Wann ist Künstliche Intelligenz (un-)fair?
Ein sozialwissenschaftliches Konzept von KI-Fairness

Frank Marcinkowski und Christopher Starke

1. Einleitung

Systeme basierend auf künstlicher Intelligenz bestimmen schon heute den Alltag vieler Menschen, sei es durch Kaufempfehlungen im Netz, die Zusammenstellung von medialen Unterhaltungsangeboten oder smarte Autokorrektursysteme im Handy. Doch auch gesellschaftlich folgenreiche Entscheidungen wie die Verteilung von öffentlichen Gütern und Leistungen oder die Genehmigung von Asylanträgen können künftig durch algorithmische Entscheidungssysteme unterstützt werden. Durch diese tiefgreifende Veränderung nahezu aller gesellschaftlichen Teilbereiche wird Künstliche Intelligenz (KI) auch zu einem Handlungsfeld von Politik. Das umfasst nicht nur die Förderung und Regulierung neuer Technologien, sondern auch die Auseinandersetzung mit diffusen Ängsten in der Bevölkerung vor potentiell negativen Auswirkungen auf Mensch und Gesellschaft. Darüber hinaus verändert sich möglicherweise auch die Politik selbst, wenn *Big Data Analytics* zur Grundlage von politischen Entscheidungsprozessen werden (Poel et al. 2018; Rieder/Simon 2016). KI-basierte Systeme erweitern schon heute digitale Anwendungen im Bereich von E-Government, wie beispielsweise die Implementierung von Distributed-Ledger-Technologien (z.B. Blockchain) in öffentlichen Verwaltungen (Kossow 2019). Wenngleich geltende rechtliche Bestimmungen den Datenzugriff regulieren, hat sich die Verfügbarkeit großer Datenmengen nicht zuletzt durch die Implementierung von Open Government Data in den vergangenen Jahren erheblich erweitert (Kersting 2017). Die Einsatzbereiche von KI-Systemen für Politik sind dabei mannigfaltig und reichen von Beratungs- und Empfehlungssystemen hin zu komplexen Entscheidungssystemen. Insbesondere im letztgenannten Fall, wenn also folgenreiche Entscheidungen über die Verteilung von Gütern und Dienstleistungen, aber auch von Gefahren und Risiken, unter Rückgriff auf KI-basierte Systeme getroffen werden, drängt sich zwangsläufig die Frage nach deren Potential für Unfairness und Diskriminierung auf (Binns 2018a, 2018b). Woran liegt es und wie kann man verhindern, dass Algorithmen bestimmte Bürger auf der Basis ihres Geschlechts, ihrer Herkunft oder ihrer Religion systematisch bevorzugen oder be-

nachteiligen, ganz gleich ob bei der Zuteilung staatlicher Transferleistungen oder vor Gericht? Begriffe wie »Fair Machine Learning«, »Data Justice« oder »Discrimination-aware Data-Mining« verweisen auf Bemühungen von Forschern aus unterschiedlichen Disziplinen, Antworten auf diese und ähnliche Fragen zu finden (Barocas/Selbst 2016; Binns 2018a; Dencik et al. 2019; Hoffmann, 2019; Taylor 2017; Veale/Binns 2017). Während der Fokus dabei zunächst auf technischen, rechtlichen und ethischen Fragestellungen liegt, rücken zunehmend auch sozialwissenschaftliche Aspekte der Thematik in den Blick (Binns et al. 2018; Grgic-Hlaca et al. 2018a, 2018b; Lee 2018; Lee/Baykal 2017). Denn technische Funktionsweisen und rechtliche Rahmenbedingungen sind nur die eine Seite der Medaille. Ebenso bedeutsam für die weitere technische Entwicklung und das durch sie induzierte Konfliktpotential erscheinen die kognitiven, emotionalen und verhaltensbezogenen Reaktionen der Endnutzer auf KI-basierte Systeme mit künstlicher Intelligenz. Deren Perzeptionen und Einschätzungen werden annahmegemäß durch eine Mehrzahl von Faktoren beeinflusst und sind nicht allein durch die technische Funktionalität determiniert. Anders formuliert: Wenn ein Algorithmus im Aggregat diskriminierungsfrei klassifiziert (faktische Fairness), heißt das nicht automatisch, dass er auf individueller Ebene tatsächlich als fair wahrgenommen wird (wahrgenommene Fairness). Die Wahrnehmungen der Bürger sind aber ein wichtiger Bezugspunkt beim Einsatz von KI in der Politik wie in anderen gesellschaftlichen Teilbereichen, nicht zuletzt im Hinblick auf die Legitimität der Entscheidungsverfahren und Entscheidungsergebnisse: Beide sind zentral für die Stabilität der demokratischen Ordnung in der digitalen Gesellschaft (Verba 2006). Im Folgenden stellen wir ein mehrdimensionales Konzept von KI-Fairness vor, das dazu dienen soll, die Reaktionen von Nutzern und Betroffenen auf den Einsatz KI-basierter Technologien in Staat und Gesellschaft systematisch erfassen und beschreiben zu können. Wesentlicher Bezugspunkt unserer Überlegungen ist die Literatur zur *Organizational Justice,* die zugleich vielfältige Überschneidungen mit demokratietheoretischen Ansätzen aufweist, beispielsweise mit den Evaluationskriterien politischer Partizipation (Kersting 2019, in diesem Band) oder mehrdimensionalen Konzepten politischer Legitimation (Scharpf 2009; Schmidt 2013). Der Beitrag lenkt den Blick über die bisher dominante Beschäftigung mit distributiver Fairness hinaus auf weitere Eigenschaften sozio-technischer Systeme, an denen das Fairnessempfinden der Nutzer Anstoß nehmen könnte.

2. Was ist künstliche Intelligenz?

Künstliche Intelligenz ist ein Sammelbegriff für technische Systeme mit Fähigkeiten, die bislang der »natürlichen« Intelligenz von Menschen vorbehalten waren. Dabei ist Intelligenz ein vielschichtiges Phänomen, das trotz des Fehlens ei-

ner allgemein verbindlichen Definition stets eine kognitive Leistungsfähigkeit beschreibt und sich unter anderem in mathematischem, sprachlichem oder räumlichem Denken, Merkfähigkeiten und Auffassungsgabe manifestiert (Kaplan 2016; Legg/Hutter 2007). Unabhängig von konkreten Anwendungsgebieten werden Maschinen (vor allem Computersysteme) als intelligent bezeichnet, bei denen man Aspekte von kognitiver Leistungsfähigkeit vorfindet. Künstliche Intelligenz im so verstandenen Sinne heißt also nicht, dass man Maschinen Bewusstsein, Geist oder einen eigenen Willen unterstellt. Obwohl mathematische Verarbeitungsfertigkeiten schon immer ein elementarer Wesenszug der »Rechner« waren, die insoweit die »Good Old-Fashioned Artifical Intelligence« (Haugeland 1985) statistischer Analyse beherrschen, sollen moderne Computersysteme darüber hinaus befähigt werden, eigenständig zu lernen, zu schlussfolgern und zu planen. Sie sollen also Entscheidungsalternativen unter der Bedingung von Unsicherheit und riskanten Folgen auswählen und gegebenenfalls auch umsetzen. Inputs in Form von digitalen Daten werden nicht mehr nur verarbeitet, sondern dienen als Grundlage für Lernprozesse, die durch Menschen unterstützt (»supervised Machine Learning«) oder auch ungestützt (»unsupervised Machine Learning«) ablaufen können. Solche KI-Anwendungen, die sich grob unter dem Stichwort *Maschinelles Lernen* zusammenfassen lassen, bilden derzeit den Schwerpunkt informationstechnologischer Forschung im Feld der KI. Sie sind auf die Verfügbarkeit über großen Datenmengen (*Big Data*) angewiesen, die als Rohstoff für Training und Lernen dienen (Thierer et al. 2017). Dabei erkennt der Computer Muster im Daten-Input, auf deren Grundlage er bei neuen und unbekannten Fällen Vorhersagen und Zuordnungen trifft. Die Ergebnisse automatisierter Klassifikation und Entscheidung sind mittlerweile in Feldern wie Bild- und Spracherkennung den Leistungen von Menschen ebenbürtig und können je nach Anwendungsgebiet bereits heute auch darüber hinausgehen.

