Lernende Systeme

DIE PLATTFORM FÜR KÜNSTLICHE INTELLIGENZ

# Künstliche Intelligenz und Diskriminierung

## Herausforderungen und Lösungsansätze

WHITEPAPER

**Susanne Beck et al.**
AG IT-Sicherheit,
Privacy, Recht und Ethik

# Inhalt

# Zusammenfassung

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

Die Problematik der potentiellen 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.

1 Ein erster Überblick über potentiell 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.

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. Ein möglicher Ansatzpunkt ist, die Erklärbarkeit und Überprüfung der Vorgänge zu fördern – etwa durch eine unabhängige Instanz, die als Stellvertreter der Bürgerinnen und Bürger agiert. Weiterhin gilt es, die Kriterien, anhand derer ein Algorithmus lernt, vorzuselektieren. Wichtig für diskriminierungsfreie KI-Anwendungen ist zudem, Gerechtigkeit als Zielgröße maschinellen Lernens zu definieren und Menschen, die durch eine KI-basierte Entscheidung benachteiligt wurden, eine effektive Rechtsdurchsetzung zu ermöglichen.

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.

# 1. Diskriminierung durch Algorithmen: Aktuelle Beispiele

Die Problematik der Diskriminierung (Definition siehe Kasten S. 7) durch Algorithmen wird im Folgenden anhand einiger realer Beispiele konkretisiert. Überwiegend handelt es sich hierbei um KI-Systeme (Definition siehe Kasten S. 7), die auf Grundlage maschineller Lernverfahren arbeiten. Fraglich ist jeweils, ob es sich tatsächlich um eine ungerechtfertigte Diskriminierung handelt und wie sich diese begründen lässt.

### ■ Fall 1: Verarbeitung von Bewerbungen

Eine große Firma setzt einen Algorithmus ein, um Bewerbungen hinsichtlich ihrer Passfähigkeit auf die ausgeschriebene Stelle vorzusortieren. Bewerbungen von Frauen werden dabei systematisch schlechter bewertet als Bewerbungen von Männern. Grund sind die Trainingsdaten, anhand derer der Algorithmus gelernt hat: Als Input verwendet wurden dafür die Bewerbungen der erfolgreich eingestellten Mitarbeiterinnen und Mitarbeiter der letzten zehn Jahre; diese waren überwiegend männlich. Der Algorithmus betrachtete die Eigenschaft „männlich" daraufhin als positiv.[4]

### ■ Fall 2: Werbung auf Online-Suchmaschinen

Eine Suchmaschinenwebsite zeigt mithilfe eines Algorithmus Werbung für weitere Websites an, darunter auch Websites, die sich auf das Auffinden von Personen spezialisiert haben. Werden in das Suchmaschinenfeld unterschiedliche Namen eingetippt, so führt dies zu unterschiedlichen Werbeergebnissen: Vornamen, die in der Wahrnehmung von Befragten häufiger Schwarzen Menschen zugeordnet werden (wie beispielsweise DeShawn, Darnell und Jermaine), führen öfter zu Werbeanzeigen, die den Namen der gesuchten Person mit einem Haftbefehl oder Straftaten in Verbindung bringen. Dies geschieht unabhängig davon, ob gegen diese Person tatsächlich ein Haftbefehl vorliegt oder vorlag. Vornamen, die in der Wahrnehmung von Befragten häufiger weißen Menschen zugeordnet werden (wie beispielsweise Geoffrey, Jill und Emma), führen deutlich seltener zu solchen Werbeeinblendungen.[5]

### ■ Fall 3: Preis- und Suchdiskriminierung im Internet

Der Preis für Leistungen im Internet (z. B. für Hotel-, Flug- oder Mietwagenbuchungen) ist in manchen Ländern nicht für alle Interessentinnen und Interessenten identisch: Jene Personen, denen eine höhere Zahlungsbereitschaft und auch -fähigkeit zugeschrieben wird,

4 Fall in Anlehnung an Amazon; für weitere Informationen siehe Dastin 2018.
5 Fall in Anlehnung an GoogleAdSense in den USA; für eine detaillierte Auswertung siehe Sweeney 2013.

erhalten die Leistung zu einem höheren Preis als andere Interessentinnen und Interessenten angeboten (Individual Pricing).[6] Die Ermittlung der Zahlungsbereitschaft hängt von vielen Informationen (z. B. Wohnort, verwendetes Gerät und andere besuchte Websites) ab, welche online (z. B. mithilfe von Cookies) gesammelt werden. So kann der Anbieter bei verschiedenen individuellen Kunden oder auch bei verschiedenen Kundengruppen die maximale Rendite erzielen.[7]

### ■ Fall 4: Planung von Polizeieinsätzen/Predictive Policing

Künstliche Intelligenz kann eingesetzt werden, um die Effizienz von Streifengängen der Polizei zu erhöhen. Der Algorithmus schlägt hierbei Routen oder Gebiete vor, in denen mit einer großen Wahrscheinlichkeit Straftaten erwartet werden. Da die Polizeibeamtinnen und -beamten allein durch ihre Präsenz vor Ort mehr Straftaten als an anderen Orten beobachten, lernt der Algorithmus wieder, dass diese Gegend ein Brennpunkt ist, und plant für diese Gegend eine erhöhte Anzahl an Streifengängen, bei welchen wiederum Straftaten beobachtet werden. Es kann also ein sich selbst verstärkender Prozess in Gang gesetzt werden.[8]

