Algorithmischer Bias: Kann ein Eingriff Neutral(er) sein?

Hausarbeit

Seminar: (Falsche) Fragen der Roboterethik WS 2019/20

Bei Lisa Schüttler M.A.

RWTH Aachen University

Daniel M. Swoboda, 378972

Informatik, B. Sc.

daniel.maximilian.swoboda@rwth-aachen.de

Abgabedatum: 28. 02. 2020

Inhaltsverzeichnis

1 Einleitung

Computer sind allgegenwärtige Begleiter der technisierten Welt und mit ihnen auch Algorithmen. In dieser nehmen Algorithmen die Rolle der Kochrezepte für Berechnungsabläufe ein, die von Computern durchgeführt werden. Jeder computergestützten Anwendung liegt ein Algorithmus zu Grunde, der deren Funktionalität beschreibt. Dabei ist es egal ob das Verhalten erst erlernt wurde (machine-learning) oder ob es durch Programmierer vorgegeben wurde (klassischer Ansatz). So vielseitig wie der Einsatz von Computern und Software ist dementsprechend auch der Einsatz von Algorithmen.

Die Anwendungsfälle von Algorithmen erstrecken sich dabei von der Kontrolle einfacher Alltagsgegenstände wie Armbanduhren, über komplexe automatische Übersetzer und Suchmaschine [Danks], bis hin zu solchen die dazu eingesetzt werden, arbeitslose Menschen zu kategorisieren [Danks] oder das Rückfallrisiko von Häftlingen zu beurteilen, um den Bewährungsprozess zu vereinfachen [Haggendorf].

Durch die vielfältigen Einsatzgebiete sind auch die potenziellen gesellschaftlichen und ethischen Implikationen von Computern und damit Algorithmen weitreichend [Bynum]. Gerade wegen ihrer weiten Verbreitung und hohen Wirkungskraft ist es interessant, den Begriff der algorithmischen Neutralität zu betrachten, wobei besonders bei Algorithmen für maschinelles Lernen, von einer gewissen Grundneutralität ausgegangen wird [Haggendorf]. Tatsächlich aber unterliegen viele Systeme nicht unerhebliche Bias, welche die Neutralität der Ausgabe beeinflussen können [Haggendorf, Danks].

Auch bei machine-learning ist algorithmischer Bias ein wichtiger Punkt in der Bewertung von computergestützten Systemen. Gerade derartige Anwendungen neigen zur Übernahme von menschlichen Bias, da sich diese in den Daten, auf denen trainiert wird, wiederfinden lassen können [Haggendorf]. Es stellt sich also ebenso wie bei Bias von Menschen die Frage nach der moralischen Verantwortung für algorithmische Bias. Ebenso stellt sich die Frage, inwiefern ein Eingriff in solche Bias die Neutralität des Algorithmus beeinflusst. Während bei Menschen ein Eingriff in die Denkweise und das Handeln die Auswirkungen von Bias reduzieren und somit neutralere Ergebnisse produzieren [Brownstein], ist dies bei einem Algorithmus, welcher durch Übertragungsprozesse von Menschen Bias übernommen hat, nicht klar [Haggendorf].

Zunächst werden in dieser Arbeit der Begriff der Bias bei Menschen eingeführt, wobei insbesondere der Fokus auf implizite Bias gelegt wird. Dann wird der Begriff des algorithmischen Bias eingeführt, Beispiele von Bias in Algorithmen vorgestellt und der Begriff der algorithmischen Neutralität diskutiert. Die

Begriffe des Bias allgemein und des algorithmischen Bias werden genutzt, um zu argumentieren, dass algorithmische Bias, ähnlich wie menschliche Bias, implizite Bias sind. Die daraus folgenden Konsequenzen für moralische Verantwortung und Neutralität werden diskutiert, bevor schlussendlich argumentiert wird, dass ein Eingriff in impliziten Bias, also auch den von Algorithmen, neutraleres Verhalten fördern kann.

2 Bias

Das Cambridge Dictionary definiert „Bias“ als „das Unterstützen oder Ablehnen einer bestimmten Person oder eines Gegenstandes auf unfaire Art und Weise, basierend auf dem Einfluss persönlicher Meinungen in der Beurteilung“ [Cambridge]. Es handelt sich also um eine nicht-objektive und nicht-logische Schlussfolgerung bedingt durch ebenso nicht-logische Überlegungen [Haselton].

