



Bias durch Künstliche Intelligenz

Soziale Ursachen in algorithmischer Generierung

Künstliche Intelligenz (KI) hält zunehmend Einzug im Alltag in allen Bereichen. Dies kann im beruflichen Umfeld sein, wie im Falle von automatisierten Bewerbungsprozessen, prädikativer Ermittlungsarbeit und algorithmischen Entscheidungssystemen für Versicherungen, für Finanzgeschäfte und Risikobewertungen. Aber auch direkt vor unseren Augen im Privaten haben wir in immer mehr Fällen damit zu tun. Beispiele hierfür wären Gesichtserkennungssysteme von Smartphones, Sprachassistenten wie Siri und Alexa, Autokorrektursysteme, Empfehlungssysteme von Streamingdiensten, die Navigation mit Google Maps oder personalisierte Werbung.

Eine Künstliche Intelligenz wertet Informationen in all diesen Fällen unter anderem aufgrund ihrer vorgelegten Trainingsdaten aus. Sind Trainingsdaten nicht repräsentativ, von schlechter Qualität, nicht in ausreichender Menge vorliegend oder beinhalten diese Unwahrheiten, dann kann dies eine Ursache für erhebliche Probleme bei der Anwendung von Künstlicher Intelligenz sein, sei dies nun in Form von Verzerrungen (Bias) oder Diskriminierung.

Geht es um den sozialen Hintergrund, läuft man bei der Nutzung Künstlicher Intelligenzen Gefahr, bestehende gesellschaftliche Probleme wie Ungleichheiten, bis hin zu Diskriminierung und Rassismus, nicht nur zu reproduzieren, sondern im gravierenden Fall noch zu verstärken. Dies betrifft dann nicht mehr nur das Individuum, sondern damit einhergehend können Künstliche Intelligenzen die politische und gesellschaftliche Meinungsbildung oder das Informationsökosystem beeinflussen.

Zunehmend sollte also die Verteilung von Rollen und Verantwortung im Falle der KI-Anwendung beleuchtet werden. Welche Verantwortung tragen Hersteller, Betreiber, Anwender und Nutzer:innen einer KI? Wie können sich diese Akteure aber auch schützen - sowohl in der Entwicklung, wie auch nach der Implementierung und während der Nutzung einer Anwendung?

In diesem Whitepaper werden die Ursachen für die Entstehung, Reproduktion und Verstärkung von Diskriminierung und Bias in Augenschein genommen. Hierfür werden auch die Rollen des Datenschutzes, der Regulierung und der Ethik betrachtet. Das Ziel dieses Whitepapers ist eine Sensibilisierung im Hinblick auf diese Themen und der Nutzung von KI-Anwendungen und das Aufzeigen von Lösungsansätzen, damit Entwicklung, Betrieb und Anwendung einer KI fairer gestaltet und im Unternehmen risikoarm eingesetzt werden können.

Inhalt

1 KI – wovon sprechen wir hier eigentlich?.....	3
1.1 Intelligenz.....	3
1.2 Künstliche Intelligenz.....	3
1.3 Maschinelles Lernen.....	4
1.4 Trainingsdaten.....	4
Datensätze.....	5
1.5 Bias.....	5
1.6 Diskriminierung.....	6
2 Wo entstehen Verzerrungen in Künstlichen Intelligenzen?.....	7
2.1 Trainingsdaten.....	7
2.2 Modellierung.....	10
2.3 Prompting.....	10
2.4 Sensibilisierung und Schulung des Personals.....	11
3 Bias: Auswirkungen durch KI-Anwendungen.....	12
3.1 Diskriminierung und soziale Ungerechtigkeit.....	12
Reproduktion sozialer Ungerechtigkeit.....	12
Verstärkung sozialer Ungerechtigkeit.....	13
3.2 Unternehmerische Risiken.....	13
3.3 Rechtliche Herausforderungen.....	13
3.4 Unternehmerische Chancen.....	14
4 Strategien zur Vermeidung und Reduzierung von Bias in künstlichen Intelligenzen.....	15
4.1 KI-Governance.....	15
Schulung und Sensibilisierung des Personals.....	16
4.2 Trainingsdaten & Diversität.....	16
4.3 Mathematische Metriken zur Fairness.....	17
4.4 Transparenz, Recht, Regulierung & Ethik.....	20
5 Beispiele: Wie wird tatsächlich gegen Bias vorgegangen?.....	25
Algorithmic Justice League (AJL).....	25
D-BIAS – Menschliche Korrektur von durch KI generierten Verzerrungen.....	25
Geschlechtsbias in KI-Modellen.....	26
6 Abschluss.....	26
6.1 Informationen zum White Paper.....	28
7 Literaturverzeichnis.....	29

