University of Groningen

Algorithmic Decision-Making, Ambiguity Tolerance and the Question of Meaning

Herzog, Lisa

Published in:
Deutsche Zeitschrift für Philosophie

DOI:
[10.1515/dzph-2021-0016](https://doi.org/10.1515/dzph-2021-0016)

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
2021

[Link to publication in University of Groningen/UMCG research database](#)

Citation for published version (APA):

Herzog, L. (2021). Algorithmic Decision-Making, Ambiguity Tolerance and the Question of Meaning. *Deutsche Zeitschrift für Philosophie*, 69(2), 197–213. <https://doi.org/10.1515/dzph-2021-0016>

Copyright

Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

The publication may also be distributed here under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license. More information can be found on the University of Groningen website: <https://www.rug.nl/library/open-access/self-archiving-pure/taverne-amendment>.

Take-down policy

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): <http://www.rug.nl/research/portal>. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.

Lisa Herzog*

Algorithmisches Entscheiden, Ambiguitätstoleranz und die Frage nach dem Sinn

<https://doi.org/10.1515/dzph-2021-0016>

Abstract: In more and more contexts, human decision-making is replaced by algorithmic decision-making. While promising to deliver efficient and objective decisions, algorithmic decision systems have specific weaknesses, some of which are particularly dangerous if data are collected and processed by profit-oriented companies. In this paper, I focus on two problems that are at the root of the logic of algorithmic decision-making: (1) (in)tolerance for ambiguity, and (2) instantiations of Campbell’s law, i. e. of indicators that are used for “social decision-making” being subject to “corruption pressures” and tending to “distort and corrupt” the underlying social processes. As a result, algorithmic decision-making can risk missing the point of the social practice in question. These problems are intertwined with problems of structural injustice; hence, if algorithms are to deliver on their promises of efficiency and objectivity, accountability and critical scrutiny are needed.

Keywords: algorithms, meaning, ambiguity, Campbell’s Law, structural justice

1 Einleitung

Algorithmische Systeme übernehmen in der heutigen Welt eine Vielzahl von Entscheidungen oder unterstützen sie. Das reicht von Empfehlungssystemen beim Online-Shopping über Entscheidungen zu Kreditvergaben bis hin zum Einsatz bei staatlichen Hoheitsaufgaben, z. B. bei Entscheidungen über Bewährungsstrafen. Wie treffen solche Systeme Entscheidungen; was ist hier anders, als wenn Menschen – in den genannten Fällen also Verkäufer, Bankangestellte oder Richterinnen und Vollzugsbeamte – Entscheidungen treffen? Wo sollte der Einsatz derartiger Systeme begrüßt werden, wo ist Vorsicht geboten?

Die Versprechen derjenigen, die auf Algorithmen setzen, klingen verlockend: Es gehe nicht nur um Zeit- und Kostenersparnisse, sondern auch um

*Kontakt: Lisa Herzog, Rijksuniversiteit Groningen, Faculty of Philosophy, Oude Boteringstraat 52, 9712 GL Groningen, Niederlande; l.m.herzog@rug.nl

größere Objektivität. Schon 1976 merkte Joseph Weizenbaum, einer der Pioniere der Künstliche-Intelligenz-Forschung, an, dass ein Computer, der Texte in Programmiersprache auswerte, nicht durch bloße Eloquenz verführt werden könne.¹ Angesichts zahlreicher Forschungsergebnisse zu den Verzerrungen („biases“) menschlicher Entscheidungen,² die z. B. zu rassistischer und sexistischer Diskriminierung führen können, wirkt die scheinbare Objektivität computerisierter Entscheidungen attraktiv. Menschliche Schwächen könnten damit – so die Hoffnung – nicht nur im Bereich der körperlichen Arbeit, sondern auch bei der Verarbeitung von Informationen und bei der Entscheidungsfindung durch Maschinen ausgeglichen werden.

Dass dabei jedoch zahlreiche Probleme auftreten können, zeigen nicht nur Berichte darüber, dass Algorithmen zur Einschätzung von Kriminalitäts-Rückfallwahrscheinlichkeiten Individuen mit weißer Hautfarbe mit einer für sie günstigeren Fehlerwahrscheinlichkeit belegten als solche mit schwarzer Hautfarbe.³ Es zeigte sich auch daran, dass Googles Suchalgorithmen eine Zeitlang bei Suchen nach Stichworten wie „black girls“ an oberster Stelle pornographische Inhalte präsentierten⁴ oder dass Stellenanzeigen für hochbezahlte Jobs öfters männlichen als weiblichen Nutzern präsentiert wurden.⁵ Darüber hinaus stellen sich zahlreiche Fragen nach der ethisch zulässigen Verwendung von Daten.⁶ Den Anstrich der technischen Objektivität und ethischen Unschuld haben Algorithmen längst verloren.⁷

In diesem Aufsatz sollen einige grundsätzliche Fragen zur Natur algorithmischen Entscheidens diskutiert werden, die den ethischen und politischen Fragestellungen an vielen Stellen vorausgehen: denn dem irreführenden Begriff von „künstlicher Intelligenz“⁸ zum Trotz funktionieren algorithmische Entscheidungen, die auf „Big Data“ beruhen, grundlegend anders als menschliche Intelligenz. Außerdem ist zu beachten, dass die Daten, auf deren Grundlage viele Entscheidungssysteme heute operieren, in Kontexten gewonnen werden, in denen andere Absichten – soweit man überhaupt von Absichten reden kann – zugrunde lagen als jene, die für die Entscheidungen, die dann mithilfe dieser Daten getroffen werden, angemessen wären.

¹ Weizenbaum (1976), 108.

² Vgl. z. B. Kahneman (2011); für eine kritische Diskussion vgl. jedoch Felin (2018).

³ Angwin et al. (2016).

⁴ Noble (2018), bes. Kap. 2.

⁵ Carpenter (2015); vgl. ähnlich auch Kim (2017), 864.

⁶ Zuboff (2019).

⁷ Für einen Überblick unterschiedlicher ethischer Herausforderungen vgl. z. B. Mittelstadt et al. (2016).

⁸ Für Kritik an diesem Begriff vgl. z. B. auch Kaplan (2017).