Bias sind in ihrer Natur grundsätzlich nicht positiv oder negativ [Danks]. Insbesondere werden sie jedoch aufgrund ihres offensichtlichen Mangels an logischer Fundierung in vielen Kontexten als negativ angesehen [Haselton]. Im Gegenteil argumentieren Psychologen allerdings, dass es sich bei Bias nicht um Mängel in der menschlichen Wahrnehmung handelt, sondern viel mehr um wichtige Hilfsmittel im kognitiven Informationsverarbeitungsprozess [Haselton]. Als solche treten sie in einer Vielzahl von Formen auf und können in verschiedene Kategorien eingeteilt werden. Eine Einteilung nach Art der Wahrnehmung durch das betroffene Subjekt wäre die in expliziten und impliziten Bias [Amodio].

Besonders in den Bereichen „Social Cognition“ und der Kognitionswissenschaften, aber auch in Soziologie und Philosophie werden implizite Bias betrachtet. Deren Auswirkungen auf das Handeln von Akteuren, bedingt durch ihre unbewusste Natur, werfen relevante Fragen zur moralischen Verantwortung auf. [Brownstein]

2.1 Impliziter Bias

Unter implizitem Bias versteht man allgemein mentale Vorgänge und Prozesse, welche Einfluss auf Wahrnehmung und Handlungen einer Person, unabhängig der bewussten Wahrnehmung, haben. Obwohl ihr Wirken nicht im Bewusstsein stattfindet, beeinflussen sie das Verhalten von Menschen. Dies gilt besonders dann, wenn es sich um Situationen spontaner Reaktionen und ohne bewusster Überlegung handelt. [Amodio]

Traditionelle Erklärungsmodelle unterscheiden implizite Bias von expliziten Bias dadurch, dass implizite Bias nicht-deklarativer und unterbewusster Natur sind [Amodio]. Somit ist der Einfluss der Bias nicht das Ergebnis von wahr-

nehmbaren Bewusstseinsvorgänge oder deklarativer Entscheidungen, sondern vielmehr ein automatischer, unterbewusster Vorgang [Brownstein]. Als solcher entzieht sich dessen Entstehung der aktiven Kontrolle von Personen.

Obwohl sich impliziter Bias also scheinbar unserem Bewusstsein entzieht und nicht immer aktiv beeinflussen lässt, so ist er dennoch kein unzähmbarer Einfluss. So kann der Effekt von impliziten Bias zum Beispiel durch mentales Training reduziert werden. [Brownstein]

2.2 Moralische Verantwortung und impliziter Bias

Durch die nur indirekte Kontrollmöglichkeit ist nicht klar wer, in welchem Ausmaß, für Aktionen, die durch Bias beeinflusst wurden, verantwortlich gemacht werden soll und kann. Dadurch ergibt sich die Frage nach der moralischen Verantwortung, bei der weiterhin unklar ist, ob die Verantwortung bereits auf Ebene der Bias-Akquisition liegt. [Brownstein]

Brownstein legt den Fokus bei der Betrachtung der moralischen Verantwortung auf die Effekte der Bias und stellt hierzu die Perspektiven Wahrnehmung, Kontrolle und Attributionismus vor [Brownstein]:

- **Wahrnehmung**:
  - Ein Argument ist, dass bewusste Wahrnehmung impliziter Bias notwendig ist um für diese moralische Verantwortung zu tragen. Hieraus folgt aber nicht, dass das Fehlen bewusster Wahrnehmung eine Person von der Verantwortung freispricht, da man sonst zu einem Verantwortungs-Skeptizismus gelangen könnte. [Brownstein]
  - Weitere Argumente besagen, dass moralische Verantwortung dann besteht, wenn man von impliziten Bias wissen *sollte*. Dies orientiert sich an der sozialen und epistemischen Umgebung der Person. [Brownstein]
  - Schlussendlich kann man Bezug zu den restlichen moralischen Vorstellungen einer Person herstellen. Würde sich ein Bias nicht mit diesen in Einklang bringen lassen, also einem impliziten Bias entspringen, so müsste man Verantwortung für diesen übernehmen. [Brownstein]
- **Kontrolle**:
  - Aus dieser Perspektive entsteht Verantwortung dann, wenn ein Akteur anders handeln könnte, also die Möglichkeit hat, das eigene Handeln zu kontrollieren. Daraus ergibt sich dann die Frage ob Handlungen beeinflusst von implizitem Bias, welche z.B. Automatismen sein können, unter der Kontrolle des Akteurs stehen. Argumente dafür sind unter anderem die Fähigkeit zur „Vorprogrammierung“ von Automa-

