Vortrag „Fairness bei algorithmischen Entscheidungsprozessen“

Vom 29.04.2024, Marc Hauer

Zusammenfassung der Fragerunde

Frage: Eine Frage in Richtung Fairness am Beispiel bedingungsloses Grundeinkommen: Gibt es denn in dieser Hinsicht Ansätze? Wenn ich an bedingungsloses Grundeinkommen denke, ist es ja im Prinzip ein Thema der Fairness: Wer soll es bekommen? Gibt es, wie beim Kindergeld, Freibeträge?

Hauer: Die Frage ist ein bisschen im philosophischen Bereich angesiedelt, es geht hier ja um eine gesamtgesellschaftliche Entscheidung aufgrund ethischer Prinzipien – und nicht mehr um algorithmische Entscheidungen. Diese Frage lässt sich schwer auf eine Metrik übertragen, weil beim bedingungslosen Grundeinkommen ja gesagt wird, dass alle es bekommen sollen. Man könnte es natürlich, so wie das Kindergeld, unter einem Quality Aspekt betrachten – das wäre die dahinterstehende Fairness Philosophie. Aber das muss durch einen gesellschaftlichen Diskurs entschieden werden.

Frage: Ein Ziel könnte ja sein, dass automatische Entscheidungen nicht schlechter als menschliche sein sollen. Haben denn Psychologen die aktuellen Fairness Maße, die bei Maschinen eingesetzt werden, denn schon an Menschen getestet?
Und eine zweite Frage: Kann man eine Korrelation herstellen, dass immerhin bei diesen Maßen, die schon bei größeren Menschengruppen getestet wurden, mindestens eine bestimmte Schwelle erreicht werden soll?

Hauer: Zum ersten Punkt, der Vergleich mit menschlichen Entscheidungen: Bei solchen Untersuchungen wird immer wieder aufgedeckt, wie hochgradig diskriminierend menschliche Personalentscheidungen sind. Daher ist der Vergleich mit menschlichen Entscheidungen schwierig – auch wenn wir es diesem Vergleich zu verdanken haben, dass auf dem Forschungsfeld so viel passiert. Ein Problem ist auch, dass die Überwachung des Personals, beispielsweise im Einstellungsprozess, verhindert werden muss; aktuell tracken viele Unternehmen die Entscheidungen, die im Einstellungsprozess getroffen werden, um möglichst viele Daten über die Entscheidung zu erhalten – und so wurden auch viele

diskriminierende, voreingenommene Unterscheidungen aufgedeckt. Der Mensch hat sich hier eigentlich eher blamiert.

Kommentar: Vielleicht wäre dann eine Rückkopplung möglich, bei welcher sich der Mensch an den Entscheidungen einer KI orientiert.

Frage: Ich habe das Gefühl, dass die Fairness ein Perpetuum Mobile ist, wo sich die Katze in den Schwanz beißt – denn auf der einen Seite soll der KI Vorgaben gemacht werden, wie sie fair sein und lernen kann; auf der anderen Seite ist Fairness aber ja extrem stark von eigenen, gesellschaftlichen Vorstellungen beeinflusst.
So wäre es beispielsweise viel fairer, das Geschlecht (in einem Bewerbungsprozess) nicht anzugeben und insofern nicht zu diskriminieren. Die KI könnte doch diese menschlichen Gedanken sehr viel besser ausgleichen und die menschliche Unfairness nivellieren. Die Frage ist: Können wir das jemals herausfiltern, wenn wir die KI ja erst einmal trainieren müssen?

Hauer: Ja, sobald in den Trainingsdaten Diskriminierung stattfindet, wird diese auch der KI beigebracht; so wird die KI diese Diskriminierung dann auch reproduzieren. Deswegen gibt es spezielle Konzepte, die Daten aufzubereiten und bewusst nach Fairness zu filtern. Diese haben aber natürlich auch wiederum Nachteile, da das reale Abbild der Gesellschaft verzerrt wird. Manchmal, beispielsweise bei Entscheidungen von Banken, müssen aber gerade die realen, wenn auch diskriminierenden Entscheidungen, berücksichtigt werden, da Banken ja wirtschaftlich agieren müssen.

