorstellen: Im Idealfall skaliert man die beste Richterin. Im schlimmsten Fall sind alle richterlichen Entscheidungen von einem rassistischen Bias geprägt.

13 Ebd.

14 Gerade bei automatisierten Entscheidungen ist es möglich, dass einmal eingeschriebene Verzerrungen sich selbst bestätigen oder gar verstärken: Die aufgrund des Bias in den Ergebnissen des Algorithmus verzerrten Entscheidungen produzieren neue Daten, welche ebenso oder gar stärker verzerrt sind – und wieder in den Algorithmus eingehen.

PRINZIPIEN DER VERTEILUNGSGERECHTIGKEIT

Wenn Entscheidungen zum gezielten Einsatz von knappen Ressourcen getroffen werden, stellt sich schnell die Frage, was gerecht oder was fair ist. Die Fairnessmetriken der KI-Forschung sehen eine Allokation von Ressourcen in der Regel dann als fair an, wenn Personen, die sich durch bestimmte geschützte Attribute unterscheiden (Geschlecht, ethnische Zugehörigkeit etc.), identische Entscheidungen mit identischer Fehlerwahrscheinlichkeit zugewiesen werden. Schaut man aus einer sozialwissenschaftlichen Perspektive auf diese Metriken, wird schnell deutlich, dass prominente Fairnessmetriken der KI-Forschung nicht gut mit den verschiedenen Ansätzen der Theorien zur Verteilungsgerechtigkeit übereinstimmen.

Nachstehende Tabelle[15] zeigt eine Zusammenfassung von zentralen Gerechtigkeitsprinzipien und eine Illustration ihrer Anwendung auf das Beispiel der HIV-Prävention. Vorhersagemetriken (d. h. Metriken zum Vergleich von Vorhersagefehlern) wenden Chancengleichheit (equality of opportunity) auf die Verteilung von Vorhersagefehlern an. Das heißt, geschützte Personengruppen sollen die gleiche Chance auf eine nicht-fehlerhafte Vorhersage haben.

Entscheidungsmetriken (d. h. Metriken zum Vergleich der Eintrittswahrscheinlichkeiten des vorhergesagten Ereignisses) implementieren Vorstellungen von Egalitarismus. Hierbei sollen Personen nicht für bestimmte Merkmale (z. B. für ihr Geschlecht) verantwortlich gemacht werden und sollten deshalb die gleichen Risikovorhersagen erhalten. Wir argumentieren, dass diese beiden Verbindungen unzureichend sind. Gerechtigkeitsprinzipien wie Equality, Desert, Need und Efficiency werden durch bestehende Fairnessmetriken nicht ausreichend erfasst. (Die englischen Fachbegriffe werden in der folgenden Tabelle erläutert.)

15 Ausführlich behandelt in Kuppler M, Kern C, Bach RL & Kreuter F. 2021. Distributive Justice and Fairness Metrics in Automated Decision-making: How Much Overlap Is There? https://arxiv.org/abs/2105.01441

Tabelle 1: Ausgewählte Prinzipien der Verteilungsgerechtigkeit und deren Illustration am Beispiel der Verteilung von Präventionsmaßnahmen für HIV-Infizierte. Adaptiert nach Kuppler u. a. 2021 in https://arxiv.org/abs/2105.01441

Gerechtigkeitsprinzip (engl. terms)	**Verteilungsregel**	**Beispiel**
Equality	Verteile R Ressourcen auf Individuum X genau dann, wenn die für X vorhandenen Ressourcen Y so sind, dass die Hinzugabe von R die gesamtgesellschaftliche Ungleichheit minimiert.	Verteile Präventionsmaßnahmen gegen Versorgungsausfall für HIV-Infizierte so, dass alle HIV-Positiven gleich gut versorgt sind.
Desert	Weise R Ressourcenanteile Individuum X genau dann zu, wenn diese auf X zutreffen.	Lasse Präventionsmaßnahmen denen zukommen, die ihre Beiträge zur Krankenversicherung bezahlt haben.
Need	Weise R Ressourcenanteile Individuum X genau dann zu, wenn X diese braucht.	Fokussiere Präventionsmaßnahmen auf diejenigen, deren Einkommen unter eine bestimmte Schwelle fällt.
Efficiency	Verteile R Ressourcen auf Individuum X genau dann, wenn die Hinzugabe von R den gesamtgesellschaftlichen Nutzen maximiert.	Fokussiere Präventionsmaßnahmen auf die am stärksten von den Maßnahmen profitierende Gruppe.
Equality of Opportunity	Weise allen X mit gleichen Attributen die gleiche Anzahl an R Ressourcenanteilen zu.	Stelle sicher, dass alle Gruppen von HIV-Infizierten die gleiche Chance haben, unterstützt zu werden.