Im Folgenden werden zunächst grundlegende Mechanismen algorithmischer Entscheidungssysteme skizziert (2). Anschließend werden zwei miteinander verwandte Probleme diskutiert, die keineswegs ausschließlich bei algorithmischen Entscheidungen auftreten, dort jedoch besonders wahrscheinlich sind. Dies ist zum einen die Frage nach angemessenen Graden der Ambiguitätstoleranz, einem Begriff, den Thomas Bauer in einem kulturwissenschaftlichen Sinne gewendet hat, der jedoch auch für die Betrachtung algorithmischer Entscheidungen relevant ist (3). Zum anderen geht es um das Problem des Einsatzes von Indikatoren, die den Erfolg von Entscheidungssystemen bewerten können. Weil Algorithmen den *Sinn* von Tätigkeiten nicht bewerten können, drohen hier verschärfte Probleme der Form, die unter dem Begriff „Campbell’s Law“ bekannt sind: Verzerrungen sowohl der Indikatoren als auch der Praktiken, die durch sie erfasst werden sollen (4). Abschließend werden diese Überlegungen mit Fragen der strukturellen Gerechtigkeit zusammengeführt, denn eine der gegenwärtigen Gefahren ist, dass bereits benachteiligte Gruppen durch die Mechanismen algorithmischen Entscheidens weitere Nachteile erleiden (5).

Mit diesen Argumenten soll – um dieses mögliche Missverständnis gleich aus dem Weg zu räumen – keineswegs allen algorithmischen Entscheidungssystemen eine generelle Absage erteilt werden. Vielmehr geht es darum, deren Charakter, und insbesondere ihre derzeitige Verschränkung mit einem „Überwachungskapitalismus“ (Shoshana Zuboff) ganz eigener Art, kritisch zu beleuchten, um sowohl Stärken als auch Schwächen klarer benennen zu können. Daraus folgt die Notwendigkeit, im Umgang mit derartigen Systemen wachsam zu sein und insbesondere Raum für menschliches Feedback zu schaffen. Die Verwender von Algorithmen müssen, genau wie andere Entscheidungsträger auch, haftbar gemacht werden können, wenn es Fehlentscheidungen gibt. Um Fehlentscheidungen jedoch überhaupt also solche erkennen zu können, ist die Einbindung unterschiedlicher Gruppen nötig – vor allem dann, wenn es um den Kernbereich gerechtigkeits-theoretisch relevanter Güter geht. Nur dann ist zu erwarten, dass das Versprechen von besseren Entscheidungen, als sie menschliche Individuen treffen könnten, eingelöst werden kann.

2 Algorithmisches Entscheiden

Algorithmen sind schrittweise Prozeduren für das Lösen von Problemen – so die ganz allgemeine Definition, die sich z. B. im Duden oder Merriam-Webster findet. In diesem Sinne sind fast alle Regelwerke, z. B. Kochrezepte, Algorithmen: In ihnen wird bereits erworbenes Wissen in Anleitungen festgehalten, um es in

Zukunft wieder abrufen zu können. Relevant für heutige Fragestellungen sind freilich *computerbasierte* Algorithmen. Aufgrund ständig steigender Rechenleistung können Computer auch lange, mit zahlreichen Iterationen versehene Algorithmen innerhalb von Sekundenbruchteilen durchlaufen. Die Übersetzung in Computercode verlangt, dass alle Angaben präzise quantifiziert oder in 0/1-Entscheidungen transformiert werden. Derartige Algorithmen können dann eine ganze Reihe von Funktionen übernehmen, z. B. Daten priorisieren, klassifizieren, nach Ähnlichkeit sortieren oder filtern.⁹

Algorithmen können relativ einfach und damit gut für Menschen nachvollziehbar sein, entsprechend der Logik von Entscheidungsbäumen im Wenn-dann-Modus, bei denen man das Verhältnis von Input und Output klar sehen kann. Sie können aber auch extrem komplex sein und für Menschen, wenn überhaupt, nur mühsam nachvollziehbar sein. Dies ist besonders dann der Fall, wenn Algorithmen sich in den Bereich des „Machine Learning“ hineinbewegen. Hier wird der Weg der Entscheidungsfindung nicht in die Algorithmen einprogrammiert; stattdessen werden sie anhand großer Datensets „trainiert“, indem sie Lösungen – z. B. die Kategorisierung eines Bildes – vorschlagen und Feedback dazu erhalten, ob sie richtig liegen. Dieses „reinforcement learning“ ist es, das üblicherweise mit dem Begriff „künstliche Intelligenz“ gemeint ist.¹⁰ Die Geschwindigkeit, mit der Algorithmen solche Prozesse erlernen können, hängt dabei maßgeblich von der Menge und Qualität der Daten und des Feedbacks ab.

Wenn man sich fragt, wie derartige Entscheidungsprozesse mit menschlichem Handeln verglichen werden können, müssen ihre Kontexte und die in ihnen möglichen Formen des Feedbacks betrachtet werden. Alltägliche Formen menschlichen Handelns, z. B. das Kochen einer Mahlzeit, sind überschaubar in dem Sinne, dass ich als tätiger Mensch meine eigenen Ziele und Präferenzen kenne und während des Prozesses ständig in der Lage bin, Feedback aufzunehmen und entsprechend zu reagieren; ich kann z. B. durch Abschmecken feststellen, dass die Sauce noch mehr Salz benötigt. In vielen arbeitsteiligen Kontexten dagegen führen Menschen nur einzelne Schritte in langen Ketten durch, die erst im Gesamtergebnis Sinn ergeben. Feedback ist oft erst am Ende, wenn das Gesamtprodukt fertig ist, möglich – es sei denn, es wird bewusst auch schon für die Einzelschritte Feedback eingeholt, z. B. durch die Qualitätskontrolle einzelner Bauteile. Dies ist jedoch nur möglich, wenn vorher klar ist, wie die Teile zusammenspielen werden und worauf es bei ihrer Erstellung ankommt. Der Erfolg der fordistischen Massenfertigung bestand genau darin, dies durch Standardisierung

⁹ Vgl. z. B. Diakopoulos (2014).

¹⁰ Vgl. z. B. Lenzen (2018), 50–57.

und Kontrolle zu ermöglichen. Doch nicht alle Tätigkeiten lassen sich „fordisieren“, und wenn dies nicht der Fall ist, ist die Bewertung der einzelnen Schritte oft nur im Kontext des Gesamtergebnisses möglich.

