Gert G. Wagner

Metriken der Ungleichheit sind uralt

Die Feststellung von Ungleichheit bedarf keiner Künstlichen Intelligenz (KI). Wir alle werden seit jeher von unseren Mitmenschen anhand von leicht erkennbaren Merkmalen, etwa dem Geschlecht, „einsortiert" und zu Ungleichen gemacht. In den frühen staatlichen Kulturen wurde die Feststellung von Ungleichheit schon vor Jahrtausenden formalisiert, etwa für die Rekrutierung von Soldaten und für die Besteuerung. Man kann aus diesen Erfahrungen lernen, wie unser Umgang mit KI-induzierter Klassifizierung und Ungleichheit aussehen sollte. Klassifizierungen wurden seit jeher nur akzeptiert, wenn sie transparent erfolgten. In neuerer Zeit gehört zur Akzeptanz auch die Möglichkeit, gegen eine Klassifizierung zu klagen. Die „Einsortierung" von Menschen durch KI muss deshalb transparent sein. Ist die Transparenz gegeben, eröffnen sich nahezu automatisch auch Klagemöglichkeiten, die nötigenfalls per Gesetz erzwungen werden müssen.

I Die lange Geschichte der Arithmetisierung

Die Klassifizierung von Ungleichheiten findet je nach Lebensbereich oder wissenschaftlicher Disziplin in unterschiedlichen Begrifflichkeiten statt. Grundlegend ist eine *Metrik*, mit deren Hilfe Ungleichheit abgebildet wird. Dies kann in Form eines *Rankings* anhand bestimmter Kriterien und Kategorien, wie etwa dem Geschlecht, geschehen oder wie beim Alter in Form einer Arithmetisierung. Christoph Markschies verbindet (in diesem Band, S. 11–27) mit Arithmetisierung die Vorstellung, „dass Leben und Lebendigkeit zählbar und daher messbar sind." „Wir erleben den Menschen", so Markschies, „im Zeitalter seiner Zählbarkeit". Arithmetisierung steigere „die präzise Wahrnehmung von Ungleichheit" und lasse diese zu einer „allzeit als Zahl beschreibbare[n] und daher messbare[n] Größe" werden (S. 22–23). Markschies grenzt also die durch Zahlenwerte reprä-

Dieses Kapitel nimmt Überlegungen aus einem früheren Aufsatz von mir auf: „Scoring ist nicht neu, sondern uralt: Aus seiner Geschichte kann man lernen, wie man heutzutage damit umgehen kann und soll", in Harald Gapski, Stephan Packard (Hg.), „Super-Scoring? Datengetriebene Sozialtechnologien als neue Bildungsherausforderung", Düsseldorf 2021, S. 91–101. Viele Anregungen beruhen auf meiner Mitarbeit an einem Gutachten: SVRV, „Verbrauchergerechtes Scoring", Gutachten des Sachverständigenrats für Verbraucherfragen, Berlin 2018. Mein Dank für instruktiven Austausch gilt Johanes Gerberding, Gerd Gigerenzer, Christian Gross, Philipp Hacker, Ariane Keitel, Felix G. Rebitschek, Christin Schäfer und Sarah Sommer.

OpenAccess. © 2022 Gert G. Wagner, publiziert von De Gruyter. BY-NC-ND Dieses Werk ist lizenziert unter einer Creative Commons Namensnennung – Nicht kommerziell – Keine Bearbeitung 4.0 International Lizenz. https://doi.org/10.1515/9783110769975-006

sentierte Ungleichheit ausdrücklich ab von einer „vorwissenschaftliche[n] Wahrnehmung von Fremdheit (beispielsweise eines Ausländers oder eines anderen Geschlechts)“ (S. 23).

Eine auf Zahlen basierende (arithmetische) Metrik wird auch als Score bezeichnet. Der vielzitierte SCHUFA-Score beispielsweise ordnet Personen einen durch Algorithmen berechneten individuellen Zahlenwert zu, der Auskunft über Kreditwürdigkeit geben soll.[1] Scores werden oft auf Basis digitaler Daten berechnet. Im Falle des SCHUFA-Scores sind es gespeicherte Finanztransaktionen einer Person, die auf einer Skala von 0 bis 100 zu einem Wert verdichtet werden, der die Kreditwürdigkeit dieser Person angibt. Moderne Scores, die auf der elektronischen Verarbeitung von Daten beruhen, lassen nicht mehr erkennen, dass die Aufstellung von Scores anhand bestimmter Merkmale sowie die Arithmetisierung von Mitmenschen uralte Unterfangen sind.

