Wegweiser Digitale Debatten

Teil 2: Algorithmenvermittelte Diskriminierung

Impressum

Innovationsbüro Digitales Leben

Die Publikation wird herausgegeben vom Innovationsbüro des Bundesministeriums für Familie, Senioren, Frauen und Jugend.

Innovationsbüro des Bundesministeriums für Familie, Senioren, Frauen und Jugend
c/o iRights.Lab GmbH
Schützenstraße 8
D-10117 Berlin

Email: kontakt@innovationsbuero.net

www.innovationsbuero.net
www.irights-lab.de
www.bmfsfj.de

Verantwortlich: Philipp Otto, Leiter des Innovationsbüros

Autor:innen: Elena Kalogeropoulos, Anne Lammers, Jaana Müller-Brehm, Michael Puntschuh

Lektorat: Julia Schrader, Annika Albert

Gestaltung und Satz: Christoph Löffler

Druck:

4.0

iRights.Lab
Think Tank für die digitale Welt

Inhalt

Inhalt

1. Einleitung

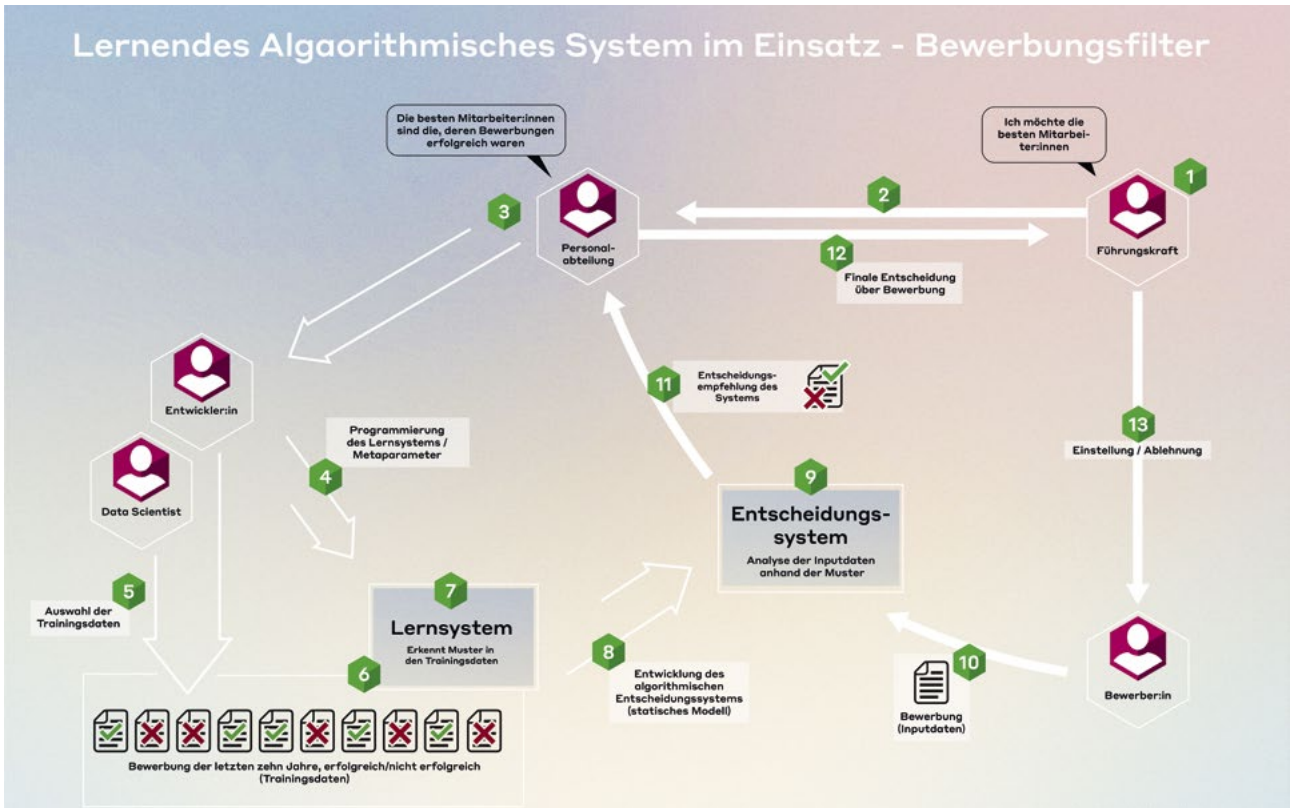
In vielen Bereichen unseres alltäglichen Lebens spielen Algorithmen und Künstliche Intelligenz (KI) mittlerweile eine bedeutende Rolle. Spracherkennungssoftware übersetzt gesprochene Worte in Befehle, Bewerbungsfilter sortieren unpassende Kandidat:innen für eine Stelle aus und Gesundheits-Apps werden zu Verhütungsmitteln, indem sie fruchtbare und unfruchtbare Tage prognostizieren. In allen diesen Beispielen kommen Algorithmen zum Einsatz, deren Anwendungen jedoch bedeutende ethische Fragen aufwerfen. Denn wenn der Smart Speaker bei der Spracherkennung nicht genau genug arbeitet, die Bewerbungsfiltersoftware Kandidat:innen nicht aufgrund ihrer Qualifikation, sondern ihres Geschlechts aussortiert und der Algorithmus für die Verhütungs-App auf fehlerhaften Daten basiert, geht das mit teils gravierenden sozialen Konsequenzen einzelner oder ganzer gesellschaftlicher Gruppen einher. Die Beschäftigung mit algorithmischen Systemen, ihren Fallstricken und Besonderheiten, hilft nicht nur verstehen, in welchen Bereichen diese zum Einsatz kommen (können), sondern sensibilisiert darüber hinaus für Folgewirkungen, die bei ihrer Anwendung auftreten können.

Mit diesem Problemaufriss ist einer der Themenschwerpunkte angesprochen, mit denen sich das Innovationsbüro Digitales Leben des Bundesministeriums für Familien, Senioren, Frauen und Jugend (BMFSFJ) beschäftigt. Das Innovationsbüro ist ein Think Tank zur positiven Gestaltung der Digitalisierung innerhalb des Ministeriums. Die hier entwickelten Maßnahmen und Strategien sind von unmittelbarer Relevanz für die Zielgruppen des BMFSFJ, denn soziale Beziehungen werden heute zunehmend digital vermittelt. Das Innovationsbüro sieht es als eine seiner Kernaufgaben, als Schnittstelle zwischen Zivilgesellschaft und öffentlicher Verwaltung diese Prozesse aktiv und zum Wohle aller mitzugestalten.

Der Wegweiser Digitale Debatten ist Teil seiner Tätigkeiten. Hier stehen zentrale digitalethische Themen im Mittelpunkt, mit denen sich das Innovationsbüro des BMFSFJ intensiv auseinandersetzt und die uns alle berühren. Denn wenn algorithmische Systeme in immer mehr alltäglichen Lebensbereichen zur Anwendung kommen, steht ein Gesellschaftsministerium ganz besonders in der Verantwortung, hier im Interesse seiner Zielgruppen zu wirken. In einem ersten Schritt ist es von zentraler Bedeutung, über Funktionsweisen, Begrifflichkeiten und die sozialen Konsequenzen bei der Anwendung derartiger Systeme zu sprechen.

Genau an diesem Punkt setzt diese Veröffentlichungsreihe an. Sie dient der Wissensvermittlung, der Veranschaulichung des Problemkomplexes sowie der kritischen Reflexion des Einsatzes von algorithmischen Systemen und KI. Damit richtet sie sich sowohl an Entscheidungsträger:innen in der öffentlichen Verwaltung als auch an Entwickler:innen, interessierte Einzelpersonen und gesellschaftlichen Interessensvertretungen. Denn die Möglichkeiten und Dilemmata solcher Technologien, die zunehmend unseren Alltag bestimmen, gehen uns alle an. Das Innovationsbüro des BMFSFJ legt mit dieser Veröffentlichung einen wesentlichen Grundstein für einen informierten und kritischen Diskurs zwischen Ministerien und Zivilgesellschaft im Bereich der digitalen Ethik.

Dieser Teil 2 der Reihe beschäftigt sich mit dem Thema der algorithmenvermittelten Diskriminierung. Um sich diesem Gegenstand zu nähern, stehen im Folgenden zunächst ganz allgemein mögliche Fehlerquellen in algorithmischen Systemen sowie eine Diskussion des (rechtlichen) Diskriminierungsbegriffs im Fokus. Anschließend werden konkrete Formen, Besonderheiten, Ursachen und Folgen algorithmenvermittelter Diskriminierungen dargelegt.



2. Lernendes algorithmisches System im Einsatz?

Bevor eine tiefergehende Beschäftigung mit algorithmenvermittelter Diskriminierung möglich ist, ist eine Vergegenwärtigung der Funktionsweise algorithmischer Systeme hilfreich. Die Grafik fasst die Erkenntnisse der Handreichung Algorithmische Systeme in vereinfachter Form zusammen. Sie stellt die Gestaltung und den Einsatz eines algorithmischen System dar, das eine lernende Komponente besitzt und dazu eingesetzt wird, in einem Unternehmen Bewerbungen zu filtern.

(1) Bevor ein solches System überhaupt eingesetzt wird, muss es gestaltet werden. Dies beginnt in der Regel damit, dass eine Führungskraft die Entwicklung eines algorithmischen Systems beschließt und die Zielvorgaben festlegt.