Auch bei Algorithmen ist die Einbettung in komplexere Ketten von Arbeitsschritten oft entscheidend für die Betrachtung – vor allem, wenn es darum geht, nicht nur Prototypen im Labor, sondern den tatsächlichen Einsatz zu betrachten. In der Literatur über Algorithmen wird vielfach darauf hingewiesen, dass sie in ihren weiteren Kontexten gesehen werden müssen; Mike Ananny zum Beispiel spricht von „networked information algorithms“ und schreibt, dass Algorithmen „embedded within the sociotechnical structures“ seien: „[T]hey are shaped by communities of practice, embodied in standards, and most visible when they fail.“¹¹ Sowohl die *Inputs* – die Qualität der Daten, mit denen die Algorithmen arbeiten – als auch der Umgang von Entscheidungsträgern, typischerweise Menschen, mit den *Outputs* müssen dabei berücksichtigt werden.

Ein Beispiel für erstere Herausforderung liefert Virginia Eubanks in ihrer Diskussion des testweisen Einsatzes eines algorithmischen Matching-Systems, mit dessen Hilfe Sozialwohnungen Obdachlosen zugewiesen werden sollten. Wie die Sozialarbeiterinnen, die die Einführung des Testsystems begleiteten, feststellten, konnten die Algorithmen nur dann sinnvolle Ergebnisse erzielen, wenn die Daten, die über obdachlose Individuen eingegeben wurden, korrekt waren – doch um dies zu erreichen, musste erst ein Mitarbeiter eine Vertrauensbasis mit den Betroffenen herstellen, da diese sonst oft absichtlich oder unabsichtlich falsche Angaben machten und z. B. Krankheiten verheimlichten.¹² Beim Umgang mit dem Output von algorithmischen Programmen wiederum besteht die Gefahr, dass diese von Menschen für autoritativ gehalten werden, selbst dann, wenn explizit kommuniziert wird, dass ein System sich noch in der Testphase befindet.¹³

Eine besonders ungünstige Konstellation kann entstehen, wenn Algorithmen auf Daten aufsetzen, die nicht für den Zweck erhoben wurden, für den sie anschließend verwendet werden. Ein alltägliches Beispiel sind Kaufempfehlungen, die – soweit man dies nachvollziehen kann – die eigenen Interessen mit denen von anderen Menschen, für die man Geschenke gekauft hat, zu vermischen scheinen. Aber auch zahlreiche andere Möglichkeiten der „Verschmutzung“ von Daten können vorliegen; so können z. B. Verzerrungen entstehen, weil die

¹¹ Ananny (2015), 6.

¹² Eubanks (2017), Kap. 4.

¹³ Dieser Effekt wird in der psychologischen Forschung „automation bias“ genannt; vgl. Skitka et al. (2000).

Tippfehler-Raten bei unterschiedlichen Bevölkerungsgruppen unterschiedlich hoch liegen.¹⁴ Feedback-Systeme können systematische blinde Flecken haben; ein Beispiel dafür sind Algorithmen zur Bewertung und Auswahl von Kandidatinnen für berufliche Positionen, bei denen über die *abgelehnten* Bewerberinnen keine weiteren Daten erhoben werden können. Insofern Algorithmen auf Mustern aus der Vergangenheit aufbauen, muss außerdem gefragt werden, ob eine derartige Fortschreibung normativ wünschenswert ist, oder vielmehr eine ungerechte Ungleichbehandlung darstellt – dies trifft z. B. auf den Fall von Männern und Frauen unterschiedlich oft angezeigten Stellenausschreibungen zu.

Zur Realität der heutigen Datennutzung und Entwicklung algorithmischer Entscheidungssysteme gehört, dass zahlreiche Datensätze das Eigentum großer Firmen sind, deren Absicht ist, mit der Vorhersage individuellen Verhaltens Gewinne zu erzielen.¹⁵ Dies wirft eine ganze Reihe von Fragen auf: zum Verhältnis von Privateigentum und Gemeinwohl, zum Recht auf Nichtwissen oder Vergessenwerden und zu den Machtverhältnissen in der digitalen Wirtschaft. Für die Fragestellung dieses Aufsatzes ist besonders relevant, dass die Gewinnorientierung ganz eigene Formen der Verzerrung in den Feedbackschleifen, die für das Training der Algorithmen essentiell sind, erzeugen kann. Wie die Mathematikerin Cathy O'Neill kritisch anmerkt, wird dabei oft Profitabilität mit Wahrheit verwechselt.¹⁶ Dies kann sich z. B. darin niederschlagen, dass Produkte, die zum Spontankauf anregen, übermäßig oft in Empfehlungssystemen erscheinen, obwohl den Kunden weit besser gedient wäre, wenn ihnen Angebote gezeigt würden, die ihren längerfristigen Präferenzen entsprächen.

Die Beispiele dafür, was bei algorithmische Entscheidungssystemen schiefgehen kann, sind dementsprechend vielfältig und reichen von trivial und komisch bis hin zu hochgradig ungerecht und tragisch. Ein Beispiel in der ersten Kategorie war der Umgang von Facebook mit den sogenannten „Superhidern“. ¹⁷ Bei der Frage, welche Beiträge in den Feeds von Nutzerinnen erscheinen sollten, hatte der ursprüngliche Algorithmus die Löschung von Beiträgen als Indikator für geringe Relevanz verwendet. Eine genauere Untersuchung ergab jedoch, dass eine Vielzahl der Löschungen von einer kleinen Zahl an Nutzern, den „Superhidern“, ausging – diese behandelten ihren Newsfeed offenbar wie ein Emailpostfach, bei dem man gelesene Beiträge löscht, ohne dass damit eine Aussage über Qualität oder Relevanz verbunden wäre. Ein anderes Beispiel, diesmal aus dem Trainingsbetrieb

¹⁴ Kim (2017).

¹⁵ Vgl. insbesondere Zuboff (2019).

¹⁶ O'Neill (2016), 12.

¹⁷ Lischka/Stöcker (2017), 6.

von Algorithmen, stellt der Fall eines Mustererkennungsalgorithmus dar, der darauf trainiert worden war, Pferde zu erkennen, und dabei eine hohe Trefferquote erreichte. Wie die Softwareingenieure bei einer Rekonstruktion des Vorgehens des Algorithmus feststellten, fokussierte er dabei stets auf die untere Ecke der Bilder – und es stellte sich heraus, dass die Pferdebilder dort mit Copyright-Zeichen versehen waren, die auf den Vergleichsbildern mit anderen Tieren fehlten. Fälle aus dem wirklichen Leben, die durchaus dramatische Konsequenzen haben können, sind die regelmäßigen Meldungen darüber, dass in Suchmaschinen Falschmeldungen oder rassistische und sexistische Inhalte an oberster Stelle erscheinen.¹⁸

Algorithmen können also nur in einem sehr limitierten Sinn als „intelligent“ beschrieben werden. Sie reproduzieren oft bestehende Probleme ihrer Umgebung und ihre Wirkung innerhalb von arbeitsteiligen Prozessen hängt stark von der Qualität des Feedbacks ab, das sie erhalten. Um potentielle Probleme besser zu verstehen, verwende ich im Folgenden zwei Konzepte, das der „Ambiguitätstoleranz“ und „Campbell’s Law“.

