Österreich Z Soziol (2019) (Suppl 1) 44:53–66
https://doi.org/10.1007/s11614-019-00347-2

ÖZS

HAUPTBEITRÄGE

Maschinelles Lernen und Diskriminierung: Probleme und Lösungsansätze

Thilo Hagendorff

Zusammenfassung Zwischen menschlichem Handeln und technischen Artefakten besteht eine permanente Wechselwirkung, welche sich unter anderem in Form von Wertübertragungsprozessen manifestiert. Besondere Beachtung findet dieser Sachverhalt in den vergangenen Jahren insbesondere im Kontext digitaler Informations- und Kommunikationssysteme und hierbei wiederum im Besonderen im Rahmen der Anwendung von Verfahren des maschinellen Lernens. Der vorliegende Aufsatz trägt Beispiele für Wertübertragungsprozesse in diesem Kontext zusammen und geht dabei insbesondere auf das Thema der algorithmischen Diskriminierung ein. Beschrieben werden Ursachen für diese Form der Diskriminierung. Anschließend werden konkrete Maßnahmen vorgestellt, anhand derer das Ziel eines nichtdiskriminierenden Einsatzes von Techniken des Maschinenlernens erreicht werden kann.

Schlüsselwörter Maschinenlernen · Diskriminierung · Anti-Diskriminierung · Werte · Algorithmen

Machine Learning and Discrimination: Problems and Solutions

Abstract Between human action and technical artefacts there is a permanent interplay, which manifests itself, among other things, in the form of value transfer processes. In recent years, this issue has received particular attention in digital information and communication systems and, in particular, in the use of machine learning applications. This paper brings together examples of value transfer processes in this context and deals with the issue of algorithmic discrimination. Causes for this form of discrimination are described. Subsequently, the paper outlines con-

T. Hagendorff (✉)
Universität Tübingen, Tübingen, Deutschland
E-Mail: thilo.hagendorff@uni-tuebingen.de

crete measures that can be used to achieve the goal of a non-discriminatory use of machine learning techniques.

Keywords Machine learning · Discrimination · Anti-discrimination · Values · Algorithms

1 Einleitung

Auf den Prozess der Entwicklung und Anwendung technischer Artefakte wirken stets soziale Werte ein – während umgekehrt die technischen Artefakte einen Einfluss auf soziale Werte zeitigen (Rammert und Schulz-Schaeffer 2002). In der techniksoziologischen und technikethischen Forschung ist dabei insbesondere die Wirkungsrichtung der Einschreibung von Werten in Techniken beleuchtet worden (Beveridge et al. 2003; Brey 2010). Dieser Prozess lässt sich gleichsam im Kontext der Technologie des maschinellen Lernens beobachten. Das Besondere hierbei ist zum einen das hohe Maß der Intransparenz, welches jene Werteinschreibungsprozesse kennzeichnet. Zum anderen haben Anwendungen, bei denen Verfahren des maschinellen Lernens zur Anwendung kommen, in viele Fällen eine große gesellschaftliche Reich- beziehungsweise Tragweite, etwa wenn es um Software geht, welche im Rechtssystem, bei großen Unternehmen oder bei Plattformen wie digitalen sozialen Netzwerken zur Anwendung kommen. Dieser Umstand macht es erforderlich, die Entwicklung und Anwendung von Verfahren des maschinellen Lernens trotz des Anscheins der damit verbundenen Objektivität (Boyd und Crawford 2012, S. 666) daraufhin zu prüfen, ob es Verzerrungen oder Voreingenommenheit (Bias) gibt und ob Formen negativer sozialer Diskriminierung bedingt oder gar forciert werden (Introna und Wood 2004).

Der folgende Aufsatz soll aus einer techniksoziologischen sowie technikethischen Perspektive beispielhaft aufzeigen, wie Diskriminierung und ungerechte Computeroutputs im Kontext von Anwendungen des maschinellen Lernens entstehen. Darüber hinaus soll er Möglichkeiten zur Verhinderung von derartiger Diskriminierung schildern. Dabei werden verschiedene Fallbeispiele beschrieben, bevor im Schlussteil des Aufsatzes eine Übersicht über bestehende technische, individuelle sowie organisatorische Maßnahmen gegeben wird, welche ergriffen werden können, um dem Problem der algorithmischen Diskriminierung zu begegnen. Das Ziel des Aufsatzes besteht schließlich darin, zu zeigen, mit welchen Rahmenbedingungen ein gerechter Einsatz von Techniken des Maschinenlernens gefördert und sichergestellt werden kann. Hierbei soll jedoch keinesfalls ein essentialistischer Begriff von Gerechtigkeit und Nicht-Diskriminierung zur Anwendung kommen. Vielmehr sollen beide Begriffe stets so verstanden werden, dass sie sich aus sozialen Praktiken ergeben und kontextspezifisch interpretiert werden müssen. Anhand dieser Tatsache kann zudem die Angewiesenheit der eher technikzentriert agierenden Computerwissenschaften auf techniksoziologische sowie technikethische Überlegungen verdeutlicht werden.

2 Diskriminierung

Der Konnex aus digitalen Technologien und sozialer Diskriminierung wird in verschiedenen wissenschaftlichen Disziplinen wie den Critical Data Studies, der Technikphilosophie und -soziologie oder den Rechtswissenschaften immer wieder intensiv diskutiert (Barocas und Selbst 2016; O'Neil 2016; Crawford und Schultz 2014; Richards und King 2014; Citron und Pasquale 2014; Kerr und Earle 2013). Das Besondere der datenbasierten Diskriminierung ist dabei, dass aufgrund der inhärenten Komplexität von Datenverarbeitungsverfahren eine besondere Intransparenz im Hinblick auf die Art und Weise der Benachteiligung vorliegt. Hinzu kommt, dass datenbasierte Diskriminierung oftmals nicht intentional gewollt ist, sondern schlicht aus einem „ungünstigen" Zusammenspiel unterschiedlichster Faktoren entsteht, welche im weiteren Verlauf des Aufsatzes noch detaillierter beschrieben werden.