tismen durch Training oder zur aktiven Überwachen ebendieser. [Brownstein]

- Eine andere Betrachtungsweise unterscheidet nur zwischen „direkter“ und „indirekter“ Kontrolle, aber geht grundsätzlich davon aus, dass jede Aktion entweder direkt oder indirekt kontrollierbar ist. Oft wird man für nur indirekt kontrollierbare Handlungen, wie z.B. Lernen von Sprachen, direkt zur Verantwortung gezogen. Dies geschieht, trotz der Tatsache, dass man nur bedingt Einfluss auf die Entwicklung solcher Dinge hat. Ähnliches könnte man nun auf implizite Bias anwenden. [Brownstein]

- **Attributionismus**: Attributionistische Ansätze stellen im Gegensatz zu den anderen beiden Perspektiven in den Vordergrund, dass eine Aktion den Akteur widerspiegeln muss, damit dieser verantwortlich sein kann. Dabei wird der Begriff des „deep self“ benutzt, also eines tiefen Selbst der innersten evaluativen Werte und Überzeugungen. Dies wäre bei den oben genannten Arten von Handlungen, die klassischerweise von impliziten Bias beeinflusst sind, nicht der Fall. [Brownstein]

Es scheint unter jeder der vorgestellten Perspektiven – zumindest teilweise – eine moralische Verantwortung für implizite Bias bei der agierenden Person zu liegen. Diese ergibt sich in jedem der drei Fälle durch die vorhandene Möglichkeit zur Reflexion bzw. zum Eingriff in Handlungen oder Automatismen.

2.3 Biasfreiheit, Biasintervention

Besteht moralische Verantwortung für implizite Bias tatsächlich, so erfordert es Gegenhandlungen in Form von Biasinterventionen um einen möglichst hohen Grad der Biasfreiheit zu erreichen. In der Realität zeigt sich der Wunsch nach Biasfreiheit auch durch die Mengen an Ressourcen, die Staaten und Unternehmen in die Bekämpfung von Diskriminierung und Vorurteile investieren [Brownstein].

Bei der Bekämpfung von impliziten Bias existieren zwei wirksame Strategien. Einerseits sind es Strategien, die versuchen die ungewollten Assoziationen, welche für die impliziten Bias verantwortlich gemacht werden, zu verändern. Andererseits wird versucht die Kontrollfähigkeit der Akteure zu bestärken, um so das Eindringen des Bias in die Handlung zu verhindern. [Brownstein]

3 Algorithmischer Bias

Algorithmen sind als Grundlage der technisierten Welt weitverbreitete Akteure. Oft, wo sie anstelle von Menschen eingesetzt werden, sollen sie die entsprechenden Nachteile menschlicher Arbeitskräfte reduzieren. Zu den Nachteilen von Menschen zählen auch ihre (impliziten) Bias, welche sich auf die Neutralität der Ergebnisse menschlicher Arbeit negativ auswirken. [Danks]

Grundsätzlich sind Algorithmen wohldefinierte Folgen von Berechnungsschritten welche eine Menge an Eingabewerten rechnerisch bearbeiten um Ausgabewerte zu erreichen [Cormen]. Aufgrund dieser mathematisch-technischen Struktur und der vermeintlich fehlenden menschlichen Komponente, wirken Algorithmen oft wie neutralere Akteure [Danks]. Tatsächlich lassen sich jedoch auch bei Algorithmen, und insbesondere bei machine-learning-Algorithmen, eine Vielzahl an Beispielen anführen, in denen die Ergebnisse algorithmischer Vorgänge systematische Bias aufweisen.