(2) Diese werden an die das System später einsetzende Stelle kommuniziert.

(3) Diese Stelle wird üblicherweise die Zielvorgaben spezifizieren: Aus der Vorgabe, die besten Mitarbeiter:innen für eine bestimmte vakante Stelle auszuwählen, definiert die Personalabteilung, was „beste:r Mitarbeiter:in“ bedeutet und nach welchen Kriterien diese gemessen werden können.

(4) Entwickler:innen des algorithmischen Systems werden diese Vorgaben dann übersetzen und daraus ableiten, wie sie ihr entsprechendes algorithmisches System gestalten wollen, welche Daten einfließen sollen und welche Parameter eine Rolle spielen sollen. In diesem Fall wird die Nutzung eines lernenden algorithmischen Systems beschlossen. Deshalb müssen noch nicht die genauen Parameter festgelegt werden, aber der:die Entwickler:innen entscheiden, welche Methode des maschinellen Lernens sie unter der Nutzung welcher Datengrundlage bzw. welches Feedbackmechanismus nutzen wollen.

(5) Gemeinsam mit Data Scientists wird dann die Datengrundlage ausgewählt und vorbereitet.

(6) In diesem Beispiel sind dies die Bewerbungen der letzten 10 Jahre. Zu diesen Bewerbungsunterlagen ist jeweils bekannt, ob die Bewerbung Erfolg hatte, was ebenfalls in das Lernsystem eingebracht wird. Das entwickelte Lernsystem wird nun eingesetzt, um die Trainingsdaten zu analysieren.

(7) Aus den Unterschieden zwischen erfolgreichen und nicht erfolgreichen Bewerbungen werden Muster abgeleitet. Dabei handelt es sich in der Regel um Eigenschaften der Bewerbungsunterlagen (z. B. verzeichnete Schulnote, Berufserfahrung), die zur Vorhersage des Bewerbungserfolgs innerhalb des Trainingsdatensatzes herangezogen werden.

(8) Diese Muster werden in ein statistisches Modell übersetzt.

(9) Das statistische Modell dient dann als Grundlage für das Entscheidungssystem, das eigesetzt werden wird, um die eigentliche Aufgabe zu erfüllen.

(10) Das algorithmische System ist nun gestaltet und im Einsatz. Die von Bewerber:innen eingesendeten Bewerbungen werden in das System eingespeist und basierend auf den erkannten Mustern hin analysiert.

(11) Das System gibt eine Entscheidungsempfehlung aus, ob die analysierte Bewerbung eher den Kriterien erfolgreicher oder nicht erfolgreicher Bewerbungen entspricht und damit, ob diese Person eingestellt werden soll oder nicht. Diese Entscheidungsempfehlung wird in den meisten Fällen nicht direkt umgesetzt, sondern erst an die Anwender:in übermittelt, hier die Personalabteilung.

(12) Die Personalabteilung betrachtet die Analyse und trifft dann auf deren Grundlage (und ggf. weiterer Daten oder Überlegungen) die eigentliche Entscheidung, die ggf. an eine verantwortliche Person weitergeleitet wird.

(13) Diese Empfehlung wird umgesetzt: Die Person wird nun eingestellt oder ihre Bewerbung abgelehnt.

3. Fehlerquellen algorithmischer Systeme

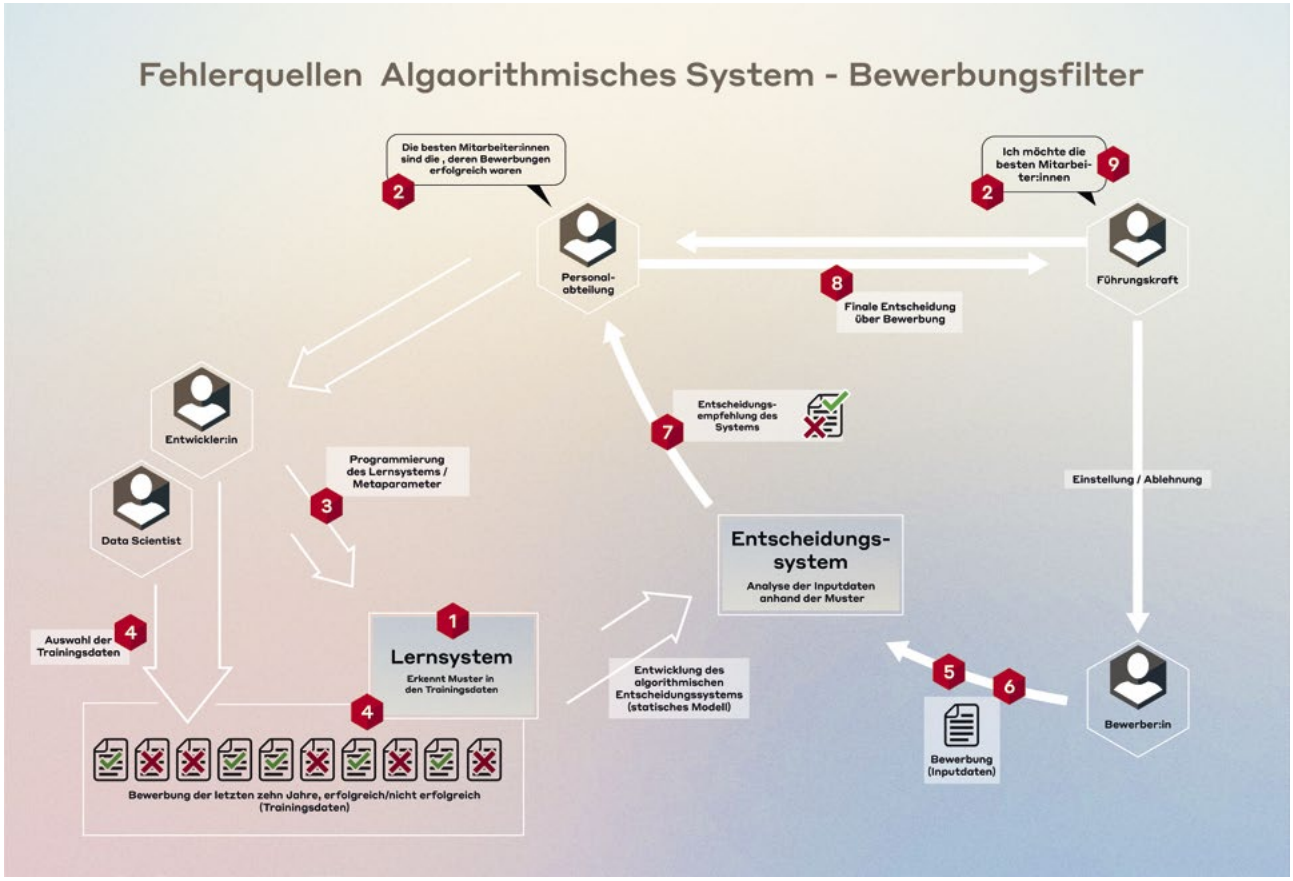
Algorithmenvermittelte Diskriminierung ist grundsätzlich eine mögliche Auswirkung eines Fehlers in diesem Gestaltungs- und Einsatzprozess des algorithmischen Systems. Insofern zeigt ein Überblick über Fehlerquellen eines algorithmischen Systems die möglichen Ursachen für algorithmenvermittelte Diskriminierung auf.

Es lassen sich folgende Fehlerquellen unterscheiden, die auch unten in der Grafik markiert sind. Die Fehlerquellen wirken sich in konkreten Anwendungen oft aufeinander aus und können zu einer Kette von Fehlern verbunden sein.

(1) Handwerkliche Fehler
Wie in jedem von Menschen gestalteten Prozess kann es zu (einfachen) handwerklichen Fehlern kommen. Solche Fehler wie Tippfehler, falsche Entscheidungen oder sonstige Versäumnisse können beispielsweise bei der Modellierung der Systeme vorkommen, indem (aus Versehen) ein unpassendes Modell ausgewählt wird oder der Code des Lernsystems fehlerbehaftet ist. Auch an anderen Stellen im Prozess kann es zu solchen Fehlern kommen, indem z. B. Absprachen vergessen werden oder Kommunikation missverständlich stattfindet.

(2) Zielvorgaben
Schon ganz zu Beginn des Gestaltungsprozesses können sich Fehler ergeben. Zielvorgaben bestimmen maßgeblich das weitere Design, weswegen Fehler an dieser Stelle große Auswirkungen haben können. Ziele können nicht ausreichend genau formuliert sein, sodass andere Personen im Prozess dann Fehler machen. Was z. B. eine:n „gute:n Mitarbeiter:in“ ausmacht, ist keineswegs eindeutig. Sind die besten Mitarbeiter:innen diejenigen, deren Bewerbungen erfolgreich waren? Oder diejenigen, die im Unternehmen besonders schnell aufgestiegen sind? Oder vielleicht diejenigen, die in der Akquise besonders stark sind? Das kann dann wiederum zu sehr unterschiedlichen und teilweise unpassenden Präzisierungen führen, was wiederum zu komplett unterschiedlichen algorithmischen Systemen führt. Ziele können aber auch schon auf der obersten Ebene problematisch formuliert sein: So kann eine Führungskraft zwar kommunizieren, die „besten“ Mitarbeiter:innen einstellen zu wollen – tatsächlich will sie aber die „effizientesten“ Mitarbeiter:innen haben. Ob dieses Ziel dann noch dem eigentlichen Zweck entspricht ist nicht selbstverständlich, es kann zu Abweichungen kommen. Problematisch sind dabei nicht nur die Fehler selbst, sondern auch das mangelnde Bewusstsein bei Führungskräften und anderen Entscheider:innen darüber, welche Auswirkungen ihre Ziel- und Aufgabendefinition hat.