Der Begriff der „Diskriminierung" steht in der Regel für negative soziale Diskriminierung und damit in einem Zusammenhang mit sozialer Ungerechtigkeit. Im Hinblick auf die Analyse großer Datenmengen kann freilich angemerkt werden, dass der wesentliche Sinn solcher Analysen darin besteht, in dem Sinn zu diskriminieren, dass verschiedene Eigenschaften oder Merkmale voneinander differenziert, Regelmäßigkeiten, Korrelationen und Muster erkannt, Cluster gebildet oder Datenpunkte sortiert werden. Eine solche grundlegende Diskriminierung zwischen verschiedenen Eigenschaften, Mustern, Clustern et cetera führt nicht zwangsläufig zu einer ungerechten sozialen Diskriminierung (Hellman 2011), zumal dann nicht, wenn keine personenbezogenen oder sensiblen Daten verwendet werden. Doch auch mit der Verarbeitung und Analyse personenbezogener Daten geht nicht automatisch eine ungerechte soziale Diskriminierung einher. Diese entsteht erst dann, wenn die aus Datenauswertungsverfahren heraus entstehenden Differenzierungen als ungerecht angesehen werden und Handlungsentscheidungen an Persönlichkeitsmerkmalen orientiert werden, welche in keinem relevanten Zusammenhang mit jener Entscheidung stehen. Klassischerweise stehen solche nicht-entscheidungsrelevanten Merkmale im Zusammenhang mit Kategorien wie Geschlecht, Alter, ethnischer oder nationaler Zugehörigkeit, Aussehen, Behinderung, Schwangerschaft und einigen mehr. Wenn Entscheidungen – beispielsweise im Bereich der Arbeitsplatzvergabe, des Versicherungswesens, der Kreditvergabe et cetera – von einem oder mehreren der beispielhaft genannten Merkmale abhängig gemacht werden, ohne dass ein angemessener Grund dafür bestünde, kann von ungerechter, negativer sozialer Diskriminierung gesprochen werden. Auch wenn im Folgenden der Einfachheit halber nur der Begriff „Diskriminierung" ohne weiteren Zusatz verwendet wird, soll damit genau auf diesen negativen Begriff von Diskriminierung referiert werden.

Im Kontext des Begriffs der „Diskriminierung" kann ferner eine Differenzierung eingezogen werden zwischen direkten und indirekten Formen der Diskriminierung bei der Verwendung von Big Data beziehungsweise der Verarbeitung personenbezogener Daten (Hajian und Domingo-Ferrer 2013a). Direkte Diskriminierung tritt dann auf, wenn Entscheidungen unmittelbar von sensiblen Informationen etwa über das Geschlecht, die sexuelle Orientierung oder die Ethnizität einer Person abhängig gemacht werden. Indirekte Diskriminierung tritt dann auf, wenn Entscheidungen in Abhängigkeit von nicht-sensiblen Informationen stehen, welche jedoch in ei-

nem starken Korrelationsverhältnis zu sensiblen Informationen stehen. Dies kann dann passieren, wenn beispielsweise von der nicht-sensiblen Information einer bestimmten Postleitzahl mit hoher Wahrscheinlichkeit auf die sensible Information der Zugehörigkeit zu einer bestimmten ethnischen Minderheit geschlossen werden kann („redlining“).

3 Wertübertragung

Im Zuge der Entwicklung und Benutzung digitaler Technologien und Plattformen – insbesondere solcher, welche mit großen Datenmengen sowie Verfahren des maschinellen Lernens arbeiten – werden, wie oben bereits angesprochen, Konglomerate sozialer Wertannahmen auf die Verfasstheit der Technik übertragen (Beveridge et al. 2003; Brey 2010; Bozdag 2013). Es gibt eine gegenseitige Konstitution zwischen Gesellschaft und Technik, wobei genaugenommen beide Seite nicht voneinander zu trennen sind (Latour 2014). Dieser Komplex ist insbesondere in der Techniksoziologie beziehungsweise in den Science and Technology Studies untersucht und beschrieben worden (Aarden und Barben 2013). Ein Problem, welches in den genannten Disziplinen ausführlich analysiert wurde, besteht darin, dass Prozesse der Übertragung von Wertannahmen beziehungsweise die Wertbeladenheit von Technik in der Phase der Technikanwendung nur noch mit Schwierigkeiten oder gar nicht mehr verhandelt werden kann (Latour 1999, S. 304). Während in genuin sozialen Interaktionen subjektive Wertannahmen, Überzeugungen, Dispositionen et cetera diskutiert werden und deren Kontingenz, Perspektivabhängigkeit oder Relativität hervorgehoben werden können, härten Wertannahmen, ohne dabei ihre Geltungskraft zu verlieren, in technischen Artefakten oder Verfahren gewissermaßen aus.

Die Frage ist nun, wie sich der Prozess der Übertragung von Wertannahmen von der sozialen Praxis in die Verfasstheit technischer Artefakte konkret gestaltet. Kurz gesagt: Wie schreiben sich Werte in Technik ein? Bei der Beantwortung dieser Frage kann im Wesentlichen auf zwei Szenarien referiert werden, wobei der Einschreibungsprozess jeweils unterschiedlich gut kausal nachvollzogen werden kann.

Das erste Szenario beschreibt den Vorgang, bei welchem Technikentwickler direkt diskriminierende Vorannahmen in technische Artefakte oder Verfahren einschreiben. Mit Friedman und Nissenbaum könnte man hier von „preexisting biases“ sprechen (Friedman und Nissenbaum 1996). Ein Beispiel für einen solchen Einschreibungsprozess von „preexisting biases“ lässt sich etwa bei Körperscannern finden – einer Technik, welche unter anderem an Flughäfen zur Sicherheitskontrolle eingesetzt wird (Bello-Salau et al. 2012; Ammicht Quinn et al. 2014). So kann es vorkommen, dass ein Körperscanner einen Fehlalarm ausgibt, wenn ein Mensch mit einer Prothese den Scanner betritt. Den fehlerhaften Alarm bedingt eine „Inkompatibilität“ der zu scannenden Person mit dem seitens der Technik vorgegebenen Normalkörperschema. In der Entwicklungsphase der Scanner haben die beteiligten Softwareentwickler bewusst oder unbewusst bestimmte Vorannahmen über „normale“ Körper in die Ausgestaltung der Scannersoftware einfließen lassen. In diesem Fall betreffen die Vorannahmen Aspekte der Physiognomie und der Form menschlicher Körper. Aus technikethischer Perspektive gesprochen stellt der Alarm, den der Körperscan-

ner ausgibt und welcher zur Detektion „auffälliger", „abweichender" Personen führt, das Resultat eines unterkomplexen Werteinschreibungsprozesses in die Technik dar.