3.1 Definition von algorithmischem Bias

Allgemein besitzt ein Algorithmus Bias bezüglich eines Standards, wenn seine Ergebnisse von ebendiesem abweichen. Solche Bias können vielfältiger Natur sein, je nachdem in welcher Domäne der Algorithmus eingesetzt wird. So existieren z.B. moralische, statistische oder rechtliche Bias in Algorithmen. Es gilt, ebenso wie bei menschlichen Bias, dass diese nicht rein logischer Natur, aber per se nicht gut oder schlecht sind. Bei der Bestimmung, ob ein Bias in einem spezifischen Anwendungsfall vorliegt, ist es wichtig, den jeweiligen Standard, nach dem dies beurteilt wird mit zu betrachten. [Danks]

Genau diese auf Standards basierende Herangehensweise ist jedoch auch ein Problem für die Betrachtung von algorithmischen Bias, insbesondere bei der Einbringung gesellschaftlicher und moralischer Faktoren. So sind reine gesellschaftliche statistische Fakten per se nicht mit Bias versehen, sofern sie die Realität widerspiegeln. Dennoch können sie als Lerngrundlage zu Bias in einem Algorithmus führen, wenn bereits in der Quelle der Daten, z.B. der Gesellschaft selbst, Bias vorliegen. [Danks]

3.2 Arten des algorithmischen Bias

Bias in computerisierten Systemen, und somit auch in Algorithmen, ist bereits seit über 20 Jahren, wie die Veröffentlichung von Friedman et al. zeigt, ein diskutiertes Thema. Dieses wird durch die immer weitere Verbreitung von Computern umso relevanter. Eine Klassifizierung verschiedener Arten algorithmischen

Bias, basierend auf ihrem Ursprung, wurde unter anderem von Danks und London vorgestellt [Danks]:

- **Trainingsdatenbias**: Diese Form von Bias tritt bei machine-learning-Algorithmen auf. Sie entsteht dann wenn bereits die Trainingsdaten, also diejenigen Daten anhand derer der Algorithmus sein eigenes Verhalten lernt, Bias nach dem Standard, nach dem auch der Algorithmus gemessen wird, enthalten. Dadurch, dass sie nicht durch aktives Zuarbeiten eingebracht werden, können sie schlecht erkennbar und subtil in der Auswirkung sein. [Danks]
- **Bias des algorithmischen Fokus**: Hierbei handelt es sich um Bias die entstehen, indem man, wieder bezüglich eines Standards, gewisse Aspekte in den Lerndaten ignoriert oder andere verstärkt betrachtet. Die entstehende Verzerrung, die vielleicht bezüglich eines Standards gewollt ist, führt zu einem nicht-neutralen Verhalten in anderen Fällen.[Danks]
- **Bias der algorithmischen Verarbeitung**: Diese Kategorie fasst Bias zusammen, die im Algorithmus selbst existiert. Meist werden sie durch die Entwickler in den Algorithmus eingebracht, zum Beispiel um den Effekt anderer Bias zu reduzieren. Sie haben bei machine-learning nur wenig Relevanz. [Danks]
- **Kontextübertragungsbias**: Durch die Anwendungen von Algorithmen außerhalb der Kontexte für die sie gedacht, oder im Fall von machine-learning für die sie trainiert wurden, kann es zu Bias-ähnlichem Verhalten kommen. Hier entstehen die Bias durch falsche Anwendung des Algorithmus durch die Benutzer. [Danks]
- **Interpretationsbias**: Diese Bias entstehen, wenn die Ergebnisse eines Algorithmus, bedingt beispielsweise durch mangelnde Information über die Funktionsweise, durch einen Nutzer falsch interpretiert werden. [Danks]

Während die letzteren vier Formen von Bias in der ein oder anderen Form durch menschliches Eingreifen entstehen, so ist gerade die erste Form von Bias interessant in einer moralischen Betrachtung. Da hier nicht effektiv durch einen Menschen eingegriffen wird, sondern der Algorithmus aus seiner Umgebung ein Bild anfertigt, wie dies Menschen auch tun, entsteht ein Bias, welcher nicht immer eindeutig oder leicht erkennbar ist.