Die historisch wahrscheinlich wirkmächtigste Arithmetisierung wird von vielen LeserInnen nicht als Metrik wahrgenommen: das *Geschlecht*. Es ist aber, wie moderne statistische Analysen zeigen, ohne weiteres quantifizierbar. Männern wird in der Regel für die Variable „Geschlecht“ der Wert null zugewiesen, Frauen der Wert eins. Was für Laien wie eine Bevorzugung von Frauen aussehen mag, ist in Wahrheit eine Diskriminierung durch StatistikerInnen: Männer bilden nämlich die „Referenzgruppe“ und die Nicht-Zugehörigkeit zu dieser Gruppe wird mit einem Score von 1 für Frauen kodiert.

Das *Geschlecht* ist aber nicht deswegen ein wichtiges Beispiel für Klassifikation durch Scoring, weil es quantifizierbar ist. Es ist von besonderem Interesse, weil die Zahlenwerte Null und Eins nicht nur anzeigen, wer Kinder gebären kann und wer nicht – was ja aussagekräftig und vernünftig sein kann, solange nicht biologisches und soziales Geschlecht verwechselt werden – sondern weil sie großen Einfluss darauf haben, wie eine Vielzahl von Lebenschancen völlig unabhängig von individuellen Fähigkeiten und Interessen verteilt werden. Die Metrik „Geschlecht“ wurde über Jahrtausende hinweg über die schlichte Anzeige von Gebärfähigkeit hinaus unzulässig auf viele Lebensbereiche übertragen: von der Papstwahl bis zum bürgerlichen Wahlrecht. Nur mühsam konnten und können die daraus resultierenden Diskriminierungstatbestände abgebaut werden.

Das Geschlecht ist überdies für den Umgang mit Arithmetisierung von besonderem Interesse, weil man seit einigen Jahren in etlichen Ländern sein amtlich festgestelltes Geschlecht ändern lassen kann. Man muss als ErwachseneR nicht mehr zwischen weiblich und männlich entscheiden, sondern kann als Geschlecht auch eine dritte Kategorie eintragen lassen. Das impliziert: der Klageweg steht

1 Vgl. das Gutachten des Sachverständigenrats für Verbraucherfragen, SVRV 2018, S. 14 f.

offen. Und dieser Weg wird sich in den folgenden Abschnitten als ein wichtiges Merkmal des Umgangs mit ungleichmachender Arithmetisierung in demokratischen Rechtsstaaten erweisen.

In Agrargesellschaften gab es eine weitere an der Demographie festgemachte Metrik, die meist auch mit dem männlichen Geschlecht verkoppelt wurde: die *Geschwisterposition*. Meist wurde Landbesitz an den erstgeborenen Sohn vererbt; Erbteilung war weniger verbreitet, und nur selten ging das Land an das jüngste Kind. Die Geschwisterposition, der man nicht entkommen kann, ist hier offensichtlich diskriminierend. Dass es für manch einen Bauernsohn oder manch eine Bauerntochter ein Glück war, den Hof nicht übernehmen zu müssen, zeigt im Übrigen die Ambivalenz pauschalisierender Klassifikationen.

Das *Lebensalter* ist eine Metrik der Ungleichheit und wird als solche genutzt, seitdem durch Geburtsurkunden das Alter einer Person nachprüfbar festgestellt werden kann. Kirchliche Initialisierungsrituale hängen ebenso vom Lebensalter ab wie die strafrechtliche Schuldfähigkeit. Seit es die Schulpflicht gibt, hängt auch der Beginn des Schulbesuchs vom Geburtsdatum ab. Heutzutage gibt es Möglichkeiten, aufgrund individueller Umstände von der Regel abzuweichen und den Schulbesuch früher oder später anzutreten, aber grundsätzlich gilt die Metrik des Alters und das damit verbundene Scoring. Mit dem Erreichen der „Regel-Altersgrenze“ können Arbeitsverträge automatisch auslaufen, da eine Altersrente als „Lohnersatz“ bezogen werden kann.