Das zweite Szenario beschreibt Einschreibungsprozesse von Wertannahmen in Technik, welche weitaus komplexer und dadurch auch intransparenter sind. Wiederum mit Friedman und Nissenbaum ließe sich hier von „emerging biases" sprechen (Friedman und Nissenbaum 1996). Hier stößt man insbesondere auf Techniken im Bereich der Big Data und des maschinellen Lernens. Diese sollen im Folgenden genauer beleuchtet werden. Beim Data-Mining beziehungsweise maschinellen Lernen werden Algorithmen eingesetzt, um anhand gewisser Parameter, einem Set an Beispielen sowie der Assistenz von Programmierern aus auftretenden Mustern und Regelmäßigkeiten in Datensätzen eigenständig Wissen zu generieren (Domingos 2012). In Anlehnung an Veale und Binns (2017) lassen sich dabei grob zwei verschiedene Bereiche oder Handlungszusammenhänge identifizieren, im Rahmen derer ungerecht operierende sowie auf Verfahren des maschinellen Lernens und Big Data basierende Systeme entstehen können.

Erstens kann dies im Kontext der Verwendung von Datensätzen entstehen, die Aufzeichnungen vergangener sozialer Ereignisse und Entscheidungen sind, welche dann als Basis genommen werden zum „Training" selbstlernender Systeme. Sobald diese Systeme eigenständig laufen und „im Feld" eingesetzt werden, perpetuieren sie die aus den Trainingsdatensätzen erlernten Disparitäten. Wenn diese Formen sozialer Diskriminierung abbilden, wird diese durch die selbstlernenden Systeme schlicht fortgesetzt. Erstmals systematisch aufmerksam gemacht wurde auf dieses Problem in einem Beitrag von Pedreshi et al. (2008). Sie schreiben:

> [...] dangerously, learning from historical data may mean to discover traditional prejudices that are endemic in reality, and to assign to such practices the status of general rules, maybe unconsciously, as these rules can be deeply hidden within a classifier. (ebd., S. 560)

Zweitens können ungerecht operierende Computersysteme dadurch entstehen, dass Methoden des Maschinenlernens so konzipiert werden, dass sie Formen sozialer Diskriminierung bedingen oder verstärken. Maschinelles Lernen ist kein automatisch ablaufender Prozess, in welchem keine Personen intervenieren. Ganz im Gegenteil ist insbesondere das „supervised learning" eine Form des Maschinenlernens, bei welcher trotz aller „Autonomie" der Algorithmen faktisch immer Personen intervenieren, um in den Prozess des maschinellen Lernens assistierend einzugreifen. Das Design sowie die konkrete Anwendungspraxis von Lernalgorithmen ist demnach unmittelbar von Personen sowie deren subjektiven Überzeugungen, Wert- sowie Vorannahmen beeinflusst. Freilich beeinflusst auch die Wahl des für einen bestimmten Zweck zu verwendenden Maschinenlernparadigmas – überwachtes, nichtüberwachtes oder verstärkendes Lernen (Mueller und Massaron 2016, S. 168 ff.) – die Outputs der jeweils verwendeten Anwendung. Hinzu kommen Aspekte wie beispielsweise subjektive Entscheidungen über Schwellenwerte, welche definieren, wann bestimmte Datenpunkte positive oder negative Werte darstellen. So etwas wie „neutrale" Entscheidungen gibt es dabei nicht (Winner 1980).

4 Beispiele für algorithmische Diskriminierung

In verschiedenen wissenschaftlichen oder journalistischen Arbeiten sind immer wieder konkrete Beispiele für problematische algorithmische Diskriminierungen insbesondere im Kontext von Big-Data-Anwendungen oder Anwendungen des maschinellen Lernens beschrieben worden. Zwar gibt es auch Ansätze, mit speziellen Analysen großer Datenmengen soziale Ungerechtigkeit zu reduzieren (Hermanin und Atanasove 2013). Die Idee dabei ist, dass durch umfassende Datenerhebungen auch Profile über Personen angelegt werden können, welche Behinderungen haben, ethnischen Minoritäten angehören, aufgrund ihres Geschlechts diskriminiert werden et cetera. Eine auf diesen Profilen basierende Datenanalyse könnte dazu eingesetzt werden, um Diskriminierung zu erkennen, etwa in Bezug auf die Bildungs-, Wohnungs- oder Arbeitsmarktsituation der Menschen. So steht zwar auf der einen Seite die Idee, Big Data in Kombination mit modernen Datenverarbeitungsverfahren dazu einzusetzen, um negative soziale Diskriminierung zu erkennen und zu reduzieren. Gerade in der Politik ist in diesem Kontext zudem die (naive) Hoffnung entstanden, bessere, auf numerischer Evidenz beruhende Entscheidungen treffen zu können (Rieder und Simon 2016).

Unabhängig davon werden informationstechnische Systeme daraufhin überprüft, ob sie, vermittelt über Verfahren des maschinellen Lernens oder andere komplexe Arten der Datenverarbeitung und Datenauswertung, Formen sozialer Diskriminierung bedingen oder fördern (Mittelstadt et al. 2016, S. 8 f.). In diesem Kontext sollen drei Beispiele für algorithmische Diskriminierung im Folgenden genauer beleuchtet werden. Es geht um die Software „COMPAS“, den Chatbot „Tay“ sowie den durch Algorithmen entschiedenen Schönheitswettbewerb „Beauty.AI“. Die Auswahl der Beispiele folgt dabei dem Umstand, dass bei den genannten Applikationen durch detaillierte Analysen der Entstehungskontext der algorithmischen Diskriminierung offengelegt werden konnte.