3 Ambiguitätstoleranz

Eine erste Dimension algorithmischer Entscheidungssystemen, die vielen praktischen Problemen zugrunde liegt, wird sichtbar, wenn man das Konzept der „Ambiguitätstoleranz“ heranzieht – zugegebenermaßen eine Übertragung, was jedoch insofern gerechtfertigt scheint, als Thomas Bauer, der das Konzept in einem weiten kulturellen Sinne verwendet, es selbst aus der Psychologie übernommen hat.¹⁹ Bauers Frage ist, wie tolerant Gesellschaften im Umgang mit Unklarheit und Uneindeutigkeit sind. Ambiguität, so Bauer, könne „nie vollständig vermieden werden“²⁰; sowohl ein Zuviel als auch ein Zuwenig seien aber problematisch, da einerseits Beliebigkeit und Gleichgültigkeit, und andererseits engstirniger Fanatismus drohten. Die Kunst sei deshalb, das richtige Maß an Ambiguität, das dem menschlichen Leben entspreche, zu erreichen – und Bauer sieht die Gefahr heute vor allem in der Tendenz, Ambiguität zu stark zu reduzieren und damit echte Vielfalt unmöglich zu machen.²¹

Bauer beschäftigt sich nicht mit algorithmischen Entscheidungssystemen. Die Frage nach dem richtigen Maß an Ambiguität ist selbstverständlich eine viel

¹⁸ Vgl. z. B. Robb (2017).

¹⁹ Bauer (2018), 14.

²⁰ Ebd., 13.

²¹ Ebd., bes. Kap. 1 u. 10.

breitere als die nach seiner Umsetzung in algorithmische Zusammenhänge. Doch scheint die These nicht allzu gewagt, dass in einem sowieso schon relativ ambiguitätsintoleranten Umfeld die Einführung algorithmischer Entscheidungssysteme diese Tendenz noch verstärken kann. Das hat zum einen damit zu tun, wie Algorithmen technisch funktionieren, zum anderen aber auch damit, wie sie in eben diesem Umfeld zum Einsatz kommen.

Auf der technischen Ebene bestehen Algorithmen aus Computercode, der in einer 0/1-Logik funktioniert – dem vielleicht schärfstmöglichen Gegensatz zu Ambiguität. Einfach programmierte Algorithmen folgen oft einer ähnlichen Logik; man denke etwa an Online-Umfragen, bei denen man auf eine Frage mit den Antwortoptionen „ja“ oder „nein“ trifft und ein Verweigern der Festlegung nicht möglich ist, wenn man zur nächsten Frage weitergehen möchte. Man könnte einwenden, dass diese technische Ebene nicht entscheidend sei, da gerade komplexere Algorithmen auch mit skalierbaren Antworten oder Wahrscheinlichkeiten arbeiten können, wenn dies denn gewollt ist. Doch zum einen ist genau dies die Frage: *Lassen* sich die Unschärfen, die bei Entscheidungen vorliegen, sinnvoll in Skalen oder Wahrscheinlichkeiten übersetzen, und *wird* dies getan? Zum anderen muss auch, wenn dies stattfindet, eine Festlegung auf bestimmte *Dimensionen* der Skalierbarkeit oder Wahrscheinlichkeit erfolgen, die der Vielschichtigkeit lebensweltlicher Vorgänge oft nicht gerecht wird.

Ein Beispiel stellt schon die Notwendigkeit der Verschriftlichung von Kommunikation dar, die in der Zusammenarbeit mit Algorithmen vorliegt – auch dann, wenn mündliche Kommunikation per Spracherkennung ins Schriftliche übersetzt wird. All die Formen der Ambiguität, die mit der Körperlichkeit mündlicher Kommunikation zu tun haben, gehen dabei verloren: die Feinheiten der Tonhöhe und Modulation, die Botschaften, die Gestik und Mimik vermitteln, etc.²² Für manche Formen der Kommunikation mag dies – wenn allen Beteiligten klar ist, worum es sich handelt und was die technischen Parameter sind – unproblematisch sein, und die effiziente Verarbeitung von Informationen wird durch Algorithmen ermöglicht. Es ist jedoch fraglich, in welchen Zusammenhängen diese Art von Kommunikation angemessen ist, und wo die höhere Ambiguitätstoleranz zwischenmenschlicher Verständigung notwendig ist. Zwischen Hotlines für technische Geräte und dem Umgang mit benachteiligten Gruppen in Einrichtungen des Wohlfahrtsstaats besteht hier z. B. ein massiver Unterschied.

²² Ein in diesem Zusammenhang interessantes Phänomen ist die Vervielfältigung von Emojis für die schriftliche Kommunikation in Messenger-Programmen, die möglicherweise diese Limitation überwinden sollen.

Eine weitere Form der Ambiguitätsintoleranz findet sich bei in Algorithmen gegossenen Formen von Eigentumsrecht, insbesondere bei Formen des Copyrights. Um Eigentum klar definieren zu können, müssen die entsprechenden Güter vereindeutigt werden; schließlich würde der Begriff des Copyrights sinnlos, wenn z. B. von einem Text so viele Varianten existierten, wie dies in vielen Fällen mündlich tradiert Texten historisch der Fall war. Dieses Problem stellt sich als ganz konkrete Herausforderung, wenn es um die im Zusammenhang mit der Reform des europäischen Copyrights viel diskutierte „Upload-Filter“ geht, die aufgrund der schier unendlichen Menge an Material algorithmisch arbeiten müssen. Um geschützte Inhalte zu erkennen, müssen diese hinreichend eindeutig sein; zwar kann man gewisse Toleranzgrenzen zulassen, doch diese müssen quantitativ festgelegt werden. Zum Beispiel könnte eine algorithmisch implementierbare Regel lauten, dass Ausschnitte aus Musikstücken, die kürzer als eine bestimmte Anzahl von Sekunden sind, im Rahmen des „fair use“ verwendet werden dürfen.