3.3 Die Neutralität des Algorithmus

Zwar wurde gezeigt, dass eine Vielzahl an Formen algorithmischen Bias existieren, dennoch gelten besonders machine-learning-Algorithmen als neutrale Algorithmen [Danks]. Dies kann darin begründet werden, dass der Algorithmus

selbst nichts anderes macht als Muster in Daten zu erkennen und Eingaben anhand dieser erkannten Muster zu kategorisieren. Nur wenn es Bias in den Daten gibt, die sich auf diese Musterfindung auswirken, so übernimmt der Algorithmus auch den Bias.

Wichtig ist bei all diesen Überlegungen die Betrachtung des zu grundliegenden Kontexts und der angewandten Standards. So ist ein Algorithmus als Akteur nie im Kontextvakuum anzutreffen, sondern in einem komplexen System aus moralischen und gesellschaftlichen Regeln, ebenso wie ein Mensch. Diese sind bei der Betrachtung der Neutralität von Algorithmen relevant. [Friedman]

4 Der implizite, algorithmische Bias

Es kann also festgestellt werden, dass sowohl menschliche als auch algorithmische Akteure Bias besitzen, wenngleich diese von diversen anderen Faktoren abhängig sind [Danks, Brownstein]. Insbesondere wurde festgestellt, dass der Kontext relevant für die Beurteilung der Bias ist [Danks, Friedman]. Es ergibt sich nun die Frage nach der Natur des Bias in Algorithmen und den daraus resultierenden Erkenntnissen über deren implizierte moralische Verantwortung. Auf welchem Level ist der Bias von Algorithmen zu erkennen und inwiefern können gegebene Rahmen für Menschen auf Algorithmen angewandt werden? Dazu soll der Fokus erneut besonders auf machine-learning-Algorithmen gelegt werden.

4.1 Abgrenzung zu explizitem algorithmischer Bias

Besonders bei traditionellen Algorithmen ist die gesamte Logik durch Programmierer, menschlich oder andersartig, festgelegt. Durch diesen Vorgang sind alle in den Algorithmus eingebrachte Bias, egal ob absichtlich oder unabsichtlich, durch einen aktiven Vorgang entstanden. Wenngleich dieser aktive Vorgang nicht immer als Ergebnis einer bewussten Entscheidung steht, so ist er doch zurückverfolgbar auf einen spezifischen Moment oder ein spezifisches Zusammenspiel von Teilen des gesamten Systems.

Nun ist die Frage der moralischen Verantwortung auch bei traditionellen Algorithmen nicht einfach zu klären, dennoch gibt der Vorgang des aktiven Einbringens, welcher zwingend erforderlich ist, einige Anhaltspunkte in der Beantwortung dieser. Davon abzugrenzen sind machine-learning Algorithmen, die an und für sich, wie bereits festgestellt, als neutrale Akteure angesehen werden können. Es ist hier der Moment des Lernens, in dem eine Übertragung des Bias nach Haggendorf vollzogen wird und in welchem wir Anhaltspunkte für die moralische Verantwortung suchen können.

4.2 Impliziter Bias im maschinellen Lernen

Aus den bereits vorgestellten Kategorien von algorithmischen Bias, sind es besonders Trainingsdaten-Bias die sich von den anderen Arten abheben. Nur hier fehlt das aktive Eingreifen in die Wirkungsweise des Algorithmus durch einen Programmierer oder Nutzer. Ähnlich dem Unterbewusstsein von Menschen, welches durch verschiedenste Lerneffekte geprägt wird, werden Muster gelernt, indem der Algorithmus diesen in Form von Trainingsdaten ausgesetzt wird.

Man kann also einen Vergleich zu dem durch verschiedenste Lerneffekte beeinflussten Unterbewusstsein von Menschen aufstellen, welches die Quelle von impliziten Bias ist. Wo menschliche Akteure Erfahrungen sammeln, welche von einer Norm abweichen und in deren Unterbewusstsein implizite Bias bilden, so sind dies bei Algorithmen die Lerndaten die dazu führen, dass unabhängig der Logik (Bewusstsein) des Algorithmus entsprechende Effekte auf die Ausgaben des Algorithmus wirken. Demnach sind Bias von machine-learning Algorithmen, welche durch Trainingsdaten verursacht wird, implizite algorithmische Bias. Diese entsprechen in ihrer Natur den menschlichen impliziten Bias.