Ein Ansatzpunkt wäre, die ADM-Prozesse in zwei Schritten zu denken: einen Vorhersage- und einen Entscheidungsschritt.[16] Das Ziel des Vorhersageschritts besteht darin, ein möglichst akkurates Modell davon zu erstellen, wie die Welt tatsächlich ist bzw. wie sie mit den gegebenen Daten abgebildet werden kann. Für den Entscheidungsschritt sollte dann explizit eine Verteilungsregel definiert werden. Diese

16 Kuppler, M, Kern, C, Bach, RL, Kreuter, F. 2022. From fair predictions to just decisions? Conceptualizing algorithmic fairness and distributive justice in the context of data-driven decision-making. Frontiers in Sociology. https://doi.org/10.3389/fsoc.2022.883999

resultiert nicht aus einem Modell, sondern aus einer gesellschaftlichen Diskussion über Werte und erwünschte Zielzustände. Die Korrektur von gesellschaftlichen Verzerrungen ist Aufgabe der Verteilungsregeln, nicht die des Vorhersagemodells. Auch heute findet dieser Entscheidungsschritt schon statt, auch ohne den Einsatz von KI. Im Kontext des Einsatzes von KI besteht allerdings die erhöhte Gefahr, dass Verzerrungen im Vorhersageschritt für die menschlichen Entscheider[17] schwer erkennbar sind und systematisch zu dem Entscheidungsschritt durchgereicht werden. Auch bei dem expliziten Einsatz von Menschen im zweiten Schritt, ist ein derartiges Vorgehen wahrscheinlich, da Menschen doch dazu neigen, vorgefertigte Annahmen[18] bzw. Vorgeschlagenes zu bestätigen.[19]

DIE ROLLE DATENGENERIERENDER PROZESSE

Wie gut ein Vorhersagemodell die Welt abbilden kann, hängt entscheidend davon ab, welche Daten zum Training der Algorithmen verwendet werden. In der Praxis beobachten wir häufig eine Diskrepanz zwischen der Art und Weise, wie die Welt tatsächlich ist und wie die Welt in den Trainingsdaten von KI-Modellen abgebildet wird. Statistische Verzerrungen entstehen z. B. durch Repräsentation/Sampling Bias (d. h. die Daten repräsentieren nicht die Gesamtpopulation, auf die das Vorhersagemodell später angewendet werden soll) oder Messfehler (d. h. es gibt systematische Diskrepanzen zwischen gemessenen und wahren Attributen).

Statistische Verzerrungen können eine robuste Anwendung von Vorhersagemodellen in der Medizin stark einschränken. Je nach Vollständigkeit der dem Training der Modelle zugrunde liegenden Daten kann die Übereinstimmung zwischen vorhergesagten und beobachteten (Krankheits-) Risiken für bestimmte Bevölkerungsgruppen sehr unterschiedlich sein.[20] So zeigen Barda et al. (2021)[21]

17 Sofern solche noch eingebunden sind. Ist die Entscheidung absolut automatisiert, entfällt die Last vollends auf die Entwickler und ggf. Prüfer (audits) des Algorithmus.

18 Klayman J. 1995. Varieties of Confirmation Bias, in: Busemeyer J, Hastie R, & Medin D L (Eds.), Psychology of Learning and Motivation, Vol. 32: 385–418. Academic Press.

19 Siehe die Ausführungen von Tourangeau R et al. 2000. The Psychology of Survey Response, Cambridge University Press, zu Acquiescence Bias.