(3) Übersetzungsfehler von Zielvorgaben in Parameter, Kriterien und Systemarchitektur
Insbesondere, wenn Ziele nicht hinreichend präzise formuliert sind, kann es im weiteren Verlauf zu Problemen kommen. So weiß z. B. eine gute Personalabteilung, was die



Zielvorgabe „erfolgreiche Mitarbeiter:innen“ bedeutet – aber Entwickler:innen, die Personalauswahlprozesse oder die inneren Strukturen des Unternehmens nicht kennen, können solche Zielvorgaben missverstehen und ihr Lernsystem entsprechend auf ein nicht passendes Fundament stellen. Auch kleine Entscheidungen können hier große Auswirkungen haben. Bedeutet die Zielvorgabe z. B., dass erfolgreiche Bewerbungen aus den letzten 10 Jahren analysiert werden oder alle Bewerbungen aus den letzten 10 Jahren? Welche Daten sollen überhaupt in das Lernsystem einfließen?

(4) Inputverzerrungen
Eng verknüpft damit ist die Gefahr von Inputverzerrungen, die das algorithmische System aus den Inputdaten übernimmt. Es gibt drei Typen von Inputverzerrungen:

a. Historischer Bias: Die Inputdaten sind historische Daten aus dem realen Einsatz, wie z. B. die Bewerbungen der letzten 10 Jahre. Kam es aber in diesem Zeitraum zu diskriminierenden Praktiken, finden sich diese in den Daten wieder und werden dann auch vom Lernsystem als Muster übernommen. So weisen Bewerbungsdaten aus dem IT-Sektor potentiell ein sexistisches Muster auf, weil laut dem Women in Digital Scoreboard der EU Männer im IT-Sektor überrepräsentiert sind.

b. Unvollständige Daten: Bei der Datenauswahl ist darauf zu achten, möglichst vollständige Daten zu wählen, die auch die Vielfalt der späteren Anwender:innen bzw. Betroffenen wiederspiegelt – ansonsten kann das System bestimmte Gruppen durchweg benachteiligen. Wird beispielsweise eine Software nur mit den Daten von hellhäutigen Menschen trainiert, sind die Ergebnisse für Menschen mit dunkleren Hauttönen weniger präzise oder gar grundsätzlich falsch und diskriminierend.

c. Fehlerhafte Daten: Die Datenqualität ist bei Trainingsdatensätzen nicht automatisch garantiert. Hier können sich Fehler einschleichen, oft über handwerkliche Fehler während des Datenerfassungsprozesses oder bei der Speicherung und Verarbeitung der Daten.

(5) Design und Interaktion Input / Probleme der Quantifizierung
Die Art und Weise, wie Inputdaten abgefragt und eingegeben werden, kann problematisch sein. Ist beispielsweise nur eine Eingabe auf Deutsch möglich, werden Bewerbungen von nicht deutschsprachigen Unterlagen nicht erfasst. Werden bei der Eingabe auch irrelevante Daten wie das Geschlecht oder der Name erfasst und später verarbeitet, kann auch das zu Verzerrungen führen. Schließlich kann die Eingabemaske für Inputdaten problematisch sein, indem sie z. B. missverständlich oder schwer zu bedienen ist. Eine ähnliche Problematik ist die Quantifizierung eigentlich qualitativer Daten, die oft bei der Eingabe von Input erforderlich ist. Wird in einer Eingabemaske beispielsweise abgefragt, ob man „gute“, „mittlere“ oder „schlechte“ Englischkenntnisse hat, können die Auswahlmöglichkeiten von unterschiedlichen Bewerber:innen unterschiedlich interpretiert werden. Eigentlich gleichwertige Qualifikationen würden dann als unterschiedlich erfasst.

(6) Sprachliche Grenzen
Die Grenzen der Sprache können eine zusätzliche Hürde beim Input darstellen. Im Deutschen ist dies beispielsweise die Tatsache, dass Berufsbezeichnungen in der Regel geschlechterspezifisch formuliert werden: Eine Bewerbung als „Ingenieurin“ wird vielleicht gar nicht erfasst oder als Tippfehler der ausgeschriebenen Stelle „Ingenieur“ gewertet. Auch ist fraglich, ob und wie ein algorithmisches System mit gendergerechter Sprache wie „Ingenieur:in“ umgehen würde.

(7) Design und Interaktion Output
Genauso wie beim Input spielt auch das Design der Ausgabe der Daten beim Output eine Rolle und kann Quelle von Fehlern werden. Wie werden die Outputdaten dargestellt und angezeigt? Welche zusätzlichen Informationen werden der Personalabteilung zur Verfügung gestellt, um das Ergebnis im Zweifel verstehen und hinterfragen zu können? Werden die Top 1% oder die besten 20 Bewerbungen angezeigt? All das kann zu unterschiedlichen und möglicherweise nicht gewünschten Auswirkungen führen.

(8) Output-Interpretation
Eine weitere Fehlerquelle ist die Interpretation des Outputs durch die Anwender:innen und die Umsetzung der Outputdaten in die tatsächliche Entscheidung. Wird eine Entscheidungsempfehlung unhinterfragt übernommen, können sich Fehler direkt auf die Betroffenen auswirken. Möglicherweise wird auch der Output fehlinterpretiert, indem beispielsweise eine Bewerbung zwar als erfolgreich dargestellt wird, aber die Personalabteilung diese Person dann für eine andere Stelle einstellt. Wie bei jeder Form der Kommunikation und Interaktion kann es auch hier zu Fehlinterpretationen kommen. Außerdem kann der gleiche Output für unterschiedliche Entscheidungen zu Rate gezogen werden. So kann eine Einstellungsempfehlung, die auch nach weiblichen und männlichen Kandidat:innen unterscheidet, dazu genutzt werden, bewusst Frauen zu bevorzugen – oder eben sie zu benachteiligen.

(9) Dilemmata
Auch wenn es sich hierbei um keine Fehlerquelle handelt, so sind Dilemmata doch auch Quelle diverser Herausforderungen und möglicher Diskriminierungen. So ist es oft nicht möglich, alle Zielvorgaben gleichzeitig oder in gleichem Maße zu erfüllen. Im Falle von Bewerbungen kann einerseits darauf hingewirkt werden, dass man möglichst wenige eigentlich unpassende Bewerber:innen doch einstellt (Ziel: Verringerung der „false positives“). Andererseits könnte aber auch der Anspruch bestehen, möglichst viele der qualifizierten Bewerbungen anzunehmen und keine potentiell erfolgreichen Mitarbeiter:innen aus Versehen abzulehnen (Ziel: Verringerung der „false negatives“). Es muss eine Abwägung zwischen sogenannten „false positives“ und „false negatives“ getroffen werden, die mit jeweils unterschiedlichen Konsequenzen verbunden ist.

4. Diskriminierung

Was ist Diskriminierung?

In Deutschland regelt **Artikel 3 des Grundgesetzes** das Diskriminierungsverbot: „Niemand darf wegen seines Geschlechtes, seiner Abstammung, seiner Rasse, seiner Sprache, seiner Heimat und Herkunft, seines Glaubens, seiner religiösen oder politischen Anschauungen benachteiligt oder bevorzugt werden. Niemand darf wegen seiner Behinderung benachteiligt werden."

Das **Allgemeine Gleichbehandlungsgesetz (AGG)** konkretisiert den Gleichheitsgrundsatz. Laut Definition des AGG liegt eine mittelbare (indirekte) Diskriminierung vor, „wenn dem Anschein nach neutrale Vorschriften, Kriterien oder Verfahren Personen wegen eines in § 1 genannten Grundes gegenüber anderen Personen in besonderer Weise benachteiligen können" und dies weder sachlich gerechtfertigt noch verhältnismäßig ist (§ 3 (2) AGG). Bei den Gründen, auf die das AGG Bezug nimmt, handelt es sich um Rasse oder ethnische Herkunft, Geschlecht, Religion oder Weltanschauung, Behinderung, Alter und sexuelle Identität. Eine unmittelbare Diskriminierung wird hingegen eine direkte, offene Benachteiligung aufgrund eines dieser Merkmale verstanden (§ 3 (1) AGG).

Das AGG spricht allerdings nicht von Diskriminierung, sondern von Benachteiligung, da nicht jede unterschiedliche Behandlung, die einen Nachteil zur Folge hat, diskriminierend sein muss. In sehr engen Grenzen sind unterschiedliche Behandlungen in Bezug auf das Berufsleben zulässig, wenn die geforderte Eigenschaft für die Ausübung der Tätigkeit wesentlich und fast unerlässlich ist.

Eine **unmittelbare Benachteiligung** liegt vor, wenn eine Person eine weniger günstige Behandlung als eine Vergleichsperson erfährt, erfahren hat oder erfahren würde. Das ist zum Beispiel der Fall, wenn eine Person mit Migrationshintergrund trotz gleicher Qualifikationen nicht zu einem Bewerbungsgespräch eingeladen wird, Personen ohne Migrationshintergrund hingegen schon.

Der Diskriminierungsschutz des AGG erstreckt sich auch auf **mittelbare Benachteiligungen**. Dabei handelt es sich dem Anschein nach um neutrale Regelungen, die Personen wegen eines AGG-Merkmals schlechterstellen.