Der Score „Lebensalter“ ist transparent und heutzutage schwer manipulierbar. Ungleichbehandlung aufgrund von Altersunterschieden wird in vielen Bereichen als vernünftig angesehen. Gleichwohl kann es unerwünschte Effekte haben und darüber hinaus diskriminierend wirken. So ist in den USA seit Jahrzehnten die automatische Auflösung von Arbeitsverträgen zu einem bestimmten Lebensalter als Altersdiskriminierung verboten. Und das Alter, das man ja nicht verändern kann und gegen dessen Feststellung in der Regel nicht geklagt werden kann, führt in vielfacher Weise zu Benachteiligungen. So hängen zum Beispiel Sportkarrieren in Mannschaftssportarten auch davon ab, ob man zu den Jüngeren *innerhalb* eines Jahrgangs gehört (und dadurch leistungsschwächer ist) oder zu den Älteren, die dann von einer „Gnade der frühen Geburt“ profitieren.

Eine weitere offenkundige Arithmetisierung ist die *Körpergröße*. Sie bestimmte zum Beispiel darüber, wer im preußischen Militär zur Elitegruppe der „Langen Kerls“ gehören durfte. Eine Mindestgröße war erforderlich – während für AstronautInnen aufgrund der Enge der bislang üblichen Raumfahrzeuge eine Maximalgröße vorgegeben ist.

Eine berüchtigte Metrik sind *Schulnoten*. Sie entscheiden über Versetzungen, die Schulwahl und schließlich das Studium sowie – nach Gusto von Personalverantwortlichen – auch über berufliche Karrieren. Schulnoten werden im All-

gemeinen akzeptiert, insoweit sie einen transparenten und nicht manipulierbaren Score darstellen – auch wenn nicht jede einzelne Schulnote transparent zustande kommt. Man kann auch gegen das Zustandekommen und die Verwendung von Schulnoten klagen. Dies mag in der Realität sehr schwer und nur selten von Erfolg gekrönt sein, aber es ist möglich. In Analogie zu den Schulnoten werden seit etlichen Jahren in etlichen Staaten potentielle ImmigrantInnen anhand von Punkten bewertet, die ihr *Humankapital* und dessen Nützlichkeit für die Zuwanderungsgesellschaft quantifizieren (sollen). Begonnen hatte dies mit einem 0/1-Scoring bei der Einwanderung in die USA, wobei als krank klassifizierte Menschen (Score 1 statt 0) nicht einwandern durften, was etwa zwei Prozent traf.[2] Dies findet heutzutage seine Fortsetzung in der Klassifizierung von hilfesuchenden geflüchteten Menschen. Asylsuchende werden dabei in Asylberechtigte, Geduldete und Abzuschiebende eingeteilt.

Haben die bisher dargestellten Beispiele für teilweise (ur-)alte arithmetische Ungleichheitsklassifikationen viele LeserInnen wahrscheinlich überrascht, dürften die folgenden Beispiele geläufiger sein. Eine sehr alte, seit Jahrtausenden von Staaten genutzte Metrik ist die *Vermögenshöhe*. In antiken Gesellschaften entschied die Höhe des Vermögens nicht nur über die Höhe der Steuerzahlung (etwa im Römischen Reich), sondern auch über den militärischen Rang: Im Alten Griechenland etwa mussten Waffen und Pferde vom Soldaten selbst gestellt werden, wodurch nur Vermögende Offiziere werden konnten – ob sie dazu geistig in der Lage waren oder nicht. Hingegen spielt in Militärsystemen mit Wehrpflicht ein Gesundheits-Scoring in Form der Musterung eine große Rolle. Ein aktuelles Beispiel für dieses Scoring ist der US-amerikanische Ex-Präsident Donald Trump, der als junger Mann aufgrund einer ärztlich attestierten körperlichen Behinderung nicht eingezogen wurde.

Die Einstufung des Millionärssohns Donald Trump als nicht wehrtauglich ist ein gutes Beispiel für ein weiteres grundsätzliches Problem: Selektive Klassifikationen, die an komplexen mehrdimensionalen Phänomenen wie der Gesundheit ansetzen, sind manipulationsanfällig und schwerer objektivierbar als beispielsweise die leichter messbaren Größen „Alter“ und „Geschlecht“. Aber natürlich können auch leicht messbare Größen, etwa das Körpergewicht, manipuliert werden, wie das „Abkochen“ vor dem offiziellen Wiegen im Sport zeigt, das Sportlern ermöglicht, ihre Einordnung in eine Gewichtsklasse zum eigenen Vorteil zu beeinflussen. Bei Scores, in die mehrere Roh-Indikatoren als Variablen

2 Siehe „Ellis Island“, *Newyorkcity.de*, https://www.newyorkcity.de/ellis-island-in-new-york/, aufgerufen am 07.12.2021; „Ellis Island Chronology“, *National Park Service*, https://www.nps.gov/elis/learn/historyculture/ellis-island-chronology.htm, aufgerufen am 07.12.2021.

einfließen (beim SCHUFA-Score u. a. die Zahlungsunfähigkeit einer Person in der Vergangenheit und die Zahl ihrer Bankkonten), kommt hinzu, dass die verschiedenen Indikatoren im Verhältnis zueinander gewichtet werden müssen, um zum eigentlichen Score zu kommen, und dass in der Regel keine Gewichtung als alternativlos und objektiv richtig gelten kann.