20 Im Allgemeinen fehlen zudem auch noch belastbare Maße für die Unsicherheit der jeweiligen Prognose.

21 Barda N, Yona G, Rothblum GN, Greenland P, Leibowitz M, Balicer R, Bachmat E, & Dagan N. 2021. Addressing bias in prediction models by improving subpopulation calibration. Journal of the American Medical Informatics Association 28(3): 549–558.

beispielsweise für ein Modell zur Einschätzung des Risikos von Herz-Kreislauf-Erkrankungen eindrucksvoll, zu welchen Verzerrungen und Vorhersagefehlern es für Teilpopulationen kommen kann. Solche Effekte können beispielsweise dann auftreten, wenn Teilgruppen in der Modellentwicklung unterrepräsentiert sind. Das ist ein Problem jedes datenbasierten Modells, das wir auch heute schon kennen und das bei KI-Anwendungen ebenso besonderer Aufmerksamkeit bedarf.[22]

Von wem haben wir welche Daten?

Vor der Anwendung eines KI-Algorithmus sollte man sich deshalb immer fragen, von welchen Populationen Informationen während der Modelltrainingsphase zur Verfügung standen. Insbesondere sollten die Personengruppen in den Trainingsdaten den Personengruppen entsprechen, für die Vorhersagen gemacht werden.[23] Nicht alle Datenquellen umfassen alle Personengruppen, für die Vorhersagen gemacht werden sollen. So gibt es beispielsweise sozial-strukturelle Unterschiede im Hinblick auf den Besitz und die Nutzung von elektronischen Geräten wie z. B. Smartphones, welche zur Generierung von Trainingsdaten genutzt werden können.[24] Um Unzulänglichkeiten einzelner Datensätze zu begegnen, werden deshalb zunehmend Daten aus verschiedensten Quellen miteinander verknüpft.[25] Allerdings kann auch diese Verknüpfung für Teilgruppen systematisch häufiger fehlschlagen. In Deutschland und Europa wird an vielen Stellen verlangt, dass Personen explizit einer Weitergabe von Daten zustimmen. Leider wird die Zustimmung oft nicht basierend auf inhaltlichen Überlegungen gegeben oder verweigert, sondern erfolgt situativ und oft ohne fest verankerte Einstellungen. Dies führt dazu, dass Personen, die nicht nachvollziehen können, was mit ihren Daten passiert, oder die wenig Vertrauen in daten-erhebende Institutionen haben,

22 Ein weiterer Grund sind Messfehler, die sich systematisch zwischen Gruppen unterscheiden. Lernt ein Algorithmus also z. B. nicht anhand des wahren Krankheitsstatus, sondern anhand von gestellten Diagnosen, die für manche Teilgruppen seltener richtig sind, so wird auch ein Algorithmus (und darauf aufbauendes ADM) diese Verzerrungen replizieren; er erkennt nicht von selbst, wenn die ihm gegebenen Informationen falsch sind.

23 Ein bekanntes Problem in der Medikamentenentwicklung, wenn z. B. (schwangere) Frauen nicht in die Studienphasen eingebunden werden, Medikamente später aber auch für sie genutzt werden.

24 Keusch F, Bähr S, Haas GC., Kreuter F, Trappmann M. 2020. Coverage Error in Data Collection Combining Mobile Surveys with Passive Measurement Using Apps: Data from a German National Survey. Sociological Methods & Research, https://doi.org/10.1177/0049124120914924

25 Die Nutzung neuer Variablen und der hohe Bedarf an Beobachtungen von KI sind weitere Gründe.

weniger häufig zustimmen.[26] Dadurch stehen für diese Personen dann später weit weniger Trainingsdaten zur Verfügung.

Wir konnten in verschiedenen Studien zeigen, dass die Zustimmungsrate zu Verknüpfungen besonders empfindlich auf bestimmte Designmerkmale reagiert, z. B. darauf, wo die Zustimmungsfrage im Fragebogen platziert ist und wie die Frage formuliert ist.[27] Für uns deutet dies darauf hin, dass die Einstellung zur Verknüpfung nicht so stark ausgeprägt ist, wie es die Vorschriften, die eine solche Zustimmung vorschreiben, vermuten lassen. Möglicherweise müssen wir uns aber auch einfach davon verabschieden, dass Individuen (in jeder Situation neu) darüber entscheiden können und sollen, welche Daten weitergegeben oder verknüpft werden sollen. Die immense Häufigkeit solcher Situationen und die Unmöglichkeit, alle zukünftig möglichen Nutzungen zu bedenken, spräche ebenfalls dafür. Eine Alternative wäre hier, stärker gesellschaftlich über angemessene Datenströme nachzudenken.

