



This book chapter was originally published by kopaed-Verlag as:
Wagner, G. G. (2021). **Scoring ist nicht neu, sondern ural: Aus seiner Geschichte kann man lernen, wie man heutzutage damit umgehen kann und soll.** In H. Gapski & S. Packard (Eds.), *Super-Scoring? Datengetriebene Sozialtechnologien als neue Bildungsherausforderung* (pp. 91–101). kopaed-Verlag.

Terms of use:



Provided by:

Max Planck Institute for Human Development
Library and Research Information
library@mpib-berlin.mpg.de

Harald Gapski / Stephan Packard (Hrsg.)

Super-Scoring?

Datengetriebene Sozialtechnologien
als neue Bildungsherausforderung

Düsseldorf – München
www.kopaed.de

Scoring ist nicht neu, sondern uralt:

Aus seiner Geschichte kann man lernen, wie man heutzutage damit umgehen kann und soll

Wir alle werden seit jeher von unseren Mitmenschen anhand von mess- bzw. leicht erkennbaren Merkmalen, etwa dem Geschlecht, ‚einsortiert‘. In frühen Staatswesen geschah dies schon vor Jahrtausenden, so für die Rekrutierung von Soldaten und für die Besteuerung. Und aus diesen Erfahrungen kann man auch lernen, wie die modernen Scores gestaltet werden sollten, damit sie in Demokratien gesellschaftlich nützlich sind und akzeptiert werden (können).

1 Einleitung

Als Scoring wird die von einem Algorithmus berechnete Zuordnung eines numerischen Wertes zu einem Menschen bezeichnet (SVRV 2018: 14f.). Heutzutage geschieht dies oft auf Basis digitaler Daten, etwa den von der SCHUFA aufgezeichneten Finanztransaktionen einer Person, die zu einem Wert verdichtet werden, der die Kreditwürdigkeit dieser Person angibt (auf einer Skala, die von null bis hundert reicht). Freilich verdecken moderne Scores, die auf der elektronischen Verarbeitung von Daten beruhen (vgl. dazu etwa Mau 2017) die Tatsache, dass das ‚Scoring‘ von Mitmenschen anhand bestimmter quantitativer bzw. kategorialer Merkmale ein uraltes Unterfangen ist. Wir alle werden seit jeher von unseren Mitmenschen anhand von mess- bzw. leicht erkennbaren Merkmalen, etwa dem Geschlecht, ‚einsortiert‘. Und in frühen Staatswesen geschah dies schon vor Jahrtausenden, so für die Rekrutierung von Soldaten und die Besteuerung (vgl. den Beitrag von Backer in diesem Band). Aus diesen Erfahrungen kann man lernen, wie die modernen Scores gestaltet werden sollten, damit sie in Demokratien gesellschaftlich nützlich sind und akzeptiert werden (können).

Im Folgenden werden die oben angerissenen Punkte näher ausgeführt (dabei wird auch auf die überschätzte Rolle von Künstlicher Intelligenz eingegangen)¹, um daraus Schlussfolgerungen für die Regulierung von Scoring in einer Demokratie abzuleiten.

2 Scores und ihre uralte Geschichte

Was ist Scoring? Im Gutachten des SVRV ist die Definition eng gezogen (SVRV 2018: 15):

Scoring ist die Zuordnung eines Zahlenwertes (des Scores) zu einem Menschen zum Zweck der Verhaltensprognose oder Verhaltenssteuerung. Die Bestimmung dieses Zahlenwertes erfolgt in der Regel auf der Grundlage einer breiten Datenbasis durch ein algorithmisches Verfahren („Computerprogramm“).

Aber: Eine Metrik, die einen Score operationalisiert, kann auch ganz einfach sein; und die Verwendung einfacher Metriken ist uralte. Bevor darauf in Abschnitt 2.2. eingegangen wird, seien vorab Beurteilungskriterien genannt.

