Lernende Systeme
DIE PLATTFORM FÜR KÜNSTLICHE INTELLIGENZ

Künstliche Intelligenz und Diskriminierung

Whitepaper von Susanne Beck et al.
Arbeitsgruppe IT-Sicherheit, Privacy, Recht und Ethik

Kurzfassung

Künstliche Intelligenz (KI) wird heute bereits in deutlich mehr Anwendungsfeldern eingesetzt als auf den ersten Blick vermutet wird. Nicht immer offensichtlich ist das damit verbundene Diskriminierungspotenzial. Obwohl auch Menschen ungerechtfertigt diskriminieren, erscheinen ihnen Entscheidungen von Computerprogrammen und Softwarelösungen oftmals faktenbasiert, objektiv und neutral. Tatsächlich aber treffen KI-basierte Systeme bisweilen problematische, diskriminierende oder ungerechtfertigt differenzierende Entscheidungen.[1] Softwaresysteme beinhalten vielfach explizit oder implizit gesellschaftliche Regelsysteme und steuern dadurch Verhalten – sei es in Form von Regelungen, von Transaktionen und Koordination oder von Zugangs- und Nutzungsrechten. Vor allem setzen sie Regelsysteme auf technischem Weg effektiv durch. Lernende Systeme bergen somit das Potenzial, bereits vorhandene Diskriminierungen nicht nur zu übernehmen, sondern sogar zu verschärfen.

So werden beispielsweise in den USA Algorithmen dazu eingesetzt, die Rückfallwahrscheinlichkeit von Angeklagten zu bestimmen.[2] Die Algorithmen ermitteln anhand verschiedener Daten einen Wert, der Richterinnen und Richtern eine Einschätzung darüber geben soll, mit welcher Wahrscheinlichkeit die Angeklagten erneut eine Straftat begehen. Der Algorithmus wird allerdings vor allem mit historischen Daten (z. B. aus Kriminalitätsstatistiken) trainiert, die nicht auf kausalen Zusammenhängen, sondern auf statistischen Korrelationen beruhen. Im Ergebnis erhalten Menschen aus Bevölkerungsgruppen, die in der Vergangenheit häufiger ins Visier der Strafverfolgungsbehörden gerieten (z. B. ethnische Minderheiten oder Gruppen mit schlechteren finanziellen Möglichkeiten), schlechtere Prognosen. Da sich das Urteil der Richterinnen und Richter unter anderem darauf stützt, werden Menschen allein aufgrund ihrer Zugehörigkeit

1 Ein erster Überblick über potenziell unbedachte negative Folgen bei der Anwendung von Künstlicher Intelligenz ist im „Atlas der Automatisierung“ von AlgorithmWatch zu finden (AlgorithmWatch 2019).
2 Für einen Überblick über diese Praxis siehe Angwin et al. 2016.

zu den aufgeführten Gruppen benachteiligt. So verstärkt die Anwendung der Algorithmen bereits vorherrschende Verzerrungen.

Die Problematik der potenziellen Diskriminierung beim Einsatz von Künstlicher Intelligenz ist Teil einer größeren Debatte um die Entwicklung und Anwendung von KI und deren Grenzen. Entsprechend wird das Thema auch in der KI-Strategie der Bundesregierung sowie in der von der Bundesregierung eingerichteten Datenethikkommission und der Enquete-Kommission „Künstliche Intelligenz" adressiert. Auf der europäischen Ebene wird in den „Ethik-Richtlinien zum Einsatz von Künstlicher Intelligenz" der High-Level Expert Group on Artificial Intelligence der Europäischen Kommission darauf hingewiesen, dass Lernende Systeme diskriminierungsfrei sein sollen. Dies haben bereits auch einige Unternehmen erkannt und sind entsprechende Selbstverpflichtungen eingegangen oder haben spezielle Ethikräte eingerichtet.

Die Unterarbeitsgruppe Recht und Ethik der Arbeitsgruppe IT-Sicherheit, Privacy, Recht und Ethik der Plattform Lernende Systeme möchte mit vorliegendem Papier einen Beitrag zu dieser Debatte leisten. So zeigen die Autorinnen und Autoren erstens die verschiedenen Facetten von Diskriminierung auf und thematisieren zweitens nicht nur technologische Lösungen[3], sondern fokussieren auch gesellschaftliche Aspekte. So wird erörtert, welche Aspekte der Diskriminierung im gesellschaftlichen Dialog behandelt werden müssen und welche Institutionen hierbei behilflich sein können. Im Mittelpunkt stehen dabei Systeme, von deren Entscheidungsvorschlägen oder Entscheidungen in erster Linie Personen und deren Zugang zu Leistungen, Gütern oder gesellschaftlichen Teilhabemöglichkeiten beeinflusst werden. Das Papier zeigt, dass nicht jede Unterscheidung per se ungerechtfertigt ist, sondern Diskriminierung dann vorliegt, wenn eine Gleich- oder Ungleichbehandlung ungerechtfertigt ist. Quellen für Diskriminierung durch Lernende Systeme sind vor allem in Input- und den Trainingsdaten, aber auch im Output der Anwendung zu finden. Die größten Herausforderungen für diskriminierungsfreie KI-Anwendungen liegen in einer mangelnden Transparenz der Algorithmen, deren stetigem Weiterlernen, der fehlenden Neutralität der Daten sowie unklaren Verantwortlichkeiten.