1 KI – wovon sprechen wir hier eigentlich?

1.1 Intelligenz

Bei der Beantwortung der Frage, was künstliche Intelligenz ist, sollte man sich vor Augen führen, wie Intelligenz definiert wird. Das Wort geht aus dem lateinischen *intelligere* hervor, was übersetzt soviel wie „erkennen“, „einsehen“ und „verstehen“ bedeutet. Und hier werden viele das Phänomen der „illusion of explanatory depth“ („IOED“)¹ bzw. „Illusion der Erklärungstiefe“ zu spüren bekommen. Der Begriff Intelligenz ist allgegenwärtig und selber hat man ein Verständnis davon, was damit ausgedrückt werden soll und stellt das eigene Wissen nicht infrage. Jedoch wird man möglicherweise schnell feststellen, dass man auf Fragen bezüglich einer detaillierteren Erklärung eben diese nicht tiefer ausführen kann und ein cognitive bias vorliegt. Bei einem „cognitive bias“ handelt es sich um einen systematischen Denkfehler, durch den Menschen Informationen verzerrt wahrnehmen, bewerten oder sich an diese erinnern. Dies geschieht oft unbewusst und abhängig von den eigenen Erwartungen, Emotionen oder sozialen Einflüssen. In der Folge kann dies Entscheidungen beeinflussen und im Falle von Künstlicher Intelligenz kann es sich auch in den Trainingsdaten widerspiegeln, wenn beispielsweise Vorurteile unbewusst in Trainingsdaten einfließen.²

In der Psychologie erschwert die Tatsache, dass verschiedene Fachgebiete sich im Hinblick auf den Begriff „Intelligenz“ in Detailfragen unterscheiden können, die Formulierung einer einheitlichen Definition von Intelligenz. Im Kern ist man sich aber einig:

- [Intelligenz ist die Fähigkeit] „zweckvoll zu handeln, vernünftig zu denken und sich mit seiner Umgebung wirkungsvoll auseinanderzusetzen.“³
- [Intelligenz ist die] „Bezeichnung für die Menge von Fähigkeiten, die zur erfolgreichen Durchführung von Lern- und Denkaufgaben notwendig ist.“⁴

Zusammengefasst können wir Intelligenz als Ergebnis aus aufeinanderfolgenden Prozessen verstehen:

1. Wahrnehmung; wir nehmen unsere Umwelt mit unseren Sinnen wahr.
2. kognitive Auseinandersetzung; wir verarbeiten die Informationen aus unserer Umwelt und versuchen diese aufgrund unserer bisherigen Erfahrungen und unseres Wissens einzuordnen und zu verstehen.
3. Handlung; aus einer möglichst objektiven Wahrnehmung unserer Umwelt entscheiden wir uns für eine Interaktion und ggfs. Problemlösung mit dieser Umwelt, die vernünftig und angepasst ist. Wichtig dabei ist auch, dass durch Erfahrungswerte Fehler aus der eigenen Vergangenheit, oder der von anderen, vermieden werden und ein aktiver Lernprozess stattfindet.

1.2 Künstliche Intelligenz

Diese Interpretation lässt verstehen, weshalb wir im Fall von Maschinen damit begonnen haben, von Künstlicher Intelligenz zu sprechen. Denn bei Künstlicher Intelligenz handelt es sich um ein Teilgebiet der Informatik, in welchem Programme und Systeme entwickelt werden, welche kognitive

1 [Waytz, Adam: Art. What scientific Term or Concept ought to be more widely known?, in: Edge, 2017](#)

2 Kahneman, Daniel u.a., *Judgement under Uncertainty. Heuristics and Biases*, Cambridge 1982

3 Wirtz, Markus Antonius (Hg.), *Lexikon der Psychologie*, 20. überarbeitete Aufl., Göttingen 2021

4 Klimke, Daniela (Hrsg.) u.a., *Lexikon zur Soziologie*. 6. Aufl., Wiesbaden 2020

Fähigkeiten des Menschen (wie z.B. logisches Denken, Lernen, Planung, Kreativität) nachahmen, um Probleme zu lösen oder auf ihre Umwelt zu reagieren.⁵ Die empfangenen Daten müssen dabei nicht mehr nur durch manuelle Eingabe wie bei einem KI-betriebenen Chatbot eingehen. Es lassen sich auch Sensoren wie Kameras und Messgeräte als Informationsquellen hinzufügen, auf deren Daten eine KI nach der Verarbeitung und Einordnung reagieren kann.