Auch die finale Entscheidung, die nach der KI stattfindet, spielt neben der Datenbasis natürlich noch eine Rolle: Durch diese kann natürlich auch wieder Fairness eingebracht werden.

Es gibt, bedingt durch die Komplexität der KI-Systeme, noch ein weiteres Problem. Wenn ich ein neuronales Netz trainiere, startet es mit einem zufälligen Ausgangszustand – und auch die Reihenfolge der Trainingsdaten ist zufällig. Die Reihenfolge beeinflusst allerdings das neuronale Netz in seinen Eigenschaften. Solche unterschiedlichen neuronalen Netze können in ihren Entscheidungen zwar insgesamt gleich gut sein, also eine gewisse Prädiktionsqualität haben, in Einzelfällen aber ganz andere Entscheidungen treffen; das ist der sogenannte Rashomon Effekt.

Unter diesen neuronalen Netzen kann man nun danach filtern, welches dieser Netze an fairsten ist – so kann man eine gewisse Qualität und Fairness gleichzeitig erreichen. Allerdings stehen wir bei diesen Entwicklungen gerade erst am Anfang und ein solches Verfahren ist sehr kompliziert und aufwendig – denn es müssen ja erst einmal viele neuronale Netze programmiert werden.

Frage: Die Frage ist dann, ob die KI wesentlich anders entscheidet als Menschen? Menschen entscheiden ja auch nach besten Wissen und Gewissen und trotzdem individuell anders. Anders gefragt: Ist eine vollständige Fairness denn überhaupt zu erreichen?
Hauer: Nein, aber es kommt auch auf die Definition an. Es kommt auch darauf an, welche Rolle der Mensch haben soll: Ist eine Entscheidung beispielsweise dann fair, wenn der Mensch eine Empfehlung von einer KI bekommt und dann doch eigenständig entscheidet?

Frage: Aus meiner Sicht werden menschliche Entscheidungen immer sehr individuell getroffen und wenig nachvollziehbar – deswegen sind die Trainingsdaten für KI von sehr schlechter Qualität. Bedeutet das nicht, dass die KI nicht mehr mit tatsächlichen Daten trainiert werden kann, sondern nur einen Grundstock braucht, die Fragen gestellt werden müssen, die eher auf theoretischen Überlegungen sind, so wie Sie das gerade präsentiert haben – und dann über Simulationsmodelle entwickelt werden, die letztlich Risiken bei der Weiterentwicklung des Modells zeigen und dadurch eine Optimierung der Fairness bekommen können – unabhängig von neuen Daten zum Lernen?
Hauer: Ja, das ist auf jeden Fall eine Überlegung wert. Wenn ich allerdings nur künstliche Daten habe, ist das Problem, dass sie eben nicht real sind und keine realen Zusammenhänge abbilden – sie bilden somit also nur Fairness ab, aber nicht die Qualität. Wenn später Entscheidungen aufgrund von echten Daten getroffen werden sollen, funktioniert das nicht, da das Training ja nur mit künstlichen Daten stattgefunden hat.
Es gibt die Überlegungen, beides zu kombinieren oder parallel laufen zu lassen; dieser Ansatz der künstlichen Daten hat auf jeden Fall seine Daseinsberechtigung und sollte verfolgt werden.
Nachfrage: Heißt „verfolgt“ iterativ verfolgt?
Hauer: Ja, zum Beispiel. Was die sehr schlechte Datenqualität angeht: Das ist der Grund, warum KI in den letzten Jahren überhaupt explodiert ist: Big Data ist deshalb zu einem so

wichtigen Begriff geworden, weil es einfach eine Unmenge an Daten gibt – das bedingt leider, dass die Daten von sehr schlechter Qualität sind, wenn auch auf andere Art schlecht. Die Hoffnung ist aktuell, dass sich die unterschiedlichen Arten der schlechten Qualität ausgleichen.