### ■ Fall 5: Einsatz eines Chatbots

Eine große Firma entwickelt einen Chatbot, der aus Konversationen lernen und das Gelernte dann später auf weitere Gespräche anwenden kann. Bereits nach einiger Zeit war die Ausdrucksweise des Chatbots von Rassismus, Sexismus sowie Antisemitismus geprägt. Diese, von den Programmiererinnen und Programmierern unbeabsichtigten Äußerungen des Chatbots sind auf eine konzertierte Aktion mehrerer Userinnen und User zurückzuführen. Sie manipulierten den Chatbot durch hasserfüllte Konversationen, sodass der Algorithmus entsprechende Formulierungen lernte.[9]

### ■ Fall 6: Der automatische Seifenspender – ein Beispiel für falsch eingestellte Hardware

Ein automatischer Seifenspender wurde entwickelt. Die Einstellung und Auswahl der Sensoren erfolgte mit einer weißen Person als Testperson. Der Seifenspender wird daraufhin produziert und eingesetzt. Als eine nicht-weiße Person ihn verwenden will, reagiert er nicht. Dies ist darauf zurückzuführen, dass die Sensoren aufgrund des einseitigen Testprozesses nur weiße Haut erkennen können.[10]

6 Ein Beispiel hierfür sind die USA. In Deutschland hingegen werden dieser Praktik durch den Datenschutz Grenzen gesetzt.
7 Für weitere Informationen siehe Mikians et al. 2012 und Bitkom 2017.
8 Weitere Informationen siehe Richter/Kind 2016.
9 Fall in Anlehnung an den Chatbot „Tay" der Firma Microsoft; für weitere Informationen siehe Misty 2016 sowie Hagendorff 2019.
10 Weitere Informationen siehe Fussell 2017.

### Fall 7: Profiling

Ein Algorithmus wurde entwickelt, um einen Wert für die Rückfallwahrscheinlichkeit einer oder eines Angeklagten einzuschätzen. Die Anwendung zeigt, dass Schwarze Menschen von diesen Algorithmen aufgrund ihrer Hautfarbe eine schlechtere Prognose erhalten als weiße Menschen. Der Grund für diese Diskriminierung ist darauf zurückzuführen, dass der Algorithmus vor allem mit historischen Kriminalitätsdaten trainiert wurde. Diese Daten beruhen allerdings auf statistischen Korrelationen und nicht auf kausalen Zusammenhängen.[11]

**Diskriminierung**

Diskriminierung stammt von dem lateinischen Wort „discriminare" (unterscheiden) ab – was keine per se negative Bedeutung hat. Mithilfe von Unterscheidungen können Kontingenzen reduziert und Sachverhalte einfacher erfasst werden. Zentral ist, ob diese Unterscheidungen gerechtfertigt sind oder nicht. Diskriminierung im negativen Sinn liegt bei einer ungerechtfertigten Ungleichbehandlung von Gleichem oder einer ungerechtfertigten Gleichbehandlung von Ungleichem vor. In vorliegendem Papier wird der Begriff „Diskriminierung" in seiner negativen Bedeutung verwendet, das heißt als ungerechtfertigte Gleich- oder Ungleichbehandlung verstanden.

**Künstliche Intelligenz (KI)**

Als Teilgebiet der Informatik versucht Künstliche Intelligenz, kognitive Fähigkeiten wie Lernen, Planen oder Problemlösen in Computersystemen zu verwirklichen. Zugleich steht der Begriff KI für Systeme, deren Verhalten gemeinhin menschliche Intelligenz voraussetzt. Da der Intelligenzbegriff jedoch nicht eindeutig festgelegt ist und sich das Verständnis für KI auch abhängig vom Stand der Technik verändert, konnte sich noch keine alleinige allgemein akzeptierte Definition durchsetzen. Ziel der Forschung ist es, moderne KI-Systeme wie Maschinen, Roboter und Softwaresysteme zu befähigen, abstrakte Aufgaben und Probleme auch unter veränderlichen Bedingungen eigenständig zu bearbeiten und zu lösen, sodass der Mensch nicht jeden einzelnen Schritt programmieren muss. Sämtliche heute technisch umsetzbaren KI-Systeme ermöglichen eine Problemlösung in beschränkten Kontexten (z. B. Sprach- oder Bilderkennung für spezifische Anwendungen, etwa das Erkennen von Tumoren auf Röntgenbildern) und zählen damit zur sogenannten schwachen KI.

11 Fall in Anlehnung an den Algorithmus COMPAS; für weitere Informationen siehe Angwin et al. 2016.

# 2. Ursachen und Formen der Diskriminierung durch Lernende Systeme

## 2.1 Mögliche Quellen von Diskriminierung

Betrachten wir zunächst die Entstehung von Diskriminierung. Oft wird angenommen, dass sie absichtlich oder durch ein Versehen von Ingenieurinnen und Ingenieuren oder Programmiererinnen und Programmierern zustande kommt. Zugleich müssen nicht alle Diskriminierungen als Fehler anzusehen sein, wenn sie gerade das erreichen, was dem Lernenden System als Ziel vorgegeben wurde. Neben fehlerhaftem Handeln gibt es weitere Einfallstore für Diskriminierung – so werden die Systeme regelmäßig mithilfe von Daten aus dem Internet trainiert und übernehmen so die dort vorhandenen Verzerrungen (siehe Definition „Bias“).

**Bias**
Ein Bias bezeichnet allgemein Verzerrungseffekte. Die Psychologie versteht darunter Einstellungen oder Stereotypen, welche die Wahrnehmung unserer Umwelt, Entscheidungen und Handlungen positiv oder negativ beeinflussen. Diese Beeinflussung kann unbewusst (impliziter Bias) oder bewusst (expliziter Bias) geschehen. In der Statistik wird ein Bias als Fehler im Rahmen der Datenerhebung und -verarbeitung (z. B. Fehler in der Stichprobenauswahl) oder die bewusste oder unbewusste Beeinflussung von Probandinnen und Probanden verstanden.