3. Was ist Fair Machine Learning?

Die Idee des *Fair Machine Learning* in der Informatik beruht auf der Beobachtung einer Paradoxie. Obwohl automatisierte Klassifikations- oder Entscheidungssysteme den Anspruch erheben, menschliche Vorurteile und Unzulänglichkeiten durch mathematische Kalkulation zu ersetzen, sind sie ihrerseits nicht davor gefeit, Diskriminierung und Ungerechtigkeit zu produzieren (Barocas/Selbst 2016). Eine der zentralen Ursachen liegt in der technischen Logik der Systeme: Intelligente Maschinen lernen aus verfügbaren Trainingsdaten, die ihrerseits Manifestationen des bisherigen Verhaltens von Personen und Institutionen sind. Daten reflektieren also nicht nur die sozialen Spannungen und die Spaltung der Gesellschaft, die sie hervorgebracht hat, sondern auch die gängige soziale Praxis fehlbarer Menschen. So ist es möglich, dass selbstlernende Algorithmen die Ungerechtigkeit der Welt

reproduzieren und zugleich deren Ursprung verschleiern, ohne dass der Maschine oder dem Programmierer willentliche Diskriminierung unterstellt werden muss. Ungerechtigkeit im so verstandenen Sinne bemisst sich an der Zielgröße einer durch KI unterstützten Entscheidung, sei es über den Leistungsanspruch eines Sozialhilfeempfängers oder die Eignung eines Stellenbewerbers im öffentlichen Dienst. Sie ist also das, was seit Aristoteles als Ergebnis- oder Verteilungsgerechtigkeit (*iustitia distributiva*) bezeichnet wird (Bien 1995). Folgerichtig hat sich die Machine-Learning-Literatur der Informatik vornehmlich mit Möglichkeiten und Grenzen der mathematischen Formalisierung sozialwissenschaftlicher Vorstellungen von gerechten Verteilungsnormen beschäftigt (Berendt/Preibusch 2014; Chouldechova 2016; Friedler et al. 2016; Gajane/Pechenizki 2018; Hu/Chen 2018). Damit ist die Frage aufgeworfen, was überhaupt unter den Begriffen »fair« und »diskriminierungsfrei« verstanden werden kann (Binns 2018b). Im Folgenden gehen wir davon aus, dass der auf Ergebnisgerechtigkeit fokussierte Fairnessbegriff der tatsächlichen Komplexität des Phänomens nicht vollumfänglich gerecht wird. Wo die Machine-Learning-Literatur darüber hinausgeht, ist die verwendete Begrifflichkeit nicht immer intuitiv und sozialwissenschaftlich nur bedingt anschlussfähig (Grgic-Hlaca et al. 2018b; Taylor 2017). Um diese Schwächen zu überwinden, schlagen wir ein mehrdimensionales Konzept von KI-Fairness vor, mit dem es gelingen soll, eine Vielzahl der im Zusammenhang mit den gesellschaftlichen Folgen Künstlicher Intelligenz diskutierten Probleme, etwa Reproduktion existierender Ungleichheiten und Stereotypen, Verletzung von Privatheit, mangelnde Transparenz automatisierter Entscheidung etc., in einem einheitlichen theoretischen Bezugsrahmen zu thematisieren.

4. Ein sozialwissenschaftliches Konzept von KI-Fairness

Im sozialpsychologischen Verständnis bezeichnet der Begriff *Fairness* die subjektiven Wahrnehmungen und Einschätzungen der Behandlung, die ein Einzelner durch andere Individuen oder Institutionen erfährt. Fairness ist demnach eine individuelle, kognitive Reaktion, die sich auf manifestes Handeln Dritter richtet: Man fühlt sich von einem Gericht fair behandelt oder auch nicht, weil es so entschieden hat, wie es entschieden hat. Die Entscheidungen und Handlungen Dritter können ihrerseits daran bemessen werden, ob sie bestimmten moralischen, sozialen, religiösen oder rechtlichen Standards entsprechen und insoweit als gerecht bzw. ungerecht gelten. Die Kriterien gerechten Handelns sind variabel und in der Regel Gegenstand einer mehr oder weniger dauerhaften Einigung über die jeweils angemessene Gerechtigkeitsvorstellung in einem gegebenen Handlungsbereich. Formale Gerechtigkeit ist demnach eine Voraussetzung für wahrgenommene Fairness. Im allgemeinen Sprachgebrauch werden die subjektive und die objektive Seite des

Phänomens häufig nicht auseinandergehalten und die Begriffe Fairness und Gerechtigkeit synonym verwendet (Knight 1998).

In der Politikwissenschaft gilt Fairness als zentraler Prädiktor für Input-, Throughput und Output-Legitimität politischen Entscheidens (Scharpf 1970; Schmidt 2013). Das politologische Verständnis des Begriffs rekurriert insoweit auf Verfahrens- (Input und Throughput) und Ergebnisgerechtigkeit (Output). Darüberhinausgehend wird Fairness in der Organisationsforschung als dreidimensionales Konstrukt behandelt (Barling/Philipps 1993; Skarlicki/Folger 1997; Tata/Bowes-Sperry 1996; Yean/Yusof 2016). Neben distributiver und prozeduraler Fairness tritt eine soziale Dimension von Fairness auf, um – über die Qualität der eigentlichen Verfahrensregeln hinaus – auf die Bedeutung des persönlichen Umgangs von Entscheidungsträgern mit Entscheidungsbetroffenen hinzuweisen. Einige Autoren schlagen darüber hinaus vor, die soziale Komponente des Konzepts in die zwei Subdimensionen interaktionale und informationelle Fairness zu untergliedern (Bies/Moag 1986; Greenberg 1993), wodurch ein vierdimensionales Modell von Fairness entsteht, das sich auch in empirischer Hinsicht als tragfähig erwiesen hat (Colquit 2001; Shapiro et al. 1994). Im Folgenden werden wir skizzieren, welche Merkmale und Eigenschaften von Systemen mit künstlicher Intelligenz aus der Perspektive jeder der vier Fairnessdimensionen in den Blick geraten und welche Forschungsfragen sich daraus ergeben.

Distributive Fairness

Das Konzept distributiver Fairness bezieht sich auf die Wahrnehmung einer fairen Verteilung von Ressourcen, also auf das Ergebnis von Entscheidungsprozessen (Cropanzano et al. 2015; Yean/Yusof 2016). Dem entspricht in demokratietheoretischer Perspektive das Konzept der Output-Legitimation (Scharpf 1970). In der Literatur werden in der Regel drei zentrale Verteilungskriterien voneinander unterschieden (Deutsch 1975): (1) *Equality* oder absolute Gleichheit beschreibt das Gleichverteilungsprinzip von Ressourcen, d.h. jeder Akteur erhält dasselbe. In demokratischen Gesellschaften wird diese Verteilungsnorm vor allem bei der Anerkennung von Grundrechten angewendet. (2) *Equity* oder relative Gleichheit beschreibt die Verteilungsnorm, bei der jeder Akteur einen leistungsgerechten Anteil der Ressourcen erhält, abhängig davon, wie viel er/sie eingesetzt hat (Adams/Freedman 1976). Diesem Verteilungsprinzip folgt beispielsweise das deutsche Rentensystem, da sich die Höhe der Rente nach der Höhe der eingezahlten Beiträge bemisst. (3) *Need* beschreibt die Verteilung von Ressourcen basierend auf dem Kriterium der Bedürftigkeit. Das bedeutet, dass diejenigen Akteure bevorzugt werden, die am dringendsten auf die zu verteilenden Ressourcen angewiesen sind. Diese Verteilungsnorm ist in demokratischen Gesellschaften eng mit sozialstaatlichen Insti-

tutionen verbunden und findet vor allem im Bereich der Grundsicherung Anwendung, beispielsweise bei der Arbeitslosenversicherung.