Doch in vielen Fällen dürfte eine derartige Quantifizierung eine völlig unangemessene Form des Umgangs mit der Wiederverwendung von Texten oder Musikstücken darstellen. Entscheidend ist vielmehr, welcher *Sinn* in dieser Wiederverwendung liegt – hier unterscheiden sich z. B. die Wiedergaben zum Zweck einer positiven Würdigung, einer theoretischen Analyse oder einer politischen Satire. Die menschliche Urteilskraft in ihrer besten Form – nicht verzerrt durch Müdigkeit, irreführende Emotionen, Ablenkung oder andere Faktoren – kann hier fallweise entscheiden. Das Rechtssystem, das seiner Natur nach stark regelbasiert ist, vertraut auf die Instanz menschlich besetzter Gerichte, um die Anwendung der allgemeinen Regeln gerade auf atypische Fälle kritisch überprüfen zu können. Algorithmen dagegen können nicht in Kategorien des „Sinns“ entscheiden – wenn überhaupt, könnten sie nur mithilfe grober Indikatoren, z. B. dem Geräusch von Lachen als Indiz für Satire, dahingehend trainiert werden, bestimmte häufig auftretende Kategorien zu unterscheiden. Doch dies führt zum zweiten hier zu diskutierenden Problem: der Frage nach dem Sinn und den Gefahren, die von der Verwendung von Indikatoren ausgehen.

4 „Campbells Gesetz“ und die Frage nach dem Sinn

Wenn Menschen interagieren, kann zwischen ihnen das Phänomen geteilter Intentionalität auftreten, das in der philosophischen Sozialontologie breit diskutiert wird.²³ Menschen können ihre Aufmerksamkeit gemeinsam auf einen Gegenstand oder eine Aufgabe richten, im Wissen darum, dass die andere Person weiß, dass man selbst weiß, etc. Derartige Phänomene sind in der „Zusammenarbeit“ mit Maschinen nicht möglich;²⁴ ihnen müssen Intentionen vorher einprogrammiert werden und sie folgen diesen dann mehr oder weniger mechanistisch. Generell ist ein Denken in Kategorien von Absicht, Sinn oder Zweck für Algorithmen und algorithmisch gesteuerten Maschinen und Entscheidungssystemen nicht möglich. Wie Nick Bostrom und Eliezer Yudkowsky lakonisch anmerken, weiß ein Toaster – auch ein intelligent programmierter – nicht, dass seine *Absicht* ist, Brot zu toasten;²⁵ deswegen „toastet“ er auch, wenn man einen entzündlichen Gegenstand hineinsteckt, der dann Feuer fängt. Natürlich könnte man einen Toaster mit Sensoren ausstatten, die das eingeführte Material überprüfen, doch dies muss *vorher* einprogrammiert werden. Algorithmen scheitern oft an Herausforderungen, bei denen bislang unverbundene Wissensbestände integriert werden müssen – etwas, das Menschen oft intuitiv tun, weil sie die „Absicht“ oder den „Sinn“ dahinter verstehen.

Wie nun gelingt es, algorithmische Systeme, die den Sinn ihres Tuns nicht verstehen, dazu zu bewegen, bestimmte Aufgaben zu erfüllen? Man macht ihnen Vorgaben darüber, wann ein Lösungsweg als Erfolg gilt, und zwar durch die Verwendung von Indikatoren. Zum Beispiel gilt bei einem einfachen Matching-Programm von Kandidatinnen für eine Stellenausschreibung das Erreichen eines gewissen Punktwerts als Erfolg, in dem Sinne, dass die Person „gut genug“ passt, um für ein persönliches Vorstellungsgespräch eingeladen zu werden. Bei einem selbstlernenden System für Mustererkennung gilt eine bestimmte Trefferquote als ausreichend dafür, dass das System bestimmte Bilder „erkennt“.

Die Verwendung von Indikatoren begegnet uns ebenfalls in zahlreichen anderen, nicht-algorithmischen Praktiken, bei denen Schwellenwerte, Prozentsätze o. Ä. Verwendung finden. Doch aus der Forschung dazu ist auch hinreichend bekannt, welche Probleme dabei auftreten können. Griffig gefasst wird

²³ Für einen Überblick vgl. z. B. Schweikard/Schmid (2013).

²⁴ Vgl. z. B. Sloman/Fernbach (2017), 137–140.

²⁵ Bostrom/Yudkowsky (2014).

eine der zentralen Herausforderungen im sogenannten „Campbell’s Law“: „The more any quantitative social indicator is used for social decision-making, the more subject it will be to corruption pressures and the more apt it will be to distort and corrupt the social processes it is intended to monitor.“²⁶ Das klassische Beispiel, das Campbell selbst diskutierte, war das einer Polizeistation, deren Beamten nach dem Verhältnis von gelösten zu eingegangenen Fällen bewertet wurden, was einen Anreiz schaffte, möglichst wenige Fälle anzunehmen.²⁷ Allgemeiner gesprochen, kann es bei der Verwendung von Indikatoren, an die Anreize geknüpft sind, dazu kommen, dass diese selbst – und damit auch die Praktiken, die sie bemessen sollen – verzerrt werden, weil die beteiligten Akteure den Prozess durchschauen und Einfluss auf die in die Indikatoren eingehenden Werte nehmen.

Es gibt zahlreiche Beispiele für dieses Problem; viele stammen aus Bereichen, in denen die Praktiken, um die es geht bzw. gehen sollte, schwer quantitativ erfassbar sind, z. B. dem Gesundheitswesen oder Bildung, Erziehung und Forschung. Wie Colin Crouch am Beispiel des britischen *National Health Service* diskutiert, kann der Versuch der Steuerung durch Indikatoren so abstruse Blüten wie Zusatzhonorare für Demenzdiagnosen treiben.²⁸ Problematisch wird es vor allem dann, wenn die Orientierung an Indikatoren die Praxis insgesamt beherrscht,²⁹ was aufgrund von Personalmangel und Kostendruck oft kaum vermeidbar ist. Die Furcht vor Sanktionen kann dann auch dazu führen, dass Zahlen bewusst gefälscht werden, oft in der Hoffnung, dadurch Schutzbefohlene, seien es Patienten oder Schülerinnen, besser versorgen zu können.³⁰ In *The Tyranny of Metrics* schildert der amerikanische Historiker Jerry Z. Muller zahlreiche Beispiele dafür, wie mit Anreizen verknüpfte Indikatoren in Bereichen wie Schulen, Universitäten, Medizin, Polizei, Militär, Wirtschaft und sogar Entwicklungszusammenarbeit und Philanthropie zu dysfunktionalen Ergebnissen führen.³¹

Dieses Problem ist alt und wohlbekannt; schon 1996 haben Harvey Goldstein und Kate Myers deshalb einen „Ethischen Code“ für Performance-Indikatoren gefordert.³² Es steht jedoch zu befürchten, dass es sich durch die Einführung von

²⁶ Campbell (1976), 49. Auch der Ausdruck „Goodhart’s law“ findet sich gelegentlich; dieses lautet: „When a measure becomes a target, it ceases to be a good measure.“

²⁷ Muller (2018), 1–3 erwähnt das Beispiel der Fernsehserie *The Wire*, in der „joking the stats“ („die Statistiken verfälschen“) ein wiederkehrendes Thema ist.