2.1 Bewertungskriterien

Wie man Scoring bewertet, hängt neben der angemessenen Aussagekraft eines Scores entscheidend von der Transparenz der eingesetzten algorithmischen Verfahren ab und der Frage, ob man einem Scoring zu vertretbaren Kosten ausweichen kann oder nicht.² Für die Akzeptanz zentral ist weiterhin, ob man das Merkmal, nach dem gescort wird, persönlich beeinflussen kann oder nicht. Soll Verhalten durch einen Score beeinflusst werden, ist natürlich eine Beeinflussung notwendig (z. B. vorsichtigeres Autofahren, um eine hohe Versicherungsprämie zu vermeiden). Kann man zur Verhaltensprognose herangezogene Merkmale nicht beeinflussen, wie das Alter, oder nur sehr schwer (wie das Geschlecht), dann kann Scoring leicht zu unerwünschter Diskriminierung führen, etwa bei Lebensversicherungen (SVRV 2018: 34 ff.), oder weil ein Score, der für einen ganz bestimmten Lebensbereich aussagekräftig ist (z. B. für die Kreditwürdigkeit) auch für andere Lebensbereiche (z. B. den Schulbesuch von Kindern) herangezogen wird.

Die Bedeutung der Kriterien Relevanz und Beeinflussbarkeit sei hier kurz anhand von Algorithmen zur Bestimmung von Kreditwürdigkeit illustriert: Kredit-Scoring-Algorithmen sind beeinflussbar, aber nur dann, wenn sie transparent sind. Und sie sind offensichtlich für das Leben eines Jeden und einer Jeden relevant: Man kann Kredit-Scores faktisch nicht ausweichen. Deswegen ist eine gesetzliche Regulierung äußerst sinnvoll. Systeme, welche die Personalauswahl und -entwicklung steuern („People Analytics“, Hammermann und Thiele 2019), sind ob ihrer Relevanz bzw. faktischer Nicht-Ausweichbarkeit weitere Kandidat(inn)en für eine gesetzliche Re-

güligung. Ein Beispiel für geringe Relevanz sind hingegen Algorithmen, die hinter Partner(innen)vermittlung-Agenturen stehen: Man kann ihnen – 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.

2.2 Geschichte des Scorings³

Das Lebensalter ist offenbar eine Scoring-Metrik und sie wird in quantifizierter Form benutzt, seitdem durch Geburtsurkunden das Lebensalter nachprüfbar feststeht.⁴ Beispielsweise hängen kirchliche Initialisierungsrituale vom Lebensalter ab (etwa die Konfirmation), ebenso wie die strafrechtliche Schuldfähigkeit. Seit es die Schulpflicht gibt, hängt auch der Beginn des Schulbesuchs vom Geburtsdatum ab. In bürokratisch gut organisierten Gesellschaften, wie zum Beispiel in Preußen, gab es vor dem Schulbesuch schon vor Jahrhunderten kein Entrinnen. 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. Beim Übergang in den Ruhestand ist dies auch wieder der Fall: Mit dem Erreichen der ‚Regel-Altersgrenze‘ können Arbeitsverträge automatisch auslaufen, da eine Altersrente als ‚Lohnersatz‘ bezogen werden kann.

Der Score ‚Lebensalter‘ ist transparent und schwer manipulierbar und er wird in vielen Bereichen als vernünftig angesehen. Er kann aber auch unerwünschte Effekte haben, insbesondere kann er diskriminieren. So ist in den USA die automatische Auflösung von Arbeitsverträgen zu einem bestimmten Lebensalter inzwischen als Altersdiskriminierung verboten. Und das Alter, das man ja nicht verändern kann, diskriminiert vielfach; so hängen zum Beispiel Sportkarrieren in Mannschaftssportarten auch davon ab, ob man zu den Jüngeren seines Jahrgangs gehört (und dadurch im Kindes- und Jugendalter leistungsschwächer ist) oder zu den Älteren, denen die „Gnade der frühen Geburt“ zuteil wird (Gerhards und Mutz 2018: 316 ff.).

Der wahrscheinlich wirkmächtigste klassische Score wird hier erst an zweiter Stelle angesprochen, da viele Leser(innen) ihn nicht als quantifizierten Score erleben: das Geschlecht. Es ist aber ohne weiteres quantifizierbar, und das wird in der Statistik und bei Regressionsanalysen tatsächlich auch gemacht (und fließt auf diese Weise beispielsweise in den SCHUFA-Score ein): Beobachtungen von Männern wird in der Regel für die Variable ‚Geschlecht‘ der Wert null zugewiesen, Frauen der Wert eins. Was für Laien wie eine verblüffende Bevorzugung der Frauen aussehen mag, ist

allerdings in Wahrheit eine Diskriminierung durch Statistiker(innen): Männer bilden nämlich die sogenannte Referenzgruppe und die Abweichung von dieser Norm wird mit einem Score von 1 für Frauen kodiert.