In der Informationstechnologie werden drei Kategorien eines Bias unterschieden (Friedman und Nissenbaum 1996):

- **Prä-existierender Bias:** Häufig wird eine in der Gesellschaft etablierte Voreingenommenheit in die Software übertragen. Das kann einerseits explizit geschehen, wenn etwa eine diskriminierende Haltung ganz bewusst eingebaut wird. Andererseits kann dies auch implizit geschehen: Aus dem Bereich des Predictive Policing ist beispielsweise bekannt, dass die Vorurteile von Beamtinnen und Beamten sich in den kontrollierten Orten und Fallzahlen niederschlagen. Ein weiteres Beispiel ist die Verwendung der Bewerbungen der letzten zehn Jahre als Trainingsdaten für die Bewerbungsverarbeitung mit KI-Systemen. Das Trainieren des Profiling-Algorithmus mithilfe historischer Daten ist ein ähnlicher Fall. Wird aus solchen Daten gelernt, entstehen verzerrte maschinelle Modelle (Kaufmann 2018).

- **Technischer Bias:** Technische Vorgaben – etwa in der Sensorik – können dazu führen, dass bestimmte Gruppen von Menschen anders behandelt werden als andere. Ein Beispiel hierfür ist der genannte Seifenspender (siehe S. 6). Weitere Beispiele wären Standards, die es nicht erlauben, bestimmte Eigenschaften zu erfassen, sowie die Übersetzung menschlicher Begriffe in mathematische Modelle, wobei sich die Bedeutung ändert. Ein technischer Bias kann aber beispielsweise auch aufgrund unterschiedlicher Bildschirmgrößen entstehen: Die Resultate auf der ersten Seite einer Suchmaschine werden mit einer größeren Wahrscheinlichkeit aufgerufen als die Ergebnisse auf den folgenden Seiten. Wie viele Resultate auf der ersten Seite angezeigt werden, hängt vom Bildschirm ab.

- **Emergenter Bias:** Diskriminierungen können auch aus dem Zusammenspiel von Software und Anwendung entstehen, wie beispielsweise durch falsche Interpretation der Ausgaben, was oft bei statistischen Werten auftritt. Auch die Verwendung einer Software aus einem bestimmten Kontext für andersgeartete Anwendungsfälle birgt dieses Problem. Solche Phänomene entstehen zum Teil erst mit der Zeit, etwa wenn sich soziale Handlungsmuster, Wertvorstellungen oder Prozesse ändern, die Technik sich dem jedoch nicht anpasst.

Einige Diskriminierungen entstehen also durch implizit oder explizit voreingenommene Eingaben oder technische Entscheidungen im Entwurf. Das lässt sich als diskriminierender Input bezeichnen. Dieser kann prinzipiell durch Analyse des Systems und der Eingabedaten diagnostiziert werden. Praktisch stehen dem aber die Größe und Komplexität der Datenmengen und des Quellcodes entgegen. Auf dieser Grundlage können anschließend Änderungen vorgenommen werden.

Manche Formen der Diskriminierung entstehen aber auch erst in der Anwendung. Diese Form der Diskriminierung lässt sich nur im Betrieb des Systems oder in Testläufen im Output feststellen. Ein Beispiel hierfür ist in der Testversion des Chatroboters „Tay" zu finden. Dieser wurde nach Veröffentlichung systematisch von organisierten Usern mit fremdenfeindlichen und diskriminierenden Konversationen „gefüttert", sodass die lernende Software hinter dem Chatroboter diese Ausdrucksweise in der Anwendung „gelernt" hat und jetzt anwendet.

An dieser Stelle deutet sich bereits eine der Herausforderungen im Kontext von Diskriminierung durch Lernende Systeme an: Viele der Daten, Rechenprozesse und damit Gründe für den diskriminierenden Output werden aufgrund des Lernens und Trainings des Systems ex ante nicht umfassend festlegbar und ex post nicht immer nachvollziehbar sein. Eine weitere Herausforderung besteht darin, dass unterschiedliche Formen des Bias und der Diskriminierung in unterschiedlichen Kombinationen auftreten können.

## 2.2 Notwendige Unterscheidungen und deren Plausibilität

Eine wichtige Voraussetzung, um beurteilen zu können, ob eine Unterscheidung gerechtfertigt oder ungerechtfertigt ist und beurteilen zu können, ob und was eingedämmt werden muss, ist es, das Konzept der Diskriminierung als solches zu definieren. So existieren nicht ausschließlich zwei Optionen – Diskriminierung vs. Nicht-Diskriminierung –, sondern viele verschiedene Abstufungen entlang eines Kontinuums. Für gewöhnlich wird unter Diskriminierung eine ungerechtfertigte Ungleichbehandlung verstanden. Aber auch die ungerechtfertigte gleiche Behandlung von Ungleichem kann diskriminierend sein, weil so beispielsweise besondere Bedarfe bestimmter Personengruppen nicht mehr abgebildet werden können (z. B. besondere Sozialleistungen).

Maschinelles Lernen (Definition siehe Kasten) entscheidet grundsätzlich mittels Gruppenzuordnungen. Bereits heute basieren viele Algorithmen auf Klassifikationen: Ob etwa ein Mensch einen Kredit erhält oder ob er auf Bewährung aus dem Gefängnis kommt, wird mittels statistischer Merkmale zumindest vorentschieden; Lernende Systeme werden diese Methode schlicht durch größere Datenmengen und selbst optimierte Rechenverfahren verbessern.