Welche Daten dürfen wir wann nutzen?

Helen Nissenbaum, Philosophin und Professorin für Informationswissenschaften der Cornell Tech University in den USA, befasst sich seit Jahren mit der Angemessenheit von Datenströmen. Unter dem Stichwort „contextual integrity" (CI) definiert sie Bedingungen, unter denen eine Datenverarbeitungspraxis angemessen ist. CI besagt, dass Datenübertragungen den Erwartungen an den Schutz der Privatsphäre entsprechen, wenn sie mit den Datenschutznormen übereinstimmen, die wiederum abhängig von der Art und den Umständen der gesammelten Informationen sowie den beteiligten Akteuren sind. Kontextuelle Informationsnormen geben für den Informationsfluss fünf Schlüsselparameter vor: (1) den Absender der Information, (2) den Empfänger der Information, (3) das Attribut oder die Art der Information, (4) das Subjekt der Information und (5) ein Übertragungsprinzip mit den Bedingungen für einen angemessenen Informationsfluss.[28] So ist es beispielsweise im Gesundheitswesen angemessen, dass Patienten (Sender und Subjekt) ihren Ärzten (Empfänger) Gesundheitsinformationen (Attribute) vertraulich zur Verfügung

26 Auch hier kann es zu systematischen Gruppenunterschieden kommen, wenn Bildung ein maßgeblicher Treiber des Verständnisses der Datennutzung ist, oder bestimmte soziale Gruppen aufgrund von Diskriminierung in der Vergangenheit vorsichtiger sind.

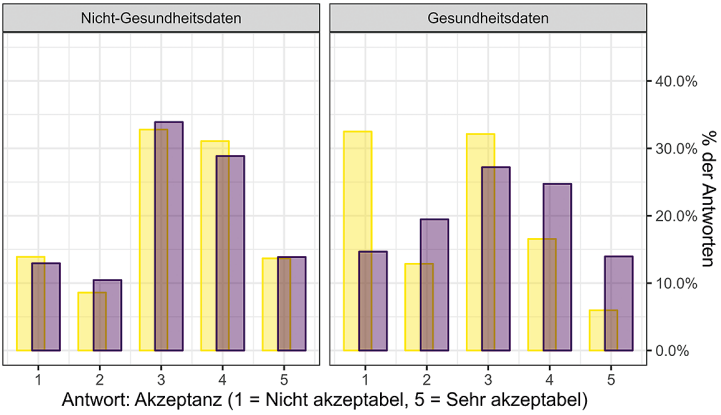
27 Sakshaug JW, Schmucker A, Kreuter F, Couper MP, Singer E. 2019. The Effect of Framing and Placement on Linkage Consent, Public Opinion Quarterly 83(S1): 289–308.

28 Nissenbaum H. 2010. Privacy in context: Technology, policy, and the integrity of social life. Stanford, Calif.: Stanford Law Books.

stellen (Übertragungsprinzip). Dieser Ablauf ist normenkonform und unproblematisch. Leitet eine Arztpraxis jedoch medizinische Informationen an einen anderen Empfänger, z. B. den Arbeitgeber eines Patienten, weiter, so ist das Prinzip der vertraulichen Übertragung verletzt. Bei der Beurteilung der Angemessenheit müssen immer alle Parameter berücksichtigt werden. Eine Datenschutzregel, die sich nur auf die Attribute oder Arten der Informationen bezieht, ist unzureichend. Die Wahrnehmung der Bevölkerung, welche Nutzung akzeptabel ist, kann sich situativ durchaus verändern. Wie die Daten von rund 600 Befragten von Gerdon et al. (2021)[29] in der nebenstehenden Abbildung zeigen, war die Bereitschaft, eigene Daten zur Bekämpfung einer Pandemie zur Verfügung zu stellen, vor der COVID-19-Pandemie deutlich niedriger als nach Beginn. Diesen Unterschied sahen wir bei anderen Datentypen und anderen Nutzungen nicht.