Das Geschlecht ist aber nicht deswegen ein Beispiel für einen interessanten Score, weil es auch quantifizierbar ist. Es ist vielmehr interessant – und durchaus mit dem geplanten *Social Credit Score* in China vergleichbar – weil es nicht nur anzeigt, wer keine Kinder gebären kann (was ja aussagekräftig und vernünftig ist, wenn nicht biologisches und soziales Geschlecht verwechselt werden), sondern auch genutzt wird, um eine Vielzahl von Lebenschancen ‚zuzuteilen‘ – völlig unabhängig von individuellen Fähigkeiten und Interessen. Die Nennung des Stichworts ‚Frauendiskriminierung‘ muss hier aus Gründen des knappen Platzes ausreichen, um die Wirkmächtigkeit und Problematik des Scores „Geschlecht“ deutlich zu machen: er wurde vom schlichten Anzeigen der Gebärfähigkeit unzulässig auf viele Lebensbereiche übertragen: von der Papstwahl bis zum bürgerlichen Wahlrecht. Nur mühsam konnten und können diese diversen Diskriminierungstatbestände abgebaut werden.⁵

In Agrargesellschaften gab es einen weiteren an der Demographie festgemachten Score, der größtenteils auch mit dem männlichen Geschlecht verkoppelt wurde: die Geschwisterposition. Meist wurde Land an den erstgeborenen Sohn vererbt; Erbteilung war seltener, und in ganz wenigen Gegenden ging das Land an das jüngste Kind. Der Score Geschwisterposition ist offensichtlich diskriminierend, da man ihm nicht entkommen kann. 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 Scores.

Ein sehr altes Scoring ist das anhand des Vermögens durch den Staat. In antiken Gesellschaften entschied es nicht nur über die Höhe der Steuerzahlung, 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 (sie schließt auch geistige Behinderungen mit ein). Ein aktuelles Beispiel für dieses Scoring ist der 45. Präsident der USA Donald Trump, der aufgrund einer Behinderung nicht eingezogen wurde.

Die Einstufung des jungen Millionärssohns Donald Trump als nicht wehrtauglich ist auch ein schönes Beispiel für das grundsätzliche Problem, dass Scores, die an komplexen Tatsachen ansetzen (wie denen der Gesundheit), manipulationsanfällig sind, da sie schwerer messbar sind als Geschlecht (das freilich als 0,1-Variable

zunehmend umstritten ist, vgl. Endnote 5), Alter und Körpergröße. Bereits das Körpergewicht ist manipulationsanfällig, wie im Leistungssport das ‚Abkochen‘ vor dem offiziellen Wiegen zeigt, anhand dessen Gewichtsklassen zugewiesen werden.

Partner(innen)-Vermittlungsagenturen arbeiten seit jeher – völlig intransparent – mit Scores, die die Passfähigkeit von Menschen beschreiben. Da die Scoring-Methoden der Vermittlungsagenturen jedoch Geschäftsgeheimnisse darstellen, ist die Skepsis bei vielen Menschen ausgeprägt. Dies gilt freilich offenkundig weniger für moderne digitale Partner(innen)-Vermittlungsplattformen. Deren Erfolge sind zwar auch überschaubar (Bauer 2015), aber da sie noch nicht vom Markt verschwunden sind, können sie zumindest nicht schlechter als konventionelles Kennenlernen bzw. *dating* sein.

Haben die bisher dargestellten Beispiele für (ur)altes Scoring viele Leser(innen) wahrscheinlich überrascht, dürften die folgenden Beispiele geläufiger sein.