Die Berechnung von Tarifen für Lebensversicherungen oder private Krankenversicherungen beruht seit jeher auf Klassifizierungen hinsichtlich gruppenspezifischer Lebenserwartungen und Krankheitskosten. Das Vorgehen der Versicherungen in diesen Bereichen macht zweierlei deutlich: (1) Wenn es sich nicht lohnt (etwa bei Versicherungssuchenden ohne nennenswerte Vorerkrankungen), wird auf eine Klassifikation aufgrund differenzierter Gesundheitsprüfungen verzichtet. (2) Es ist möglich, auf eine Klassifizierung, auch wenn sie durchaus aussagekräftig sein mag (etwa geschlechtsspezifische Lebenserwartungen), zu verzichten, wenn sie als unfair und diskriminierend bewertet wird. So hat der deutsche Gesetzgeber nach Geschlecht differenzierte Krankenversicherungstarife verboten und „Unisex-Tarife" erzwungen, ohne dass der Versicherungsmarkt deswegen zusammengebrochen wäre.

Früherkennung von Krankheiten beruht auf einem Scoring von bestimmten biologischen Krankheitsmarkern (etwa anhand von bildgebenden Verfahren und Blutwerten). Dies verdeutlicht, wie wichtig es ist, dass Scores aussagekräftig sind und dass sie nicht zu oft fälschlich Alarm schlagen. Dies ist bei kleinen und sehr kleinen Schadenswahrscheinlichkeiten, wie etwa der Gefahr, an bestimmten Krebsarten zu erkranken, jedoch häufig der Fall. Um Menschen eine rationale Entscheidung darüber zu ermöglichen, ob sie sich „scoren" lassen wollen, ist nicht nur Transparenz entscheidend, sondern auch eine für Laien verständliche Darstellung der Aussagekraft und der Wirkungen des Scorings. Dazu gehört ein Vergleich der möglichen Schäden durch Nebenwirkungen, die bei einem Verzicht auf Scoring (hier: Früherkennung) nicht auftreten würden, mit dem erhofften Nutzen, der ja keineswegs sicher eintritt.

In der Medizin erhofft man sich, dass künftig die vielen Variablen, die den verschiedenen Aspekten des Gesundheitszustandes einer Person – etwa durch bildgebende Verfahren oder Labortests und genetische Analysen – einen Zahlenwert zuordnen, mithilfe einer Vielzahl statistischer Methoden (*data integration tools*), zu einem aussagekräftigeren Bild zusammengefasst werden können als dies durch menschliche Diagnose der Fall ist.[3] Die in Deutschland diskutierte

3 Indhupriya Subramanian et al., „Multi-omics Data Integration, Interpretation, and Its Application", *Bioinformatics and Biology Insights*, 14, 2020, S. 1–24.

„Elektronische Patientenakte“ bietet eine reichhaltige Datenbasis für medizinisches Scoring mit diagnostischen und therapeutischen Anwendungen.

Neuartig ist das elektronische Scoring des persönlichen Fahrverhaltens durch „Telematik-Optionen“ für spezielle Tarife bei der Kfz-Versicherung (*Pay As You Drive*). Eine Klassifizierung von FahrzeughalterInnen nach ihrem Risikoprofil ist jedoch keineswegs neu. Seit jeher werden Kfz-Haftpflicht-Versicherungen danach differenziert, ob es sich um eineN ErstversicherteN handelt bzw. wie die Schadenshäufigkeit einer versicherten Person in der Vergangenheit aussah. Das ist nichts anderes als Scoring – und da es wahrscheinlich das Fahrverhalten tatsächlich positiv beeinflusst und transparent ist, wird es selbst von FahrzeughalterInnen mit hohen Prämien allgemein akzeptiert. Im Kommen sind ebenfalls automatisierte Scoring-Verfahren bei Personalbewertungen in Unternehmen und für die Personalauswahl (*People Analytics*). Die Bewertung von Menschen aufgrund ihrer (vermuteten) Leistungsfähigkeit ist in Schulen und Hochschulen seit dem 19. Jahrhundert üblich. Eine Ausdifferenzierung von Noten mit Hilfe von vielen Indikatoren für Lernwilligkeit, -fähigkeit und -erfolg in Form von personalisierten „Learning Analytics“ ist also nichts grundsätzlich Neues, sondern schlicht die Ausnutzung von digital verfügbaren Informationen. In spezialisierten Arbeitsmärkten spielen derartige Scores seit längerem eine Rolle, etwa im professionellen Fußball oder auf dem Arbeitsmarkt für HochschullehrerInnen, die anhand ihrer Publikationsleistung gerankt werden, gemessen etwa anhand des h-Indexes für die Zitationshäufigkeit ihrer Publikationen.