**Maschinelles Lernen**
Maschinelles Lernen ist eine Schlüsseltechnologie der KI. Auf Basis einer großen Anzahl an Beispieldaten entwickeln Maschinen dabei Regeln und Modelle, die auf neue, unbekannte Situationen und Daten angewendet werden können. Es ist grob zu unterscheiden zwischen überwachtem und unüberwachtem Lernen.

- Beim **überwachten Lernen** werden den Algorithmen im Trainingsprozess sowohl der gewünschte Output als auch eindeutige Trainingsdaten vorgegeben. Soll der Algorithmus beispielsweise lernen, einen Hund von einem Wolf zu unterscheiden, erhält er Beispielbilder von Hunden und Wölfen. Das System erhält im Trainingsprozess Feedback, ob die Eingabe richtig zugeordnet wurde und zur richtigen Ausgabe – Hund oder Wolf – geführt hat. So lernt das System und verbessert das Modell, mittels dessen eine Zuordnung oder Vorhersage getroffen wird. Anschließend wird das System in Betrieb gebracht und die gelernten Regeln und Modelle werden auf unbekannte Daten angewendet.

- Beim **unüberwachten Lernen** werden die Rohdaten ohne vorgegebenes Prognoseziel übergeben. Der Lernalgorithmus entwickelt selbstständig Klassifikatoren, nach denen er die Eingabemuster einteilt. Ziel ist es, in einem großen, unstrukturierten Datensatz interessante und relevante Muster zu erkennen oder die Daten kompakter zu repräsentieren. Ein Beispiel ist die Segmentierung von Kundendaten nach Zielgruppen, die man auf ähnliche Weise adressieren möchte.

Bereits die obige Definition von Diskriminierung (siehe Kasten Seite 7) zeigt, dass jedoch nicht die Differenzierung als solche problematisch ist – unsere Gesellschaft basiert notwendigerweise auf Unterscheidungen (z. B. Schulnoten, Berufsqualifikationen, Steuerklassen) und auch auf allgemein akzeptierten Ungleichbehandlungen von benachteiligten Gruppen, die deren Benachteiligung entgegenwirken sollen (ein Beispiel ist die Einführung einer Quote für Frauen in Aufsichtsräten). Entscheidend ist also vielmehr, ob die jeweilige Ungleichbehandlung (bzw. Gleichbehandlung von Ungleichem) gerechtfertigt ist oder nicht. Wird diese Diskussion nach der Legitimität und Legalität von Ungleichbehandlung geführt, so ist die Verortung auf diesem Kontinuum von Diskriminierung keinesfalls eine einfache Aufgabe und eindeutig, sondern kann konfliktbehaftet und umstritten sein.

Bei der Beurteilung der Legalität und Legitimität der Ungleichbehandlung fordert maschinelles Lernen unsere bestehenden Intuitionen und Normen heraus:

- **Auf Beurteilungen grundsätzlich verzichten:** An einem Ende des Spektrums wäre denkbar, jede Beurteilung durch Lernende Systeme als ungerechtfertigt zu sehen, da diese immer auf statistischen Werten und entsprechenden Kategorien basiert. Man könnte fordern, nicht als Teil einer statistisch signifikanten Gruppe, sondern als Einzelfall betrachtet zu werden. Alles andere wäre dann ungerechtfertigt.

- **Angemessen differenzieren:** Eine zweite Möglichkeit wäre, Beurteilungen durch maschinelles Lernen grundsätzlich zu akzeptieren. Zu diskutieren wäre dann jedoch die Angemessenheit jeder konkreten Differenzierung. In diesem Sinn wäre zweifellos die Klassifikation durch die Software zur Beurteilung der erneuten Straffälligkeit diskriminierend, weil weiße Hautfarbe dazu führt, dass man mit höherer Wahrscheinlichkeit als nicht rückfallgefährdet gesehen wird. In vielerlei Hinsicht wird man auf bestehenden Konsens zu akzeptablen Unterscheidungen zurückgreifen können, doch werden durch Lernende Systeme neue Korrelationen entstehen, über deren Bewertung neu zu diskutieren sein wird.

- **Bewertungen begründen:** Schließlich spielt auch eine Rolle, ob die Begründung bzw. Begründbarkeit der Entscheidung die Bewertung als ungerechtfertigte Diskriminierung beeinflusst. Es könnte sein, dass zwar die statistische Entscheidung akzeptiert, dann aber zumindest eine nachvollziehbare Begründung bzw. eine auf den Einzelfall Bezug nehmende Begründbarkeit verlangt wird.

## 2.3 Entscheidungen über Differenzierung vs. Diskriminierung

Für die Frage, was eine Diskriminierung darstellt, sind folgende Ansatzpunkte denkbar:

- **Sozialer Kontext:** Man spricht nur dann von Diskriminierung, wenn eine Gruppe durch Merkmale ausgezeichnet wird, die bereits als diskriminierend gelten (z. B. Hautfarbe, Geschlecht, Herkunft, Religion, sexuelle Orientierung oder Behinderung).

- **Datenstruktur:** Als diskriminierend gilt auch, wenn maschinelle Lernverfahren eine Gruppe von Menschen systematisch und ungerechtfertigt benachteiligen, die sich durch bestimmte Muster in heterogenen Daten auszeichnet (z. B. Aktivitäten im Internet oder in Social Media, Kreditkarten- und Bankdaten, Bewegungsprofile, Ratings oder Reisen).