Besondere mediale Prominenz erlangte eine Recherche von „ProPublica“, in welcher Julia Angwin et al. (2016) herausgefunden haben, dass in der amerikanischen Strafverfolgung dunkelhäutige Menschen durch den Einsatz von „legal tech“ benachteiligt werden. Im Speziellen ging es bei der Recherche um die Software „COMPAS“ („Correctional Offender Management Profiling for Alternative Sanctions“). Bei „COMPAS“ handelt es sich um ein Programm, welches eingesetzt wird, um unter anderem die Rückfallgefahr von Straftätern zu berechnen. Die Software benachteiligt jedoch dunkelhäutige Menschen, indem sie für diese eine höhere Rückfallgefahr berechnet als für vergleichbare hellhäutige Menschen. Der Grund für diesen offensichtlichen Bias in der Software erklärt sich aus dem zur Anwendung gebrachten Verfahren des Maschinenlernens. „COMPAS“ errechnet für jeden Straftäter einen Risikowert. Dieser Wert errechnet sich auf der Grundlage einer Befragung, welche jeder Inhaftierte ausfüllen muss. Die Antworten werden dann in eine Datenbank eingespeist und anschließend mit Eigenschaften von in der Vergangenheit verurteilten Straftätern abgeglichen. Dabei kategorisiert die Software zwar nicht die Hautfarbe der Angeklagten. Was sie allerdings kategorisiert, sind Aspekte wie soziale Netzwerke, Arbeitslosigkeit, Einkommen, Wohnsituation, Beziehungsstatus der Eltern, Straftaten von Freunden und Verwandten et cetera. Das Risiko für Straffälligkeit

wird dann unter anderem an diesen Faktoren festgemacht. Genau hinsichtlich dieser Faktoren besteht jedoch von vornherein bereits eine Benachteiligung von dunkelhäutigen Menschen. Letztlich lernt der Algorithmus auch hier gesellschaftlich etablierte Ungleichheiten – und er perpetuiert diese dann, indem er jene Bevölkerungsgruppen benachteiligt, welche ohnehin bereits sozial benachteiligt sind. Das Problematische daran ist jedoch, dass die Benachteiligung plötzlich nur noch sehr schwer sichtbar gemacht werden kann.

Ein zweites Beispiel wäre „Tay", ein intelligenter „Chat-Bot". Entwickelt wurde „Tay" von Microsoft, und man konnte etwa über Twitter nach eigenem Belieben mit „Tay" chatten. Was Microsoft bei der Anwendung von „Tay" nicht antizipiert hatte, war, dass es kurz nach dem Start der Software eine konzertierte Aktion von Nutzern der Plattform „4chan" gab. Diese verabredeten sich untereinander, um sich mit „Tay" zu „unterhalten", allerdings in einer Art der Kommunikation, welche gefärbt war von Rassismus, Sexismus, Antisemitismus et cetera. Die konzertierte Aktion endete darin, dass „Tay", welcher im Grunde nichts anderes als ein komplexer Lernalgorithmus war, sich den Rassismus, Sexismus, Antisemitismus et cetera aneignete. Dies bedeutete, dass „Tay" automatisch Aussagen wie etwa „Hitler did nothing wrong!" und dergleichen mehr generierte (Misty 2016).

Ein drittes Beispiel wäre „Beauty.AI", ein durch Algorithmen entschiedener Schönheitswettbewerb (Levin 2016). Eine Software wurde dazu eingesetzt, um Hautfalten, Augenposition, Gesichtssymmetrie et cetera zu bewerten. An dem Schönheitswettbewerb nahmen 6000 Menschen aus über 100 Ländern teil, wobei das Endergebnis dergestalt war, dass unter den 44 Gewinnern aus den verschiedenen Altersklassen quasi keine dunkelhäutigen Menschen waren. Gleiches galt für asiatisch aussehende Frauen und Männer. Der Grund für diesen offensichtlich diskriminierenden Ausgang des Wettbewerbs lag darin, dass mit Deep Learning eine Methode des maschinellen Lernens eingesetzt wurde, welche vorerst über einen bestimmten Trainingsdatensatz Attraktivitätsstandards lernen musste. Das Problem dabei war allerdings, dass innerhalb des eingesetzten Trainingsdatensatzes kaum Bilder von Menschen waren, welche ein anderes als das amerikanische beziehungsweise europäische Schönheitsideal verkörperten. Freilich hatten die Entwickler des „Beauty.AI"-Codes diesem nicht einprogrammiert, helle Haut gegenüber dunkler Haut als ein Zeichen von Schönheit zu sehen. Dennoch eignete der Algorithmus sich genau dies an und präferierte so Menschen mit heller Haut gegenüber Menschen mit dunkler Haut. Obwohl der zur Anwendung gebrachte Lernalgorithmus vorerst quasi „neutral" agiert, eignet er sich aus einem bestimmten Trainingsdatensatz an, bestimmte Eigenschaften – in diesem Fall von Gesichtern – zu präferieren. Dies bedeutet, dass die Software einen Bias aus dem Trainingsdatensatz übernimmt, während der Trainingsdatensatz wiederum ein Abbild des Bias der Technikentwickler ist. Das Problem dabei ist jedoch wiederum, dass dieser komplexe Übertragungsprozess in der Phase der Technikanwendung kaum mehr nachvollzogen werden kann.

5 Anti-Diskriminierung

Der folgende Abschnitt soll sowohl technisch als auch organisatorisch gelagerte Methoden der Anti-Diskriminierung im Kontext des maschinellen Lernens beleuchten. Freilich kann hierbei keine klare Trennlinie zwischen rein technisch und rein organisatorisch ausgerichteten Ansätzen gezogen werden. Dennoch lassen sich Methoden differenzieren, welche eher auf die Einwirkung auf die technische Verfasstheit von Machine-Learning-Anwendungen zielen, und Methoden, welche eher organisatorische Maßnahmen vorsehen, um die Wahrscheinlichkeit des Auftretens algorithmischer Diskriminierung zu verhindern.

5.1 Technische Ansätze

Unter den beiden Schlagwörtern „DADM" („discrimination-aware data-mining") sowie „FATML" („fairness, accountability and transparency machine learning") haben sich interdisziplinäre wissenschaftliche Forschungsgemeinschaften, welche insbesondere aus Vertretern der Informatik, aber auch der IT-Industrie bestehen, herausgebildet, welche Phänomene algorithmischer Diskriminierung sowie entsprechende Lösungsansätze beforschen. So ist in verschiedenen wissenschaftlichen Forschungszweigen die Tatsache, dass der Einsatz von komplexen Algorithmen respektive von Methoden des maschinellen Lernens Formen sozialer Diskriminierung perpetuiert oder sogar verstärkt, immer wieder betont worden. Besonders problematisch wird dies, wenn Verfahren des Maschinenlernens in Bereichen wie etwa dem Scoring (Khandani et al. 2010), dem Polizeiwesen (Brennan et al. 2008) oder der Kreditvergabe (Mahoney und Mohen 2007; Rothmann et al. 2014) zur Anwendung kommen. Zwar gibt es in diesem Zusammenhang eine Reihe an Forschungsarbeiten darüber, wie über statistische oder mathematische Verfahren konkrete Lösungen für das Problem ungerechter Diskriminierungen erarbeitet werden können (Calders und Verwer 2010; Dwork et al. 2011; Kamiran und Calders 2012; Zafar et al. 2017). Dennoch ist keines der entwickelten Verfahren perfekt, zumal nicht im Einsatz in der „echten Welt" mit entsprechend „chaotischen" und schwer zu bereinigenden Datensätzen (Kleinberg et al. 2016). Die Auswahl zwischen den verschiedenen mathematischen Verfahren für einen fairen Einsatz der Technik des maschinellen Lernens muss kontextabhängig passieren (Musik 2011, S. 350), da verschiedene Verfahren verschiedene Vor- und Nachteile besitzen.