Was Künstliche Intelligenzen von konventionellen Systemen und Programmen unterscheidet, ist hier vor allem die Lernfähigkeit aus vergangenen Prozessen. Eine KI kann die eigene Interpretation und Reaktion auf eine Information anpassen, um ein besseres, effizienteres Ergebnis herbeizuführen. Dabei spricht man von programmierten Abläufen, auf denen sich die KI bewegen kann.⁶

In diesem Whitepaper wird der Begriff „Künstlichen Intelligenzen“ im Sinne von „KI-Systemen“ und „KI-Anwendungen“ verstanden. Da der Schwerpunkt auf der soziologischen Betrachtungsweise liegt, wird der Begriff „Künstliche Intelligenzen“ bevorzugt verwendet.

1.3 Maschinelles Lernen

Eine Alternative zu programmierten Abläufen bietet maschinelles Lernen (machine learning). Dabei handelt es sich wiederum um ein Teilgebiet der Künstlichen Intelligenz. Hier wird darauf abgezielt, Systemen die Fähigkeit zu verleihen, aus Daten zu lernen und die Effizienz und Leistung der eigenen Ergebnisse zu verbessern, ohne dass dies explizit programmiert werden muss. Wie auch beim Menschen braucht es für eine Verbesserung eines Prozesses zunächst ein vorheriges Ergebnis und daraufhin einen Erfahrungswert. Erfahrung wird im Falle von machine learning durch Trainingsdaten simuliert. So werden Trainingsdaten genutzt, um einer KI die Möglichkeit zu geben, Muster und Zusammenhänge in Datensätzen zu erkennen und aus diesen dann eine ständig angepasste Lösungsfindung hervorzubringen.⁷

1.4 Trainingsdaten

Die europäische KI-Verordnung definiert Trainingsdaten wie folgt: „Daten, die zum Trainieren eines KI-Systems verwendet werden, wobei dessen lernbare Parameter angepasst werden.“⁸ Trainingsdaten werden einer KI in Datensätzen vorgelegt. Ein Beispiel hierfür wären eine grosse Anzahl an geschäftlichen E-Mails, welche von einer KI zum Trainieren analysiert werden. In der Folge ist die KI in der Lage, E-Mails in der erlernten Förmlichkeit und Professionalität zu formulieren. Die KI wird daraufhin alle Informationen aus diesen Datensätzen auswerten und je nach Aufgabenstellung für eine Lösung heranziehen. Dabei haben Qualität und Quantität der Daten eine ausschlaggebende Rolle inne, da sowohl der Mangel dieser zu einem schlechteren Ergebnis führen kann, als auch die „Überfütterung“ durch Daten. Zu wenig Daten können zu einem ungenauen Ergebnis führen, da mehr Unbekannte nicht berücksichtigt werden können. Gibt man einer KI aber zu viele Datensätze läuft man Gefahr, dass diese genaue Ergebnisse zu bekannten Daten liefern kann, seine Flexibilität und Lernfähigkeit bei neuen, unbekanntem Daten aber einbüsst („overfitting“).⁹

5 [Roscher, Karsten u.a., Künstliche Intelligenz und maschinelles Lernen, in: Fraunhofer IKS, 27.11.2024](#)

6 [Was ist künstliche Intelligenz und wie wird sie genutzt?, in: Europäisches Parlament, 14.09.2020](#)

7 [Prof. Dr. Schmid, Ute, Maschinelles Lernen, in: BIDT, 06.09.2022](#)

8 [Artikel 3: Begriffsbestimmungen, in: EU Artificial Intelligence Act, 02.02.2025](#)

9 [Was ist Overfitting?, in: AWS, 16.04.2025](#)

Dabei müssen Trainingsdaten nicht nur in Form von Text vorliegen. Je nachdem, um was für ein KI-Modell es sich dabei handelt, können Trainingsdaten in Form von Text, Audio, Bildern oder numerischen Werten vorliegen.