Zu beachten ist, dass das Versprechen von maschinellem Lernen (speziell bei unüberwachtem Lernen) auch darin liegt, neue Muster in Daten zu erkennen, die sich Menschen bislang nicht erschlossen haben. Es geht also nicht nur um die Repräsentation einer Bevölkerungsgruppe durch neue Merkmale. Beispielsweise entstehen bei der algorithmischen Überwachung vielmehr neue Gruppen von Daten-Verdächtigen (Matzner 2016); durch völlig neue Korrelationen entstehen neue Unterscheidungen und Verknüpfungen, neue Gleich- und Ungleichbehandlungen – und wir werden intuitiv gar keine Vorstellung mehr davon haben, ob dieser Verdacht gerechtfertigt ist. Es ist eine Form des Verdachts, der aufgrund der Charakteristika des maschinellen Lernens überhaupt entstehen kann – und sich deutlich von der uns bekannten Form des Verdachts unterscheidet. So wird das Verhalten eines Menschen durch einen anderen Menschen beispielsweise als „verdächtig" eingestuft, wenn sich dieser unkonventionell verhält, sich sonderbar bewegt oder unerwartete Handlungen unternimmt. Eine Software hingegen bestimmt „verdächtige" Personen beispielsweise über Kreditkartentransaktionen, getätigte Onlinekäufe sowie Bewegungspunkte, welche aus Daten mobiler Endgeräte erstellt werden.

Mithilfe dieser Software entstehen sehr große Datenströme, deren Analyse neue Muster und Gruppen zu Tage fördern kann. Diese Datenströme könnten auch für ein Social-Scoring-Verfahren eingesetzt werden, wie es beispielsweise in einigen Städten in China getestet wird. Social-Scoring-Verfahren beschreiben das automatisierte Sammeln und Auswerten von sehr vielen Daten über das Verhalten von Bürgerinnen und Bürgern zur Ermittlung eines Punktwertes. Dieser beschreibt, ob die jeweilige Person ein „guter" und teilweise auch „linientreuer" Bürger ist oder ob er oder sie sich unerwünscht verhält. Ein hoher Punktwert geht mit Annehmlichkeiten wie dem einfacheren Zugang zu bestimmten Leistungen (wie beispielsweise Flugreisen) einher, während ein niedriger Score Nachteile mit sich bringen kann. Neben dem Aspekt, dass das Sammeln von Daten einen massiven Eingriff in die Privatsphäre darstellt, ist die Einteilung von Menschen in Klassen ethisch höchst umstritten.

Die Klassifizierung eines Menschen als „verdächtig" ist ein Fall, der vielen bekannt ist und bei welchem die menschliche Intuition zur Beurteilung des Ergebnisses unter Umständen hilfreich sein kann. Dies ist allerdings nur ein Beispiel von vielen. Was aber geschieht, wenn wir dieses Gedankenexperiment auf andere, komplexere und nicht so leicht greifbare Zusammenhänge anwenden? Hier kann eine Software Korrelationen in Daten erkennen und diese Daten als sinnvoll wahrnehmen; auch wenn sich diese Zusammenhänge dem menschlichen Verständnis und somit auch der Möglichkeit der Nachvollziehbarkeit entziehen. In solchen Fällen kann sich der Mensch nur schwer auf seine Intuition zur Beurteilung der Sinnhaftigkeit und Plausibilität der gebildeten Zusammenhänge verlassen. Letztendlich droht eine implizite Bedeutungsverschiebung mancher Begriffe – die aber vom Menschen nicht als diese wahrgenommen oder erkannt werden kann. Hier besteht Forschungsbedarf hinsichtlich der Frage, ab wann Nachteile, die durch diese Bedeutungsverschiebungen entstehen, statt als Fehler oder Einzelfälle als eine neue Form der Diskriminierung gewertet werden sollten.

Würden solche Formen von Benachteiligung als Diskriminierung gewertet, könnte das dem Versuch widersprechen, die Bewertung durch maschinelles Lernen so spezifisch und persönlich wie möglich zu machen. Entsprechend sollte im Rahmen eines gesellschaftlichen Dialogs abgewogen werden, ob es Anwendungsbereiche gibt, in denen die nicht immer nachvollziehbare Gruppen- und Musterbildung möglicherweise unerwünscht ist, etwa weil die Zuordnung von Personen aufgrund einer Wahrscheinlichkeit nicht zulässig oder unangemessen wäre, wie beispielsweise im Bereich des Zugangs zu bestimmten staatlichen Leistungen.

Damit könnte durch die Forderung nach Gleichbehandlung zugleich eine Grenze für Personalisierung und Spezifizierung gefordert werden. Neben der Vermeidung von Diskriminierung kommt das auch dem Solidaritätsprinzip entgegen, das in vielen Bereichen, zum Beispiel in der gesetzlichen Renten- und Arbeitslosenversicherung, wirksam ist. Das Solidaritätsprinzip bedeutet, dass der oder die Einzelne nicht allein für sich selbst verantwortlich ist, sondern auch für die anderen Mitglieder der (Solidar-)Gesellschaft. Eine Versicherung beruht im Sinne eines ähnlichen Prinzips darauf, dass trotz verschiedener Tarife Menschen „ungleich" behandelt werden: Diejenigen, die den Schutz nicht brauchen, zahlen mehr als jene, die ihn brauchen. Ein fiktiv angenommener perfekter Prädiktionsalgorithmus würde das Prinzip und damit die Querfinanzierung durch die „ungleich" Behandelten aushebeln. Solche Prinzipien sind auch bei der Preisgestaltung vieler Produkte und Dienstleistungen oder dem Betrieb von Verlagen oder Plattformen wirksam. Dies ist bei der Diskussion über die Angemessenheit von Differenzierungen zu beachten.