Ein Vorschlag, wie er beispielsweise von Faisal Kamiran et al. (2013) stammt, zielt speziell auf Trainingsdatensätze ab und umfasst Maßnahmen wie etwa die gezielte Bereinigung derselben, sodass im Zuge des Trainings von Lernalgorithmen Formen sozialer Diskriminierung gar nicht erst vom Computer erlernt werden können. Ferner geht es darum, Algorithmen in der Lernphase nur mit solchen Daten beziehungsweise Aufzeichnungen vergangener sozialer Handlungszusammenhänge in Kontakt kommen zu lassen, welche in keinem Zusammenhang mit ungerechter Diskriminierung stehen. Problematisch ist jedoch, dass im Zuge einer solchen „pre-processing discrimination prevention" die gezielte Eliminierung von Attributen gleichsam dazu führen kann, dass die Qualität und damit die Nützlichkeit der Datenanalysen sinkt (Hajian und Domingo-Ferrer 2013b, S. 243).

Ein weiteres Konzept, nämlich das der „individual fairness“, sieht vor, dass Personen, welche unabhängig von denjenigen Attributen, welche sich auf sensible beziehungsweise geschützte Informationen beziehen, gleich sind, auch gleiche Ergebnisse bei maschinellen Entscheidungsfindungsprozessen erhalten müssen. Eine Weiterentwicklung dieses Ansatzes stellt das Konzept der „counterfactual fairness“ dar (Kusner et al. 2017), bei welcher, grob gesagt, geprüft wird, ob beispielsweise eine Frau dieselben Ergebnisse maschineller Entscheidungsfindungsprozesse erhielte, wenn sie in den verwendeten Datenbanken als Mann verzeichnet wäre.

Ebenfalls gibt es die Idee, über die Verwendung von Methoden des „unsupervised learning“ Algorithmen zu entwickeln, welche den fairen Einsatz von Big-Data-Anwendungen fördern. Dabei werden Datensätze auf Muster hin abgesucht, welche bei einer späteren Verwendung zu Diskriminierung führen könnten. Diese „explorative data analysis“ (EDA) geht der „confirmatory data analysis“ (CDA) voraus und dient dem Zweck, unerwartete und irreführende Muster, Trends oder Abweichungen in Datensätzen zu erkennen, um letztlich auch die Ergebnisse der „confirmatory data analysis“ besser interpretieren und bewerten zu können (Behrens 1997). Über eine explorative Datenanalyse kann ferner die Struktur der vorliegenden Datensätze erkannt werden, wichtige Variablen können identifiziert sowie Fehler oder fehlende Datenpunkte ausgemacht werden. Diesem Vorgehen liegt jedoch stets die Annahme zugrunde, dass Verzerrungen oder Diskriminierung sich aus den jeweils verwendeten Datensätzen ergeben, was nicht in allen Fällen zutrifft. Unabhängig davon kann eben zur explorativen Datenanalyse das Verfahren des „unsupervised learning“ eingesetzt werden, da genau mit dieser Methode unbekannte Muster und Strukturen in Datensätzen identifiziert werden können. Ein solcher Ansatz kann unilateral eingesetzt werden. Es sind keine Drittorganisationen oder Akteure neben demjenigen Akteur, welcher die Datenanalyse durchführt, erforderlich.

Sämtliche der auf technische Verfahren abzielenden „discrimination prevention methods“ können in drei Kategorien eingeteilt werden, nämlich „pre-processing“, „in-processing“ sowie „post-processing“ (Hajian und Domingo-Ferrer 2013b, S. 247). Das „pre-processing“ schließt sämtliche Methoden ein, bei welchen Datensätze mit verschiedenen Verfahren so bereinigt oder manipuliert werden, dass anschließend ansetzende Verfahren des maschinellen Lernens keine diskriminierenden Outputs erzeugen sollen. Basis dafür ist ein Datensatz möglichst ohne Bias. Das „in-processing“ zielt auf eine Manipulation von Data-Mining-Algorithmen, und zwar in einer solchen Weise, dass die zur Anwendung kommenden Entscheidungsmodelle keine ungerechten Outputs liefern (Calders und Verwer 2010). Beim „post-processing“ schließlich geht es darum, die aus dem Trainingsprozess resultierenden Maschinenlernmodelle so zu modifizieren, dass technische Diskriminierung möglichst verhindert wird.

5.2 Organisatorische Ansätze

Neben den insbesondere auf mathematische Verfahren fokussierten „fairness-aware machine learning techniques“ ist es wichtig, sich ebenfalls auf organisatorische Ansätze zu fokussieren. Hier sind Vorschläge, wie konkrete Lösungen für Antidiskriminierungsmaßnahmen aussehen können, allerdings eher rar gesät.

Dennoch hat die Forschung auch hier inzwischen zumindest einige wenige Konzepte ausgearbeitet (Veale und Binns 2017; Citron und Pasquale 2014; Hajian und Domingo-Ferrer 2013b; Musik 2016).