PartnerInnen-Vermittlungsagenturen arbeiten beim *Dating* seit jeher – völlig intransparent – mit Scores, die die wechselseitige „Passfähigkeit“ von Menschen beschreiben. Da die Scoring-Methoden der Vermittlungsagenturen jedoch Geschäftsgeheimnis bleiben, sind viele Menschen skeptisch. Dies gilt offenkundig weniger für moderne digitale PartnerInnen-Vermittlungsplattformen, aber auch deren Erfolgsbilanz ist überschaubar. Sie sind aber noch nicht vom Markt verschwunden und dürften deshalb nicht schlechter abschneiden als konventionelles *Dating* – mit oder ohne Einschaltung einer analogen Agentur. Die Intransparenz der Algorithmen und des Scorings moderner digitaler PartnerInnen-Vermittlungsplattformen wird – ebenso wie bei traditionellen HeiratsvermittlerInnen – gesellschaftlich als wenig problematisch wahrgenommen, denn niemand wird in modernen westlichen Kulturen gezwungen, PartnerInnen über derartige Plattformen oder traditionelle HeiratsvermittlerInnen zu finden.

II Bewertungen

Gleichheit und Ungleichheit sind keine naturwissenschaftlichen Gegebenheiten, sondern in der Regel unter pragmatischen Gesichtspunkten entstandene soziale Konstruktionen. Markschies weist in seinem Beitrag zu diesem Band zurecht darauf hin, dass auch unsere persönliche Identität ein Konstrukt ist; er schreibt:

> Ich stelle an mir selbst jeden Morgen Gleichheit fest, obwohl ich von meinem Ich am voraufgehenden Abend durch erhebliche Ungleichheit unterschieden bin. Ich weiß, obwohl ich anders bin als gestern, dass ich dieselbe Person bin. Aristoteles verbindet diese Erfahrung der Selbigkeit mit der Seele und macht die Seele für das basale Bewusstsein der Selbigkeit verantwortlich. (S. 25)

Es sei angefügt: Da wir mit Mikroben zusammenleben, die in unserem Körper angesiedelt sind und sich mit ihm austauschen, sehen wir nicht nur anders aus, sondern unser Körper hat sich tatsächlich über Nacht verändert.

Wenn wir als Gesellschaft empirischer Ungleichheit dennoch Gleichheit herstellen wollen, müssen wir dafür *geeignete* Regeln finden, etablieren und auch mit rechtsstaatlichen Klagemöglichkeiten versehen. Keineswegs sind in allen Bereichen perfekte und flächendeckende Regulierungen etwa in Form von Verboten notwendig. Zum Teil werden klassifizierbare Ungleichheiten gar nicht zur Geltung gebracht, weil es sich nicht „rechnet", wie Kaufleute gerne sagen, sie in diskriminierender Weise zu nutzen. Dies zeigt der Verzicht auf aufwendige Gesundheitsprüfungen durch Versicherungen bei Menschen, die nicht erkennbar bereits erkrankt sind.