Im Zuge der fortschreitenden Digitalisierung werden in nahezu allen gesellschaftlichen Teilbereichen digitale Daten erzeugt, gesammelt, analysiert und verwertet (Dencik et al. 2019). Im politischen System bilden digitale Daten die Grundlage von E-Government, um Transaktionen innerhalb und zwischen staatlichen Institutionen sowie die Kommunikation zwischen Bürgern und Staat durch digitale Technologie zu unterstützen. KI-basierte Klassifikationssysteme werden derzeit auch bei der Implementation von Verteilungspolitiken genutzt (Poel et al. 2018). Aktuelle Beispiele betreffen etwa die Verteilung von geflüchteten Menschen zwischen Gebietskörperschaften (Bansak et al. 2018) oder die Allokation von Gesundheitsdienstleistungen (Wahl et al. 2018). Im Rahmen derartiger Entscheidungsprozesse kann es vorkommen, dass Algorithmen willentlich oder unwillentlich gegen die geltenden rechtlichen und sozialen Gerechtigkeitsnormen verstoßen. Ein Beispiel für willentliche Diskriminierung wäre, dass individuelle Grundrechte durch Algorithmen nicht auf Basis des *Equality*-Prinzips, sondern unter Rückbezug auf das *Equity*-Prinzip zuerkannt werden. Das würde bedeuten, dass lediglich denjenigen Bürgern, die zur wirtschaftlichen Produktivität des Landes beitragen, auch bestimmte Freiheitsrechte (z.B. Mobilität) zugesprochen werden. Ein vergleichbarer Gedanke ist etwa im Social-Credit-System in China angelegt (Campbell 2019). Unwillentliche Diskriminierung kann entstehen, wenn die Entscheidungen von Algorithmen auf Grundlage verzerrter Trainingsdaten getroffen werden, in denen realweltlich existierende Diskriminierungen (z.B. in Bezug auf Geschlecht, Ethnizität, sozialer Status) bereits enthalten sind. Das kann zur Folge haben, dass bestimmte gesellschaftliche Gruppen systematisch bevorteilt bzw. benachteiligt werden (Barocas/Selbst 2016; Hoffmann 2019; Taylor 2017).

Während sich soziologische oder politikwissenschaftliche Analysen algorithmischer Ungleichbehandlung auf De-facto-Diskriminierungen im Aggregat richten und an objektiven Tatbeständen festmachen lassen, lenkt das hier vorgeschlagene Fairnesskonzept den Blick auf die individuelle Wahrnehmung dieser Ergebnisse. Dabei ist zu vermuten, dass automatisiert getroffene Entscheidungen über die Verteilung von Ressourcen von den Anwendern und Betroffenen immer dann als unfair wahrgenommen werden, wenn die eigene individuelle Fairnessvorstellung von der verwendeten Verteilungsnorm des Algorithmus abweicht. Dabei ist es für das persönliche Empfinden zunächst einmal unerheblich, ob die eigene Fairnessvorstellung mit der gesellschaftlich geltenden Verteilungsnorm übereinstimmt oder nicht. Problemverschärfend wirkt, dass die Betroffenen in der Regel gar nicht wissen, welches Ergebnis der Algorithmus in der Gesamtpopulation erzeugt hat, geschweige denn, welche Verteilungsnorm der Entscheidung zugrunde liegt. Die individuelle Bewertung von distributiver (Un)Fairness automatisierter Entscheidungen dürfte daher zuvorderst durch das Ergebnis determiniert sein, das man

selbst erhalten hat. Im ungünstigsten Fall werden Betroffene immer dann davon ausgehen, einen ungerechten Anteil der zu verteilenden Lasten und zu gewinnenden Leistungen abbekommen zu haben, wenn die eigenen Erwartungen enttäuscht worden sind. Um zu erkennen, dass diese Erwartungen unrealistisch waren und mithin das Ergebnis gerechter als zunächst gedacht ist, wäre ein Mindestmaß an Information über andere Fälle nötig, zu denen sich Betroffene in Beziehung setzen können. Wo diese Informationen nicht verfügbar sind, kann sich der Verdacht von Unfairness technischer Systeme erhärten. Diese Überlegungen verweisen auf zusätzlichen Forschungsbedarf. Zum einen gilt es zu untersuchen, unter welchen Bedingungen sich Betroffene als »gute bzw. schlechte Verlierer« automatisierter Entscheidungsverfahren erweisen. Darüber hinaus interessieren mögliche Effekte wahrgenommener (Un)Fairness auf individuelle Emotionen und Verhaltensweisen. Hier rücken unter anderem Konzepte wie Angst oder Wut, aber auch Kategorien wie Technikakzeptanz, Selbstwirksamkeit, Vertrauen oder Protestbereitschaft in den Fokus sozialwissenschaftlicher Betrachtungen. Wenngleich erste empirische Ergebnisse zur Beantwortung der skizzierten Fragestellung vorliegen, steckt die sozialwissenschaftliche Forschung zu distributiver KI-Fairness noch in den Kinderschuhen (Binns et al. 2018; Grgic-Hlaca et al. 2018b; Lee 2018; Lee/Baykal 2017).

Prozedurale Fairness

Prozedurale Fairness bezieht sich auf den Entscheidungsprozess, der zu einem bestimmten Ergebnis führt (Thibaut/Walker 1975). Zahlreiche empirische Forschungsarbeiten zeigen, dass die Wahrnehmung von Verfahren für Meinungsbildung und Verhalten wichtiger sein können als die Bewertung des Ergebnisses selbst (z.B. Tyler et al. 1997). In seinen einflussreichen Arbeiten nennt Leventhal (1980) sechs Kriterien, nach denen die Fairness von Verfahrensregeln gemeinhin bewertet wird: (1) Konsistenz, (2) Neutralität, (3) Genauigkeit, (4) Revidierbarkeit, (5) Ethik, (6) Repräsentativität. KI-basierte Entscheidungsprozesse weisen die Besonderheit auf, dass die Vorgänge innerhalb der »black box« für eine qualifizierte Fairnessbewertung der Betroffenen und Anwender kaum zugänglich sind. Deren Urteile werden sich also an den sichtbaren prozeduralen Regeln orientieren und sich darüber hinaus auf subjektive Perzeptionen stützen. In einem ersten Zugriff verweisen die Leventhal-Kriterien auf sechs Ansatzpunkte für die Fairnesswahrnehmung KI-basierter Verfahren, die im Übrigen in ähnlicher Form auch bei der Evaluation politischer Partizipation eine zentrale Rolle spielen (Kersting 2019, in diesem Band).

- *Konsistenz*: die Entscheidungsregeln sollten konsequent angewendet werden, unabhängig vom Entscheidungsträger, von den Betroffenen sowie vom Zeitpunkt der Entscheidung. Aus der sozialpsychologischen Forschung (z.B. Evans

1989) wissen wir, dass Menschen aufgrund von individuellen (z.B. Empathiefähigkeit) sowie situativen Faktoren (z.B. Müdigkeit, Emotionen) nicht immer konsistent entscheiden. Hingegen sollten derartige menschliche *biases* bei automatisierten Systemen keine Rolle spielen. Dies bedeutet jedoch nicht, dass KI-Systeme zwangsläufig zu zuverlässigeren Entscheidungen gelangen, geschweige denn, dass sie von den Betroffenen als konsistent wahrgenommen werden. Ein Grund liegt in der Besonderheit komplexer KI-Systeme, wie neuronale Netzwerke selbstständig aus vorherigen Entscheidungen lernen zu können. Sie sind daher in der Lage, ihre eigenen Verfahrensregeln ständig zu optimieren (Schmidhuber 2015). Es gilt also zu klären, ob und unter welchen Bedingungen automatisierte Entscheidungsprozesse von den Betroffenen als konsistent wahrgenommen werden. Darüber hinaus stellt sich die Frage, welche individuellen Faktoren (z.B. Vertrauen in Technik) die Konsistenzwahrnehmung beeinflussen und welche Auswirkungen dies auf relevante Zielvariablen hat.