²⁸ Crouch (2016), 1–2

²⁹ Ebd., bes. 73–76.

³⁰ Vgl. z. B. Aviv (2014). Für eine Diskussion vgl. Muller (2018), Kap. 1.

³¹ Ebd., bes. Kap. 7–13.

³² Goldstein/Myers (1996), zit. in Crouch (2016), 82.

algorithmischen Entscheidungssystemen massiv verschärfen könnte. Ein erster Grund für diese Befürchtung liegt darin, dass in algorithmischen Systemen zahlreiche Gegenkräfte, die andernfalls oft vorliegen, ausgeschaltet werden. Diese reichen von gesundem Menschenverstand über spontan empfundene Empathie mit anderen Individuen bis hin zu einem Ethos beruflicher Verantwortlichkeit, das bestimmte Verhaltensformen als unvereinbar mit der eigenen beruflichen Identität ablehnt. Bei menschlichen Akteuren kann die „economy of esteem“³³ unter Berufskollegen dazu beitragen, die Orientierung an den Werten der jeweiligen Profession hochzuhalten und somit ein Gegengewicht zu den Anreizen, sich rein an der Indikatorenoptimierung zu orientieren, schaffen.³⁴

Ein weiteres Problem, das die Problematik im Fall von algorithmischen oder algorithmisch unterstützten Systemen verschärfen kann, ist das Fehlen von qualitativ hochwertigem Feedback darüber, ob die gewählten Indikatoren die Praxis überhaupt sinnvoll widerspiegeln – oder die Schwierigkeit, Feedback an jene zurückzuspielen, die die algorithmischen Systeme entwerfen. Dies liegt unter anderem daran, dass es sich hier oft um weltweit agierende Firmen handelt, deren Programmierer außerdem mit der Materie, um die es geht, nicht unbedingt vertraut sind. Ein Beispiel ist die Nutzung von Qualitätsindikatoren in der Wissenschaft. In einem nationalen, öffentlichen Wissenschaftssystem sind Diskussionen über die Sinnhaftigkeit unterschiedlicher Indikatoren möglich und Impulse für Veränderungen können aus der Wissenschaftsgemeinde an die entsprechenden Organisationen zurückgespielt werden. Im Vergleich dazu scheint höchst fraglich, ob es gelingen könnte, Veränderungen bei der Erstellung von Indikatoren wie den weltweit verwendeten „science citation indices“ zu erreichen. Die dafür verwendeten Algorithmen sind nicht transparent und gehören privaten Firmen, so dass schon das Durchschauen der dahinterstehenden Logik schwierig ist. Da diese Systeme weltweit eingesetzt werden, gibt es außerdem ein verschärftes Problem kollektiven Handelns über Fach- und Ländergrenzen hinweg.

Eine zweite Schwierigkeit hängt mit dem oben diskutierten Problem der Ambiguitätsintoleranz zusammen. Algorithmen sind besser dafür geeignet, klar definierbare oder quantifizierbare Daten auszuwerten, z. B. im Zusammenhang mit Jobbewerbungen: Wie viele Monate Berufserfahrung hat ein Kandidat, hat er einen Universitätsabschluss, etc.? Wie O'Neill diskutiert, gibt es aber erwiesenermaßen auch viele schwieriger zu erfassende Faktoren, die für Erfolg im Beruf

³³ Brennan/Pettit (2005).

³⁴ Vgl. ähnlich auch Crouch (2018), 94. Zu „Professionen“ in einem moralisch aufgeladenen Sinn und ihrer Rolle in demokratischen Gesellschaften vgl. z. B. Dzur (2008).

entscheidend sind, z. B. soziale Fähigkeiten.³⁵ Es ist schwierig, dafür Proxys zu finden, die in die Software zur Bewerberinnenauswahl eingebaut werden könnten. Auch die Verwendung von Daten, die bei schon vorhandenen Mitarbeitern erhoben werden, ist nur bedingt hilfreich, weil diese ohne den Vergleich mit den Daten abgelehnter Kandidatinnen nicht aussagekräftig sind. Da zahlreiche Bewerberinnen außerdem nie erfahren, warum sie abgelehnt wurden, fehlt zudem die Möglichkeit, Feedback einzuholen und das System gegebenenfalls zu verbessern. Um einen Job zu finden, kommt es in einem von algorithmischen Systemen beherrschten Umfeld zunehmend weniger darauf an, was die eigentlichen Qualifikationen sind; es geht vielmehr um die Metaqualifikation „to learn what the machines are looking for“.³⁶

Bei selbstlernenden Programmen kommt hinzu, dass oft nicht einmal klar ist, welche Faktoren in die Auswertungen eingehen, und ob diese sinnvollerweise eingesetzt werden *sollten*. Ist z. B. die Tatsache, dass Leserinnen japanischer Mangas Datenauswertungen zufolge für bestimmte Aufgaben die „besseren“ Mitarbeiterinnen sind – eine Erkenntnis, die auf Daten beruhte, die in ganz anderen Kontexten erhoben wurden –, ein Grund dafür, dass sie bei der Bewerberauswahl einen Bonus haben sollten? Und wenn dies bekannt wird, ist dann nicht damit zu rechnen, dass nur noch Bewerbungen mit der Angabe der entsprechenden Interessen verschickt werden?³⁷

An diesem Beispiel zeigt sich eine grundsätzliche Spannung beim Einsatz derartiger Programme: Um zu verhindern, dass sie „ausgetrickst“ werden, sollten die Faktoren, die in sie eingehen, nicht bekannt werden; derartige Intransparenz steht aber oft im Widerspruch mit der Fairness und dem Respekt, die jenen geschuldet werden, die die Programme nutzen müssen.³⁸ Dies ist besonders dann der Fall, wenn es um staatliche Einrichtungen geht, die die Pflicht haben, Bürgerinnen gegenüber fair und transparent vorzugehen.