Diskriminierung basiert demnach auf einer tatsächlichen oder vermuteten Gruppenzugehörigkeit von Menschen und der damit verbundenen Zuschreibung von Eigenschaften, die eine negative Bewertung beinhalten (auch scheinbar positive). Die „dominante" Gruppe hat dabei die Deutungsmacht und definiert, welche Gruppe „anders" als die Norm ist, d. h. nicht zur Gruppe/Gesellschaft „dazugehört" (nicht zwingend die Mehrheit ist).

Auf Europäischer Ebene sichert die **Europäische Menschenrechtskonvention (EMRK)** in ihrem **Artikel 14**[1] die diskriminierungsfreie Ausübung der in der Konvention garantierten Menschenrechte und Grundfreiheiten zu. Dabei verbietet die Menschenrechtskonvention in Ansehung der Menschenrechte jedwede Diskriminierung, gleich aus welchem Grund.

Allerdings verbietet Artikel 14 EMRK nur eine Diskriminierung im Hinblick auf die gewährleisteten Menschenrechte und Grundfreiheiten. Mit dem im Jahr 2000 in Rom verabschiedeten **12. Protokoll zur Europäischen Menschenrechtskonvention** ändert sich das, indem ein allgemeines Diskriminierungsverbot verabschiedet wird. Dieses Zusatzprotokoll ist am 1. April 2005 in Kraft getreten. Deutschland, Liechtenstein und Österreich haben es unterzeichnet, aber bisher noch nicht ratifiziert. (Stand: 28. November 2019).

Wie äußert sich Diskriminierung?

Diskriminierung kann sich durch verschiedene Charakteristika auszeichnen. Dazu können u. a. gehören:

- bewusste Ignorierung
- Willkür
- physische, psychische, verbale Gewalt
- eingeschränkter Zugang zu Rechten, Bildung, Arbeitsmarkt, Wohnungsmarkt
- begrenzte (politische) Repräsentation und Partizipation
- keine, geringe sowie abwertende Darstellung z. B. in Literatur oder Medien

1 Europäische Menschenrechtskonvention, Abschnitt I - Rechte und Freiheiten (Art. 2-18), Art. 14, Diskriminierungsverbot: *„Der Genuss der in dieser Konvention anerkannten Rechte und Freiheiten ist ohne Diskriminierung insbesondere wegen des Geschlechts, der Rasse, der Hautfarbe, der Sprache, der Religion, der politischen oder sonstigen Anschauung, der nationalen oder sozialen Herkunft, der Zugehörigkeit zu einer nationalen Minderheit, des Vermögens, der Geburt oder eines sonstigen Status zu gewährleisten."*

5. Algorithmenvermittelte Diskriminierung

a. Definition

Algorithmische Systeme haben positive wie negative soziale, wirtschaftliche, administrative und rechtliche Konsequenzen. Diese Konsequenzen können mit Diskriminierungen einhergehen. Algorithmenvermittelte Diskriminierung ist demnach eine bestimmte Form der Perpetuierung von Diskriminierung. Dabei hängen die Funktionsweise bzw. Wirkungen der Diskriminierung eng mit dem Einsatz algorithmischer Systeme zusammen. Unter algorithmenvermittelter Diskriminierung werden alle Formen der direkten und indirekten Diskriminierung gefasst, bei der algorithmische Systeme eingesetzt werden, um die Ungleichbehandlung von Menschen festzustellen, zu beeinflussen oder zu informieren.

Algorithmenvermittelte Diskriminierung ist damit eine Auswirkung, die einen Fehler im algorithmischen System zur Ursache hat. Die verschiedenen möglichen Fehlerquellen sind oben beschrieben.

Neben dem Begriff „algorithmenvermittelte Diskriminierung“ gibt es viele weitere Begriffe, die deckungsgleich sind oder sich in Nuancen unterscheiden. Dazu zählen die Begriffe „algorithmenbasierte Diskriminierung“, „Diskriminierung durch Algorithmen“, „automatisierte Diskriminierung“, „Machine Bias“ und „AI Bias“. Trotz der kleinen Unterschiede können die Begriffe in den meisten Fällen synonym verwendet werden.

b. Besonderheiten algorithmenvermittelter Diskriminierung

Algorithmenvermittelte Diskriminierung trennscharf zu erfassen und zu beschreiben ist aufgrund ihrer mannigfaltigen Ausprägungen schwierig. Um sich aber dem Konzept und insbesondere den sich daraus ergebenden Herausforderungen anzunähern, kann es helfen, sich mit den Besonderheiten algorithmenvermittelter Diskriminierung zu beschäftigen. Diese Tabelle liefert einen Überblick über die wichtigsten Eigenschaften algorithmenvermittelter Diskriminierung, insbesondere im Kontrast zu menschenvermittelter Diskriminierung. Die Besonderheiten werden im Folgenden näher beschrieben.

Versteckt	Algorithmische Systeme werden (oft) in nicht öffentlich zugänglichen Prozessen eingesetzt.
Opak	Algorithmische Systeme sind nicht immer automatisch für Menschen einsehbar und verständlich.
Indirekt	Algorithmische Systeme ermöglichen Diskriminierung auf Basis von Ersatzidentitäten (Proxy-Variablen).
Systemisch	Algorithmenvermittelte Diskriminierung wird teils erst im Gruppenbezug erkennbar.
Intersektional	Algorithmenvermittelte Diskriminierung kann (nur) gegen eine Schnittmenge verschiedener Gruppen gerichtet sein.
Scheinbar objektiv	Entscheidungen algorithmischer Systeme „wirken“ objektiv.
Expliziert	Algorithmenvermittelte Diskriminierung spiegelt bestehende, teils versteckte Diskriminierungsstrukturen wider. Letztere werden damit technisch expliziert und für eine öffentliche Debatte zugänglich.

c. Einzelerklärung der Besonderheiten algorithmenvermittelter Diskriminierung

Versteckt: Algorithmenvermittelte Diskriminierung ist häufig versteckt, weil sie oft in nicht-öffentlichen Räumen stattfindet. Plattformen im Internet sind meistens in privater Hand und entsprechend sind die gesetzlichen Anforderungen an Transparenz und Kontrolle niedriger, als es im öffentlichen Raum wäre. Diese Unternehmen haben auch Anreize, mögliche Diskriminierungen versteckt zu halten, solange sie keine Lösung erarbeitet haben. Deshalb ist schon auf dieser Ebene algorithmenvermittelte Diskriminierung schwerer beobachtund erfassbar.

Opak: Wenn ein Mensch eine Entscheidung über einen anderen Menschen trifft, kann man die Frage nach dem „Warum?“ zumindest stellen. Bei algorithmischen Systemen sind aber die Funktionsweise und die Ursachen für einen bestimmten Output nicht automatisch nachvollziehbar. Selbst für Fachleute gibt es technische und architektonische Hindernisse, algorithmenbasierte Entscheidungsprozesse im Detail zu verstehen. Diese Hürde ist zusätzlich größer für Menschen, die

bereits technische Grundlagen nicht beherrschen. Gerade für Betroffene stellt das ein Problem dar.

Indirekt: Eine algorithmenbasierte Diskriminierung kann auf einem Merkmal basieren, das nicht zu den geschützten Merkmalen gehört, wie z. B. Lieblingsfarbe oder sprachlicher Stil einer Bewerbung. Diese Merkmale korrelieren aber stark mit anderen Merkmalen, die geschützt sind, wie die Angabe zum Geschlecht. Lieblingsfarbe kann dann eine Ersatzidentität (auch „Proxy-Variable") für das Geschlecht sein. Diskriminierung wirkt dann indirekt, was für Wahrnehmbarkeit, Erfassung und Nachweis der Diskriminierung Schwierigkeiten erzeugt.

Systemisch: Algorithmische Systeme werden oft eingesetzt, um große Datenmengen zu verarbeiten. Ihre Auswirkungen betreffen damit oft viele Menschen auf einmal. Dabei kommt es nicht zur Diskriminierung von Einzelpersonen, weil ein algorithmisches System keine „individuelle Beziehung" zu einem Individuum aufnehmen kann. Stattdessen kommt es zu Diskriminierungen aufgrund eines Merkmals, das viele Menschen teilen. Verknüpft mit der Indirektheit kann es zu Diskriminierungen von großen Menschengruppen auf Basis von Eigenschaften kommen, die bisher nicht explizit vor diesem Hintergrund betrachtet wurden, wie z. B. Wohnort, Vornamen oder bestimmtes Sportverhalten. Diese systemische Wirkung erschwert es zudem für Einzelne, die Diskriminierung auch wahrzunehmen.

Intersektional: Da algorithmische Systeme oft verschiedene Kategorien von Merkmalen gleichzeitig erfassen, kann Diskriminierung stärker intersektional wirken: Eine Person wird dann nicht mehr aufgrund ihrer Zuordnung zu einer Gruppe, sondern zu mehreren Gruppen gleichzeitig diskriminiert. Es kommt dann zu einer Diskriminierung z. B. gegen dunkelhäutige Frauen, also einen Schnittbereich, anstatt gegen alle Frauen oder alle Menschen mit dunkler Hautfarbe. Auch diese Eigenschaft erschwert die Wahrnehmung der Diskriminierung durch das Individuum.