- *Neutralität*: Die persönlichen Prozesspräferenzen der Entscheidungsträger sollten die faire Entscheidung nicht beeinflussen. Neutralität bezieht sich somit auf die vermeintliche Unvoreingenommenheit automatisierter Entscheidungssysteme. Das zentrale Argument lautet hier, dass Algorithmen keine eigenen Interessen verfolgen und somit objektiv entscheiden können. Diese positive Sichtweise wird jedoch von einigen Forschern grundlegend bezweifelt. Bollier (2010, 13) fragt: »Can the data represent an ›objective truth‹ or is any interpretation necessarily biased by some subjective filter or the way that data is ›cleaned‹?« (siehe auch Boyd/Crawford 2012). Es ist also durchaus nicht zwingend, dass KI-basierte Systeme automatisch als objektiv und unvoreingenommen wahrgenommen werden. Da die Neutralität des Prozesses jedoch einerseits auf den eingespeisten Daten beruht und andererseits erst im Ergebnis evident wird, steht der Neutralitätsaspekt in engem Zusammenhang mit *Interaktionaler Fairness* auf der einen und *Distributiver Fairness* auf der anderen Seite.
- *Genauigkeit*: Faire Entscheidungen sollten auf möglichst vollständigen und korrekten Informationen beruhen. Damit ist im Falle von automatisierten Entscheidungen die Zuverlässigkeit und Gültigkeit des Dateninput angesprochen (Grgic-Hlaca et al. 2018b). Auch hierüber entfaltet sich eine kontroverse Debatte in der Literatur. Einerseits sind KI-basierte Entscheidungssysteme in der Lage, Zusammenhänge in großen Datenmengen zu identifizieren, die für menschliche Akteure schlichtweg zu komplex sind (Berendt/Preibusch 2014). Andererseits sind bestimmte Phänomene kaum reliabel und valide messbar und verzerren somit mögliche Schlussfolgerungen (Grgic-Hlaca et al. 2018b). Unabhängig von der wissenschaftlichen Debatte über die Genauigkeit von Daten stellt sich die Frage, wie Endnutzer Urteile über die Gültigkeit der Datengrundlage bilden. So wird ein/e Betroffene/r eines automatisierten Stellenbesetzungs-

verfahrens das Ergebnis dann als fair wahrnehmen, wenn er/sie Anlass hat zu glauben, dass die Informationen, die zur Entscheidungsfindung herangezogen wurden, korrekt, zuverlässig, gültig und vollständig waren. In diesem Zusammenhang gilt es zu erforschen, worauf ein solches Urteil gründet.

- *Revidierbarkeit*: Im fairen Entscheidungsprozess ist sichergestellt, dass fehlerhafte oder unangemessene Entscheidungen rückgängig gemacht werden können. Revidierbarkeit bezieht sich somit auf die Fähigkeit des Entscheidungsprozesses, sich selbst zu korrigieren. Für die Fairnesswahrnehmung KI-basierter Prozesse hat dies zwei zentrale Konsequenzen. Zum einen stellt sich die Frage, wie Betroffene angesichts der reklamierten »Intelligenz« technischer Systeme ihre eigenen Möglichkeiten einschätzen, Revision gegen das Ergebnis einzulegen. Diese Einschätzung kann durchaus von der tatsächlichen Revidierbarkeit des automatisierten Verfahrens abweichen, beispielsweise aufgrund von mangelndem Wissen oder mangelndem Vertrauen in den Berufungsprozess. Zum anderen dürfte es für Bürger wichtig sein, wie das Zusammenspiel von Mensch und Maschine konkret ausgestaltet ist. Bisherige Forschungen deuten auf eine hohe Akzeptanz computergestützter Systeme hin, wenn sie dem Menschen zuarbeitet (Cockerill et al. 2004). Was aber sieht das Verfahren vor, wenn Mensch und Maschine in ihren Urteilen abweichen? Einerseits wäre es denkbar, einen anderen Algorithmus hinzuzuziehen und erneut über dieselbe Angelegenheit zu entscheiden. Andererseits besteht die Möglichkeit, die Entscheidung auf ein Gremium aus menschlichen Akteuren zu übertragen, welches über die Autorität verfügt, Entscheidungen von KI-basierten Systemen zu überstimmen. Im zweiten Fall ergeben sich jedoch neue Herausforderungen für die Wahrnehmung von prozeduraler Fairness, vor allem im Hinblick auf die Nachvollziehbarkeit der Begründung, mit der eine für den Betroffenen günstige KI-basierte Entscheidung überstimmt wird.
- *Ethik*: der Entscheidungsprozess sollte auf grundlegenden ethischen Werten beruhen. Ethisch relevante Verfahrensaspekte beziehen sich zunächst auf die Frage, welche Entscheidungsprozesse überhaupt durch Systeme künstlicher Intelligenz unterstützt werden dürfen, welche sogar komplett an automatisierte Systeme übertragen werden können und welche weiterhin in den Händen von Menschen liegen sollten. Die Meinungen der Bürger hierüber haben annahmegemäß direkte Effekte auf ihre Einschätzung der Verfahrensgerechtigkeit. Sie gehen derzeit noch weit auseinander und sind je nach Einsatzbereich unterschiedlich ausgeprägt. Beispielsweise lehnten bei einer repräsentativen Bevölkerungsumfrage 77 % der Befragten die Vorstellung ab, dass Maschinen mit künstlicher Intelligenz Vorstellungsgespräche führen, in der gleichen Umfrage haben aber 49 % der Befragten kein Problem damit, dass Maschinen Zeitungsberichte für den Wirtschaftsteil verfassen (YouGov 2018). Zudem stellen sich wichtige Fragen nach den moralischen Entscheidungsregeln KI-basierter

Systeme: »With the rapid development of artificial intelligence have come concerns about how machines will make moral decisions, and the major challenge of quantifying societal expectations about the ethical principles that should guide machine behavior« (Awad et al. 2018, 59). Für die Erforschung von individuellen Fairnesswahrnehmungen führt das zu mindestens zwei weiteren Fragen: Erstens, in welchen Bereichen und zu welchem Grade sollten Maschinen in Entscheidungsprozesse integriert werden? Zweitens, welche moralischen Entscheidungsregeln sollten den Prozess leiten?

- *Repräsentativität* bedeutet, dass die frei geäußerten Meinungen aller relevanten Parteien im Entscheidungsprozess zum Ausdruck gebracht werden können. Thibaut und Walker (1975) sprechen in diesem Zusammenhang von Prozesskompetenz. In Bezug auf KI-basierte Prozesse bedeutet das vor allem, dass für verschiedene Identitäten, Kulturen, Ethnien und Sprachen, die Gegenstand des Verfahrens sind, aussagekräftige Daten verfügbar sind und berücksichtigt werden – also Datenrepräsentativität im statistischen Sinne gewährleistet ist (Binns 2018b). Darüber hinaus bezieht sich der Repräsentativitätsaspekt auf die wahrgenommenen Möglichkeiten der Betroffenen, den Prozess beeinflussen zu können, was Thibaut und Walker (1975) als Entscheidungskompetenz bezeichnen. Es lässt sich argumentieren, dass durch die Implementierung selbstlernender Systeme insbesondere die Mündigkeit der Bürger geschwächt wird. Umgekehrt könnte Repräsentativität aber auch eine große Stärke solcher Systeme sein, indem für die Entscheidung große Datenmengen aus vielen verschiedenen Datenquellen (z.B. Bevölkerungsumfragen, Internetsuchanfragen) herangezogen werden, die womöglich die Meinungen der Anspruchsgruppen umfassender abbilden, als das auf konventionellen Wegen der Verfahrensbeteiligung möglich wäre. Wie und in welchem Umfang sich Individuen und soziale Gruppen in algorithmischen Entscheidungsprozessen tatsächlich repräsentiert fühlen und wovon das jeweils abhängt, ist derzeit noch weitgehend unerforschtes Gelände.