Die Tarifierung von Lebensversicherungen und privaten Krankenversicherungen basiert seit jeher auf Scores hinsichtlich gruppenspezifischer Lebenserwartungen und Krankheitskosten. Diese Versicherungen machen aber auch deutlich, dass es möglich ist, auf bestimmtes Scoring zu verzichten, wenn es als unfair bzw. diskriminierend eingeschätzt wird. Und diese Einschätzung kann sich im historischen Zeitverlauf ändern – Diskriminierung ist keine mathematisch-statistische Kategorie. Die Einschätzung dessen, was fair ist, hat sich in den letzten Jahrzehnten hinsichtlich geschlechtsspezifischer Tarife bei privaten Kranken- und Lebensversicherungen verändert, für die die deutsche Gesetzgebung inzwischen Unisex-Tarife erzwungen hat.

Neuartig – und deswegen neben den Kredit-Scores im SVRV-Gutachten untersucht – ist Verhaltens-Scoring für Bonusprogramme von Krankenkassen und -versicherungen und insbesondere unmittelbares elektronisches Scoring des Fahrverhaltens für spezielle Tarife bei der Kfz-Versicherung (z. B. *Pay as you drive*, Telematik-Optionen). Das Beispiel Fahrverhalten zeigt sehr schön, dass eine Klassifizierung von Fahrzeughalter(inne)n nach ihrem Risikoprofil keineswegs neu ist: Seit jeher werden Kfz-Haftpflicht-Versicherungen danach differenziert, ob es sich um eine(n) Erstversicherte(n) handelt und wie die Schadenshäufigkeit einer versicherten Person in der Vergangenheit aussah. Das ist nichts anderes als Scoring – und da es wahrscheinlich in der Tat das Fahrverhalten positiv beeinflusst und dieses Scoring transparent ist, wird es allgemein akzeptiert. Selbst von Fahrzeughalter(inne)n mit hohen Prämien.

Früherkennung von Krankheiten beruht auf einem Scoring von bestimmten biologischen Krankheitsmarkern (etwa anhand bildgebender Verfahren oder Blutwerten).

An diesem Beispiel wird deutlich, wie wichtig es ist, dass Scores wirklich aussagekräftig sind und vor allem nicht zu oft fälschlich Alarm schlagen, was bei kleinen und sehr kleinen Schadenswahrscheinlichkeiten (wie etwa der Gefahr, an bestimmten Krebsarten zu erkranken) häufig der Fall ist (Gigerenzer et al. 2017). Um Menschen rationale Entscheidungen darüber zu ermöglichen, ob sie sich scores lassen wollen und dadurch als Nebenwirkung Schäden in Kauf nehmen, die bei einem Verzicht auf Scoring (hier: Früherkennung) nicht auftreten können, ist nicht nur Transparenz entscheidend, sondern auch eine geeignete Darstellung der Aussagekraft und der Wirkungen des Scorings, sodass statistische Laien sie verstehen können. Dazu gehört ein Vergleich der möglichen Schäden durch Nebenwirkungen mit dem erhofften Nutzen, der ja keineswegs sicher eintritt. Und zu diesen Statistik-Laien zählen nicht nur Patient(inn)en, sondern leider bislang auch fast alle Ärzte und Ärztinnen (Gigerenzer et al. 2016, 2017). Ein Beispiel für eine aussagekräftige Aufbereitung des Wissens über Früherkennung in der Medizin sind „Faktenboxen“ (AOK Gesundheitskasse).⁶ Die in Deutschland in der Diskussion befindliche „elektronische Patientenakte“ kann offenkundig auch für Scoring als Basis für medizinische Diagnosen und Therapien herangezogen werden (Gigerenzer et al. 2016).⁷

Im Kommen begriffen sind automatisierte Scoring-Verfahren für die Bewertung von Personal hinsichtlich Beförderungen in Unternehmen bzw. für die Personalauswahl („People Analytics“, Hammermann und Thiele 2019). In spezialisierten Arbeitsmärkten spielen derartige Scores seit längerem eine Rolle, etwa im professionellen Fußball (Gerhards und Mutz 2018: 322 ff.) oder auf dem Arbeitsmarkt für Hochschullehrer(innen), die anhand ihrer Publikationsleistung gerankt werden (etwa anhand des h-Indexes für die Zitationshäufigkeit ihrer Publikationen). Die Bewertung von Menschen anhand ihrer (vermuteten) Leistungsfähigkeit ist zudem 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“ (Ebner 2019) ist also nichts grundsätzlich Neues, sondern schlicht die Ausnutzung von digital verfügbaren Informationen für das alte Instrument der Benotung. In etlichen Staaten werden potentielle Immigrant(inn)en anhand von Punkten bewertet, die ihr Humankapital und dessen Nützlichkeit für die Zuwanderungsgesellschaft quantifizieren (sollen).