Datensätze

Die oben beschriebenen Trainingsdaten basieren auf Datensätzen. Dabei ist ein „Trainingsdatensatz“ dem Wortlaut nach nur einer von drei verschiedenen gebräuchlichen Datensätzen¹⁰:

- Trainingsdatensatz: Dieser wird für das Training eines KI-Modells verwendet, indem es Muster und Zusammenhänge zu interpretieren und darauf basierende Ausgaben zu generieren lernt.
- Validierungsdatsatz: Mit diesem Datensatz wird das Modell während des Trainings überwacht mit dem Ziel, Parameter anzupassen. Durch die Anpassungen der Parameter kann beispielsweise ein KI-Modell effizienter gemacht und Fehler können behoben werden.
- Testdatensatz: Nach dem Training wird dieser Datensatz zur Bewertung der erlernten Leistung verwendet, um sicherzustellen, dass das Modell auch mit unbekanntem Daten umgehen kann. In dieser Phase wird also geprüft, ob ein Modell nach dem Training mit der richtigen Menge an Daten „gefüttert“ wurde, um auch reale Anwendungsdaten verarbeiten zu können.

1.5 Bias

Bias gibt es nicht erst seit der Künstlichen Intelligenz und beschreibt erst einmal die systematische Verzerrung der Wahrnehmung bzw. von Urteilen.¹¹ Das Wort Bias kommt aus dem Englischen und wird mit „Voreingenommenheit“ oder „Verzerrung“ übersetzt. „Systematisch“ beschreibt hier ein Verhalten, welches einem System folgt und diesem entspricht.¹²

Was heisst das aber genau? Ein Bias tritt in einem systematischen Zusammenhang nicht zufällig oder vereinzelt auf. Ein systematischer Bias wird durch Strukturen, Prozesse, Denk- und Datenmuster innerhalb eines Systems verursacht oder begünstigt. Im Falle der KI sind also für entstehende Bias deren Entwurf, Aufbau, Training und Einsatz der KI die Ursache.

Alltägliche Beispiele für Bias könnten Wolkenformationen am Himmel oder die Muster auf der Oberfläche des Mondes sein - aufgrund unseres Wissens, unserer Denkweise, unseres Umfeldes, auch durch unsere Herkunft, assoziieren wir Bilder und Muster anders und sehen darin unterschiedliche Dinge.

Welche Arten von Bias gibt es? Laut Forschern gibt es mittlerweile alleine über 180 Formen der kognitiven Verzerrung¹³, weshalb erst einmal auf die Hauptkategorien von Bias/Verzerrungen eingegangen wird:

- Kognitiver Bias: Diese Art von Verzerrung beschreibt systematische Denkfehler, welche entweder durch begrenzte Informationsverarbeitung, Heuristik oder soziale Einflüsse entstehen können. Diese Verzerrung entsteht durch Menschen, die dazu neigen, Informationen durch die eigene Interpretation zu verzerren, gerade wenn sie die eigenen Ansichten, Glauben und Überzeugungen stützen. Dies nennt man dann einen Confirmation

¹⁰ [Wuttke, Laurenz, Training-, Validierung- und Testdatensatz, in: datasolut, 10.07.2022](#)

¹¹ [Hintze, Christian u.a., Bias, in: DocCheck Flexikon, 19.07.2025](#)

¹² [Systematisch, in: Duden, 14.04.2023](#)

¹³ [Desjardins, Jeff, 24 cognitive biases that are warping your perception of reality, in: World Economic Forum, 30.11.2021](#)

Bias. Manchmal werden Informationen auch verzerrt in Bezug auf den Kontext, mit oder ohne äussere Einwirkung, interpretiert, hier spricht man vom Framing-Effekt.¹⁴ So können kognitive Bias' bereits im frühen Stadium der Erhebung der Trainingsdaten entstehen.