Interaktionale Fairness

Im Unterschied zur Verfahrensgerechtigkeit geht es bei der Interaktionsgerechtigkeit nicht um die Wahrnehmung der prozeduralen Regeln selbst, sondern darum, wie sie im konkreten Fall angewendet werden (Bies 2001; Tyler/Bies 1989). Dem entsprechen in demokratietheoretischer Perspektive Kategorien wie Respekt oder Zivilität, die namentlich in der Theorie deliberativer Demokratie betont werden (Herbst 2010). Greenberg (1990, 1993) hat diese Dimension als *interpersonal fairness* bezeichnet, weil sie auf Erfahrungen im unmittelbaren Kontakt von Beteiligten und Betroffenen beruht. Wenn das Konzept auf die Interaktion mit technischen Systemen angewendet wird, ist damit unterstellt, dass sich Menschen auch vom »Ver-

halten« intelligenter Maschinen in ihrem Selbstwertgefühl verletzt fühlen können. Ansätze für eine solche Argumentation ließen sich etwa in der Literatur zur *Media Equation Theory* oder dem *Computer as Social Actors* Paradigma finden (Nass/Moon 2000), wonach Computer nach den gleichen sozialen Maßstäben beurteilt und behandelt werden, wie andere Menschen auch. Da hier der Platz fehlt, ein solches Argument auszuarbeiten, gehen wir zunächst davon aus, dass sich die Urteile über interpersonale Fairness auf das jeweilige Entscheidungssystem insgesamt beziehen, in dem Künstliche Intelligenz und Menschen in spezifischer Weise zusammenwirken. Daher verwenden wir den übergreifenden Begriff *interaktionale Fairness* (Bies/Moag 1986). Die »soziale Seite« (Greenberg 1993) der Fairness umfasst folglich alle Wahrnehmungen, die sich auf den persönlichen Umgang der Entscheidungsträger mit den Adressaten einer Entscheidung beziehen. Das betrifft zunächst die Frage, ob sich Betroffene als gleichberechtigte Partner wertgeschätzt und respektiert fühlen, oder ob sie – gerade umgekehrt – den Eindruck gewinnen, dass es den Vorgesetzten an Empathie und Wohlwollen mangelt. Besonders negative Effekte auf diese Dimension von Fairness beinhalten solche Erfahrungen, die als Verletzung der eigenen Identität und Würde erlebt werden. Als Beispiele für Verletzungen des »sacred self«, die als Unfairness wahrgenommen werden können, nennt Bies (2001) die Erfahrung von Geringschätzung, etwa durch abfällige Bemerkungen oder ungerechtfertigte Beschuldigungen, die Erfahrung, das Versprechen nicht eingehalten und Vertrauen missbraucht wird, Verletzungen der Privatsphäre, etwa durch Verwendung vertraulicher Information oder das Stellen unangemessener Fragen und schließlich alle Formen der Respektlosigkeit in Sprache und Verhalten des Gegenübers.

Vor dem Hintergrund erscheinen zwei Eigenschaften KI-basierter Entscheidungssysteme besonders geeignet, bei Betroffenen das Gefühl despektierlicher Behandlung zu erregen. (1) Die Intelligenz technischer Systeme beruht immer darauf, dass sie Strukturen und Zusammenhänge in einer für Menschen nicht zu bewältigenden Informationsfülle entdeckt und für die Identifikation bestimmter Fälle verwendet. Wird KI für die Identifikation oder Voraussage von individuellen Verhaltensweisen genutzt, die als unerwünscht gelten oder regelrechte Normverstöße darstellen, heißt das zugleich, dass die Systeme zunächst einmal die gesamte Population unter Generalverdacht stellen. Das gilt etwa für die Aufdeckung von Sozialbetrug durch Leistungsbezieher, falsche oder unvollständige Angaben im Rahmen von Asylverfahren, Früherkennung von kriminellen Wiederholungstätern oder die Identifikation potentieller Problemfälle in einem Bewerberfeld. Binns (2018b) spricht in diesem Zusammenhang auch von statistischer Diskriminierung. Vorliegende Forschungen weisen darauf hin, dass ein flächendeckendes Aussetzen der Unschuldsvermutung von Betroffenen als Respektlosigkeit verstanden wird, mit entsprechenden Konsequenzen für die wahrgenommene Interaktionsgerechtigkeit (Bies 1993). Damit erhebt sich die Frage, ob andere Regeln im Umgang mit

der Informationsbasis (etwa Vorauswahl begründeter Verdachtsfälle, Zufallsauswahl einer Stichprobe) die wahrgenommene Fairness verstärken, ohne die Leistungsfähigkeit des technischen Systems einzuschränken. (2) Weil die Funktionalität KI-basierter Systeme mit der Menge an Informationen (Daten) steigt, die für die Analyse zur Verfügung stehen, ist die Wahrscheinlichkeit groß, dass darunter auch solche Informationen sind, deren Nutzung Betroffene als Verletzung ihrer »informationellen Privatheit« begreifen. Damit ist hier der Anspruch des einzelnen auf Kontrolle darüber gemeint, wer über welche persönlichen Informationen und zu welchem Zweck Kenntnis erhält (Stone/Stone 1990; Westin 1967). Verschiedene Autoren haben Verletzungen der informationellen Privatheit mit Einschränkungen der *interaktionalen Fairness* in Verbindung gebracht (Bies 1993, 1996; Leventhal 1980). Welche Information in dem Sinne als privat gelten, kann von Fall zu Fall unterschiedlich sein, potentiell aber vieles betreffen: etwa Angaben über Einkommen und Vermögen, sexuelle und religiöse Orientierungen, politische Präferenzen, medizinische Fakten. Zu einem besonderen Problem wird in diesem Zusammenhang das lange Gedächtnis des Internet. Sollte etwa ein Algorithmus Informationen über früheren Alkohol- oder Drogenkonsum, die sich über Datenspuren auf sozialen Netzwerken erschließen lassen, in aktuelle Beurteilungen einer Person einbeziehen, werden Betroffene das in mehrfacher Hinsicht als unfair betrachten: Erstens, weil ihnen ein möglicherweise weit zurückliegender und einmaliger Vorgang heute noch zugerechnet wird; Zweitens, weil ihnen damit unterstellt wird, keine persönliche Entwicklung durchgemacht zu haben, und drittens, weil eine Verletzung der Privatsphäre vorliegt, wenn es sich um Informationen aus privaten Kontexten handelt, die irgendwann einmal unvorsichtigerweise veröffentlicht wurden. Ein einschlägiges Forschungsprogramm sollte sich auf die Frage konzentrieren, welches spezifische Gewicht verschiedenen Faktoren (z.B. die Art der benutzten Information, die Form der Datensammlung, informierte Einwilligung) auf die wahrgenommene Fairness haben. Die wachsende Literatur über *Online Privacy Concerns* (Baruh et al. 2017) kann genutzt werden, um künftige Analysen der Interaktionsgerechtigkeit von KI-gestützten Studien zu informieren.