Heißt das, jeder kann an die Daten ran?

Aus den Prinzipien der CI können auch Regeln für die Datenverarbeitungspraxis abgeleitet werden. Das sogenannte „Five Safes Framework" besteht aus einer Reihe von Grundsätzen, die es Datendiensten ermöglichen, einen sicheren Forschungszugang zu Daten anzubieten.



Siehe auch: Gerdon, Nissenbaum, Bach, Kreuter, Zins 2021. Harvard Data Science Review. https://doi.org/10.1162/99608f92.edf2fc97

29 Gerdon F, Nissenbaum H, Bach RL, Kreuter F, & Zins S. 2021. Individual Acceptance of Using Health Data for Private and Public Benefit: Changes During the COVID-19 Pandemic, Harvard Data Science Review : HDSR, 3(Spec. Iss. 1), 1–27, https://doi.org/10.1162/99608f92.edf2fc97

Das britische Office of National Statistics und einige andere Datenanbieter arbeiten seit den 2010er Jahren an diesem Rahmenwerk mit.[30] Die Five Safes haben sich als Best Practice für den Datenschutz etabliert und erfüllen gleichzeitig die Anforderungen an offene Wissenschaft und an Transparenz. Auch in Deutschland nutzen bereits einige Forschungsdatenzentren diese Prinzipien und erlauben Forschenden einen kontrollierten Zugang zu sensiblen oder vertraulichen Daten, so dass sie auf sichere und verantwortungsvolle Weise auf Datensätze zugreifen und diese nutzen können.

- **Sichere Daten:** Die Daten werden so behandelt, dass die Vertraulichkeit gewahrt bleibt.

- **Sichere Projekte:** Forschungsprojekte werden auf ihre Angemessenheit geprüft.

- **Sichere Personen:** Forscher:innen sind geschult und autorisiert, Daten sicher zu nutzen.

- **Sichere Einstellungen:** Eine sichere Umgebung verhindert die unbefugte Nutzung.

- **Sichere Ergebnisse:** Geprüfte und genehmigte Ergebnisse, die nicht vertraulich sind.

In den USA nutzt die „Administrative Data Research Facility" (ADRF) der Coleridge Initiative eine sichere Umgebung innerhalb der Amazon AWS GovCloud zum Hosten vertraulicher Daten. Sie wurde vom U.S. Census Bureau eingerichtet, um die Nutzung von administrativen und anderweitig sensitiven personenbezogenen Daten zu erleichtern. Die ADRF folgt ebenfalls dem Rahmenwerk der Five Safes zum Schutz von Daten und hat bereits über 100 vertrauliche Datensätze von Bundes-, Landes- und Kommunalbehörden sowie von akademisch Forschenden in die Cloud-Umgebung eingespeist. Datenanbieter können ihre Daten in dieser Umgebung bereitstellen und den Zugriff und die Nutzung mit einer speziell dafür entwickelten App steuern und verfolgen. Eine breiter aufgestellte Verfügbarkeit von (hochwertigen) Daten hat auch den Vorteil, dass sich Sampling- oder andere

30 Ein Beispiel der Umsetzung in Australien finden sich hier https://www.abs.gov.au/about/data-services/data-confidentiality-guide/five-safes-framework

Datenerhebungsstrategien, die systematisch Teile der Bevölkerung ausschließen (z. B. Kranke an der Teilnahme von Umfragen[31] oder Ältere bei der Teilnahme an Datenspenden[32]), weniger leicht auf die Modellentwicklung durchschlagen.

FAZIT

- KI hat ein hohes Potenzial zur Effizienzsteigerung und globalen Verbesserung von Entscheidungen. Unintendierte Konsequenzen sind denkbar und insbesondere bei naiver Anwendung durchaus wahrscheinlich.

- Metriken zur Beurteilung von Fairness sind wichtig; wichtiger ist jedoch, gesellschaftliche Klarheit darüber zu bekommen, welche Verteilungsgerechtigkeit angestrebt werden soll.

- Auch bei klaren Verteilungszielen kann variierende Datenqualität zu Fehlern bei dem der Verteilungsentscheidung zugrundeliegenden Vorhersagemodell führen.