Die gesellschaftliche Bewertung von Scores und arithmetischen Klassifizierungen hängt neben der statistischen Aussagekraft eines Scores entscheidend von der Transparenz der eingesetzten Verfahren ab – und davon, ob man einem Scoring zu vertretbaren Kosten ausweichen kann.[4] Für die Akzeptanz klassifizierender Systeme ist es darüber hinaus wichtig, dass eine Person einen Einfluss auf die Ausprägung des Merkmals hat, nach dem sie klassifiziert wird. Wenn Scoring eine Verhaltensänderung bewirken soll (etwa vorsichtigeres Autofahren aufgrund gestaffelter Versicherungsprämien), ist die Möglichkeit der Beeinflussung ohnehin notwendig. Kann man zur Verhaltensprognose herangezogene Merkmale nicht beeinflussen (wie das Alter) oder nur mit hohen Kosten und Opfern (wie das Geschlecht), führt ein Scoring leicht zu unerwünschter Diskriminierung. Probleme entstehen auch, wenn eine Klassifikation, die für einen Lebensbereich aussagekräftig ist, beispielsweise die Kreditwürdigkeit, auch über

4 Vgl. SVRV 2018, S. 140 ff.

einen ganz anderen Bereich entscheidet, etwa den Schulbesuch von Kindern, wie das im „Sozialen Scoring“ in China geschehen soll.

Die Bedeutung der Kriterien Relevanz und Beeinflussbarkeit für die Beurteilung von Scores lässt sich gut anhand von Algorithmen zur Bestimmung der Kreditwürdigkeit illustrieren. *Credit Scores* sind von großer praktischer Relevanz für die allermeisten Menschen und kaum jemand kann ihnen ausweichen. Sie sind aber auch durch das Verhalten des Einzelnen beeinflussbar, vorausgesetzt, dass ihr Zustandekommen transparent ist. Deswegen ist eine gesetzliche Regulierung äußerst sinnvoll. Systeme, die die Personalauswahl und -entwicklung steuern (*People Analytics*), sind aufgrund ihrer gesellschaftlichen Relevanz und ihrer Nicht-Ausweichbarkeit weitere Kandidaten für eine gesetzliche Regulierung. Ein Beispiel für geringe Relevanz sind hingegen – wie bereits angesprochen – Algorithmen, die Agenturen für PartnerInnenvermittlung benutzen: Man kann diesen Algorithmen – zumindest bislang – ohne großen Schaden ausweichen, und positiv beeinflussen kann man seinen Score leicht durch Falschangaben (etwa für Geschlecht und Alter) – selbst dann, wenn die Entstehung des Scores nicht wirklich transparent ist.

Wilfried Hinsch diskutiert in seinem Beitrag zu diesem Band (S. 67–87) einen in der Literatur zur Arithmetisierung und Scoring bislang wenig beachteten Punkt: In einer Welt mit unvollkommenen Informationen (und dies ist unsere Welt) kann man sich nur anhand unvollkommener Indikatoren, die den Zustand der Welt unvollkommen anzeigen, entscheiden. „Statistische Diskriminierung“ durch arithmetische Klassifikationen und Score-Berechnungen aufgrund der Zugehörigkeit zu statistischen Referenzgruppen ist in einem gewissen Maße unvermeidlich. Daher gilt es, Fairnesskriterien für die Akzeptabilität selektiver Entscheidungen auf der Grundlage von Scores zu entwickeln und darüber hinaus rechtliche Klagemöglichkeiten zu schaffen, um sich gegen konkrete Klassifizierung und Scoring wehren zu können.[5]

III Algorithmen prüfen und kontrollieren

Es wird oft behauptet, dass die Arithmetisierung und algorithmische Klassifikation von Menschen im Zeitalter von Big Data und Künstlicher Intelligenz etwas ganz Neues sei, weil diese Klassifikation undurchschaubar werde. Scoring-Algo-

5 Vgl. zur Anpassung des Rechtsystems an neuartige digitale Klassifizierungs- und Scoring-Systeme: Johannes Gerberding, Gert G. Wagner, „Gesetzliche Qualitätssicherung für ‚Predictive Analytics‘ durch digitale Algorithmen“, *Zeitschrift für Rechtspolitik*, 52/4, 2019, S. 116–119.

rithmen seien quasi-autonome, „selbstlernende" Wesen, die sich ohne menschliches Zutun verbessern und letztlich – so wird suggeriert – in unkontrollierter Weise selbst programmieren. So wurde in der *Neuen Zürcher Zeitung* in einem Beitrag des Literaturwissenschaftlers Manfred Schneider die Hypothese aufgestellt, die Undurchschaubarkeit der KI führe uns in die dunkle Zeit des Mittelalters zurück, da wir nicht mehr kontrollieren könnten, was uns steuert und worauf wir deswegen glaubend vertrauen müssten. Die „als großer Fortschritt gefeierte künstliche Intelligenz" ziehe, so Schneider, ganze Teile unserer Welt „ins Unsichtbare", und die von Algorithmen gesteuerte Hypermoderne werde jenem Mittelalter ähnlich, das man das „dunkle Zeitalter" nannte; die „Kirchenväter der künstlichen Intelligenz" sorgten dafür, dass kein Unglaube aufkomme.[6] Schneider schreibt weiter:

> Microsoft-Präsident Brad Smith predigt: „Eine Ethik der KI muss den Faktor Mensch in den Mittelpunkt stellen. Es muss verhindert werden, dass die KI anonym Entscheidungen über uns trifft, die aus einer ‚Blackbox' kommen und nicht überprüfbar sind." Und weiter heisst es: „Bei Microsoft arbeiten wir daran, der KI beizubringen, uns ihre Ergebnisse zu erklären." Das aber überfordert unseren Glauben. Wie sollen Algorithmen, die riesige Datenmengen nach Vorgabe der Programmierer in intelligenter Fortschreibung verarbeiten, zugleich ihre Ergebnisse erklären? Aufgrund der Komplexität der algorithmischen Systeme ist das einstweilen gar nicht möglich (Schneider a. a. O.).

Freilich: Die ohne jeden Beweis von Manfred Schneider hingeschriebene Behauptung, dass „einstweilen" Algorithmen ihre Ergebnisse nicht erklären könnten, ist falsch und einfach zu widerlegen.

Auch wenn ein Algorithmus auf „künstlicher Intelligenz", also schwer zu verstehenden Computerprogrammen und großen Datenbasen beruht, ist er deswegen noch keineswegs undurchschaubar. Um einen Algorithmus zu testen, muss man weder seinen logischen Bauplan (das Computerprogramm) noch seine Datenbasis kennen. Man muss lediglich genau hinschauen, zu welchen Ergebnissen er im praktischen Gebrauch führt. Dazu muss man wiederum nichts anderes tun, als den Algorithmus mit Beispieldaten zu füttern und festzustellen, welche Ergebnisse er bei dem gegebenen Input auswirft. Liefert der Algorithmus für einzelne Personen nachvollziehbare oder nicht nachvollziehbare Ergebnisse und werden durch diese Ergebnisse womöglich bestimmte Personengruppen in problematischer Weise bevorzugt oder benachteiligt?

---

6 Manfred Schneider, „KI schafft eine neue Dunkelheit in der hypermodernen Welt – wir kehren zurück in selbstverschuldete Unmündigkeit", *Neue Züricher Zeitung,* 02.08.2021, https://www.nzz.ch/meinung/kuenstliche-intelligenz-neue-dunkelheit-in-der-hypermodernen-welt-ld.1635147, aufgerufen am 07.12.2021.

Ein Clou dieses Vorgehens zum Testen eines Algorithmus ist, dass ein womöglich rechtlich geschütztes Geschäftsgeheimnis nicht verletzt wird. Denn man muss den Programmiercode eines Algorithmus gar nicht kennen oder verstehen, um den Algorithmus auf die beschriebene Weise zu überprüfen. Mit Produkten, die handfester sind als Computer-Algorithmen, verfahren wir so seit Jahrzehnten: Der Limonadenhersteller muss seine Geheimrezeptur nicht verraten, damit die Limonade getestet werden kann. Um festzustellen, ob die Limo schmeckt oder Bauchweh verursacht, trinkt man sie einfach. Und zwar nicht nur unter Laborbedingungen – die, wie bei den Dieselautos geschehen, manipuliert werden können – sondern unter Alltagsbedingungen.

Bislang ist allerdings keinE Algorithmen-EntwicklerIn verpflichtet, es der kritischen Öffentlichkeit leicht zu machen, eigene Produkte zu testen.[7] Das Befüllen der „Black Box" eines Algorithmus mit Testdaten kann deshalb ein mühsames Unterfangen sein. Anders lägen die Dinge, wenn es einen rechtlichen Anspruch auf Durchführung von Tests gäbe. Dann müssten die Algorithmen-EntwicklerInnen eine Schnittstelle vorsehen, über die Testdaten eingespeist werden können.[8] Schwierig wäre dies nicht – die Gesetzgebung muss das „nur" wollen und Richtlinien für Algorithmen durchsetzen, denen wir nicht oder nur schwer ausweichen können.