Informationelle Fairness

Die vierte Dimension von Fairness betrifft die Frage, ob Entscheidungsträger die Logik ihres Handelns erklären oder die Betroffenen über die Grundlagen ihrer Entscheidung im Unklaren lassen. Dem korrespondieren in demokratietheoretischer Perspektive Kategorien wie Transparenz oder Rechenschaftspflicht (Mulgan 2003). Während Bies und Moag (1986) undurchsichtiges Entscheiden als Mangel an Respekt verstehen und mithin konzeptuell als Subdimension der Interaktionsgerechtigkeit behandeln, haben neuere Forschungen im organisatorischen Kontext gezeigt, dass »informational justice« nicht nur einen empirisch distinkten Faktor

darstellt, sondern auch andere Zielvariablen beeinflusst, als beispielsweise Wertschätzung oder Freundlichkeit (Colquit 2001). Danach bildet der Anspruch auf adäquate Erklärung einer einmal getroffenen Entscheidung einen integralen Bestanteil des Gerechtigkeitsempfindens von Betroffenen. Empirische Evidenz für diese Annahme liegt vor allem aus dem Bereich der Managementtheorie und Organisationsforschung vor (Greenberg 1990; Tyler/Bies 1989). So lässt sich etwa zeigen, dass abgelehnte Stellenbewerber den Besetzungsprozess als fairer wahrnehmen, wenn ihnen eine Begründung für ihre Nichtberücksichtigung gegeben wird, als wenn die Erklärung ausbleibt (Bies/Shapiro 1988). Die weitergehende Forschung richtet sich auf die Frage, welchen Kriterien eine Erklärung genügen muss, um als adäquat anerkannt zu werden. Soweit sich das beim jetzigen Stand der Forschung sagen lässt, bilden Rechtzeitigkeit, Aufrichtigkeit und Verständlichkeit die Mindestbedingungen für Ansprüche an eine zufriedenstellende Erklärung organisationaler Entscheidungen (Shapiro et al. 1991, 1994). Der hiermit angesprochene Zusammenhang von verständlicher Information und Fairness lässt sich unmittelbar auf den Einsatz von Künstlicher Intelligenz im organisatorischen Kontext übertragen. Entsprechende Forschungen können an die entstehende Literatur zu »explainable artificial intelligence« oder »transparency in machine learning« (Annay/Crawford 2016; Burrell 2016; Holzinger 2018; Wachter et al. 2017) anknüpfen, die bisher allerdings noch wenig empirisch ausgerichtet ist und vornehmlich auf Technikakzeptanz und -vertrauen als Zielvariablen abzielt.

Die Problematik adäquater Erklärung für das Verhalten selbstlernender Algorithmen ist vielschichtig. Zunächst haben die Anwender selbst wenig Interesse daran, die exakte Funktionsweise des Algorithmus und dessen Datengrundlagen offen zu legen, soweit ihre kommerziellen Interessen betroffen sind. Die rechtlichen Möglichkeiten, vollständige Transparenz einzuklagen, sind auch im Einzugsbereich der Europäischen Datenschutz-Grundverordnung beschränkt (Wachter et al. 2017). Darüber hinaus gibt es technische Hürden, weil die verwendeten mathematischen (meist nicht-linearen) Modelle wenig intuitiv, schwer rekonstruierbar und im Einzelfall extrem komplex sein können. Schließlich ist davon auszugehen, dass das informationstechnische Verständnis des durchschnittlichen Bürgers in der Regel nicht ausreicht, um die technische Funktionsweise der Systeme zu verstehen, die über sie entscheiden. Allgemeinverständliche Erklärungen algorithmischer Entscheidungen setzen also voraus, dass deren Logik und Datengrundlage den Endnutzern und interessierten Gruppen in nicht-technischen Begriffen erläutert werden können. Dazu werden bisher häufig graphische Oberflächen verwendet. Ein Beispiel ist die sogenannte *Layer-Wise Relevance Propagation (LRP)*, mit deren Hilfe Bild- und Gesichtserkennungssysteme in Form einer Wärmekarte anzeigen, welche Bildpunkte oder Bildbereiche für die Vorhersage besonders wichtig waren (Bach et al. 2015). Die *Local Interpretable Model-Agnostic Explanation (LIME)* liefert ebenfalls eine graphische Darstellung der individuellen Wichtigkeit einzel-

ner Merkmale, etwa Symptome einer Krankheitsgeschichte oder Wörter in einem Text, für die Vorhersage (Ribeiro et al. 2016). Die kontrafaktische Methode kommt demgegenüber ohne graphische Instrumente aus; hierbei folgt der Mitteilung einer Klassifikationsentscheidung unmittelbar eine Aussage darüber, was in der Realität hätte anders sein müssen, damit eine andere Entscheidung herausgekommen wäre (Sharma et al. 2019). Die Entwicklung von Verfahren, die über bisherige Lösungen des Erklärungsproblems hinausgehen, bildet einen Schwerpunkt der informationstechnischen Forschung. Davon erhofft man sich, dass selbstlernende Systeme der Zukunft besser in der Lage sein werden, ihre Logik in verständliche und nützliche Erklärungsdialoge für den Endverbraucher zu übersetzen. Juristen und Sozialwissenschaftler fordern von »verantwortlicher« Künstlicher Intelligenz, dem Endnutzer verständlich zu machen, auf welcher Grundlage eine automatisierte Entscheidung/Klassifikation zustande gekommen ist, wo mögliche »Fehler im System« und Ansatzpunkte für Einspruch liegen könnten und was sich ändern müsste, um beim nächsten Mal ein anderes Ergebnis zu bekommen (Wachter et al. 2017). Welche der genannten Methoden im Einzelfall diese Forderung erfüllen kann, ist beim jetzigen Stand der Forschung und Theoriebildung kaum zu sagen. Im Übrigen ist davon auszugehen, dass die Kriterien einer adäquaten Erklärung mit einer Reihe von Randbedingungen variieren, etwa der Art der technisch zu lösenden Aufgabe, der formalen Regeln des soziotechnischen Verfahrens oder dem organisatorischen Kontext. In einem weiteren Schritt wäre dann zu klären, welcher dieser Faktoren (Befähigung zum Verstehen, Widersprechen, Verändern) tatsächlich und mit welchem Gewicht auf die wahrgenommene informationelle Fairness eines technischen Systems einzahlt. Auch in dieser Dimension eröffnet sich ein weithin unbestelltes Feld für die empirische Erforschung von KI-Fairness.

5. Fazit

In Ergänzung existierender Ansätze im Fair Machine Learning fokussiert der vorliegende Beitrag die subjektive Dimension von Fairness, wodurch der Einzelne ins Zentrum des Forschungsinteresses rückt. Diese Perspektive ist von der Überzeugung getragen, dass ein demokratischer und sozialverträglicher Weg in die Digitale Gesellschaft nur in Kenntnis von und im Einklang mit den Überzeugungen, den Präferenzen und den Wahrnehmungen der Betroffenen beschritten werden kann, nicht aber in Unkenntnis oder gar gegen sie. Anders formuliert, die gesellschaftlichen Hoffnungen auf eine bessere Zukunft mit Künstlicher Intelligenz werden sich nur dann erfüllen, wenn die betroffenen Menschen das Gefühl haben, dass die Welt dadurch nicht nur effizienter, sondern auch gerechter (oder zumindest nicht ungerechter) wird. Dies wird vor allem dann der Fall sein, wenn (1) die Ergebnisse automatisierter Entscheidungsprozesse, (2) die Architektur und Verfahrensre-