# 3. Herausforderungen

Zu diskutieren ist aber nicht nur, welche Differenzierungen angemessen und welche unangemessen sind – denn alle Konstellationen vorab zu entscheiden ist nicht nur unmöglich, sondern in einer demokratischen Gesellschaft auch Aufgabe eben dieser Gesellschaft und nicht einzelner Expertinnen und Experten. Wichtig ist also auch, wer konkret darüber entscheiden darf. Das kann nicht der oder die jeweilige Programmierende (oder das entsprechende Team) allein sein, da hier möglicherweise eine in der Gesellschaft etablierte Voreingenommenheit unreflektiert übernommen wird (prä-existierender Bias) und weil hier möglicherweise die technische Machbarkeit im Vordergrund steht. Nicht überzeugend ist es auch, keine Entscheidung zu treffen und schlicht zu akzeptieren, dass mit den Trainingsdaten, beispielsweise in Form von Dokumenten, Videos oder Bildern, die im Internet frei verfügbar sind, die Systeme vorhandene Diskriminierungen aufgreifen und gegebenenfalls verschärfen. Eine moralische Bewertung kann nicht durch die Systeme selbst vorgenommen werden. So ist beispielsweise „Gerechtigkeit" eine normative und damit kontrafaktische Größe. Lernende Systeme, die auf Mustererkennung basieren, können nichts Kontrafaktisches entwickeln. Wenn sie Daten aus der Realität für ihr Lernen nutzen, können sie nur die ethisch nicht tragbare „Normativität des Faktischen" reproduzieren.

Es sind somit gesellschaftliche Institutionen zu suchen oder zu schaffen, die zur Entscheidung über angebrachte Kategorisierungen berufen sind. Auf welcher Ebene eine solche Institution angesiedelt werden sollte, ob bestehende Institutionen genutzt werden können oder welche Akteure einbezogen werden müssen, muss in einem gesellschaftlichen Diskussionsprozess präzisiert werden. Auch die Bundesregierung hat diese Notwendigkeit erkannt und kündigt in ihrer Strategie Künstliche Intelligenz an, die Einrichtung oder den Ausbau von staatlichen und privaten Institutionen zur Kontrolle algorithmischer Entscheidungen zu prüfen (Bundesregierung 2018: 40).

In dem Bemühen, ungerechtfertigte Diskriminierungen (Definition siehe Kasten S. 7) durch Lernende Systeme zu vermeiden, sehen sich Entwicklerinnen und Entwickler, Anwenderinnen und Anwender wie auch die Gesellschaft mit ganz unterschiedlichen Herausforderungen konfrontiert.

- **Unmöglichkeit der Transparenz:** Die häufig geforderte Transparenz von maschinellen Entscheidungen ist aus faktischen und rechtlichen Gründen problematisch. Faktisch sind die Algorithmen so komplex, basiert das Deep Learning (Definition siehe Kasten, S. 15) auf so vielen Daten, dass ein Nachvollziehen und Bestimmen der Daten und Kategorisierungen durch einen Menschen prinzipiell realistisch kaum möglich erscheint. Entsprechend ist die Möglichkeit, Transparenz herzustellen, vom verwendeten Lernverfahren abhängig. So sind die Ergebnisse klassischer Lernverfahren oft sehr gut interpretierbar (Fraunhofer 2018: 12, 30). Neben der Frage, ob Transparenz technisch überhaupt hergestellt werden kann, ist ebenso relevant, wie Transparenz für Bürgerinnen

und Bürger ohne besonderes Fachwissen hergestellt werden kann. Rechtlich gelten Algorithmen zudem als Firmengeheimnis, sodass es bisher zulässig ist, wenn Firmen wie Google oder Facebook ihre Algorithmen nicht offenlegen.

**Deep Learning**
Deep Learning (tiefes Lernen) bezeichnet das maschinelle Lernen mit großen künstlichen neuronalen Netzen. Es handelt sich hierbei um Knotenschichten (sogenannte Neuronen), die durch eine Software realisiert und numerisch gewichtet miteinander verbunden sind. Diese Gewichtung kann während des Trainingsprozesses angepasst werden, sodass die Ergebnisse sich verbessern können. Je komplexer das Netz (gemessen an der Anzahl der Schichten, der Verbindungen zwischen Neuronen sowie der Neuronen pro Schicht), desto komplexere Sachverhalte können verarbeitet werden. Deep Learning hat in vielen Bereichen bereits bemerkenswerte Durchbrüche erzielt und wird etwa in der Verarbeitung natürlicher Sprache oder beim Erkennen von Objekten eingesetzt.

- **Unkontrollierbarkeit des Weiterlernens:** Systeme, die im laufenden Betrieb weiterlernen, können sich einer Vorabkontrolle oder der Möglichkeit strengerer Vorgaben entziehen. Auch wenn das Lernen der Software begrenzt und gelenkt wird, kann nie mit vollständiger Sicherheit ausgeschlossen werden, dass unerwünschte Lernvorgänge stattfinden. Dies kann beispielsweise dann geschehen, wenn auf Grundlage eines eigentlich nicht repräsentativen Falls generalisierte Rückschlüsse gezogen und diese auf die Bearbeitung anderer Fälle angewandt werden.

- **Fehlende Neutralität der Daten:** Die bedingte Kontrollierbarkeit der Weiterentwicklung basiert auch darauf, dass die Entwicklerinnen und Entwickler nur beschränkten Einfluss auf die Daten haben. Schon für das Training müssen viele Daten bereitstehen, sodass meist auf Daten aus dem Internet zurückgegriffen wird statt auf eigene, gefilterte Daten. Diese Daten und die, an denen das System später weiterlernt, sind nicht neutral. Sie spiegeln bestehende gesellschaftliche Diskriminierungen bzw. verschärfen diese möglicherweise noch, weil im Internet nur ein Teil – und möglicherweise in mancherlei Hinsicht extremer Teil – der gesellschaftlichen Werte und Kategorisierungen zu finden ist.