Schon 1976 warnte der KI-Pionier Weizenbaum davor, dass die Verwendung von Algorithmen dem Trunkenbold aus dem bekannten Witz gleichen könnte, der unter der Straßenlampe nach seinem verlorenen Schlüsselbund suche, weil es dort heller sei.³⁹ „We can count, but we are rapidly forgetting how to say what is worth counting and why“, so seine düstere Prophezeiung.⁴⁰ Betrachtet man

³⁵ O'Neill (2015), 132.

³⁶ Ebd., 115.

³⁷ Vgl. ähnlich Lenzen (2018), 180.

³⁸ Vgl. auch Eubanks (2017), 184–185.

³⁹ Weizenbaum (1976), 127–130.

⁴⁰ Ebd., 16.

die Macht- und Interessenskonflikte zwischen denen, die die algorithmischen Programme verwenden und jenen, die unter den Verzerrungen des Sinns von Praktiken leiden, dann ist nicht zu erwarten, dass für diese Probleme schnelle Lösungen gefunden werden; und, wie Eubanks anmerkt: Sind algorithmische Systeme einmal eingeführt, ist es höchst schwierig, sie wieder abzuschaffen.⁴¹ Dies bedeutet jedoch nicht, dass man in Passivität verfallen sollte – im Umgang mit algorithmischen Entscheidungssystemen sind durchaus bessere und schlechtere Lösungen möglich!

5 Schluss: Sinn und Gerechtigkeit

Die geschilderten Probleme, die durch Ambiguitätsintoleranz und Verzerrungen durch Indikatoren entstehen können, würden auch für algorithmische Entscheidungssysteme Herausforderungen darstellen, die in perfekt gerechten Gesellschaften eingeführt würden. Doch dies ist nicht die Situation, in der wir uns befinden. Zwei Faktoren sind besonders relevant: zum einen die schon erwähnte Rolle von Privatunternehmen mit den entsprechenden Eigentumsverhältnissen und Geschäftsgeheimnissen, zum anderen die Tatsache, dass unsere heutigen Gesellschaften von zahlreichen Formen struktureller Ungerechtigkeit durchzogen sind, die Positionen von Privilegien oder Benachteiligungen schaffen.⁴² Zur Phänomenologie derartiger struktureller Ungerechtigkeiten gehört auch, dass sich unterschiedliche Formen von Benachteiligung oft gegenseitig verstärken, wenn z. B. finanziell benachteiligte Gesellschaftsmitglieder geringere Chancen auf politische Ämter haben oder bei der Kreditaufnahme höhere Zinsen zahlen müssen.

Es ist zu erwarten, dass die geschilderten Probleme in besonders hohem Maß zuungunsten bereits benachteiligter Gruppen wirken. Bessergestellte Bewerber*innen, Patient*innen oder Käufer*innen können sich individuelle, menschliche Beratung leisten; sie haben das Selbstbewusstsein und das soziale Kapital, bei möglichen Fehlentscheidungen Widerspruch einzulegen. Sie sind wahrscheinlich auch oft besser in der Lage, schon aus Gründen der verfügbaren Zeit, sich mit derartigen Problemen auseinanderzusetzen und die dahinterliegen-

⁴¹ Eubanks (2017), 186. Ein in akademischen Kreisen bekanntes Beispiel ist die oben genannte Verwendung von automatisch generierten Zitationsindizes, über deren höchst begrenzte Aussagekraft weithin geklagt wird, die aber dennoch als Proxys in der Bewertung von Wissenschaftlerinnen weiterhin regelmäßig herangezogen werden.

⁴² Vgl. insbes. Young (2011).

den Mechanismen zu verstehen. Safiya Umoja Noble schildert das eindrückliche Beispiel einer schwarzen Frisörin in einem überwiegend weißen Ort, die sich durch die Einführung von *Yelp*-Bewertungen massiv benachteiligt sah. Sie hatte einen festen Kundenstamm in der schwarzen Community und war auf dessen besondere Bedürfnisse spezialisiert; gegenüber größeren „weißen“ Friseursalons erschien sie in den *Yelp*-Algorithmen jedoch quasi als unsichtbar, u. a., weil ihre Stammkundinnen – die sie um Bewertungen bat – von den algorithmischen Systemen in ihrer Relevanz herabgestuft wurden, weil sie nur für einen einzigen Friseursalon Bewertungen abgegeben hatten, die deswegen als möglicherweise gefälscht bewertet wurden. Die Frisörin schilderte außerdem den Druck, bei *Yelp* Werbung zu schalten, der sie gegenüber finanzstärkeren Salons oder Ketten benachteiligte.⁴³ Dieses Beispiel macht deutlich, wie Algorithmen, die auf den „Mainstream“ hin programmiert sind, Minderheiten benachteiligen können.

Könnte der Algorithmus den *Sinn* unterschiedlicher Friseursalons verstehen, dann wäre klar, dass hier unterschiedliche Kategorien vorliegen. Doch mangels eines derartigen Verständnisses wird der Maßstab der Mehrheit angelegt und unterstellt, dass ein bestimmter Salon *diesen* Maßstab *schlecht* erfülle. Auch in Bezug auf die Frage nach angemessener Ambiguitätstoleranz und dem Umgang mit Abweichungen von etablierten Kategorien steht zu befürchten, dass sich strukturelle Ungerechtigkeiten fortschreiben könnten, wenn etwa bei Bewerberinnen nach der Anzahl der eigenen Kinder, aber nicht der Anzahl anderer zu betreuender Angehöriger gefragt wird, weil der Algorithmus nicht versteht, dass der Sinn der Abfrage ist, die Arbeitsbelastung im privaten Umfeld bei der Evaluation angemessen zu berücksichtigen. Es ist keine technische Unmöglichkeit, auch derartige, fallspezifische Faktoren zu berücksichtigen. Doch dies verlangt ein entsprechendes Bewusstsein bei jenen, die Programme schreiben und verwenden – und die Bereitschaft, algorithmische Vorschläge auch an der einen oder anderen Stelle zurückzuweisen und sich von besseren Gründen, die dem Einzelfall gerecht werden, leiten zu lassen.

Die *positiven* Potentiale von algorithmischen Entscheidungssystemen, auch bei der Bekämpfung von struktureller Ungerechtigkeit, sollen mit diesen Überlegungen nicht geleugnet werden. Doch diese Potentiale können nur verwirklicht werden können, wenn mit den geschilderten Risiken bewusst umgegangen wird. Hierzu werden in der Literatur unterschiedliche Vorschläge diskutiert; sie reichen von der Forderung nach mehr Diversität und Perspektivenvielfalt unter Programmiererinnen über einen „Hippokratischen Eid“ für IT-Experten⁴⁴ bis hin

⁴³ Noble (2018), 173–200.