geln sozio-technischer Systeme, (3) der Umgang dieser Systeme mit den Betroffenen – vor allem im Hinblick auf Respekt und Diskretion und schließlich (4) die Auskünfte der Systeme über ihr eigenes Tun von den Endnutzern als fair wahrgenommen werden. Sozialwissenschaftliche Forschungen zur KI-Fairness hätten zu zeigen, unter welchen Bedingung sich solche Überzeugungen einstellen und unter welchen nicht. Die hierzu im Text formulierten Vermutungen können nicht mehr als ein Anfang sein. Ein so verstandenes Konzept von KI-Fairness ist offenbar nicht ein für alle Male zu bestimmen, vielmehr ist KI-Fairness immer spezifisch. Die jeweils herangezogenen Beurteilungsdimensionen und die individuelle Wahrnehmung dieser Kategorien können je nach Einsatzbereich und Anwendungsfall, aber auch von Person zu Person erheblich variieren. Damit werden die Umrisse eines umfangreichen sozialwissenschaftlichen Forschungsprogramms erkennbar, das im Kern von folgenden Fragestellungen geleitet wird: Von welchen internen und externen Faktoren sind individuelle Fairnesswahrnehmungen in den Dimensionen Ergebnis, Verfahren, Interaktion und Information abhängig? Auf welche gesellschaftlich relevante(n) Zielvariable(n) wirkt welche Dimension von Fairness? Wie wirken die Dimensionen bei der Gesamtbeurteilung eines technischen Systems zusammen? Eignet sich ein Additionsmodell, in dem die Fairness des Systems umso höher eingeschätzt wird, desto höher die Werte jeder Einzeldimension ausfallen, zum Verständnis von KI-Fairness? Oder muss man sich das Zusammenwirken der Einzeldimensionen als einen Kompensationsprozess vorstellen, indem beispielsweise ein wahrgenommener Mangel an Verfahrensgerechtigkeit durch eine besonders positive Beurteilung der Ergebnisgerechtigkeit ausgeglichen werden kann? Empirisch gestützte Antworten auf diese Fragen können dabei helfen, Systeme mit Künstlicher Intelligenz so zu gestalten und zu implementieren, dass sie der oben erhobenen Forderung nach sozial- und demokratieverträglicher Technik genügen.

Literaturverzeichnis

Adams, J. Stacy/Freedman, Sara (1976): Equity Theory Revisited: Comments and Annotated Bibliography. In: Advances in Experimental Social Psychology 2 (9), S. 43–90.

Ananny, Mike/Crawford, Kate (2016): Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. In: New Media & Society 20 (3), S. 973–989.

Awad, Edmond et al. (2018): The Moral Machine Experiment. In: Nature 563 (7729), S. 59–64.

Bach et al. (2015): On Pixel-wise Explanations for Non-Linear Classifier Decisions by Layer-wise Relevance Propagation. In: PLoS ONE 10 (7).

Bansak, Kirk et al. (2018): Improving Refugee Integration through Data-Driven Algorithmic Assignment. In: Science 359 (6373), S. 325–329.

Barocas, Solon/Selbst, Andrew (2016): Big Data's Disparate Impact. In: California Law Review 104 (1), S. 671–729.

Barling, Julian/Phillips, Michelle (1993): Interactional, formal, and distributive justice in the workplace: An exploratory study. In: The Journal of Psychology 127 (6), S. 649–656.

Baruh, Lemi et al. (2017): Online privacy concerns and privacy management: A meta-analytical review. In: Journal of Communication 67 (1), S. 26–53.

Berendt, Bettina/Preibusch, Sören (2014): Better Decision Support through Exploratory Discrimination-Aware Data Mining: Foundations and Empirical Evidence. In: Artificial Intelligence and Law 22 (2), S. 175–209.

Bien, Günther (1995): Gerechtigkeit bei Aristoteles. In: Höffe (Hg.): Aristoteles. Die Nikomachische Ethik. Berlin, S. 135–164.

Bies, Robert J. et al. (1988): Causal accounts and managing organizational conflict: Is it enough to say it's not my fault? In: Communication Research 15 (4), S. 381–399.

Bies, Robert J./Moag, Joseph S. (1986): Interactional justice: Communication criteria for fairness. In: Lewicki et al. (Hg.): Research on negotiation in organizations. Greenwich, S. 43–55.

Bies, Robert J./Shapiro, Debra L. (1988): Voice and justification: Their influence on procedural fairness judgments. In: The Academy of Management Journal 31 (3), S. 676–685.

Bies, Robert J. (1993): Privacy and procedural justice in organizations. In: Social Justice Research 6 (1), S. 69–86.

Bies, Robert J. (1996): Beyond the hidden self: Psychological and ethical aspects of privacy. In organizations. In: Messick/Tenbrunsel (Hg.): Codes of Conduct: Behavioral Research into Business Ethics. New York, S. 104–116.

Bies, Robert J. (2001): International (in)justice: The sacred and the profane. In: Greenberg/Cropanzano (Hg.): Advances in organization justice. Stanford, S. 89–118.

Binns, Reuben (2018a): Fairness in Machine Learning: Lessons from Political Philosophy. In: Journal of Machine Learning Research 19 (81), S. 1–11.

Binns, Reuben (2018b): What Can Political Philosophy Teach Us About Algorithmic Fairness? In: IEEE Security & Privacy 16 (3), S. 73–80.

Binns, Reuben et al. (2018): ›It's Reducing a Human Being to a Percentage‹: Perceptions of Justice in Algorithmic Decisions. In: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, (27.05.2019).

Bollier, David (2010): The promise and peril of big data. URL: www.aspeninstitute.org/sites/default/files/content/docs/pubs/The_Promise_and_Peril_of_Big_Data.pdf (27.05.2019).

Boyd, Danah/Crawford, Kate (2012): Critical Questions for Big Data. In: Information, Communication & Society, 15 (5), S. 662-679.

Burrell, Jenna (2016): How the machine ›thinks‹: Understanding opacity in machine learning algorithms. In: Big Data & Society 3 (1), S. 1–12.

Campbell, Charlie (2019): How China Is Using ›Social Credit Scores‹ to Reward and Punish Its Citizens. URL: http://time.com/collection/davos-2019/5502592/china-social-credit-score/(27.05.2019).

Chouldechova, Alexandra (2016): Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. In: FATML, URL: http://arxiv.org/abs/1610.07524 (27.05.2019).

Cockerill, Kristan et al. (2004): Assessing Public Perceptions of Computer-Based Models. In: Environmental Management 34 (5), S. 609–619.

Colquitt, Jason A. (2001): On the dimensionality of organizational justice: A construct validation of a measure. In: Journal of Applied Psychology 86 (3), S. 386–400.

Colquitt, Jason A. et al. (2001): Justice at the millennium: A meta-analytic review of 25 years of organizational justice research. In: Journal of Applied Psychology 86 (3), S. 425–445.

Cropanzano, Russell et al. (2015): Three Roads to Organizational Justice. In: Research in Personnel and Human Resources Management 20, S. 1–113.

Dencik, Lina et al. (2019): Exploring Data Justice: Conceptions, Applications and Directions. In: Information, Communication & Society 22 (7), S. 873–881.

Deutsch, Morton (1975): Equity, Equality, and Need: What Determines Which Value Will Be Used as the Basis of Distributive Justice? In: Journal of Social Issues 31 (3), S. 137–149.

Evans, Jonathan St. B T. (1989): Bias in human reasoning: Causes and consequences. Hillsdale.

Friedler, Sorelle A. et al. (2016): On the (Im)Possibility of Fairness. URL: http://arxiv.org/abs/1609.07236 (27.05.2019).

Gajane, Pratik/Pechenizkiy, Mykola (2017): On formalizing fairness in prediction with machine learning. In: ArXiv E-Prints, arXiv:1710.03184. URL: https://arxiv.org/abs/1710.03184v3 (27.05.2019).

Greenberg, Jerald (1990): Organizational justice: Yesterday, today, and tomorrow. In: Journal of Management 16 (2), S. 399–432.

Greenberg, Jerald (1993): The social side of fairness: Interpersonal and informational classes of organizational justice. In: Cropanzano (Hg.): Justice in the workplace: Approaching fairness in human resource management. Hillsdale, S. 79–103.