So steht etwa der Vorschlag im Raum, maschinelles Lernen stärker mit der Idee sozialer Gerechtigkeit in Einklang zu bringen, indem von Organisationen, welche personenbezogene Daten verarbeiten und für maschinelles Lernen einsetzen, verlangt wird, sensible Informationen über Personen wie deren Geschlecht, Ethnizität, sexuelle Orientierung, politische Überzeugungen, Angaben zum Gesundheitszustand oder ähnliche Informationen schlicht nicht zu speichern und damit auch nicht in Datenverarbeitungsprozesse zu integrieren – ähnlich wie dies auch die Idee des Erforderlichkeits- und Verhältnismäßigkeitsgrundsatzes im Datenschutzrecht vorsieht. Dieses Vorgehen entspricht dem Konzept der „fairness through unawareness". Hierbei dürfen Attribute, welche sich auf sensible beziehungsweise geschützte Informationen beziehen, nicht explizit in maschinelle Entscheidungsprozesse einfließen. Das Problem dabei ist jedoch, dass die Streichung dieser Attribute mitnichten eine Garantie dafür bietet, dass faire Computerentscheidungen getroffen werden. Dies erklärt sich dadurch, dass nicht-sensible mit sensiblen, gelöschten Attributen korrelieren können, sodass auch über die Verarbeitung nicht-sensibler Attribute Diskriminierung entstehen kann.

Eine weitere Idee im Rahmen organisatorischer Ansätze besteht darin, Drittorganisationen in Form von Nichtregierungsorganisationen oder eventuell auch Behörden heranzuziehen, welche prüfend, beratend oder sanktionierend die Datenverarbeitungspraxis bei Unternehmen oder staatlichen Institutionen überwachen (Tutt 2017). So können Organisationen, welche sensible Informationen in ihre Datenverarbeitungsverfahren einbeziehen, vertrauenswürdige Drittorganisationen angegliedert werden. Diese speichern die sensiblen Informationen ebenfalls auf eigenen Speichermedien und können bei Bedarf modellartig nachvollziehen, ob im Zuge der originalen Datenverarbeitungsmaßnahmen ungerechte Diskriminierung stattfand. Dabei werden nicht nur die zum Einsatz kommenden Algorithmen evaluiert und auditiert, sondern ferner auch Trainingsdatensätze sowie Methoden der Datensammlung. Differenziert werden müsste außerdem, ob Drittorganisationen nur die Aufgabe haben, mögliche Diskriminierung beim Einsatz von Verfahren des maschinellen Lernens zu detektieren oder ob sie weitergehende Berechtigungen besitzen und so präventiv agieren können, um von vornherein Diskriminierung zu verhindern. In Abhängigkeit davon, ob Drittorganisationen mit erst- oder letztgenannter Aufgabe betraut werden, sind unterschiedlich stark ausgeprägte technische Kompetenzen und Expertisen bei selbigen erforderlich.

Ein weiterer Vorschlag sieht vor, dass über kollaborative Online-Plattformen gleich eine Vielzahl an Einzelpersonen oder Drittorganisationen über experimentelle Datenverarbeitungsprozesse Techniken des fairen Maschinenlernens entwickeln und fördern kann (Veale und Binns 2017). Die Idee dahinter ist, dass Kollektive an sachkundigen Personen bestimmte Verfahren und Anwendungen des maschinellen Lernens modellieren und experimentell testen können, um gleichsam Ansätze und Lösungen für diskriminierungsfreie Maschinenlerntechnologien kollektiv entwickeln und diskutieren zu können. Alles dies würde im Rahmen einer offenen, möglicherweise mit Wikipedia vergleichbaren Plattform geschehen, welche wiederum

von Nichtregierungsorganisationen oder anderen unabhängigen Akteuren administriert wird. Vorbilder für solche Kollaborationsplattformen wären neben Wikipedia Dienste wie „Git“ oder „Stack Exchange“.

6 Fazit

Bei all den genannten Beispielen aus dem Bereich der „discrimination prevention methods“ muss betont werden, dass das, was als fair, gerecht oder ethisch richtig ausgezeichnet wird und Zielgrößen des Handelns darstellt, stets in Abhängigkeit von bestimmten sozialen Kontexten gesehen werden muss. Wenn es darum geht, faire, nichtdiskriminierende Verfahren des maschinellen Lernens zu entwickeln und einzusetzen, muss erkannt werden, dass Konzepte von Fairness, Diskriminierung et cetera stets situationsspezifisch sowie kulturell und historisch variabel sind. Universell gültige Kriterien, wie Maschinenlern- beziehungsweise Computersysteme gestaltet sein müssen, um nicht sozial ungerecht zu sein, lassen sich nicht definieren. Softwareentwickler sind daher immer angehalten, ihr Handeln als eingebettet in einen bestimmten sozialen und kulturellen Kontext oder Sektor zu sehen (Brey 2010), welcher sowohl als solcher reflektiert und hinterfragt als auch als Grundlage genommen werden muss, um überhaupt entscheiden zu können, was jeweils fair, gerecht und nichtdiskriminierend ist (Bijker und Pinch 2005). Universelle, über mathematische Verfahren verbriefte Richtlinien, welche sich in Form technischer Handlungsschritte manifestieren, um Systeme diskriminierungsfrei zu machen, können nicht gefunden werden. Anders gesagt: Reine Statistik kann nicht beantworten, was Gerechtigkeit und Fairness ist. Umso aufwendiger ist es, sich die ethischen und gesellschaftswissenschaftlichen Kompetenzen und Wissensbestände anzueignen, um diese dann mit den technischen und mathematischen Kompetenzen in Verbindung zu bringen. Eine solche Aufgabe benötigt große zeitliche und intellektuelle Kapazitäten.

Anders als es für Laien vielleicht den Anschein machen mag, ist das Maschinenlernen keine Technik, welche „von alleine“ läuft, welche autonom agiert. Faktisch sind, obwohl das Maschinenlernen in einer unmittelbaren Verbindung mit der Idee der Automatisierung steht, die Einflüsse durch subjektive Vorannahmen, kulturelle Werte oder gesellschaftliche Normen nicht minder gering als bei anderen Technologien. Wenn Softwareingenieure beziehungsweise Programmierer Anwendungen entwickeln, welche auf Verfahren des maschinellen Lernens basieren, besteht ein Ziel darin, die „statistical power“ der jeweiligen Anwendung zu erhöhen, um damit eine möglichst gute Generalisierungsleistung derselben zu erreichen. Eine zentrale Aufgabe ist dabei die „feature selection“ (Domingos 2012), bei welcher es nicht um den Einsatz weiterer mathematischer oder statistischer Methoden geht, sondern um die subjektive Entscheidung für oder gegen die Angemessenheit der Beschaffenheit von Datensätzen, Lernalgorithmen und erzielten Ergebnissen. Und aus diesen Einflüssen menschlicher Akteure auf Computer beziehungsweise Softwareanwendungen erklärt sich auch die Tatsache, dass Algorithmen diskriminieren können.