In vielen weiteren Bereichen spielen Scores inzwischen eine Rolle (Mau 2017). Auf von Menschen vergebenen Scores, nämlich Punkten und Likes, beruhen Social-Media-Plattformen und z. B. die Zimmer-Vermittlung Airbnb sowie Kundenbewertungen von Teams und Einzelpersonen im Servicebereich, z. B. in Restaurants. Auf mathematischen Verfahren beruhen Scores, die etwa personalisierte Gutscheine in Supermärkten oder Microtargeting im Online-Handel steuern. Inwieweit Robo

Advisor für Geldanlagen auf komplexen Datenanalysen oder einfachen Entscheidungsregeln beruhen, ist im Allgemeinen nicht bekannt.

2.3 Ein Exkurs zum Black-Box-Scoring durch Künstliche Intelligenz

Da immer wieder behauptet wird, dass Scoring im Zeitalter von Big Data und künstlicher Intelligenz undurchschaubar würde, ja Scoring-Algorithmen „selbstlernend“ wären (vgl. etwa SVRV 2017: 21f.), also aus sich heraus besser werden und sich – so wird suggeriert – selbst programmieren, ist hier ein Exkurs über die Testfähigkeit von Scoring-Algorithmen angebracht (SVRV 2018: 26ff). Auch wenn ein Scoring-Algorithmus auf Künstlicher Intelligenz, also schwer zu verstehenden Computerprogrammen und großen Datenbasen beruht, ist er keineswegs undurchschaubar. Um einen Algorithmus zu testen, muss man nicht seinen Bauplan (Daten und Computerprogramm) kennen. Man muss nur genau hinschauen, wie er sich im alltäglichen Einsatz verhält. Dazu muss man nichts anderes tun, als den Algorithmus mit Beispieldaten zu füttern um dann nachzuschauen, wie er sich verhält: Liefert der Algorithmus Unsinn oder plausible Ergebnisse, und wie sehen die im Einzelnen aus? Durch Tests kann man auch herausfinden, ob der Algorithmus verbotenerweise diskriminiert, also bestimmte Personengruppen benachteiligt. Der Clou am Testen ist, dass das Geschäftsgeheimnis der Algorithmen-Entwickler(innen) nicht direkt beeinträchtigt wird. Man muss den benutzten Programmiercode nicht kennen oder gar verstehen. 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 unter Laborbedingungen – wie bei den Dieselaautos, wodurch geschummelt werden konnte – sondern unter Alltagsbedingungen.

Bislang muss allerdings kein(e) Algorithmen-Entwickler(in) es der kritischen Öffentlichkeit leicht machen, ihre Produkte zu testen (SVRV 2018: 132ff). Das Befüllen der Black Box eines Algorithmus mit Testdaten kann deshalb ein mühsames Unterfangen sein. Anders lägen die Dinge, sobald es einen rechtlichen Anspruch auf Durchführung von Tests gäbe. Dann müssten die Algorithmen-Entwickler(innen) eine Schnittstelle vorsehen, über die Testdaten eingespeist werden können (Gerberding und Wagner 2019). Da das Testen von Computer-Algorithmen bislang allerdings nur in Fachkreisen diskutiert wird, hat der Sachverständigenrat für Verbraucherfragen (SVRV) vorsichtshalber bei Experten der Gesellschaft für Informatik (Fachgruppe Rechtsinformatik 2018) eine Studie in Auftrag gegeben. Die Studie hat ausgelotet, ob ein Test machbar ist und wie der Algorithmentest in geltendes Recht eingefügt

werden könnte. Sie kam zu einem eindeutigen Ergebnis: Testen ist unter realistischen Bedingungen umsetzbar und unter den vielen gesetzlichen Herausforderungen, die sich aus den neuen vollautomatisierten Entscheidungsprozessen ergeben, stellt der Algorithmestest eines der kleineren Probleme dar. Deshalb haben die Experten konkrete Vorschläge gemacht, an welchen Stellen die Gesetzgebung tätig werden müsste, um einen Anspruch auf Algorithmestests zu verankern. Schwierig wäre dies nicht – die Gesetzgebung muss das nur *wollen* und Richtlinien für Algorithmen durchsetzen, denen wir nicht oder nur schwer ausweichen können.⁸