Für die Entwicklung von diskriminierungsfreien Lernenden Systemen benennen die Autorinnen und Autoren folgende Ansatzpunkte:

Erklärbarkeit und Überprüfung

KI-Entscheidungen sollten nachvollziehbar sein. Ist dies der Fall, würden bestimmte Formen der Diskriminierung durch KI möglicherweise akzeptiert werden. Aber hier stellen sich neben den dargestellten technischen Herausforderungen weitere Probleme: Die Transparenz der Systeme ist kein Selbstzweck, auch Firmengeheimnisse sind wichtig für den technologischen Fortschritt. Es müsste – evtl. durch eine unabhängige Institution – geklärt werden, in welchem Maß und gegenüber welchen Akteuren Transparenz hergestellt wird.

Denkbar wäre eine unabhängige Instanz, die als Stellvertreter für die potenziell diskriminierten Bürgerinnen und Bürger – die als gesellschaftlich Benachteiligte ihre Rechte meist nur schwer geltend machen können – die Outputs Lernender Systeme kontrolliert und bewertet. Sie soll die Ergebnisse und von den Systemen selbst gegebenen Erklärungen mithilfe von klar definierten Instrumenten und Prinzipien auf Plausibilität überprüfen. Dabei stellt sich zwar das Problem der erheblichen Geschwindigkeit, mit der sich Systeme verändern oder neue

3 Für einen ersten Überblick zu den technologischen Lösungen für ethische KI-Anwendungen kann z. B. der Report des Dagstuhl Seminars 16291 (Abiteboul et al. 2016) oder das Projekt Data Responsibly herangezogen werden.

Systeme angewendet werden. Gerade deshalb ist aber daran zu denken, bereits die externe Überprüfung der (Trainings-)Daten, der verwendeten Methoden sowie eine laufende Optimierung der Testfälle zu ermöglichen. Dies geschieht allerdings unter der Gegebenheit, dass maschinelles Lernen nur Korrelationen, aber keine Kausalbezüge beobachten und erfassen kann. Auf der Grundlage dieser Korrelationen erfolgt dann das maschinelle Lernen. Diese Tatsache erschwert die unabhängige Beobachtung und Überprüfung durch eine dritte, neutrale Instanz.

Nötig sind darüber hinaus laufende Schulungen und Fortbildungen für Mitarbeiterinnen und Mitarbeiter in beispielsweise Unternehmen oder der öffentlichen Verwaltung, die die Systeme anwenden. Auch wäre unabhängig von einer derartigen prüfenden Institution eine permanente nachträgliche Beobachtungspflicht der Hersteller oder Betreiber der Systeme zu fordern, da bestimmte Formen der Diskriminierung eben erst in der Anwendung sichtbar werden. Sollten sich später Diskriminierungen zeigen, wären Hersteller oder Betreiber zur Nachbesserung bzw. zur Tragung der Schäden, die durch diese Diskriminierungen entstehen, zu verpflichten.

Selektion der Kriterien

Um Diskriminierung zu vermeiden, sollten von der Gesellschaft im jeweiligen Kontext als diskriminierend bewertete Merkmale (wie beispielsweise die ethnische Zugehörigkeit) aus dem Input für maschinelle Lernverfahren gestrichen werden. Bestehen bleibt das Problem, dass viele Merkmale als Stellvertreter für andere dienen können (Harcourt 2010). So verzichtete die Software zur Beurteilung der Rückfallwahrscheinlichkeit eines Straftäters zwar auf „race“ als Eingabevariable, wirkt aber genau anhand dieser Variable diskriminierend. Dies geschieht über stellvertretende Variablen (wie beispielsweise Wohnort, finanzielle Situation etc.) über welche Rückschlüsse auf die ursprüngliche diskriminierende Variable gezogen werden können. Inwieweit Versuche, die Daten im Vorhinein entsprechend zu bereinigen, erfolgversprechend wären, ist umstritten (Kilbertus et al. 2017, Doshi-Velez/Kim 2017). Insofern ist auch zu diskutieren, wie bloßer diskriminierender Output zu bewerten ist, wenn er nachweisbar nicht auf diskriminierendem Input basiert.

Generell setzt dieser Ansatz Einigkeit darüber voraus, welche Kriterien diskriminierend sind bzw. welche für uns derzeit noch nicht vorstellbaren Korrelationen akzeptabel sind und welche nicht. Wie problematisch diese Festsetzung ist, zeigen die bereits lange andauernden Debatten um Art. 3 GG, welcher die Gleichheit vor dem Gesetz und die Gleichberechtigung der Geschlechter garantiert sowie Diskriminierung, Bevorzugung und Benachteiligung aufgrund verschiedener Merkmale wie beispielsweise Geschlecht, Herkunft, Glauben und Behinderung verbietet. Die vermehrte Nutzung Lernender Systeme wird diese Debatten verstärken und die Notwendigkeit verschärfen, sich zu einigen. Denkbar wäre auch hier, Institutionen zu schaffen, die für diese inhaltliche Entscheidung legitimiert werden, statt zu versuchen, alle Kategorisierungen vorab zu beurteilen.