- Fehlende Daten können oft durch die Verknüpfung verschiedenster Datenquellen ausgeglichen werden. Verlinkung bedarf häufig einer informierten Zustimmung.

- Informierter Zustimmung liegt in der Praxis oft nicht tatsächlich Informiertheit zugrunde. Experimentell lassen sich Zustimmungsraten leicht manipulieren. Zustimmungen sollten nicht als akkurate Abbildung der Einstellung der Betroffenen interpretiert werden.

- Contextual integrity (Kontext der Datenerhebung und der angemessene Fluss von Informationen) ist eine ernstzunehmende Alternative zur informierten Einwilligung und könnte Kernprinzip des Datenschutzes werden.

- Prinzipien des angemessenen Informationsflusses lassen sich praktisch in Konzepten sicherer Datennutzung (wie die Five Safes) implementieren.

31 Schnell R & Trappmann M. 2006. Konsequenzen der Panelmortalität im SOEP für Schätzungen der Lebenserwartung. https://www.uni-due.de/~hq0215/documents/schnell_tote_100306.pdf (27.3.2022).

32 Schnell R & Smid M. 2020. Methodological Problems and Solutions in Sampling for Epidemiological COVID-19 Research. Survey Research Methods, 14(2): 123–129.

AUTORINNEN UND AUTOREN

Barth, Rico (rico.barth@cape-it.de), Open Source Business Alliance, Chemnitz

Dössel, Olaf (olaf.doessel@kit.edu), Mitglied der Berlin-Brandenburgischen Akademie der Wissenschaften (BBAW) und Leiter des Instituts für Biomedizinische Technik, Karlsruhe Institut für Technologie (KIT)

Fröhlich, Holger (holger.froehlich@scai-fraunhofer.de), Universität Bonn, Leibniz-Institut für Präventionsforschung und Epidemiologie – (BIPS)

Ganten, Peter (ganten@osb-allianz.de), Vorsitzender Open Source Business Alliance – Bundesverband Digitale Souveränität e. V., CEO Univention GmbH

Haufe, Stefan (haufe@tu-berlin.de), Technische Universität Berlin

Intemann, Timm (intemann@leibniz-bips.de), Leibniz-Institut für Präventionsforschung und Epidemiologie – (BIPS), Bremen

Kern, Christoph (christoph.kern@stat.uni-muenchen.de), Ludwig-Maximilians-Universität München

Kreuter, Frauke (frauke.kreuter@stat.uni-muenchen.de), Ludwig-Maximilians-Universität München | University of Maryland

Lippert, Christoph (Christoph.Lippert@hpi.de), Hasso-Plattner-Institut, Potsdam

Pigeot, Iris (pigeot@leibniz-bips.de), Nationale Forschungsdateninfrastruktur für Gesundheitsdaten (NFDI4Health) und Leibniz-Institut für Präventionsforschung und Epidemiologie – (BIPS) und Nationale Forschungsdateninfrastruktur für Gesundheitsdaten (NFDI4Health)), Bremen

Prause, Guido (guido.prause@mevis.fraunhofer.de), Fraunhofer-Institut für Digitale Medizin MEVIS, Bremen

Rutert, Britta (rutert@bbaw.de), bis Juli 2022 wissenschaftliche Mitarbeiterin der Interdisziplinären Arbeitsgruppe „Zukunft der Medizin: Gesundheit für alle“ der BBAW

Schäffter, Tobias (tobias.schaeffter@ptb.de), Mitglied der BBAW und Leiter der Physikalisch-Technischen Bundesanstalt (PTB), Berlin

Schenk, Patrick Oliver (patrick.schenk@stat.uni-muenchen.de), Ludwig-Maximilians-Universität München

Schöck, Fabian (fabian.schoeck@siemens-healthineers.com), Siemens Healthineers, Erlangen

Schwabe, Daniel (daniel.schwabe@ptb.de), Physikalisch-Technische Bundesanstalt, Berlin

Strech, Daniel (daniel.strech@charite.de), Charité – Universitätsmedizin Berlin

Urban, Manuela (urban@osb-alliance.com), Open Source Business Alliance, Chemnitz

Wright, Marvin N. (wright@leibniz-bips.de), Leibniz-Institut für Präventionsforschung und Epidemiologie – (BIPS), Bremen