Scheinbar objektiv: Mit dem Einsatz algorithmischer Systeme wird oft ein scheinbar „objektiver" Entscheidungsprozess auf Grundlage von Daten verbunden. Man vertraut den Entscheidungen (bzw. Entscheidungsempfehlungen) solcher Systeme eher und Diskriminierungen können so quantitativ begründet werden. Dieses als „Mathwashing" bezeichnete Phänomen führt dazu, dass Anwender:innen algorithmische Systeme weniger in Frage stellen. Soziale Ungleichheiten werden so unter dem Schleier eines „objektiven" Prozesses reproduziert.

Expliziert: Die algorithmenvermittelte Diskriminierung hat eine letzte Eigenschaft, die auch einen großen Vorteil mit sich bringt: Sie ist expliziert. Während menschenbasierte Diskriminierung ihren Ursprung in den Köpfen von Menschen hat, in die man nicht hineinsehen kann, werden algorithmische Systeme bewusst entwickelt, gestaltet und eingesetzt und ihre Funktionsregeln, Datengrundlage und Arbeitsweise sind festgeschrieben. Soll beispielsweise ein einfaches, nicht-lernendes algorithmisches System eingesetzt werden, um bewusst gegen Frauen zu diskriminieren, müsste dies auch so explizit in den Code geschrieben werden. Wird ein lernendes System eingesetzt, das auf Basis historischer Daten trainiert wird, so können die bestehenden diskriminierenden Praktiken aufgedeckt werden, wenn das System diese später auch widerspiegelt. Beide Beispiele zeigen, wie algorithmische Systeme auch dazu beitragen können, Diskriminierungen aufzudecken oder transparent zu machen.

d. Beispiele algorithmenvermittelter Diskriminierung

Um besser zu illustrieren, wie sich algorithmenvermittelte Diskriminierung auswirkt und welche Ursachen und Formen diese haben kann, werden im Folgenden vier Fälle von algorithmenvermittelter Diskriminierung beschrieben.

i. Amazons Bewerbungsfilter – offene Türen, versperrte Wege? (Fehlerquelle: Übersetzungsfehler, Inputverzerrung – historischer Bias, Design und Interaktion Input)

Im Jahr 2014 begann Amazon, eine Bewerbungsfiltersoftware einzusetzen, um passende Kandidat:innen für neue Einstellungsrunden bestimmen zu können. Schon kurze Zeit später jedoch stand fest, dass die Software Personen aufgrund ihres Geschlechts diskriminierte – sie sortierte Bewerbungen, in denen bestimmte Stichworte wie „women", zum Beispiel in „women's chess club captain", vorkamen, systematisch aus. Auch Bewerber:innen, die rein weibliche Colleges besucht hatten, wurden aussortiert. Was war geschehen? Bei der Software handelte es sich um ein lernendes algorithmisches System, das dazu programmiert wurde, bisherige erfolgreiche Bewertungen auf dort wiederkehrende Begriffe zu analysieren. Allerdings „lernte" das System dabei nicht, wie erwartet, auf Wörter zu achten, die auf eine gute Qualifikation oder Ausbildung hinwiesen. Stattdessen stellte die Software beim Betrachten der vorherigen erfolgreichen Bewerbungen fest, dass bevorzugt männliche Kandidaten eingestellt wurden. Die Software übernahm diese Feststellung als wesentliches Kriterium für eine erfolgreiche Anstellung. Ohne, dass das in der Entwicklung der Software explizit vorgegeben wurde, führte das algorithmische System den historischen Bias in der Einstellungspraxis auf Grundlage der Kategorie „Geschlecht" fort. Ursache war, dass das Geschlecht als Kriterium auch nicht explizit ausgeschlossen wurde und es sich zumindest indirekt in den Trainingsdaten wiederfindet.

Aufgrund dieses und weiterer Probleme mit der Software musste Amazon ihre Verwendung letztendlich einstellen und nutzt seither nur noch eine stark abgespeckte Version, die zum Beispiel in der Lage ist, doppelte Bewerber:inneneintragungen in einer Datenbank zu entdecken. Ungeachtet dieser Erfahrung gibt es jedoch eine Reihe an weiteren Groß- sowie Kleinunternehmen, die auf Bewerbungsfiltersoftware setzen. Für die Kandidat:innen ist unterdessen häufig nicht ersicht-

lich, dass ihre Bewerbung von einem algorithmischen System aussortiert wurde, geschweige denn, auf Grundlage welcher Kriterien dies geschehen ist.

ii. Medizinische Versorgung abhängig von der Hautfarbe? Wie ein algorithmisches System nicht-weiße Patient:innen benachteiligt (Fehlerquellen: Zielvorgaben, Übersetzungsfehler, Inputverzerrung – historischer Bias)

Im Oktober 2019 veröffentlichte das Wissenschaftsmagazin Science einen Artikel, in dem von der systematischen Benachteiligung von Afroamerikaner:innen durch das US-amerikanische Gesundheitssystem berichtet wurde. Konkret heißt es dort, dass eine häufig verwendete Software, die unter anderem von Krankenhäusern und Versicherungen eingesetzt wird, auf einem algorithmischen System basiert, das weißen Patient:innen sehr viel häufiger als Afroamerikaner:innen teure medizinische Behandlungen genehmigt. Lediglich 17,7 Prozent der schwarzen Patient:innen bekamen einen Zuspruch für teure medizinische Leistungen. Dabei müsste, nach Berechnung der Autor:innen, dieser Anteil bei 46,5 Prozent und damit ähnlich hoch wie bei weißen Patient:innen liegen. Die Software wird eingesetzt, um Patient:innen zu identifizieren, die von derartigen Behandlungen am ehesten profitieren. Die Forscher:innen hatten festgestellt, dass dem algorithmischen System dabei eine Input-Verzerrung zugrundeliegt, weil die Ausgangsdaten unvollständig waren. Denn um die Anspruchsberechtigung für medizinische Behandlungen prüfen zu können, nutzten die Softwareentwickler:innen eine Proxy-Variable als Grundlage für die Berechnung. Dabei handelt es sich um die Höhe der medizinischen Ausgaben einer Person im Laufe eines Jahres. Der Hintergedanke bestand darin, dass Menschen mit höheren Behandlungskosten mehr medizinische Hilfe benötigen als solche mit geringeren. In den USA jedoch gelten Schwarze hinsichtlich der Gesundheitsleistungen als unterversorgt, da sie wesentlich seltener medizinische Hilfe in Anspruch nehmen. Die Ursachen dafür liegen hauptsächlich im Rassismus begründet: Schwarze in den USA verdienen häufig weniger und haben gleichzeitig längere Arbeitszeiten, was Arztbesuchen im Weg stehen kann. Darüber hinaus spielen der Wohnort, die verfügbaren Transportmittel sowie ein geringes Vertrauen in das Gesundheitssystem eine Rolle, weil Afroamerikaner:innen zum Beispiel direkte Diskriminierung durch Ärzt:innen erfahren haben. Die Anwendung dieser Software führt damit sehr eindrücklich die sozialen Konsequenzen vor Augen, die sich ergeben können, wenn algorithmische Systeme auf Grundlage fehlerhafter oder unvollständiger Daten operieren.

iii. Gut gemeint, schlecht umgesetzt? Wenn mit algorithmischen Systemen bestehende strukturelle Benachteiligungen fortgeschrieben werden (Fehlerquellen: Input-Verzerrung – historischer Bias, Output-Interpretation, Dilemmata)

Der österreichische Arbeitsmarktservice (AMS) testet derzeit eine Software zur Bestimmung von Arbeitsmarktchancen Arbeitsloser, die ab Mitte 2020 in ganz Österreich zum Einsatz kommen soll. Schon während der Testphase allerdings hagelt es Kritik an diesem Programm: Es sei „ein Paradebeispiel für Diskriminierung“, heißt es von Expert:innen. Dieser Unmut richtet sich vor allem dagegen, dass dem Algorithmus Kriterien zur Bewertung der Arbeitsmarktchancen zugrundeliegen, die die Integrationschancen einzelner von Variablen und Eigenschaftsmerkmalen abhängig machen, auf die das Individuum keinerlei Einfluss nehmen kann. Ein besonders markantes Beispiel stellt das Kriterium „Geschlecht“ dar, das der Algorithmus in Bezug auf die Arbeitsmarktchancen negativ bewertet. Als Konsequenz werden Frauen hinsichtlich ihrer Arbeitsperspektiven schlechter eingestuft. Das System gruppiert die Arbeitslosen in drei verschiedene Sparten – gute Chancen, mittlere Chancen, schlechte Chancen – und aufgrund der spezifischen Bewertungspraktik landen Frauen überdurchschnittlich oft in der mittleren oder schlechteren Gruppe. Dass es für das Kriterium „Geschlecht“ Punktabzug gibt, liegt laut Entwickler:innenseiten nicht am algorithmischen System, sondern am aktuellen Zustand auf dem Arbeitsmarkt, der Frauen diskriminiere. Die Software spiegele diesen Ist-Zustand lediglich wider. An dieser Stelle offenbart sich die Input-Verzerrung des Programms: Es schreibt einen historisch begründeten Geschlechterbias des Arbeitsmarktes fort. Doch die Diskussionen rund um diese Software bleiben nicht hier stehen. Vielmehr liegt hier ein Beispiel dafür vor, dass die von algorithmischen Systemen gelieferten Aussagen nicht notwendigerweise unumstritten oder eindeutig interpretierbar sind. Die Output-Interpretationen können im Gegenteil in ganz unterschiedliche Richtungen gehen. So sehen die einen in der Überrepräsentation von Frauen in der mittleren Gruppe eine Chance, würden für diese Gruppen doch die größten Unterstützungsleistungen bereitgestellt werden. Andere hingegen wenden ein, dass „es durch diesen Algorithmus zu einer Festschreibung und Verstärkung von Ungleichheiten und Benachteiligungen kommt“. Ganz konkret wird darauf hingewiesen, dass der Algorithmus, wie er momentan genutzt wird, ausschließlich Aussagen über die Vergangenheit machen könne, da nur vergangene Daten verwendet werden: „Der Algorithmus”, so Paola Lopez, Mathematikerin an der Universität Wien, „berechnet also nicht die `Chancen´, die ein Individuum am Arbeitsmarkt hat, sondern die strukturelle Benachteiligung, die Menschen mit gleichen Dateneinträgen in der Vergangenheit widerfahren ist.“

iv. Einfach vergessen? Wie technische Parameter zu Diskriminierungen führen können (Fehlerquellen: Handwerkliche Fehler, Input-Verzerrung – unvollständige Daten, Design und Interaktion Input)