Gerechte Behandlung als Ziel maschinellen Lernens

Eine weitere Möglichkeit wäre, eine gerechte Behandlung selbst zum Ziel maschineller Lernverfahren zu machen. Dann ginge es nicht mehr darum, möglichst effiziente oder genaue Klassifikationen zu ermöglichen, sondern eben möglichst gerechte. Das wird unter dem Stichwort der „Fairness“ in der Forschung zu Künstlicher Intelligenz thematisiert. Allerdings schlägt sich in diesen Versuchen eine Problematik nieder, die die informatische Forschung ganz grundsätzlich beschäftigt. Unsere Vorstellungen davon, was „gerecht“ oder

„fair" ist, lassen sich in ihrer Komplexität nicht formalisieren und damit ohne Weiteres zum Lernziel von maschinellem Lernen machen. Ein entsprechender Versuch von Kleinberg et al. (2016) zeigt drei Möglichkeiten, wie Fairness formal ausgedrückt werden könnte:

- Die Vorhersage soll **„well calibrated"** sein: Wenn ein Algorithmus vorhersagt, dass eine bestimmte Eigenschaft mit einer bestimmten Wahrscheinlichkeit, z. B. 0.1, auf eine Gruppe zutrifft, dann sollten ein der Wahrscheinlichkeit entsprechender Anteil der Gruppe auch diese Eigenschaft haben, hier also ein Zehntel.

- Wenn es mehrere Gruppen unter den Klassifizierten gibt, z. B. Männer und Frauen, dann könnte man fordern dass es eine **„balance for the positive class"** gibt: Die durchschnittliche Wahrscheinlichkeit für eine Eigenschaft, die den Menschen zugewiesen wird, die die Eigenschaft tatsächlich besitzen, sollte in jeder Gruppe gleich sein. Das stellt sicher, dass keine Gruppe übermäßig häufig falsch-positiv klassifiziert wird.

- Analog hierzu kann man fordern, dass es eine **„balance for the negative class"** gibt: Die durchschnittliche Wahrscheinlichkeit für eine Eigenschaft, die den Menschen zugewiesen wird, die die Eigenschaft nicht besitzen, sollte in jeder Gruppe gleich sein. Das stellt sicher, dass keine Gruppe übermäßig häufig falsch-negativ klassifiziert wird.

Kleinberg et al. zeigen nun aber gerade, dass es nicht möglich ist, diese drei intuitiv richtigen Charakterisierungen alle gleichzeitig perfekt zu erfüllen. Das verdeutlicht die inhärente technische Grenze des Ansatzes, Fairness in maschinelles Lernen zu integrieren.

Effektiver Rechtsschutz und Rechtsdurchsetzung

Neben den oben aufgeführten Lösungsansätzen ist es darüber hinaus wichtig, dass die Betroffenen selbst in die Lage versetzt werden, ihre Rechte zu verteidigen. Dies bedeutet nicht, dass sie zu Expertinnen und Experten auf dem Gebiet der Künstlichen Intelligenz ausgebildet werden müssen. Vielmehr sollten die Betroffenen über ihre Rechte informiert und in die Lage versetzt werden, diese auch faktisch einfordern zu können. Dies schließt die Möglichkeit ein, seine Rechte vor Gericht einzufordern. Die finanziellen Aufwendungen hierfür sollten möglichst gering gehalten werden. Eventuell sollte auch der Abschluss einer Versicherung gegen Diskriminierung durch Lernende Systeme ermöglicht werden. Staatlichen Behörden kommt die Aufgabe zu, einer rechtswidrigen Diskriminierung durch Selbstlernende Systeme entgegenzuwirken. Bei all diesen Maßnahmen sollte allerdings auf ein angemessenes Maß an Regulierung geachtet und Überregulierung vermieden werden.

Impressum
Herausgeber: Lernende Systeme – Die Plattform für Künstliche Intelligenz | Geschäftsstelle | c/o acatech | Karolinenplatz 4 | D-80333 München | kontakt@plattform-lernende-systeme.de | www.plattform-lernende-systeme.de |
Folgen Sie uns auf Twitter: @LernendeSysteme | Stand: Juni 2019 | Bildnachweis: r.classen / Shutterstock

Diese Kurzfassung entstand auf Grundlage des Whitepapers *Künstliche Intelligenz und Diskriminierung – Herausforderungen und Lösungsansätze*, München, 2019. Die Autorin und Autoren sind Mitglieder der Arbeitsgruppe IT-Sicherheit, Privacy, Recht und Ethik der Plattform Lernende Systeme. Die Originalfassung der Publikation ist online verfügbar unter: https://www.plattform-lernende-systeme.de/publikationen.html

GEFÖRDERT VOM

Bundesministerium für Bildung und Forschung

acatech
DEUTSCHE AKADEMIE DER TECHNIKWISSENSCHAFTEN