Grgic-Hlaca, Nina et al. (2018): Beyond Distributive Fairness in Algorithmic Decision Making: Feature Selection for Procedurally Fair Learning. In: The Thirty-Second AAAI Conference on Artificial Intelligence, S. 51–60.

Grgic-Hlaca, Nina et al. (2018): Human Perceptions of Fairness in Algorithmic Decision Making. In: Proceedings of the 2018 World Wide Web Conference, S. 903–912.

Haugeland, John (1985): Artificial Intelligence: The Very Idea. Cambridge.

Herbst, Susan (2010): Rude democracy: Civility and incivility in American politics. Philadelphia.

Hoffmann, Anna L. (2019): Where Fairness Fails: Data, Algorithms, and the Limits of Antidiscrimination Discourse. In: Information, Communication & Society 22 (7), S. 900–915.

Holzinger, Andreas (2018): Explainable AI (ex-AI). In: Informatik-Spektrum 41 (2), S. 138–143.

Hu, Lily/Chen, Yiling (2018): Welfare and distributional impacts of fair classification. In: ArXiv E-Prints, arXiv:1807.01134. URL: https://arxiv.org/abs/1807.01134 (27.05.2019).

Kaplan, Jerry (2016): Artificial Intelligence: What Everyone Needs to Know. Oxford.

Kersting, Norbert (2017): Open Data, Open Government und Online Partizipation in der Smart City. Vom Informationsobjekt über den deliberativen Turn zur Algorithmokratie? In: Buhr/Hammer/Schölzel (Hg.): Staat, Internet und digitale Gouvernementalität. Wiesbaden, S. 87–104.

Kersting, Norbert (2019): Online Partizipation: Evaluation und Entwicklung– Status Quo und Zukunft. In: Hofmann et al. (Hg.): Politik in der digitalen Gesellschaft. Bielefeld, S. 105–121.

Knight, Jack (1998): Justice and Fairness. In: Annual review of Political Science 1, S. 425–449.

Kossow, Niklas (2019): Blockchain: viel Potential, begrenzte Umsetzbarkeit. In: Skutta et al. (Hg.): Digitalisierung und Teilhabe. Baden-Baden, S. 97–112.

Kroll, Joshua A. et al. (2017): Accountable algorithms. In: University of Pennsylvania Law Review 165 (3), S. 633–705.

Lee, Min Kyung (2018): Understanding Perception of Algorithmic Decisions: Fairness, Trust, and Emotion in Response to Algorithmic Management. In: Big Data & Society 5(1), S. 1–16.

Lee, Min Kyung/Baykal, Su (2017): Algorithmic Mediation in Group Decisions. In: Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing, S. 1035–1048.

Legg, Shane/Hutter, Marcus (2007): A Collection of Definitions of Intelligence. In: Frontiers in Artificial Intelligence and Applications 157, S. 17–24.

Leventhal, Gerald. S. (1980): What should be done with equity theory? New approaches to the study of fairness in social relationships. In: Gergen et al. (Hg.): Social exchange: Advances in theory and research. New York, S. 27–55.

Mulgan, Richard (2003): Holding Power to Account. Accountability in Modern Democracies. Basingstoke.

Nass, Clifford/Moon, Youngme (2000): Machines and mindlessness: Social responses to computers. In: Journal of Social Issues 51 (1), S. 81–103.

Poel, Martijn et al. (2018): Big Data for Policymaking: Great Expectations, but with Limited Progress? In: Policy and Internet 10 (3), S. 347–367.

Ribeiro et al. (2016): Why Should I Trust You? Explaining the Predictions of Any Classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, S. 1135–1144.

Rieder, Gernot/Simon, Judith (2016) Datatrust Or, the Political Quest for Numerical Evidence and the Epistemologies of Big Data. In: Big Data & Society 3 (1), S. 1–6.

Scharpf, Fritz W. (1970): Demokratietheorie zwischen Utopie und Anpassung. Konstanz.

Schmidhuber, Jürgen (2015): Deep Learning in Neural Networks: An Overview. In: Neural Networks 61, S. 85–117.

Schmidt, Vivien A. (2013): Democracy and Legitimacy in the European Union Revisited: Input, Output and ›Throughput.‹ In: Political Studies, 61 (1), S. 2–22.

Shapiro, Debra L. et al. (1991): Explanations: When are they judged adequate? In: Academy of Management Proceedings (1), S. 395–399.

Shapiro, Debra. L. et al. (1994): Explanations: What factors enhance their perceived adequacy? In: Organizational Behavior and Human Decision Processes 58 (3), S. 346–368.

Sharma, Shubham et al. (2019): CERTIFAI: Counterfactual explanations for robustness, transparency, interpretability, and fairness of artificial intelligence models. In: ArXiv E-Prints, arXiv:1905.07857. URL: https://arxiv.org/abs/1905.07857 (27.05.2019).

Skarlicki, Daniel P./Folger, Robert (1997): Retaliation in the workplace: The roles of distributive, procedural, and interactional justice. In: Journal of Applied Psychology 82 (3), S. 434–443.

Stone, Emily F./Stone, Lawrence D. (1990): Privacy in organizations: Theoretical issues, research findings, and protection mechanisms. In: Rowland/Ferris (Hg.): Research in personnel and human resources management (8). Greenwich, S. 349–411.

Tata, Jasmine/Bowes-Sperry, Lynn (1996): Emphasis on distributive, procedural, and interactional justice: Differential perceptions of men and women. In: Psychological Reports 79 (3), S. 1327–1330.

Taylor, Linnet (2017): What Is Data Justice? The Case for Connecting Digital Rights and Freedoms Globally. In: Big Data & Society 4 (2), S. 1–14.

Thibaut, John/Walker, Laurens (1975): Procedural Justice: A Psychological Analysis. Hillsdale.

Thierer, Adam et al. (2017): Artificial Intelligence and Public Policy. URL: https://www.mercatus.org/publications/artificial-intelligence-public-policy (27.05.2019).

Tyler, Tom R. et al. (1997): Social Justice in a Diverse Society. Boulder.

Tyler, Tom R./Bies, Robert J. (1990): Beyond formal procedures: The interpersonal context of procedural justice. In: Carroll (Hg.): Applied social psychology and organizational settings. Hillsdale, S. 77–98.

Veale, Michael/Binns, Reuben (2017): Fairer Machine Learning in the Real World: Mitigating Discrimination without Collecting Sensitive Data. In: Big Data & Society 4 (2), S. 1–17.

Verma, Sahil/Rubin, Julia (2018): Fairness Definitions Explained. ACM/IEEE International Workshop on Software Fairness. S. 1–7.

Verba, Sidney (2006): Fairness, equality, and democracy: three big words. In: Social Research: An International Quarterly of Social Sciences 73 (2), S. 499–540.

Wachter, Sandra et al. (2017): Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation. In: International Data Privacy Law 76, S. 79–90.

Wachter, Sandra et al. (2018): Counterfactual explanations without opening the black box: Automated decisions and the GDPR. In: Harvard Journal of Law and Technology 31 (2), S. 841–887.

Wahl, Brian et al. (2018): Artificial Intelligence (AI) and Global Health: How Can AI Contribute to Health in Resource-Poor Settings? In: BMJ Global Health 3(4).

Westin, Alan F. (1967): Privacy and Freedom. New York.

Yean, Tan Fee/Yusof, Ab Aziz (2016): Organizational Justice: A Conceptual Discussion. In: Procedia-Social and Behavioural Sciences 219, S. 798–803.

YouGov (2018): Künstliche Intelligenz: Deutsche sehen eher die Risiken als den Nutzen. URL: https://yougov.de/news/2018/09/11/kunstliche-intelligenz-deutsche-sehen-eher-die-ris/ (27.05.2019).