Auch simple technische Systeme können zu diskriminierenden Auswirkungen führen. Ein fast schon skurriles Beispiel wurde durch Chukwuemeka Afigbo öffentlich, Facebooks Head of Platform Partership für Afrika und den Nahen Osten. Er veröffentlichte ein Video auf Twitter, das ihn und einen Freund von ihm in einem öffentlichen WC zeigt. Wenn Afigbos Freund seine Hand unter den automatischen Seifenspender hält, so kommt, wie erwartet, Seife auf seine Hand und er kann sich die Hände waschen. Wenn hingegen Afigbo seine Hand unter den Seifenspender hält, bekommt er keine Seife – egal, wie viel er seine Hand bewegt. Der Unterschied: Die Haut- und damit auch die Handflächenfarbe von Afigbo ist dunkel, während sein Freund eine sehr helle Hautfarbe hat.

Grund dafür ist die Funktionsweise der Seifenspender: Diese senden über eine Lampe Infrarotlicht aus. Haut reflektiert dieses Licht, das dann wiederum ein Sensor empfängt und die Hand erkennt. Der Sensor sendet dann ein Signal, Seife auszugeben. Dunkle Haut absorbiert aber relativ viel Infrarotlicht und reflektiert weniger als helle Haut. Um das Problem zu lösen, hätte man lediglich die Sensitivität des Sensors höher einstellen müssen. Scheinbar wurde aber der Seifenspender auf hellhäutige Hände kalibriert und auch nur an diesen getestet. Der Fehler fiel daher nicht auf, weil bereits in der Design- und Testphase implizit nur von hellhäutigen Nutzer:innen ausgegangen war. Vermutlich war auch das Entwickler:innen-Team nicht divers aufgebaut, sodass dieser Bias nicht erkannt wurde. Außerdem hätten die Auswirkungen auch ganz einfach zumindest verringert werden können: Indem der Seifenspender auch über einen Druckknopf Seife ausgegeben hätte.

6. Anhang

a. Kurze Einführung zur Nutzung der Arbeitsblätter

Als Ergänzung zu den vermittelten Inhalten finden Sie hier Arbeitsblätter. Diese wurden vom Innovationsbüro entwickelt. Das Ziel der Arbeitsblätter ist es, anhand konkreter Beispiele das Gelernte anzuwenden und zu problematisieren. Das gemeinsame Arbeiten an den Arbeitsblättern ist unbedingt zu empfehlen, weil nicht alle Schlussfolgerungen selbsterklärend sind. Wir raten dazu, auch wenn die Arbeitsblätter hier frei verfügbar sind, diese gemeinsam mit Expert:innen zu algorithmenvermittelter Diskriminierung zu besprechen.

Diese Arbeistblätter können anhand folgender Leitfragen diskutiert werden:

- Worum geht es in dem Beispiel? Wer wird diskriminiert?
- Welche Besonderheiten können identifiziert werden?
- Wo liegt die Fehlerquelle? Gibt es Unterschiede im Vergleich zur Diskriminierung durch Menschen?
- Welche Ideen für Vermeidungsstrategien gibt es? Welche spezifischen Herausforderungen gibt es bei der möglichen Verringerung dieser Form der Diskriminierung?
- Welche Schlussfolgerungen ziehen Sie für sich?

Zusammengefasst diskutieren die Arbeitsblätter folgende Themen:

Arbeitsblatt 1:
Zielkonflikte bei der Rahmensetzung algorithmischer Systeme; grundsätzliche Frage der Angemessenheit des Einsatzes eines algorithmischen Systems in diesem Bereich

Arbeitsblatt 2:
Rolle von Inputdaten; Proxy-Variablen und ihre Relevanz (Lieblingsfarbe als Proxy-Variable)

Arbeitsblatt 3:
Relevanz von Nachvollziehbarkeit algorithmischer Systeme; Relevanz der Zielvorgaben und der weiteren Einbettung/Nutzung des Systems; Spannungsfeld zwischen der Abbildung bestehender Verhältnisse und zukünftiger Ideale

Algorithmenvermittelte Diskriminierung

Arbeitsblatt 1: Zielvorgaben und Einsatzkontext bei „einfachen“ algorithmischen Systemen

Der österreichische Arbeitsmarktservice (AMS, entspricht der deutschen Agentur für Arbeit) hat ein einfaches algorithmisches System entwickeln lassen, um die Arbeitsmarkt-Integrationschancen von vorgemerkten Arbeitssuchenden quantifizieren zu können. Arbeitssuchenden sollen so je nach Integrationschance bestimmte Integrationsmaßnahmen angeboten werden, indem die Sachbearbeiter:innen im AMS eine entsprechende Empfehlung bekommen. Grundsätzlich wird in drei Gruppen unterschieden: Gruppe A mit hohen Integrationschancen, Gruppe B mit mittleren und C mit niedrigen Integrationschancen. Wie genau die Fördermittel zwischen diesen Gruppen verteilt sein sollen und insb. was mit Gruppe C passiert, ist noch unklar.

Unten ist eine vereinfachte Version dieses Algorithmus zu sehen. Dabei werden persönliche Merkmale und der bisherigen Erwerbsverlauf mit eingezogen. Dazu zählen u. a.: Geschlecht, Alter, Staatsbürgerschaft, Ausbildung, Betreuungspflichten, gesundheitliche Einschränkungen, der bisherige Beruf und das Ausmaß der Beschäftigung. Der Koeffizient eines bestimmten Merkmals gibt die Änderung der Integrationschance im Vergleich zur Basisgruppe an, wenn sich die Ausprägung dieses (und nur dieses) Merkmals ändert. Die Gleichung ist in Variablen formuliert (z. B. „GESCHLECHT_WEIBLICH“, sogenannte Dummy-Variablen); der Wert einer Variable beträgt „1“, wenn die genannte Ausprägung auf die Person zutrifft und „0“, wenn das nicht der Fall ist. Je höher der Wert BE_INT, desto höher die Integrationschance. Die Merkmale und die entsprechenden Koeffizienten entstammen einer statistischen Analyse bestehender, realer Daten und sollen so die realen Integrationschancen abbilden.

BE_INT = f (0,10
– 0,14 x GESCHLECHT_WEIBLICH
– 0,13 x ALTERSGRUPPE_30_49
– 0,70 x ALTERSGRUPPE_50_PLUS
+ 0,16 x STAATENGRUPPE_EU
– 0,05 x STAATENGRUPPE_DRITT
+ 0,28 x AUSBILDUNG_LEHRE
+ 0,01 x AUSBILDUNG_MATURA_PLUS
– 0,15 x BETREUUNGSPFLICHTIG
– 0,67 x BEEINTRÄCHTIGT
+ 0,17 x BERUFSGRUPPE_PRODUKTION
– 0,74 x BESCHÄFTIGUNGSTAGE_WENIG)

- Machen Sie sich zunächst Gedanken zu diesem Algorithmus selbst und seinem diskriminierenden Potential, seinen Vor- und Nachteilen.
- Lesen Sie danach den Artikel „Was der neue AMS-Algorithmus für Frauen wirklich bedeutet“.
- Hat sich Ihre Meinung geändert?