Frage: Sie haben verschiedene Formen von Fairness vorgestellt – *equity* und *equality*. Das sind ja zwei widersprüchliche Definitionen. Ihrem Vortrag habe ich entnommen, dass es noch weitere Definitionen von Fairness gibt. Heißt das nicht, dass das Problem von Fairness grundsätzlich nicht lösbar ist, wenn es solch widersprüchliche Definitionen gibt?
Meine Vermutung ist, dass sich Menschen, wenn sie versuchen, sich fair zu verhalten, aufgrund ihrer unterschiedlichen Begriffe von Fairness oft nicht einig werden; für mich läuft das darauf hinaus, dass sehr viele unterschiedliche Definitionen der Fairness herangezogen werden müssen, um eine umfassende Entscheidung treffen zu können.
Ich bin ein Freund der Wertedarstellung der Menschen, die 130 verschiedene Werte des Menschen umfasst.
Gibt es Versuche, Fairness darzustellen, indem man diese 130 Werte näherungsweise darstellt?
Hauer: Die Forschung ist hier ehrlicherweise etwas hintendran. Die zwei hier dargestellten Modelle findet man am häufigsten in der Literatur. Es gibt Tools, die genau diese nach einem vorgegebenem Datensatz berechnen. Dabei gibt es aber mehrere Probleme: Zum einen erklärt das Tool das Ergebnis nicht – es kann sein, dass ein Fairnessmaß für meine konkrete Anwendung und ihren sozialen Kontext völlig irrelevant ist und die KI aber suggeriert, dass es wichtig sei. Außerdem haben wir das Problem, dass es vielleicht für einen bestimmten Sachverhalt überhaupt kein passendes standardisiertes Fairnessmaß gibt, dass das abbildet, was in einem konkreten Einzelfall das Beste wäre.
Nachfrage: Ich habe das Gefühl, das geht in die Richtung dessen, was ich gesagt habe – dass viele zusätzliche Faktoren berücksichtig werden.
Außerdem kann man ja beobachten, dass menschliche Werte situationsspezifisch unterschiedlich sind; das heißt, man muss die Fairness nicht nur mit der Anzahl der 130 Werte, sondern auch mit der Anzahl der Situationen multiplizieren; da ist man schnell bei 30.000 bis 50.000. Das ist für die derzeitige KI nicht machbar. Meine Vermutung ist, so wie Sie das gerade auch angedeutet haben: Indem man sehr viele weitere Parameter

einbezieht, werden wir in die richtige Richtung gehen.

Noch einmal die Frage: Wird das irgendwo bereits versucht?

Hauer: Es ist schwierig, ein Tool zu erstellen, dass das alles kann: Denn es sollte ja einerseits die Metriken über die eingehenden Daten berechnen, was noch gut machbar ist. Allerdings wird die Qualität hier nicht pauschal definiert.

Für ein bestimmtes Produkt kann es wichtig sein, dass die Ergebnisse für Männer und Frauen gleich schnell berechnet werden müssen. Und solche Qualitätsdefinitionen gibt es sehr viele – das Problem ist, dass für ein Tool immer viele verschiedene Qualitätsdefinitionen gleichzeitig berechnet werden und teilweise auch miteinander verrechnet werden müssen. Wir kommen hier also in einen mehrfach unendlichen Raum hinein – von daher besteht so gesehen also keine Chance.

Trotzdem gibt es vermehrt den Wunsch nach einem Tool, das eine Art Liste zur Auswahl verschiedener Parameter anbietet.

Frage: Ich komme aus dem medizinischen Fachbereich, in dem die Genauigkeit im ersten Schritt vor der Fairness statt. Wäre es nicht fair, für die einzelnen Fragestellungen eben diese genau definierbaren Faktoren wie Qualität und Genauigkeit zu nehmen und die Fairness außer Acht zu lassen?

Fairness wäre doch, eine Fragestellung oder eine Gruppe von Menschen nach denselben Kriterien zu beurteilen. Denn durch die Zuschreibung einer Fairness von außen findet ja auch wieder eine Bewertung statt.

Hauer: Ich möchte gerne mit einer Gegenfrage antworten: Wie sehr wird Ihrer Erfahrung nach bei Medikamentenstudien der unterschiedliche Hormonhaushalt von Frauen berücksichtigt?