- **Unklare Verantwortlichkeiten:** Generell wird im Kontext von Lernenden Systemen das Problem diffundierender Verantwortlichkeiten diskutiert. Diese Unklarheiten schreiben sich auch bei der Problematik der Diskriminierung fort – das gilt sowohl für die Verantwortlichkeit ex ante, für die Auswahl und angemessene Bewertung der Trainingsdaten als auch für die permanente Verantwortlichkeit der Überwachung der Entscheidungsergebnisse sowie schließlich für eine nachträgliche Haftung für eindeutig diskriminierende Entscheidungen.

# 4. Ansatzpunkte für diskriminierungsfreie Lernende Systeme

Unsere Rechtsordnung, die auf dem Prinzip der Menschenwürde und den Menschenrechten beruht, bildet ein verlässliches Fundament für die Entwicklung und Anwendung von Lernenden Systemen, die keine Einzelpersonen oder Gruppen diskriminieren. Klarzustellen ist auf gesellschaftlicher Ebene, dass Künstliche Intelligenz nicht per se neutraler oder objektiver entscheidet als der Mensch und somit vorhandene Diskriminierungen beseitigt. Wichtig scheint, ein Bewusstsein für die Diskriminierungsrisiken zu schaffen, die mit dem Einsatz Lernender Systeme einhergehen können. Begrenzen lassen sich diese durch die im Folgenden aufgeführten Ansätze.

Es gibt technische Ansätze, mit denen versucht wird, ethische Prinzipien im Designprozess der Software zu integrieren, etwa Value Sensitive Design (z. B. Friedman et al. 2008, van den Hoven 2013) oder Constructive Technology Assessment (z. B. Schot/Rip 1997, Genus 2006). Der technische Ansatz allein ist jedoch nicht ausreichend, da aufgrund der gesellschaftlichen Einbettung der Anwendungsbereiche der Systeme, aber auch aufgrund des Entwicklungsprozesses, Mechanismen oder Institutionen einer gesellschaftlichen Regulatorik benötigt werden. Dazu gehören Mechanismen, die Menschen ohne vertieftes Fachwissen in die Lage versetzen, Abhängigkeiten von Entscheidungen Lernender Systeme zu vermeiden und die Entscheidungen zu hinterfragen und einschätzen zu können, etwa durch entsprechende Bildungs- und Aufklärungsinitiativen.

## 4.1 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 potentiell 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 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. zum Tragen der Schäden, die durch diese Diskriminierungen entstehen, zu verpflichten.

## 4.2 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.

## 4.3 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 sollte 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.

## 4.4 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.

# Über dieses Whitepaper

**Autorinnen und Autoren**
Prof. Dr. Susanne Beck, Leibniz Universität Hannover
Dr. Armin Grunwald, Karlsruher Institut für Technologie (KIT)
Kai Jacob, SAP
Prof. Dr. Tobias Matzner, Universität Paderborn

Die Autorinnen und Autoren sind Mitglieder der Arbeitsgruppe IT-Sicherheit, Privacy, Recht und Ethik der Plattform Lernende Systeme. Als eine von insgesamt sieben Arbeitsgruppen thematisiert sie Fragen zur Sicherheit (Security), Zuverlässigkeit (Safety) und zum Umgang mit Privatheit (Privacy) bei der Entwicklung und Anwendung von Lernenden Systemen. Sie analysiert zudem damit verbundene rechtliche sowie ethische Anforderungen und steht in engem Austausch mit allen weiteren Arbeitsgruppen der Plattform Lernende Systeme.

**Redaktion**
Stephanie Dachsberger, Geschäftsstelle der Plattform Lernende Systeme
Johannes Melzer, Geschäftsstelle der Plattform Lernende Systeme

## Die Plattform Lernende Systeme

Lernende Systeme im Sinne der Gesellschaft zu gestalten – mit diesem Anspruch wurde die Plattform Lernende Systeme im Jahr 2017 vom Bundesministerium für Bildung und Forschung (BMBF) auf Anregung des Fachforums Autonome Systeme des Hightech-Forums und acatech – Deutsche Akademie der Technikwissenschaften initiiert. Die Plattform bündelt die vorhandene Expertise im Bereich Künstliche Intelligenz und unterstützt den weiteren Weg Deutschlands zu einem international führenden Technologieanbieter. Die rund 200 Mitglieder der Plattform sind in Arbeitsgruppen und einem Lenkungskreis organisiert. Sie zeigen den persönlichen, gesellschaftlichen und wirtschaftlichen Nutzen von Lernenden Systemen auf und benennen Herausforderungen und Gestaltungsoptionen.

# Literatur

**Abiteboul et al. (2016):** Data, Responsibly (Dagstuhl Seminar 16291). http://drops.dagstuhl.de/opus/volltexte/2016/6764/pdf/dagrep_v006_i007_p042_s16291.pdf (abgerufen am 28.03.2019).

**Aggarwal (2018):** Law and Autonomous Systems Series: Algorithmic Credit Scoring and the Regulation of Consumer Credit Markets. https://www.law.ox.ac.uk/business-law-blog/blog/2018/11/law-and-autonomous-systems-series-algorithmic-credit-scoring-and (abgerufen am 28.03.2019).