⁴⁴ O'Neill (2016), 205.

zu unterschiedlichen Formen von „Audits“ für algorithmische Entscheidungssysteme.⁴⁵ Manche Algorithmen, die jetzt noch intransparente „Black Boxes“ sind und möglicherweise diskriminierende oder dysfunktionale Ergebnisse liefern, können durch derartige Schritte möglicherweise im Lauf der Zeit so stark verbessert werden, dass ihr Einsatz unproblematisch würde und möglicherweise sogar menschlicher Entscheidungsfindung überlegen wäre.

In anderen Bereichen allerdings, insbesondere wenn hohe Ambiguität vorherrscht und der Kern der sozialen Praktiken mit Indikatoren nicht gut erfasst werden kann, sollte auch der Verzicht auf algorithmische Systeme in Betracht gezogen werden, und stattdessen in die Erforschung und das Austesten der bestmöglichen menschlichen Entscheidungsfindung, z. B. durch Deliberation, investiert werden. Dies sind – gerade, wenn es um staatliche Hoheitsaufgaben geht – genuin politische Aufgaben, in denen auch harte Machtkämpfe zu erwarten sind. Zu hoffen, dass technischer Fortschritt von alleine auch Fragen nach dem Sinn von Tätigkeiten und nach Gerechtigkeit beantworten könnte, wäre hier fehlgeleiteter Utopismus. Nur sorgfältig durch menschliche Vernunft überwachbare und korrigierbare algorithmische Systeme, in Bereichen, in denen algorithmisches Entscheiden nachgewiesenermaßen zu guten Ergebnissen führt, können das Versprechen der „künstlichen Intelligenz“ auf effizientere und vielleicht sogar gerechtere Entscheidungen einlösen.

Literatur

- Angwin, J., Larson, J., Mattu, S., u. Kirchner, J. (2016), Machine Bias. There's software used across the country to predict future criminals. And it's biased against blacks, in: *Pro Publica*, 23.5.2016.
- Ananny, M. (2015), Toward an Ethics of Algorithms: Convening, Observation, Probability, and Timeliness, in: *Science, Technology, & Human Values* 41.1, 1–25.
- Bauer, T. (2018), *Die Vereindeutigung der Welt. Über den Verlust an Mehrdeutigkeit und Vielfalt*, Ditzingen.
- Bostrom, N., u. Yudkowsky, E. (2014), The ethics of artificial intelligence, in: Frankish, K., u. Ramsey, W. M. (Hg.), *The Cambridge Handbook of Artificial Intelligence*, Cambridge, 316–334.
- Brennan, G., u. Pettit, P. (2005), *The Economy of Esteem. An Essay on Civil and Political Society*, New York.
- Campbell, D. T. (1976), Assessing the Impact of Planned Social Change, in: *Occasional Paper 8*, Public Affairs Center, Dartmouth College, Hanover, N. H.

⁴⁵ Ebd., 212; Campolo et al. (2017); Diakopoulos (2014).

- Campollo, A., Sanfilippo, M., u. Whittaker, M. (2017), *AI Now 2017 Report*, New York, URL: https://ainowinstitute.org/AI_Now_2017_Report.pdf (1.2.2021).
- Carpenter, J. (2015), Google's algorithm shows prestigious job ads to men, but not to women. Here's why that should worry you, in: *The Washington Post*, 6.7.2015.
- Crouch, C. (2016), *The Knowledge Corrupters. Hidden Consequences of the Financial Takeover of Public Life*, Cambridge.
- Diakopoulos, N. (2014), *Algorithmic Accountability Reporting: On the Investigation of Black Boxes*, URL: http://towcenter.org/wp-content/uploads/2014/02/78524_Tow-Center-Report-WEB-1.pdf (1.2.2021).
- Dzur, A. W. (2008), *Democratic Professionalism. Citizen Participation and the Reconstruction of Professional Ethics, Identity, and Practice*, University Park, Ill.
- Eubanks, V. (2017), *Automating Inequality. How High-Tech Tools Profile, Police, and Punish the Poor*, New York.
- Felin, T. (2018), The Fallacy of Obviousness, in: *Aeon Magazine*, 5.7.2018.
- Goldstein, H., u. Myers, K. (1996). Freedom of information: towards a code of ethics for performance indicators, in: *Research Intelligence* 57, 12–16.
- Kahneman, D. (2011), *Thinking, Fast and Slow*, London.
- Kaplan, J. (2017), AI's PR Problem, in: *MIT Technology Review*, 3.3.2017.
- Kim, P. T. (2017), Data-Driven Discrimination at Work, in: *William & Mary Law Review* 58.3, 857–936.
- Lenzen, M. (2018), *Künstliche Intelligenz. Was sie kann & was uns erwartet*, München.
- Lischka, K., u. Stöcker, C. (2017), *Digitale Öffentlichkeit. Wie algorithmische Prozesse den gesellschaftlichen Diskurs beeinflussen*, URL: <https://doi.org/10.11586/2017028> (1.2.2021).
- Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., u. Floridi, L. (2016), The ethics of algorithms: Mapping the debate, in: *Big Data & Society* 3.2, 1–21.
- Muller, J. Z. (2018), *The Tyranny of Metrics*, Princeton, N. J.
- Noble, S. (2018), *Algorithms of Oppression. How Search Engines Reinforce Racism*, New York.
- O'Neill, C. (2015), *Weapons of Math Destruction. How Big Data Increases Inequality and Threatens Democracy*, New York.
- Robb, A. (2017), Anatomy of a Fake News Scandal, in: *Rolling Stone*, 16.11.2017.
- Schweikard, D. P., u. Schmid, H. B. (2013), Collective Intentionality, in: Zalta, N. E. (Hg.), *The Stanford Encyclopedia of Philosophy* (Summer 2013 Edition), URL: <https://plato.stanford.edu/archives/sum2013/entries/collective-intentionality/> (1.2.2021).
- Skitka, L. J., et al. (2000), Automation Bias and Errors: Are Crews Better Than Individuals?, in: *International Journal of Aviation Psychology* 10.1, 85–97.
- Sloman, S., u. Fernbach, P. (2017), *The Knowledge Illusion: Why We Never Think Alone*, New York.
- Weizenbaum, J. (1976), *Computer Power and Human Reason*, San Francisco, Calif.
- Young, I. M. (2011), *Responsibility for Justice*, Oxford.
- Zuboff, S. (2019), *Das Zeitalter des Überwachungskapitalismus*, Frankfurt u. New York.