Letztlich muss das Ziel bei der Entwicklung digitaler Informations- und Kommunikationssysteme darin bestehen, eine Art der Algorithmengestaltung zu finden, in welcher die ungerechtfertigte Diskriminierung von Personen oder Personengruppen

so gut es geht vermieden wird. Und wenn doch aufgedeckt werden kann, dass Algorithmen Formen ungerechter sozialer Diskriminierung fördern, dann ist es nicht unbedingt richtig, nur zu fragen, was auf der Ebene der Technik an Fehlern gemacht wurde. Vielmehr sollte man fragen, wo man sozusagen im „ideologischen Setting" einer Gesellschaft ansetzen kann (Mager 2012), um negative Diskriminierung zu bekämpfen. Denn schließlich ist dieses ideologische Setting das, was letztlich der Technik vorausgeht. Wenn von „algorithmischer Diskriminierung" die Rede ist, dann mag das suggerieren, dass Technik diejenige Instanz ist, welche diskriminiert. Doch es gibt, so betrachtet, für sich genommen keine diskriminierende Technik, sondern stets eine Verwobenheit von diskriminierender Technik und diskriminierenden Menschen, welche ihren Vorstellungen entsprechend Technik entwickeln und nutzen. Das Maschinenlernen bildet hier keine Ausnahme. Gerade beim „supervised learning" ist das Maschinenlernen kein Prozess, welcher von alleine, nur für sich läuft. Faktisch interveniert, bei aller „Autonomie" der Algorithmen, an verschiedenen Stellen der Mensch, um bei dem Prozess des maschinellen Lernens zu assistieren. Unter anderem diese Interventionen sind die Einfallstore für Maschinen, welche soziale Diskriminierung erlernen und widerspiegeln können. Und genau weil dies so ist, sind in den Computerwissenschaften in den letzten Jahren verschiedene Ideen und Konzepte entwickelt worden, um den Diskriminierungsrisiken des maschinellen Lernens zu begegnen.

Ein wichtiger Schritt wäre jedoch darüber hinaus, dass die Computerwissenschaften enger mit der Techniksoziologie sowie der Technikethik zusammenarbeiten, damit die Erkenntnisse und Wissensbestände beider wissenschaftlichen Disziplinen vereint werden können. Gerade der Themenkomplex der algorithmischen Diskriminierung im Rahmen von Anwendungen des maschinellen Lernens kann nur dann angemessen und vollumfänglich adressiert werden, wenn sowohl die genuin technische als auch die soziologische beziehungsweise technikethische Dimension des Komplexes beachtet werden. Damit die Zusammenarbeit gelingen kann, ist ein Interesse der genannten Fachbereiche füreinander sowie die Findung allseits verständlicher Vokabulare erforderlich.

Literatur

Aarden, Eric, und Daniel Barben. 2013. Science and Technology Studies. In *Konzepte und Verfahren der Technikfolgenabschätzung*, Hrsg. Georg Simonis, 35–50. Wiesbaden: Springer VS.

Ammicht Quinn, Regina, Maria Beimborn, Thilo Hagendorff, Anja Königseder, Michael Nagenborg, Magdalena Schuler, und David Schumann. 2014. *Forschungsprojekt KRETA. Abschlussbericht. Tübingen.*, 1–52.

Angwin, Julia, Jeff Larson, Surya Mattu, und Lauren Kirchner. 2016. Machine Bias. There's software used across the country to predict future criminals. And it's biased against blacks. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing. Zugegriffen: 18. Jan. 2018.

Barocas, Solon, und Andrew D. Selbst. 2016. Big data's disparate impact. *California Law Review* 104:671–732.

Behrens, John T. 1997. Principles and procedures of exploratory data analysis. *Psychological Methods* 2(2):131–160.

Bello-Salau, H., A.F. Salami, und M. Hussaini. 2012. Ethical analysis of the full-body scanner (FBS) for airport security. *Advances in Natural and Applied Science* 6:664–672.

Beveridge, Ross J., Bruce A. Draper, und David Bolme. 2003. *A statistical assessment of subject factors in the PCA recognition of human faces*. Computer Vision and Pattern Recognition Workshop, IEEE., 1–9.

Bijker, Wiebe E., und Trevor J. Pinch. 2005. The social construction of facts and artifacts. Or how the sociology of science and the sociology of technology might benefit of each other. In *The social construction of technological systems. New directions in the sociology and history of technology*, Hrsg. Wiebe E. Bijker, Thomas P. Hughes, und Trevor J. Pinch, 17–50. Cambridge: MIT Press.

Boyd, Danah, und Kate Crawford. 2012. Critical questions for Big Data. Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society* 15(5):662–679.

Bozdag, Engin. 2013. Bias in algorithmic filtering and personalization. *Ethics and Information Technology* 15(3):209–227.

Brennan, Tim, William Dieterich, und Beate Ehret. 2008. Evaluating the predictive validity of the compas risk and needs assessment system. *Criminal Justice and Behavior* 36(1):21–40.

Brey, Philip. 2010. Values in technology and disclosive computer ethics. In *The cambridge handbook of information and computer ethics*, Hrsg. Luciano Floridi, 41–58. Cambridge, Massachusetts: Cambridge University Press.

Calders, Toon, und Sicco Verwer. 2010. Three naive Bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery* 21(2):277–292.

Citron, Danielle K., und Frank Pasquale. 2014. The scored society. Due process for automated predictions. *Washington Law Review* 89:1–33.

Crawford, Kate, und Jason Schultz. 2014. Big data and Due process. Toward a framework to redress predictive privacy harms. *Boston College Law Review* 55(93):93–128.

Domingos, Pedro. 2012. A Few Useful Things to Know About Machine Learning. *Communications of the ACM* 55(10):78–87.

Dwork, Cynthia, Moritz Hardt, Toniann Pitassi, Omer Reingold, und Richard Zemel. 2011. Fairness Through Awareness. *arXiv:1104.3913*:1–24.

Friedman, Batya, und Helen Nissenbaum. 1996. Bias in computer systems. *ACM Transactions on Information Systems* 14(3):330–347.

Hajian, Sara, und Josep Domingo-Ferrer. 2013a. A Methodology for Direct and Indirect Discrimination Prevention in Data Mining. *IEEE Transactions on Knowledge and Data Engineering* 25(7):1445–1459.