Betrachtet man nun unter diesen Annahmen die verschiedenen Perspektiven der moralischen Verantwortung bezüglich impliziter Bias nach Brownstein, so erhält man ein Framework, welches sich auch auf Algorithmen anwenden lässt.

4.3 Manueller Eingriff, erhöhte Neutralität?

Unter den Perspektiven Wahrnehmung, Kontrolle und Attributionismus ergeben sich nun klare Fälle für die argumentiert werden kann, dass eine moralische Verantwortung für implizite algorithmische Bias gegeben ist. So kann aus der Perspektive Wahrnehmung argumentiert werden, dass ein Algorithmus der wahrnehmbar gegen eine gegebene und intendierte moralische Vorstellung arbeitet, oder dessen Ergebnis nicht seinem vermeintlichen Ziel in Einklang steht eine moralische Verantwortung für die offensichtlichen Bias impliziert. Bezüglich der Perspektive der Kontrolle, lässt sich argumentieren, dass nur indirekt Kontrolle darüber besteht, welche Muster gelernt werden, dennoch aber nicht nur das Lernen, sondern auch diverse Ausgabeschritte vorhanden sind. Gerade diese sind explizit kontrollierbar und implizieren somit moralische Verantwortung. Schlussendlich kann aus der Perspektive des Attributionismus argumentiert werden, dass die inneren Überzeugungen eines Menschen dem intendierten Ziel eines Algorithmus entsprechen, und jeglicher Bias der diesem entgegenwirkt eine moralische Verantwortung impliziert.

Es existieren viele Fälle, in denen moralische Verantwortung für implizite Bias gegeben ist. Somit liegt die Forderung nach Biasfreiheit oder zumindest

Biasintervention nahe. Gerade bei menschlichen Akteuren wird Biasfreiheit in vielen sozialen und gesellschaftlichen Kontexten gefordert, besonders wenn die gegebenen impliziten Bias in eben diesem Kontext negativ angesehen werden [Brownstein]. Somit wäre, unabhängig davon wer die moralische Verantwortung trägt, die Entfernung von impliziten Bias aus dem Algorithmus ebenso relevant ist, wie die Biasintervention bei Menschen. Ein Eingriff in implizite Bias eines Algorithmus ist also auch zu bewerten, wie dies beim Eingriff in Bias eines menschlichen Akteurs der Fall wäre. In der Regel handelt es sich somit also um eine Aktion, welche die Neutralität erhöht. Ähnlich einem Menschen wird dieses Ziel dadurch erreicht, dass bestimmte Faktoren in der Funktionsweise des Algorithmus modifiziert werden [Danks].

5 Zusammenfassung

Algorithmen sind allgegenwärtige Akteure in der Welt des 21. Jahrhunderts. Sie bestimmen das Verhalten aller computerisierten Systeme und stellen somit eine wichtige Grundlage jeder technischen Anwendung dar. Besonders im Bereich der machine-learning-Algorithmen wurden in den letzten Jahren massive Fortschritte erzielt und somit eine Reihe von Anwendungen realisiert deren Verhalten nicht mehr explizit durch Entwickler, Designer oder Programmierer vorgegeben ist. Stattdessen lernen, diese zunächst neutralen Algorithmen, anhand von Beispielen die als aussagekräftig gelten verschiedene Muster in den Daten. Anhand dieser Muster können sie Aussagen über neue Eingabedaten treffen. Es ist offensichtlich, dass bereits klassische Algorithmen betroffen von Bias sein können. Die Menschen die sie entwickeln, sind selbst oft mit Bias versehen und bringen absichtlich oder unabsichtlich ihre eigenen Bias ein. Jedoch sind auch machine-learning-Algorithmen von Bias Problemen betroffen, und zwar genau dann, wenn die Lerndaten bereits mit Bias – unabhängig der Quelle – versehen sind.