3 Bewertung

Was kann man aus dem in Abschnitt 2 gegebenen Überblick lernen? Scores sind in der menschlichen Geschichte eine sehr alte Angelegenheit (etwa das Lebensalter), die oftmals diskriminierend wirkten (insbesondere der 0,1-Score des Geschlechts). Diskriminierung findet immer dann statt, wenn ein Score (etwa das Geschlecht) nicht nur für Bewertungen eingesetzt wird, für die ein Score ursächlich ist (das Geschlecht für das Gebären von Kindern), sondern der Score auf andere Lebensbereiche übertragen wird (etwa, wenn das Geschlecht bestimmte Positionen begründet bzw. verhindert). Dieses Beispiel macht deutlich, dass die Definition von Diskriminierung nicht nur an gesellschaftlichen Konventionen hängt, sondern Diskriminierung umso wahrscheinlicher wird, je mehr ein Score sachfremd angewendet wird; so, wenn etwa die SCHUFA-Bonität über den beruflichen Aufstieg entscheidet und überdies diese Bonität nicht selbst verursacht ist, sondern von der Wohngegend abhängt.

Eine notwendige, wenn auch nicht hinreichende Bedingung für die gesellschaftliche Akzeptanz von Scores ist deren Aussagekraft bzw. Nützlichkeit (etwa für die Entscheidung, welche Schule ein(e) Schüler(in) besuchen sollte und kann), deren Transparenz (um sie zu beeinflussen, was bei Schulnoten gegeben ist) und in modernen demokratischen Gesellschaften die Möglichkeit, auf dem Gerichtsweg gegen einen Score vorzugehen (was z. B. bei Schulnoten sehr schwer ist, aber grundsätzlich möglich).

Der Super-Score in China soll anhand weniger Messungen (insbesondere der Bonität) für viele Lebensbereiche benutzt werden, also auch für solche, die kausal nichts mit dem Score zu tun haben (wie etwa traditionell das Geschlecht über Lebenschancen entscheidet, obwohl es nicht kausal mit diesen Bereichen verknüpft ist). Auch wenn der Super-Score transparent sein wird (damit die Menschen ihr Verhalten danach ausrichten können; dies ist ja das Ziel des Scores) und ggf. im Rahmen des chinesischen Rechtssystems dagegen geklagt werden kann, wird der Super-Score

gewissermaßen definitionsgemäß diskriminierend wirken, da er im Prinzip auf alle Lebensbereiche angewendet werden soll, ohne dass ein kausaler Zusammenhang zwischen allen diesen Lebensbereichen und dem Score bestehen wird.