- **Algorithmischer Bias:** Diese Verzerrung basiert, wie der Name schon vermuten lässt, auf dem Algorithmus. Dies kann ebenfalls durch fehlerhafte Trainingsdaten, eine fehlerhafte Modellierung der KI oder aufgrund von entsprechenden Optimierungszielen durch die Entwickler geschehen. Dabei werden ungleiche Ausgaben für verschiedene Gruppen getroffen, also erhält man zu einer Gruppe oder einem Individuum eine andere Ausgabe, wie zu einer vergleichbaren Gruppe oder einem Individuum, ohne dass dies aus den dafür vorgesehenen Parametern geschehen sollte.
- **Statistischer Bias:** Eine statistische Verzerrung findet auf Basis der verwendeten Daten oder Methoden statt. Dies kann passieren, wenn die Trainingsdaten nicht repräsentativ sind, dabei spricht man von einem „Selection Bias“, oder die Datensätze systematisch fehlerhaft sind, dies wäre der Definition nach ein Sampling Bias. Auch durch ausser Acht lassen einzelner Variablen kann es zu einer Verzerrung kommen, wenn die vorliegenden Variablen zu viel Aussagekraft in einer Gleichung besitzen, die nicht der Realität entspricht. In diesem Fall hat man es mit einem Omitted Variable Bias zu tun.

Zu Beginn kann eine Verzerrung durch menschliches Zutun verursacht werden, also einem kognitiven Bias. Je nachdem, wie eine KI entwickelt, modelliert und trainiert wird, kann aus diesem kognitiven Bias ein algorithmischer Bias werden, wenn die KI durch ihre Ausgabe (Mit-)Verursacher einer Verzerrung ist. In der Folge kann ein statistischer Bias in den Daten die ersten beiden Bias' verstärken, wenn beispielsweise historische Verzerrungen mit in die KI-Modelle einfließen und nicht entdeckt und behoben werden.

1.6 Diskriminierung

Das Lexikon der Psychologie definiert Diskriminierung wie folgt: „[...] die Diskriminierung von Menschen aufgrund ihrer Gruppenzugehörigkeit. Soziale Diskriminierung kann z. B. im Arbeitsumfeld entstehen, wenn ältere Personen, Menschen mit Behinderung, Frauen oder Angehörige ethnischer Minderheiten bei Einstellung, Beförderung oder Entlohnung benachteiligt werden, nur weil sie der jew. Kategorie angehören.“¹⁵ Das Staatslexikon ordnet Diskriminierung in drei Elemente unter: „a) eine Unterscheidung, die mit einer b) negativen Wertung verbunden ist und zumeist eine c) soziale Benachteiligung zur Folge hat.“¹⁶

Wo steht Künstliche Intelligenz im Hinblick auf Diskriminierung? Der Aussage aus dem Staatslexikon nach kann man die Abfolge wie eine Checkliste behandeln:

- a) Unterscheidet KI Menschen aufgrund ihrer Eigenschaften wie Alter, Hautfarbe, sexueller Orientierung, Glauben, sozialer Herkunft usw.?*
- b) Ist damit eine negative Wertung verbunden?*
- c) Hat dies eine soziale Benachteiligung zur Folge?*

14 [Stapel, Helmut, Framing: Wenn das eigene Denken durch gezielte Kommunikation fremdbestimmt wird, in: GEO, 21.04.2022](#)

15 Wirtz, Markus Antonius (Hg.), Lexikon der Psychologie, 20. überarbeitete Aufl., Göttingen 2021

16 [Diskriminierung, in: Staatslexikon, 08.06.2022](#)

Wichtig anzumerken ist hier, dass eine Künstliche Intelligenz von Natur aus nicht diskriminierend, aber auch nicht antidiskriminierend ist. Eine KI steht Individuen neutral gegenüber und verarbeitet durch ihre Nutzung in entsprechenden Fällen Informationen über die Person. Darauf hin gibt sie Ergebnisse aus, die sich auf die individuellen Informationen und Eingaben, die Trainingsdaten und ihre Modellierung stützen. Alle drei Quellen können, bewusst oder unbewusst, zu Diskriminierung führen.

So hat man mit Künstlicher Intelligenz ein Mittel, welches aufgrund seiner Natur selbst nicht zu Diskriminierung in der Lage ist, jedoch aufgrund seiner Natur wiederum trotzdem diskriminierend wirkt?. In einem stark vereinfachten Beispiel lässt sich Künstliche Intelligenz mit einem Lautsprecher oder Social Media vergleichen: Die Nutzung dieser Mittel zur weiteren Verbreitung der eigenen Ansichten und Ideologien - ob diese bewusst oder unbewusst diskriminierend sein können ist hier egal - macht diese Mittel nicht zu eigenständig Diskriminierenden. Sie können Diskriminierung aufgrund der jeweiligen Nutzung reproduzieren und verstärken, aber nicht selbst diskriminierend sein.