**AlgorithmWatch (2019):** Atlas der Automatisierung. Automatisierte Entscheidungen und Teilhabe in Deutschland. https://atlas.algorithmwatch.org/wp-content/uploads/2019/04/Atlas_der_Automatisierung_von_AlgorithmWatch.pdf (abgerufen am 28.03.2019).

**Angwin et al. (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 (abgerufen am 28.03.2019).

**Bundesregierung (2018):** Strategie Künstliche Intelligenz der Bundesregierung. www.bmbf.de/files/Nationale_KI-Strategie.pdf (abgerufen am 28.03.2019).

**Dastin (2018):** Amazon scraps secret AI recruiting tool that showed bias against women. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G (abgerufen am 28.03.2019).

**Doshi-Velez/Kim (2017):** Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608.

**European Group on Ethics in Science and New Technologies (2018):** Artificial Intelligence, Robotics and ‚Autonomous' Systems.
European Commission Directorate-General for Research and Innovation.

**Fraunhofer Institut (2018):** Zukunftsmarkt Künstliche Intelligenz – Potenziale und Anwendungen.

**Friedman/Nissenbaum (1996):** Bias in computer systems.
ACM Transactions on Information Systems 14, Nr. 3 (1996), p. 330–347.

**Friedman et al. (2008):** Value Sensitive Design and Information Systems. In: Himma/Tavani (Eds.): The Handbook of Information and Computer Ethics, p. 69–101.

**Fussell (2017):** Why can't this Soap Dispenser identify dark Skin? https://gizmodo.com/why-cant-this-soap-dispenser-identify-dark-skin-1797931773 (abgerufen am 28.03.2019).

**Genus (2006):** Rethinking Constructive Technology Assessment as democratic, reflective, discourse. Technological Forecasting and Social Change 73, Nr. 1 (2006), p. 13–26.

**Hagendorff (2019):** Rassistische Maschinen? Übertragungsprozesse von Wertorientierungen zwischen Gesellschaft und Technik. In: Rath/Krotz/Karmasin, Maschinenethik. Normative Grenzen autonomer Systeme, Springer VS, p. 121–134.

**Harcourt (2010):** Risk as a Proxy for Race.
University of Chicago Public Law & Legal Theory Working Paper No. 323 (2010).
https://papers.ssrn.com/abstract=1677654 (abgerufen am 28.03.2019).

**Friedman et al. (2016):** Inherent Trade-Offs in the Fair Determination of Risk Scores.
arXiv preprint arXiv:1609.05807

**Matzner (2016):** Beyond data as representation: the performativity of Big Data in surveillance. Surveillance & Society 14, Nr. 2 (2016), p. 197–210.

**Mareile (2018**): The co-construction of crime predictions: Dynamics between digital data, software and human beings. In: Fyfe/Gundhus/Rønn, Moral Issues in Intelligence-led Policing. Routledge, p. 143–160.

**Mikians et al. (2012):** Detecting price and search discrimination on the internet.
Proceedings of the 11th ACTM Workshop on Hot Topics in Networks, p. 79–84.

**Misty (2016):** Microsoft creates AI Bot – Internet immediately turns it racist. https://socialhax.com/2016/03/24/microsoft-creates-ai-bot-internet-immediately-turns-racist/ (abgerufen am 28.03.2019).

**Niki et al. (2017):** Avoiding discrimination through causal reasoning.
Advances in Neural Information Processing Systems 30 (2017), p. 656–666.

**Richter/Kind (2016):** Predictive Policing. Büro für Technikfolgenabschätzung, Themenkurzprofil Nr. 9 (2016). https://www.tab-beim-bundestag.de/de/pdf/publikationen/themenprofile/Themenkurzprofil-009.pdf (abgerufen am 28.03.2019).

**Schot/Rip (1997):** The past and the future of constructive technology assessment.
Technological Forecasting and Social Change 54, Nr. 2–3 (1997), p. 251–268.

**Sweeney (2013):** Discrimination in Online Ad Delivery. Communications of the Association of Computing Machinery (CAM) 56, Nr. 5 (2013), p. 44–54.

**Van den Hoven (2013):** Value Sensitive Design and Responsible Innovation. In: Owen/Bessant/Heintz (Eds.): Responsible Innovation. Managing the Responsible Emergence of Science and Innovation in Society, p. 75–83.

GEFÖRDERT VOM

Bundesministerium
für Bildung
und Forschung

acatech
DEUTSCHE AKADEMIE DER
TECHNIKWISSENSCHAFTEN

# Impressum

**Herausgeber**
Lernende Systeme –
Die Plattform für Künstliche Intelligenz
Geschäftsstelle | c/o acatech
Karolinenplatz 4 | 80333 München
www.plattform-lernende-systeme.de

**Gestaltung**
PRpetuum GmbH, München

**Stand**
Juni 2019

**Bildnachweis**
r.classen / Shutterstock | Titelbild

Bei Fragen oder Anmerkungen zu dieser Publikation kontaktieren Sie bitte Johannes Winter (Leiter der Geschäftsstelle):
kontakt@plattform-lernende-systeme.de

Folgen sie uns auf Twitter: @LernendeSysteme

**Empfohlene Zitierweise**
Susanne Beck et al.: Künstliche Intelligenz und Diskriminierung – Whitepaper aus der Plattform Lernende Systeme, München 2019.

Dieses Werk ist urheberrechtlich geschützt.
Die dadurch begründeten Rechte, insbesondere die der Übersetzung, des Nachdrucks, der Entnahme von Abbildungen, der Wiedergabe auf fotomechanischem oder ähnlichem Wege und der Speicherung in Datenverarbeitungsanlagen, bleiben – auch bei nur auszugsweiser Verwendung – vorbehalten.