Literaturangaben

- AOK Gesundheitskasse. AOK-Faktenboxen. <https://www.aok.de/pk/nordost/medizin-versorgung/aok-faktenboxen/> <Zugriff 16.02.2019>
- Bauer, Thomas (2015): Unstatistik des Monats: Liebestrunken – Vermittlungsbörse schießt statistisches Eigentor. In: Max-Planck-Institut für Bildungsforschung. <https://www.mpib-berlin.mpg.de/de/presse/2015/12/unstatistik-des-monats-liebestrunken-vermittlungsbörse-schießt-statistisches-eigentor> <Zugriff 27.11.2020>
- Datenethikkommission (2019): Gutachten der Datenethikkommission (Kurzfassung). https://www.bmjv.de/SharedDocs/Downloads/DE/Themen/Fokusthemen/Gutachten_DEK_Kurzfassung.pdf?__blob=publicationFile&v=2 <Zugriff 15.02.20>
- Duncan, Laramie; Shen, H.; Gelaye, B.; Meijsen, J.; Ressler, K.; Feldman, M.; Peterson, R.; Domingue, B. (2019): Analysis of polygenic risk score usage and performance in diverse human populations. In: Nature Communications, 10/3328. <https://doi.org/10.1038/s41467-019-11112-0> <Zugriff 15.02.20>
- Ebner, Martin (2019): Learning Analytics – eine Einführung. In: Bildung und Beruf. Februar 2019, S. 46-49.
- Gerberding, Johannes; Wagner, Gert G. (2019): Gesetzliche Qualitätssicherung für „Predictive Analytics“ durch digitale Algorithmen. In: Zeitschrift für Rechtspolitik, 52/4. S. 116-119.
- Gerhards, Jürgen; Mutz, Michael (2018): Die empirische Vermessung der schönsten Nebensache der Welt: Fußball und sozialwissenschaftliche Forschung. In: Innovation und Wissenstransfer in der empirischen Sozial- und Wirtschaftsforschung. Hrsg. v. Marcel Erlinghagen, Karsten Hank und Michaela Kreyenfeld. Frankfurt und New York. S. 315-335.
- Fachgruppe Rechtsinformatik der Gesellschaft für Informatik (2018): Technische und rechtliche Betrachtungen algorithmischer Entscheidungsverfahren. <https://www.svr-verbraucherfragen.de/dokumente/technische-und-rechtliche-betrachtungen-algorithmischer-entscheidungsverfahren/> <Zugriff 15.02.20>
- Gigerenzer, Gerd; Rebitschek, Felix G.; Wagner, Gert G. (2017): Kritische Voraussetzungen für ein digitales Gesundheitswesen in Deutschland. In: Wirtschaftsdienst, 97/10. S. 691-694.

- Gigerenzer, Gerd; Gerberding, Johannes; Groß, Christian; Keitel, Ariane ; Rebitschek, Felix G.; Sommer, Sarah; Wagner, Gert G. (2019): Verbraucher-Scoring – fair und korrekt? In: Konsumentenpolitisches Jahrbuch 2019, Schriftenreihe Verbraucherrecht, Verbraucherpolitik, Bd. 49. Hrsg. v. Maria Reiffenstein und Beate Blaschek. Wien. S. 77-111
- Gigerenzer, Gerd; Schlegel-Matthies, Kirsten; Wagner, Gert G. (2016): Digitale Welt und Gesundheit. eHealth und mHealth – Chancen und Risiken der Digitalisierung im Gesundheitsbereich. Veröffentlichung des Sachverständigenrats für Verbraucherfragen. <https://www.svr-verbraucherfragen.de/wp-content/uploads/Digitale-Welt-und-Gesundheit.pdf> <Zugriff 15.02.20>
- Hammermann, Andrea; Thiele, Christopher (2019): People Analytics: Evidenzbasierte Entscheidungsfindung im Personalmanagement. In: IW-Report 35/19. https://www.iwkoeln.de/fileadmin/user_upload/Studien/Report/PDF/2019/IW-Report_2019_People_Analytics.pdf <Zugriff 15.02.20>
- Herzog, Stefan M.; Hertwig, Ralph (2019): Kompetenzen mit „Boosts“ stärken – Verhaltenswissenschaftliche Erkenntnisse jenseits von „Nudging.“ In: Verbraucherbildung: Ein weiter Weg zum mündigen Verbraucher. Hrsg. v. Christian Bala et al. <https://www.verbraucherforschung.nrw/sites/default/files/2019-12/bzv10-02-Herzog-Hertwig-Kompetenzen-mit-Boosts-staerken.pdf>. S. 19-40 <Zugriff 15.02.20>
- Mau, Steffen (2017): Das metrische Wir. Berlin.
- Sachverständigenrat für Verbraucherfragen (2017): Digitale Souveränität. https://www.svr-verbraucherfragen.de/wp-content/uploads/Gutachten_Digitale_Souver%C3%A4nit%C3%A4t_.pdf <Zugriff 15.02.20>
- Sachverständigenrat für Verbraucherfragen (2018): Verbrauchergerechtes Scoring. https://www.svr-verbraucherfragen.de/wp-content/uploads/SVRV_Verbrauchergerechtes_Scoring.pdf <Zugriff 15.02.20>

Anmerkungen

- ¹ Viele Anregungen für den vorliegenden Beitrag beruhen auf der Mitarbeit des Autors an einem Gutachten zu verbrauchergerechtem Scoring des Sachverständigenrats für Verbraucherfragen (SVRV 2018); insbesondere die Empfehlungen des vorliegenden Aufsatzes gehen unmittelbar auf dieses Gutachten zurück (vgl. auch Gigerenzer et al. 2019). Besonderer Dank für die Arbeit am Gutachten und einiger nachfolgender Beiträge geht an Johannes Gerberding, Gerd Gigerenzer, Christian Gross, Ariane Keitel, Felix G. Rebitschek und Sarah Sommer. Für den vorliegenden Beitrag, der in Abschnitt 2 weit über das Scoring-Gutachten hinausgeht, ist nur der Autor dieses Beitrags verantwortlich.