2 Wo entstehen Verzerrungen in Künstlichen Intelligenzen?

Im Folgenden werden die Ursachen für Verzerrungen aus Sicht der Anwender erläutert.

2.1 Trainingsdaten

Verzerrungen können unterschiedliche Ursachen bei der Nutzung von Künstlichen Intelligenzen haben. Einer der wichtigsten Punkte für das Funktionieren einer Künstlichen Intelligenz und die Qualität ihrer Ausgaben sind die Trainingsdaten. Entsprechend ist dies auch einer der wichtigsten Punkte, wenn es darum geht, Bias und Diskriminierung zu verhindern. Grundlegend gilt - in allen Bereichen, nicht nur in sozialen - dass die Qualität der Trainingsdaten die Qualität einer KI massgeblich steigert. Sind Datensätze, was Bias und Diskriminierung angeht, vorbelastet, unvollständig oder nicht repräsentativ, können sie Verzerrungen übernehmen, reproduzieren und verstärken. Verwendet man also veraltete und unausgewogene Datensätze, kann eine KI keine objektive oder korrekte Analyse und Antwort daraus generieren.

Dabei handelt es sich um statistische Diskriminierung. Die Ausgaben einer KI-Anwendung reproduziert und/oder verstärkt diskriminierende Gesellschaftsverhältnisse, weil ihr möglicherweise individuelle Daten fehlen, der Kontext nicht gegeben ist, die Daten fehlerhaft oder nicht repräsentativ sind oder vergangenheitsorientiert sind. Für die Entstehung einer statistischen Diskriminierung benötigt es keine explizit diskriminierenden Variablen. Das bedeutet, dass Daten und Variablen im Einzelnen aus ethischer Sicht keinerlei Rückschlüsse auf Diskriminierung geben. In entsprechenden Datensätzen aufgrund oben genannter Gründe aber eine verzerrte Interpretation und Ausgabe durch eine KI stattfinden kann. Ein bekanntes Beispiel hierfür ist die Aufforderung an eine KI, ein Bild von Geschäftsführern zu generieren. In den meisten Fällen wurden dabei lange Zeit Männer abgebildet, nicht selten suggerierten KI-Modelle zusätzlich, dass es sich dabei um weisse Männer handeln würde. Historisch betrachtet mag diese Ansicht Fakt sein, jedoch haben sich die Verhältnisse in Vorständen

über die Zeit verändert. Entsprechend sind solche Ausgaben eines KI-Modells nicht repräsentativ oder verhältnismässig.¹⁷

Das hier zu sehende Bild wurde eigens für die Veranschaulichung erstellt. Es wurden offensichtlich erste Schritte hin zu mehr Diversität und Gleichberechtigung getan. Unter den acht weissen Personen sind drei Frauen. Es bleibt mit Spannung zu erwarten, welche Schritte als nächstes angegangen werden. Es wurden drei weitere Bilder mit der gleichen Aufforderung erstellt, welche keine inhaltlichen Veränderungen erzeugten.

Des weiteren kann es zu Problemen führen, wenn man einzelne Daten oder Variablen nicht angemessen berücksichtigt oder ganz aussen vor lässt. Gibt man einer KI-Anwendung eine Aufgabe und stellt nicht alle Daten in der Anfrage (im „Prompt“) und Variablen in der Modellierung der Anwendung zur Verfügung, versucht diese, die Aufgabe dennoch zu lösen und dafür die unvollständige Sammlung an Informationen heranzuziehen. Dies führt zu einem „Omitted Variable Bias“ (OVB).¹⁸



Erstellt am 19.03.25 mit DALL-E mit der Aufforderung, "ein Bild einer Gruppe von Geschäftsführern zu erstellen"

Dies lässt sich an einem vereinfachten Beispiel erklären:

Man analysiert, was und in welchem Umfang das Einkommen von Personen beeinflusst. Um ein repräsentatives Ergebnis zu erlangen würde man hierfür beispielsweise Bildung, Fähigkeiten und Intelligenz als Daten bzw. Variablen heranziehen. Lässt man Fähigkeiten und Intelligenz nun aber weg und lediglich Bildung in der Gleichung, ergibt sich aus Sicht der Anwendung ein überproportionaler Einfluss von Bildung auf das Einkommen.

Werden die Datensätze nicht bereinigt und korrigiert, kann durch die einseitige Repräsentation von Einflüssen auf das Gehalt eine Verzerrung entstehen. Intelligenz und individuelle Fähigkeiten haben ebenfalls positive Einflüsse auf das Gehalt, jedoch wird durch ihr Entfernen aus der Gleichung deren Einflüsse auf das Gehalt alleine der Bildung zugeschrieben.