Das ist eine statistische Größe und es kommt ja darauf an, möglichst viele Daten in die Studie einlaufen zu lassen – aber diese Daten müssen genau definiert sein. Das ist ja der Vorteil, dass wir durch KI nicht mehr über Statistik sprechen, bei der es ja ganz andere Ungenauigkeiten und Größenordnungen von Gruppen gibt, die eingebunden oder ausgeschlossen werden sollen.

Hauer: Trotzdem – bei den Standardmedikamenten hat man, soweit ich weiß, sehr wenig darauf geachtet, dass Männer und Frauen aufgrund ihres unterschiedlichen Hormonhaushalts sehr unterschiedlich reagieren, und dass der zyklusbedingt schwankende

Hormonhaushalt von Frauen hier auch große Unterschiede machen kann. Hier geht es also wiederum um Qualität vs. Fairness: Ich kann die Qualität über diesen ganzen Korps berechnen, aber ob die Nebenwirkungen dann speziell bei Frauen, oder speziell zyklusbedingt, häufiger auftreten, das wird nicht automatisch berücksichtigt – was uns wieder zurück zur Fairness Definition bringt.

Das ist ja ein Lerneffekt. Meiner Meinung nach braucht man, mit Blick in die Zukunft von KI, keine Statistik mehr; hier kann man zukünftig viel mehr in die Breite gehen, so dass auch diese Fragen abgedeckt werden können.

Das hat meiner Meinung nach weniger mit KI zu tun als mit der Datengröße. Und es ist auch historisch gewachsen: Heute werden bei der Entwicklung von Medikamenten dort, wo es absehbar ist, unterschiedliche Bevölkerungsschichten, unterschiedliche Genetik, unterschiedliche Herkünfte, sowie auch Männer und Frauen inkludiert. Aber ich glaube, das hat mit KI und Fairness nichts zu tun.

Hauer: Wenn ich Sie richtig verstanden habe, läuft es auf genau das Problem hinaus, das hier auf der Vortragsfolie zu sehen ist: Es ist möglich, sich ausschließlich auf die Qualität zu konzentrieren und eine gesamte Population zu betrachten, Männer und Frauen also nicht zu differenzieren – man kann aber auch Frauen und Männer als zwei getrennte Populationen betrachten und somit einfach zu einem anderen Ergebnis kommen. Im zweiten Fall ist die Qualität eigentlich viel besser, doch in Hinblick auf Fairness im Sinne von Gleichbehandlung wird eine ungerechte Entscheidung getroffen.

Aber es ist doch fair, zu sagen, dass einfach durch die Betrachtung einer großen Datenmenge mehr Gruppen berücksichtigt werden. Es muss hier vorher definiert werden, welche Informationen wichtig sind und nach welchen Informationen gefiltert werden soll. Nach meinem Verständnis ist das Fairness, da nichts aus der Angst, zu diskriminieren, weggelassen wird.

Hauer: Das ist das sogenannte Simpson Paradox. In wenigen Worten erklärt: Die Parameter, die die KI berücksichtigt, sind von Hand ausgewählt. Die KI hat kein Verfahren, selbst zu entscheiden, welche Parameter relevant sind. Wenn jetzt, wie Sie gesagt haben, sehr viele Parameter verwendet werden, kann es sein, dass dabei Parameter verwendet werden, die gar keine Rolle spielen.

Frage: Bei den ganzen Diskussionen habe ich den Eindruck gewonnen, dass nicht immer so ganz klar ist, welcher Aspekt unter den Begriff „Fairness“ und welcher unter den Begriff „Qualität“ fällt. Zum Beispiel bei dieser Medizinstudie: Die Thematik des weiblichen Zyklus und dessen Einfluss auf die Medikamentenwirksamkeit ist für mich eine reine Qualitätsfrage – es ist in diesem Zusammenhang aber auch der Begriff „Fairness“ gefallen. Gibt es denn eine klare Trennlinie bei diesen beiden Begriffen?