Hajian, Sara, und Josep Domingo-Ferrer. 2013b. Direct and indirect discrimination prevention methods. In *Discrimination and privacy in the information society. Data mining and profiling in large databases*, Hrsg. Bart Custers, Toon Calders, Bart Schermer, und Tal Zarsky, 241–256. Berlin: Springer.

Hellman, Deborah. 2011. *When is discrimination wrong?* Cambridge: Harvard University Press.

Hermanin, Costanza, und Angelina Atanasove. 2013. Making „Big Data“ Work for Equality. http://www.opensocietyfoundations.org/voices/making-big-data-work-equality-0. Zugegriffen: 9. Jan. 2014.

Introna, Lucas D., und David Wood. 2004. Picturing Algorithmic Surveillance. The Politics of Facial Recognition Systems. *Surveillance & Society* 2(2/3):177–198.

Kamiran, Faisal, und Toon Calders. 2012. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems* 33(1):1–33.

Kamiran, Faisal, Toon Calders, und Mykola Pechenizkiy. 2013. Techniques for discrimination-free predictive models. In *Discrimination and privacy in the information society. Data mining and profiling in large databases*, Hrsg. Bart Custers, Toon Calders, Bart Schermer, und Tal Zarsky, 223–239. Berlin: Springer.

Kerr, Ian, und Jessica Earle. 2013. Prediction, preemption, presumption. How big data threatens big picture privacy. *Stanford Law Review Online* 66:65–72.

Khandani, Amir E., J. Kim Adlar, und Andrew W. Lo. 2010. Consumer credit-risk models via machine-learning algorithms. *Journal of Banking & Finance* 34(11):2767–2787.

Kleinberg, Jon M., Sendhill Mullainathan, und Manish Raghavan. 2016. Inherent trade-offs in the fair determination of risk scores. *arXiv:1609.05807*:1–23.

Kusner, Matt J., Joshua R. Loftus, Chris Russell, und Ricardo Silva. 2017. Counterfactual Fairness. *arXiv:1703.06856*:1–21.

Latour, Bruno. 1999. *Pandora's hope. Essays on the reality of science studies*. Cambridge: Harvard University Press.

Latour, Bruno. 2014. Technology is society made durable. *The Sociological Review* 38(1):103–131.

Levin, Sam. 2016. A beauty contest was judged by AI and the robots didn't like dark skin. https://www.theguardian.com/technology/2016/sep/08/artificial-intelligence-beauty-contest-doesnt-like-black-people. Zugegriffen: 10. Sept. 2016.

Mager, Astrid. 2012. Algorithmic Ideology. How capitalist society shapes search engines. *Information, Communication & Society* 15(5):769–787.
Mahoney, John F., und James M. Mohen. 2007. *Method and system for loan origination and underwriting (US7287008 B1)*
Misty, Adrienne. 2016. Microsoft creates AI bot—Internet immediately turns it racist. https://socialhax.com/2016/03/24/microsoft-creates-ai-bot-internet-immediately-turns-racist/. Zugegriffen: 17. Jan. 2018.
Mittelstadt, Brent D., Patrick Allo, Mariarosaria Taddeo, Sandra Wachter, und Luciano Floridi. 2016. The ethics of algorithms. Mapping the debate. *Big Data & Society* 3(2):1–21.
Mueller, John P., und Luca Massaron. 2016. *Machine learning for dummies*. New Jersey: John Wiley & Sons.
Musik, Christoph. 2011. The thinking eye is only half the story. High-level semantic video surveillance. *Information Polity* 16:339–353.
Musik, Christoph. 2016. *Ground Truth Studies. A Socio-Technical Framework.*, 1–7.
O'Neil, Cathy. 2016. *Weapons of math destruction. How big data increases inequality and threatens democracy*. New York: Crown Publishers.
Pedreshi, Dino, Salvatore Ruggieri, und Franco Turini. 2008. Discrimination-aware data mining. In *Proceeding of the 14th ACM SIGKDD international conference on knowledge discovery and data mining—KDD 08*, Hrsg. Ying Li, Bing Liu, und Sunita Sarawagi, 560–568. New York: ACM Press.
Rammert, Werner, und Ingo Schulz-Schaeffer. 2002. *Technik und Handeln. Wenn soziales Handeln sich auf menschliches Verhalten und technische Artefakte verteilt.*, 1–37.
Richards, Neil M., und Jonathan H. King. 2014. Big Data Ethics. *Wake Forest Law Review* 49:393–432.
Rieder, Gernot, und Judith Simon. 2016. Datatrust. Or, the political quest for numerical evidence and the epistemologies of Big Data. *Big Data & Society* 3(1):1–6.
Rothmann, Robert. 2014. *Jaro Sterbik-Lamina und Walter Peissl.*, 1–86. Wien: Credit Scoring in Österreich.
Tutt, Andrew. 2017. An FDA for Algorithms. *Administrative Law Review* 83:83–123.
Veale, Michael, und Reuben Binns. 2017. Fairer machine learning in the real world. Mitigating discrimination without collecting sensitive data. *Big Data & Society* 4(2):1–17.
Winner, Langdon. 1980. Do artifacts have politics? In: modern technology: problem or opportunity? *Daedalus* 109(1):121–136.
Zafar, Muhammad B., Isabel Valera, Manuel G. Rodriguez, und Krishna P. Gummadi. 2017. Fairness beyond disparate treatment & disparate impact. Learning classification without disparate mistreatment. *arXiv:1610.08452*:1–10.

Thilo Hagendorff studierte Philosophie, Kulturwissenschaften und Deutsche Literatur in Konstanz und Tübingen. Er promovierte 2013 mit einer soziologischen Arbeit zum Thema *Sozialkritik und soziale Steuerung*. Seit 2013 ist er wissenschaftlicher Mitarbeiter am Internationalen Zentrum für Ethik in den Wissenschaften und seit 2014 Dozent an der Universität Tübingen. Thilo Hagendorff studied philosophy, cultural studies and German literature in Konstanz and Tübingen. In 2013, he completed his Ph.D. with a sociological work entitled *Sozialkritik und soziale Steuerung* (Social Criticism and Social Control). Since 2013, he has been working as a research associate at the International Centre for Ethics in the Sciences and Humanities. Since 2014, he has also been a lecturer at the University of Tübingen.