- ² Vgl. SVRV (2018: 140 ff.), der schreibt: „Es bedarf nicht jedes Scoring-Verfahren in gleichem Umfang einer tiefgehenden Prüfung. Im Mittelpunkt stehen solche Scoring-Verfahren, die für das Wohl der Verbraucher(innen) erheblich sind. Besonders relevant sind sensible Lebensbereiche, etwa, wenn der einem Menschen zugeordnete Score direkte Auswirkungen darauf hat, ob dieser einen Kredit bekommt und zu welchen Bedingungen, wenn der Scorewert die Höhe der Versicherungsbeiträge beeinflusst oder das Gesundheitsverhalten gezielt steuert. Bislang wurden weniger sensible Bereiche (wie Betrugserkennung im Online-Handel) von den hoch relevanten nicht deutlich abgegrenzt. Eine solche Abstufung sollte jedoch vorgenommen werden, um die notwendige Regulierungstiefe und -dichte eines verbrauchergerechten Scorings angemessen zu bestimmen.“ Die Datenethikkommission (2019) schlägt deswegen bezüglich Algorithmen im Allgemeinen eine „Kritikalitätspyramide“ (2019: 177 ff.) vor, die unter der Überschrift „Risikoadaptierter Regulierungsansatz“ (ebd.:173) diskutiert wird.
- ³ Aus Platzgründen nicht betrachtet werden Metriken und Scores von Staaten, die zum Beispiel hinsichtlich der Erfolge ihrer Schulsysteme (Stichwort: PISA) oder ihrer Kreditwürdigkeit (d.h. dem Risiko, dass Staatsanleihen nicht bedient und zurückgezahlt werden können) gerankt werden. Sterne und Punkte für Unternehmen werden ebenso wenig betrachtet wie Hochschul-Rankings.
- ⁴ Auf ausführliche Literaturverweise wird in diesem Abschnitt meist verzichtet, da Informationen gegeben werden, die Allgemeingut sind und hier lediglich im Lichte von Scoring etwas aus dem Rahmen fallend interpretiert werden.
- ⁵ Hier fehlt der Platz, um näher darauf einzugehen, wie die klassischen Scores Alter und Geschlecht inzwischen in westlichen Gesellschaften behandelt werden. Das amtlich festgestellte Alter kann man nicht – auch nicht auf dem Klageweg – verändern; das Geschlecht lässt sich rein juristisch (und durch eine Geschlechtsanpassung) freilich verändern. Auch muss man in verschiedenen Ländern als Erwachsene(r) nicht mehr zwischen den Kategorien weiblich und männlich entscheiden, sondern kann als Geschlecht auch eine dritte Kategorie eintragen lassen.
- ⁶ Bessere Informationen (*boosting*) sind eine Alternative zu *nudges* (Stupser), die gewissermaßen hinter dem Rücken der Betroffenen Gutes in die Welt bringen wollen. Die AOK Gesundheitskasse stellt z. B. die Onlinedienstleistung der „Faktenboxen“ bereit (<https://www.aok.de/pk/nordost/medizin-versorgung/aok-faktenboxen/> <Zugriff 16.02.20>).
- ⁷ Nicht näher eingegangen werden soll hier auf genetische Marker, etwa in Form von „Polygenic Risk Scores“ (Duncan et al. 2019), die freilich bislang in der Praxis der Medizin noch kaum eine Rolle spielen.
- ⁸ Diesen Überlegungen ist 2019 auch die Datenethikkommission der Bundesregierung mit ihrer „Kritikalitätspyramide“ gefolgt (Abschnitt 3, „Algorithmische Systeme“, insbesondere Abbildung 2).