Hauer: Ja, es gibt eine klare Grenze. Die Qualität im mathematischen Sinne, von der man auch bei der Operationalisierung von Qualität spricht. Es geht hier um die Prädiktionsqualität: Ich gehe davon aus, das bestimmte Personen in eine gewisse Gruppe fallen, und vergleiche das Ergebnis mit der Realität. Hier geht es darum, wie exakt die Vorhersage ist.

Wenn derselbe Datensatz in die zwei Gruppen „Männer“ und „Frauen“ trenne, kann man dann zum Beispiel feststellen, dass die Prädiktionsqualität sich je nach Gruppe unterscheidet – also beispielsweise bei Frauen nur 10% korrekt vorhergesagt wurden, bei den Männern ab 99%. Wenn es im Datensatz allerdings nur 3 Frauen und 1.000 Männer hatte, ist meine Gesamtpositivrate, also die Qualität der Vorhersage, immer noch sehr gut.

Und hier kommt die Fairness ins Spiel: Sobald die Teilgruppen verglichen werden, stellt sich erst heraus, dass – trotz guter Gesamtqualität – eine sensitive Gruppe sehr schlecht dasteht.

Frage: Eine pragmatische Frage: Man könnte ja im Prinzip argumentieren, dass die Datenqualität in Bezug auf Fairness in den *large language* Modellen am besten sein müsste – denn hier wurde ja die größte Datenmenge für das Training verwendet. Ich könnte ein solches Entscheidungsproblem, wie die, über die wir gesprochen haben, durch entsprechendes *prompt engineering* einem *large language* Modell eingeben und das Ergebnis messen.

Wurde das schon einmal versucht?

Hauer: Ja, das ist tatsächlich schon lange Standardvorgehen. Das hat sogar schon bei Übersetzungssystemen angefangen. Zum Beispiel wurde verglichen, wie „Arzt“ und „Ärztin“ ins Englische übersetzt wird – mit „doctor“ wird dabei stets dasselbe Ergebnis erzielt.

Außerdem ist auch folgender *prompt* interessant: Übertrage das Verhältnis von „Mann“ zu „Arzt“ auf „Frau“ – das Ergebnis war dann nicht „Ärztin“, sondern „Krankenschwester“.

Darüber wurde bereits viel geforscht und es gibt spannende Paper darüber.

Frage: Mich hat die Frage sehr beeindruckt, was schlimmer ist: Schuldige frei laufen zu lassen oder Unschuldige einzusperren. Das bedeutet nämlich, dass wir über gesellschaftlichen Kontext sprechen. Wenn wir bedenken, dass die Künstliche Intelligenz stark von den USA und als zweiten starken Spieler von China beeinflusst wird, dann frage ich mich natürlich, was das für Entscheidungen bedeutet und was das wiederum für Ihre Forschung bedeutet, und wie Sie diesen Kontext berücksichtigen.

Hauer: Zum Glück nimmt uns hier der Gesetzesgeber sehr viel ab – denn das europäische Recht sorgt ja vehement dafür, dass Entscheidungen im sensitiven Bereich nicht von einer Maschine getroffen werden dürfen – KI darf hier maximal beraten. Das heißt, eine Einstellungsentscheidung beispielsweise darf nicht von einer KI getroffen werden; hier darf nur eine Beratung stattfinden und die finale Entscheidung wird im Anschluss durch einen Menschen getroffen. Wenn sich ein Unternehmen dem widersetzt und das auffällt, wird es sehr teuer.

Nachfrage: Aber wenn ich Vertrauen in die KI habe und mit ihr bereits gute Erfahrungen gemacht habe, dann höre ich ja sicherlich relativ leicht auf den KI-Berater.

Hauer: Ja, das ist ein großes Problem. Das wird aktuell verstärkt im Medizin Bereich diskutiert – hier werden solche Beratungssysteme stark kritisiert, weil die Gefahr besteht, dass sich ein Arzt trotz besseren Wissens zu einer von der KI empfohlenen Behandlung überreden lässt und diese sich schlussendlich als die falsche herausstellt.

Das Problem ist also immerhin bewusst und es wird regulatorisch in Europa stark dafür gesorgt, dass dieses Bewusstsein in den Köpfen bleibt.