Die Form von Bias in machine-learning-Algorithmen ähneln in ihrer Entstehung und Wirkung stark den impliziten Bias von Menschen. Hierbei handelt es sich um unterbewusste Einstellungen die durch Erlebnisse und andere Lernvorgänge eingeprägt wurden und nicht direkt kontrolliert werden können. Auf Basis der Übetragungsprozesse nach Haggendorf und der Definition von Trainingsdatenbias nach Danks et al. wurde argumentiert, dass es sich bei algorithmischen Bias, wie er bei machine-learning auftritt, um implizite Bias handelt, welche gleichwertig den impliziten Bias von Menschen ist. Aus diesem Grund lässt sich anhand der Perspektiven nach Brownstein ein Framework ableiten, mit dem das Vorhandensein moralische Verantwortung von Algorithmen bestimmt werden

kann. So lässt sich für jede der Perspektiven ein ebenso passendes Szenario entwickeln, welches sich auf machine-learning Algorithmen anwenden lässt.

Das resultierende Framework ermöglicht es, zu bestimmen, wann moralische Verantwortung für implizite Bias vorliegt und ob ein Eingriff in den Algorithmus die Neutralität des Algorithmus erhöht. Für viele Fälle lässt sich sowohl aus Sicht der Verantwortung, als auch was die Erhöhung der Neutralität betrifft, dies mit ja beantworten. Jedoch ist es nicht möglich, anhand des Framework zu bestimmen wer Verantwortungsträger für Bias ist und wie der Bias genau reduziert werden kann. Zumindest im Bereich der Biasintervention für Algorithmen liefern u. a. Danks et al. verschieden Ansätze. Gerade aber die Frage nach den Verantwortungsträger bleibt in diesem Rahmenwerk vollständig ungeklärt.

Ebenso ist nach dieser Feststellung noch die Frage der Bewertung von Bias relevant, insbesondere wenn diese in einem sozialen Kontext betrachtet werden. So lässt das gegebene Framework keinerlei Beurteilung darüber zu, ob Bias in einem Algorithmus gut oder schlecht ist. Nach Friedman et al. und Danks et al. ist dies allerdings, wenn überhaupt bestimmbar, hochgradig abhängig von den gegeben Kontexten und sozialen bzw. gesellschaftlichen Umfeldern in denen der Algorithmus eingesetzt wird.

Insgesamt erhält man zwar ein Rahmenwerk, welches genutzt werden kann um die Frage nach dem Vorhandensein moralischer Verantwortung zu klären, das gewisse Ansätze zur Neutralitätsbewertung und Biasintervention nahelegt, jedoch dabei keine Aussagen darüber trifft, wer tatsächlich die Verantwortung übernehmen soll, noch eine moralische Bewertung des Bias zulässt.

Quellen

[Amodio] David M. Amodio, Kyle G. Ratner 2011: „A memory systems model of implicit social cognition“. *Current Directions in Psychological Science* 20(3), 143-148.

[Brownstein] Michael Brownstein 2019: „Implicit Bias“. Edward N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy*. URL = https://plato.stanford.edu/archives/fall2019/entries/implicit-bias/. Aufgerufen am 27. 02. 2020

[Bynum] Terrell Bynum 2018: „Computer and Information Ethics“. Edward N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy*. URL = https://plato.stanford.edu/archives/sum2018/entries/ethics-computer/. Aufgerufen am 27. 02. 2020

[Cambridge] Cambridge Dictionary Redaktion o.J.: „Bias“. *Cambridge Online Dictionary*. URL = https://dictionary.cambridge.org/de/worterbuch/englisch/bias. Aufgerufen am 27. 02. 2020

[Cormen] Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest, Clifford Stein 2009: *Introduction to Algorithms*, 1-30.

[Danks] David Danks, Alex John London 2017: „Algorithmic Bias in Autonomous Systems“. *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, 4691-4697.

[Friedman] Batya Friedman, Helen Nissenbaum 1996: „Bias in Computer Systems“. *ACM Transactions on Information Systems* 14(3), 330-347.

[Haggendorf] Thilo Haggendorf 2019: „Rassistische Maschinen? Übertragungsprozesse von Wertorientierung zwischen Gesellschaft und Technik“. M. Rath et al. (Hsg.), *Maschinenethik,* 121-134. Wiesbaden.

[Haselton] Martie G. Haselton, Daniel Nettle, Paul W. Andrews 2015: „The Evolution of Cognitive Bias“. D. M. Buss (ed.), *The Handbook of Evolutionary Psychology*, 724-746.