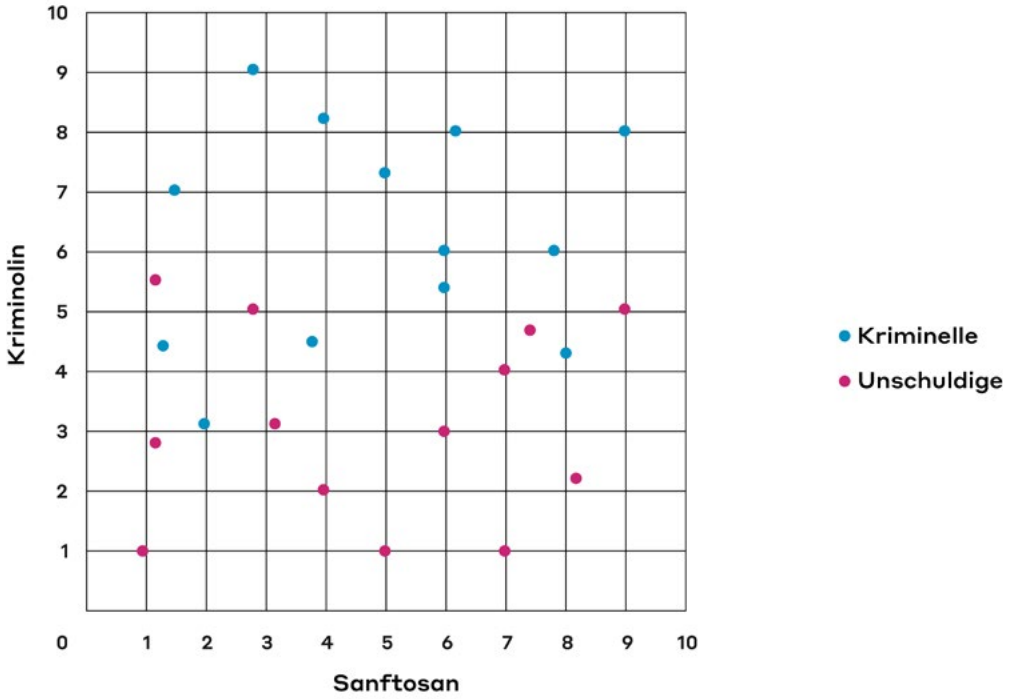
Machen Sie sich zunächst Gedanken zu diesem Algorithmus selbst und seinem diskriminierenden Potential, seinen Vor- und Nachteilen. Lesen Sie danach den Artikel „Was der neue AMS-Algorithmus für Frauen wirklich bedeutet“. Hat sich Ihre Meinung geändert?

Die hier gemachten Angaben sind verkürzt, entsprechend aber ansonsten dem tatsächlichen Arbeitsmarktassistenzsystem – auch „AMS-Algorithmus“. Dessen Einsatz ist aktuell gestoppt.

Innovationsbüro Digitales Leben - Wegweiser Digitale Debatten

Algorithmenvermittelte Diskriminierung

Arbeitsblatt 2: Dilemmata bei Support-Vectoring-Machines



Im Staat Safeland wurden bei Untersuchungen Korrelationen zwischen der Menge der beiden Hormone Kriminolin und Sanftosan im Blut mit dem individuellen Kriminalitätsrisiko entdeckt. Die Regierung erwägt nun, auf Basis dieser Erkenntnisse mögliche Maßnahmen einzuleiten und sie in Justizprozessen zu nutzen. Um die entsprechende Entscheidung zu treffen, wurde eine repräsentative Stichprobe genommen. Die Daten der Stichprobe sind oben als Diagramm dargestellt, wobei jeder Punkt eine Person darstellt mit ihren jeweiligen Hormonwerten.

Ihre Aufgabe ist es nun zu bestimmen, wo die Trennlinie zwischen denjenigen verlaufen soll, die potentielle Kriminelle sind, und denjenigen, die als Unschuldige gelten. Ziehen sie dafür eine **gerade Linie** (Neigung egal), die möglichst gut die Kriminellen von den Unschuldigen trennt.

Neben den allgemeinen Diskussionsfragen, sprechen Sie bitte ebenfalls über:

- Warum ziehen Sie die Linie so und nicht anders? Wie viele und welche Fehler nehmen Sie dabei in Kauf?
- Würde Ihre Entscheidung anders ausfallen, wenn die Regierung unterschiedliche Maßnahmen an diese Trennung koppeln würde (z. B. präventive Maßnahmen anstatt Einsatz in Gerichtsprozessen)?

Hintergrund:

Eine Support Vector Machine (SVM) ist ein Verfahren zur automatischen Klassifikation (Einordnung) von Daten zu bestimmten Klassen. Dazu teilt die Support-Vector-Machine einen multi-dimensionalen Datenraum in Form einer Ebene - einer sogenannten Hyperplane - auf, die die Daten in zwei Klassen separiert. In einem vereinfachten Beispiel mit zwei Dimensionen bedeutet dies, dass Punkte in einem Koordinatensystem die zu bewertenden Daten darstellen und die SVM durch diese Punkte eine Gerade ermittelt, die die Punkte in zwei Klassen trennt.

Dieses Arbeitsblatt basiert auf einer Übung von Prof. Katharina Zweig (2019, „Ein Algorithmus hat kein Taktgefühl“, S. 152ff.). Das korrespondierende reale Beispiel ist COMPASS, ein System im Einsatz in den USA.

Innovationsbüro Digitales Leben - Wegweiser Digitale Debatten

Algorithmenvermittelte Diskriminierung

Arbeitsblatt 3: Input-Verzerrung bei neuronalen Netzen

Sie arbeiten in der Personalabteilung der Firma Supertech AG. Sie wollen einen Bewerbungsfilter erstellen, um aus allen Bewerbungen, die an Ihre Firma gesendet werden, diejenigen herauszufiltern, die vielversprechend sind. Ihnen stehen dafür zwei Datensets von in der Vergangenheit eingereichten Bewerbungen zur Verfügung (Trainingsdatenset A und Trainingsdatenset B).

Erstellen Sie zunächst ein Muster für die Bewerbungen auf Basis bereits getroffenen Entscheidungen in Trainingsdatensets A. Schauen Sie dafür, welche Eigenschaften diejenigen teilen, die im Trainingsdatenset A aufgenommen wurden oder diejenigen, die abgelehnt wurden. Geben sie dabei den Eigenschaften jeweils ein Gewicht (z. B. sehr wichtig, etwas wichtig, unwichtig). Notieren Sie Ihre Erkenntnisse in dem dafür vorgesehenen Feld.

Trainingsdatenset A

Person 1 – abgelehnt Lieblingsfarbe: **rot** Note Informatik: **4** Berufserfahrung: **2 Jahre**	**Person 2 – angestellt** Lieblingsfarbe: **blau** Note Informatik: **1** Berufserfahrung: **2 Jahre**	**Person 3 – angestellt** Lieblingsfarbe: **blau** Note Informatik: **2** Berufserfahrung: **2 Jahre**
Person 4 – abgelehnt Lieblingsfarbe: **orange** Note Informatik: **3** Berufserfahrung: **2 Jahre**	**Person 5 – angestellt** Lieblingsfarbe: **grün** Note Informatik: **1** Berufserfahrung: **2 Jahre**	**Muster A** *Lieblingsfarbe:* *Note Informatik:* *Berufserfahrung:*

Haben Sie das Muster erstellt, wenden Sie es auf das Trainingsdatenset B an. Wenn es passt, müssen Sie Ihr Muster nicht verändern. Bemerken Sie aber Fehler, dann überprüfen Sie, ob Ihre Gewichtungen stimmen und passen Sie diese an. Haben Sie Ihre Mustergewichtung angepasst, notieren sie diese in dem dafür vorgesehenen Feld.

Trainingsdatenset B

Person 6 – abgelehnt Lieblingsfarbe: **rot** Note Informatik: **3** Berufserfahrung: **1 Jahr**	**Person 7 – angestellt** Lieblingsfarbe: **grün** Note Informatik: **2** Berufserfahrung: **1 Jahr**	**Person 8 – angestellt** Lieblingsfarbe: **grün** Note Informatik: **3** Berufserfahrung: **2 Jahre**
Person 9 – abgelehnt Lieblingsfarbe: rot Note Informatik: 1 Berufserfahrung: 2 Jahre	**Person 10 – abgelehnt** Lieblingsfarbe: blau Note Informatik: 3 Berufserfahrung: 2 Jahre	**Muster A+B** *Lieblingsfarbe:* *Note Informatik:* *Berufserfahrung:*

Wenden Sie das angepasste Muster nun auf das Anwendungsset an.

Anwendungsdatenset

Person 11 Lieblingsfarbe: **rot** Note Informatik: **3** Berufserfahrung: **2 Jahre**	**Person 12** Lieblingsfarbe: **rot** Note Informatik: **2** Berufserfahrung: **1 Jahr**	**Person 13** Lieblingsfarbe: **blau** Note Informatik: **2** Berufserfahrung: **1 Jahr**

Neben den allgemeinen Diskussionsfragen, sprechen Sie bitte ebenfalls über:

- Welche der Personen wird eingeladen und welche nicht? Warum? Ist das problematisch?
- Welche Unterschiede sehen Sie in den drei Datensets? Zu welchen Problemen kann das führen?