Hinzu kommt, dass die EntwicklerInnen KI-basierter Algorithmen inzwischen erkannt haben, dass es für die Akzeptanz von KI-Anwendungen nützlich ist, wenn sie erklärbar sind. *Explainable Artificial Intelligence* ist zu einem boomenden Forschungszweig geworden.[9] Viele Anwendungen algorithmischer Scorings, die geheimnisvoll aussehen, benötigen allerdings gar keine KI. Oft reichen einfache Heuristiken als Entscheidungsregeln aus, was wahrscheinlich beim SCHUFA-Score der Fall ist und ein Grund dafür sein könnte, warum seine simple Rezeptur nicht vollständig offengelegt wird.

7 Vgl. SVRV 2018, S. 132ff. Auch der Abschlussbericht der Enquete-Kommission „Künstliche Intelligenz" des Deutschen Bundestags ist in diesem Punkt wenig ergiebig und bleibt recht allgemein (Enquete-Kommission, *Künstliche Intelligenz – Gesellschaftliche Verantwortung und wirtschaftliche soziale und ökologische Potenziale*, 2020, S. 37ff.).

8 Vgl. Gerberding, Wagner, „Gesetzliche Qualitätssicherung", a. a. O.

9 Vgl. Wojciech Samek, Klaus-Robert Müller, „Towards Explainable Artificial Intelligence", in: Wojciech Samek, Grégoire Montavon, Andrea Vedaldi et al. (Hg.), *Explainable AI: Interpreting, Explaining Visualizing Deep Learning*, Cham 2019.

IV Schluss

Metriken der Ungleichheit und einfache arithmetische Klassifikationen (beispielsweise anhand des Geschlechts, Alters oder Vermögens) sind in der menschlichen Geschichte eine sehr alte Angelegenheit. Viele dieser Klassifizierungen wirken diskriminierend, weil Menschen nicht aufgrund individueller Eigenschaften und Fähigkeiten, sondern aufgrund von statistischen Gruppenmerkmalen klassifiziert werden.

Diskriminierung findet statt, wenn eine Klassifizierung nicht nur für Bewertungen in Lebensbereichen eingesetzt wird, für die sie erkennbar relevant ist, wie das Geschlecht etwa für das Gebären von Kindern, sondern auch dort, wo sie ohne Bedeutung sein sollte; das Geschlecht etwa bei der Verteilung wirtschaftlicher oder politischer Positionen. Diskriminierung wird umso wahrscheinlicher, je mehr eine Klassifizierung sachfremd angewendet wird, etwa wenn die SCHUFA-Bonität über den beruflichen Aufstieg entscheidet.

Notwendige, wenn auch nicht hinreichende Bedingungen für die gesellschaftliche Akzeptanz von arithmetischen Klassifizierungen sind deren Aussagekraft, etwa für Entscheidungen über den Schulbesuch von Kindern, und ihre Transparenz. Dadurch wird vielfach nicht nur ermöglicht, den eigenen Score im eigenen Interesse zu verändern, sondern vor allem auch die Möglichkeit eröffnet, auf dem Gerichtsweg gegen eine Klassifizierung bzw. einen Score vorzugehen.

Die Beispiele in diesem Beitrag zeigen zum Teil uralte nicht-digitale Klassifikationsschemata und „Scores". Der Umgang mit ihnen zeigt, dass wir Klassifizierungen nicht schutzlos ausgeliefert sind. Gesellschaftliche Regulierungen, die Unfairness und Diskriminierung verhindern, sind möglich, aber keineswegs selbstverständlich und einfach zu erreichen. Um den durch Big Data und KI auf die Spitze getriebenen Möglichkeiten der Klassifikation und der Ungleichheit nicht schutzlos ausgeliefert zu sein, sollten wir vorsichtshalber Iyad Rahwans (2018) Ratschlag folgen: „*Treat AI like a wild animal.*"[10] Markschies ist gleichwohl zuzustimmen, dass das empirisch vorfindliche spannungsreiche Verhältnis von Ungleichheitswahrnehmung und Gleichheitspostulaten dafür sorgen wird, dass es uns „wegen der durch die Arithmetisierung unserer Welt- und Menschenwahrnehmung gestiegenen Wahrnehmung von Ungleichheit nicht bange sein [sollte]" (in diesem Band, S. 26). Zurecht fügt er ebenfalls an, dass dieses spannungsreiche Verhältnis bewusst aufrechterhalten werden sollte: „In Zeiten der

10 Iyad Rahwan im Gespräch mit Sean O'Neill, in: *New Scientist*, Band 240, Nr. 3198, 2018, S. 42–43.

verschärften Wahrnehmung von Ungleichheiten müssen die Wahrnehmung von Gleichheit und die Theorien über Gleichheit gepflegt werden“ (S. 26).