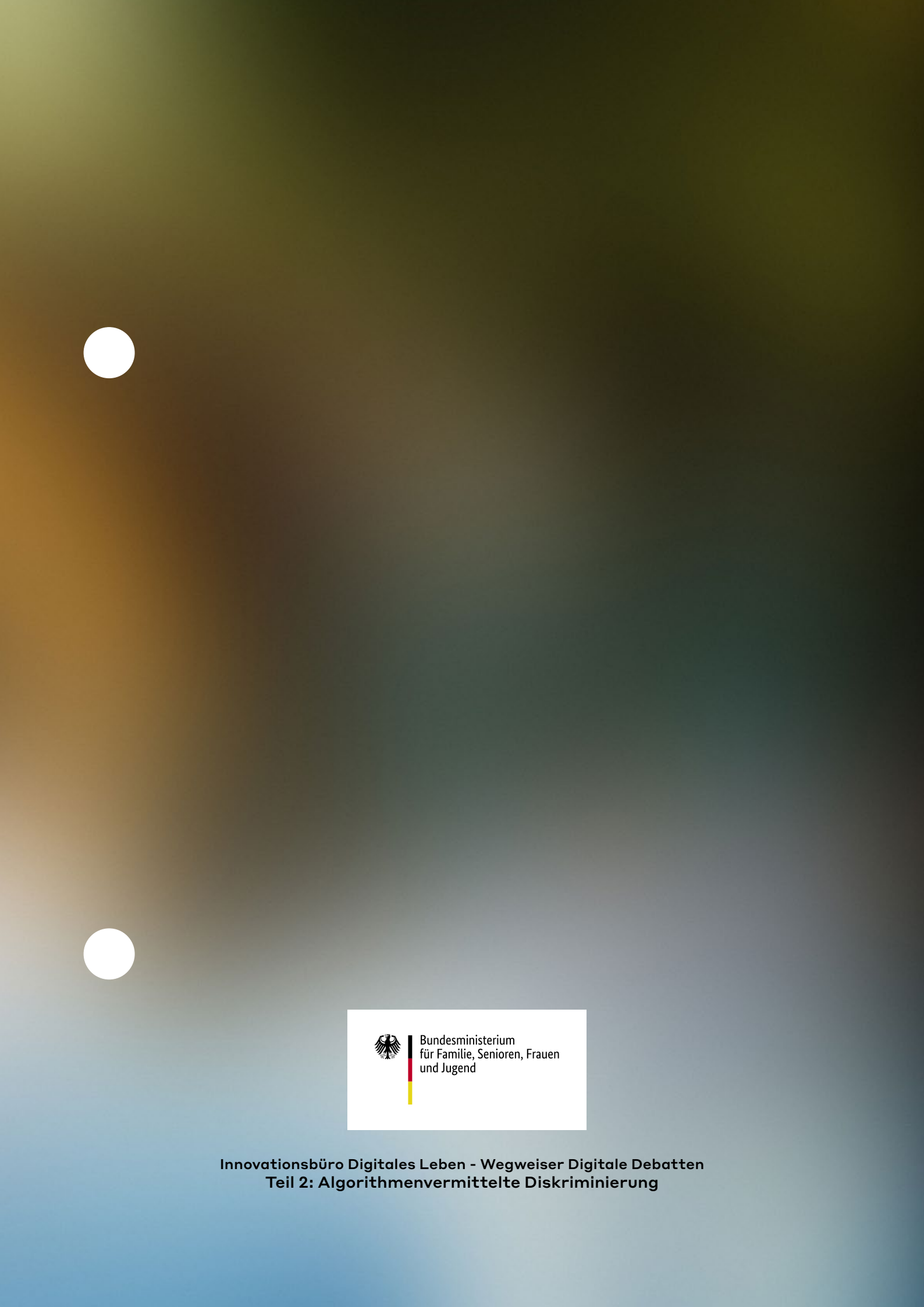
Hintergrund:

Künstliche neuronale Netze simulieren nach dem Vorbild des Gehirns ein Netzwerk aus miteinander verbundenen Neuronen. Sie lernen aus Erfahrung, indem sie die Verbindungsstärke der simulierten Neuronenverbindungen verändern. Zunächst lernt das Netz in der Trainingsphase anhand vorgegebenen Materials: Bestehende Beispiele werden aufgenommen und analysiert. Für jedes Beispiel ist bekannt, was die gewünschte Ausgabe sein soll. Stimmt die Ausgabe des Netzes für ein Beispiel mit dem gewünschten Output überein, dann braucht nichts weiter getan zu werden. Weichen tatsächliche und gewünschte Ausgabe voneinander ab, dann müssen die Verbindungsstärken bzw. Gewichte im Netz so verändert werden, dass sich der Fehler bei der Ausgabe verringert. Je größer der Betrag des Gewichtes ist, desto größer ist der Einfluss eines Neurons auf ein anderes Neuron. Das Wissen eines neuronalen Netzes ist in diesen Gewichten „gespeichert“. Dieser Trainings-Prozess erfolgt im Idealfall so lange, bis alle Beispiele richtig berechnet werden. Danach wird das System eingesetzt.

Das korrespondierende reale Beispiel ist der Bewerbungsfilter bei Amazon.

c. Literatur / Quellen

- AlgorithmWatch (2019): Automating Society – Taking Stock of Automated Decision-Making in the EU, online unter URL: https://algorithmwatch.org/en/automating-society/ [Abruf: 2020-07-06]
- Angwin, Julia/Larson, Jeff/Mattu, Surya/Kirchner, Lauren (2016): Machine Bias. There's software used across the country to predict future criminals. And it's biased against blacks, online unter URL: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing [Abruf: 2020-07-06]
- Augsten, Stephan (2018): Definition „Waterfall Modell". Was ist das Wasserfallmodell?, online unter URL: https://www.dev-insider.de/was-ist-das-wasserfallmodell-a-680501/ [Abruf: 2020-07-06]
- Bertelsmann Stiftung/iRights.Lab (Hrsg.) (2019): Algo. Rules – Regeln für die Gestaltung algorithmischer Systeme, online unter URL: https://www.bertelsmann-stiftung.de/fileadmin/files/BSt/Publikationen/GrauePublikationen/Algo.Rules_DE.pdf [Abruf: 2020-07-06]
- Beuth, Patrick / Breihut, Jörg (2019): Diskriminierender Algorithmus. Patienten-Software benachteiligt Millionen Afroamerikaner, online unter URL: https://www.spiegel.de/netzwelt/apps/usa-algorithmus-benachteiligt-afroamerikanische-patienten-a-1293382.html [Abruf: 2020-07-06]
- Deutscher Bundestag (2018): Sachverständige klären Begriffe rund um „Künstliche Intelligenz", online unter URL: https://www.bundestag.de/dokumente/textarchiv/2018/kw42-pa-enquete-ki-573436 [Abruf: 2020-07-06]
- Cooper, Yvette (2018): Amazon ditched AI recruiting tool that favored men for technical jobs, online unter URL: https://www.theguardian.com/technology/2018/oct/10/amazon-hiring-ai-gender-bias-recruiting-engine [Abruf: 2020-07-06]
- Deutscher Bundestag (2019): Diskriminierungsfreiheit und Daten im Fokus, online unter URL: https://www.bundestag.de/dokumente/textarchiv/2019/kw23-pa-enquete-ki-644010 [Abruf: 2020-07-06]
- Europäischer Gerichtshof für Menschenrechte (2020): Die Europäische Menschenrechtskonvention, online unter URL: https://www.echr.coe.int/Documents/Convention_DEU.pdf [Abruf: 2020-07-06]
- Europarat: Details zum Vertrag-Nr.177, Protokoll, Nr. 12 zur Konvention zum Schutze der Menschenrechte und Grundfreiheiten, online unter der URL: https://www.coe.int/de/web/conventions/full-list/-/conventions/treaty/177 [Abruf: 2020-11-03]
- Holl, Jürgen/Kernbeiß, Günter/Wagner-Pinter, Michael (2018): Das AMS-Arbeitsmarktchancen-Modell, online unter URL: http://www.forschungsnetzwerk.at/downloadpub/arbeitsmarktchancen_methode_dokumentation.pdf [Abruf: 2020-07-06]
- Kompetenzzentrum Öffentliche IT (2012): Trend- und Themensammlung, online unter URL: https://www.oeffentliche-it.de/trendschau [Abruf: 2020-07-06]
- Obermeyer, Ziad / Powers, Brian / Vogeli, Christine / Mullainathan, Sendhil (2019): Dissecting racial bias in an algorithm used to manage the health of populations, online unter URL: https://science.sciencemag.org/content/366/6464/447 [Abruf: 2020-07-06]
- Orwat, Carsten (2019): Diskriminierungsrisiken durch Verwendung von Algorithmen, Antidiskriminierungsstelle des Bundes, online unter URL: https://www.antidiskriminierungsstelle.de/SharedDocs/Downloads/DE/publikationen/Expertisen/Studie_Diskriminierungsrisiken_durch_Verwendung_von_Algorithmen.html?nn=6575434 [Abruf: 2020-07-06]
- Pavey, Harriet (2017): Automatic soap dispenser sparks ‚racism' outrage after footage shows it doesn't work for darkskinned people, https://www.standard.co.uk/news/world/automatic-soap-dispenser-sparks-racism-outrage-after-footage-shows-it-doesnt-work-for-darkskinned-a3615096.html [Abruf: 2020-07-06]
- Scheer, Judith (2019): Algorithmen und ihr Diskriminierungsrisiko, online unter URL: https://www.berlin.de/sen/lads/_assets/ueber-uns/materialien/algorithmendiskriminierungsrisiko_bf.pdf [Abruf: 2020-07-06]
- Schulz, Wolfgang/Dreyer, Stephan (2018): Was bringt die Datenschutz-Grundverordnung für automatisierte Entscheidungssysteme?, online unter URL: https://algorithmenethik.de/2018/04/13/studie-dsgvo/ [Abruf: 2020-07-06]
- Stadt Wien (2020): Formen von Diskriminierung, online unter URL: https://www.wien.gv.at/verwaltung/antidiskriminierung/definition/formen.html [Abruf: 2020-07-06]
- Wimmer, Barbara (2019): Was der neue AMS-Algorithmus für Frauen wirklich bedeutet, online unter: https://futurezone.at/netzpolitik/was-der-neue-ams-algorithmus-fuer-frauen-wirklich-bedeutet/400617302 [Abruf: 2020-07-06]
- Zweig, Katharina (2019): Ein Algorithmus hat kein Taktgefühl, München: Heyne Verlag.



Bundesministerium
für Familie, Senioren, Frauen
und Jugend

Innovationsbüro Digitales Leben - Wegweiser Digitale Debatten

Teil 2: Algorithmenvermittelte Diskriminierung