Hier spricht man auch von algorithmischen Verzerrungen, welche ihren Ursprung in verfälschten, begrenzten oder unvollständigen Datensätzen zum Trainieren, in subjektiven Programmierentscheidungen oder ebenso subjektiven Ergebnisinterpretationen haben.¹⁹

So legte eine Studie offen, dass Gesichtserkennung-Software verschiedener namhafter Entwickler bei Personen mit dunkler Haut eine grössere Fehlerquote bei der Erkennung des Geschlechts auswies, als es bei hellhäutigen Personen der Fall war.²⁰

17 <https://www.heise.de/hintergrund/Neue-Tools-zeigen-wie-voreingenommen-KI-Bildgeneratoren-sind-7744035.html>

18 [Definition Omitted Variable Bias – University of California, Berkeley, 2015](#)

19 [Jonker, Alexandra u.a., Was ist algorithmische Verzerrung?, in: IBM, 20.09.2024](#)

20 [Buolamwini, Joy u.a., Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification, in: proceedings.mlr, 15.01.2018](#)

Ein weiteres Beispiel liefern Algorithmen von Polizeibehörden, wie COMPAS²¹ (Correctional Offender Management Profiling for Alternative Sanctions) aus den USA oder „Sharing Data to Improve Risk Assessment“²² (ehemals „Homicide Prediction Project“) aus Grossbritannien. Solche Algorithmen werden eingesetzt, um das Rückfallrisiko von Straftätern einzuschätzen. Das Problem im Falle der USA hierbei stellte zum Zeitpunkt der Studie dar, dass schwarze Straftäter merklich häufiger als Risiko eingestuft wurden, als dies bei anderen Hautfarben der Fall war. Dem zugrunde liegt das Problem einer nicht real existierenden Überrepräsentation einzelner ethnischer Gruppen in Kriminalstatistiken. Aufgrund einer historisch rassistischen Vorgehensweise von Polizeibehörden in den USA waren Nicht-Weisse häufiger Ziel von polizeilichen Untersuchungen, weswegen es öfter zu Vermerken oder Verhaftungen kam. Auf dem Papier deuten solche Statistiken eine höhere Bereitschaft von Kriminalität einzelner Gruppen an, welche in der Realität jedoch nicht existiert.²³

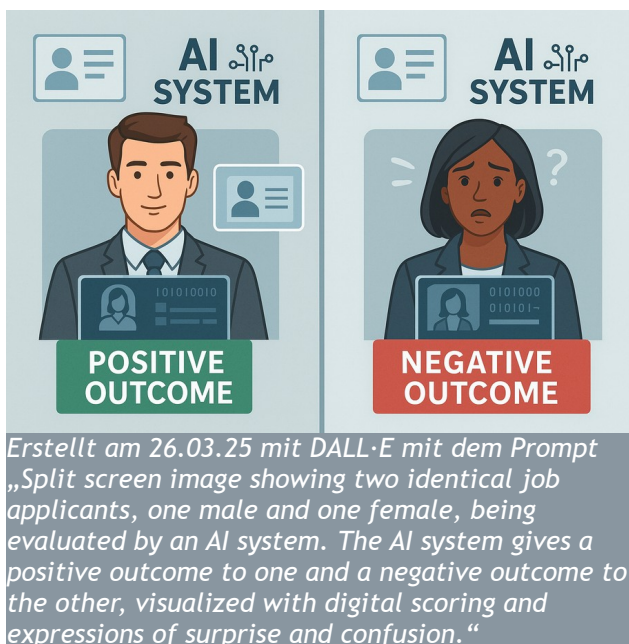
Im britischen Fall wurden personenbezogene Daten von 100.00 bis 500.000 Personen zur Entwicklung des KI-Tools verwendet. Dies schliesst, entgegen einer Stellungnahme des Ministry of Justice in Grossbritannien, Personenkategorien wie Verdächtige, Opfer, Zeugen, vermisste Personen und Personen mit Bedenken bezüglich deren Schutzes, mit ein, aber auch Kategorien von personenbezogenen Daten wie Informationen über potentielles Strafverhalten, sensible Gesundheitsdaten (psychische Gesundheit, Suchtverhalten, Selbstverletzungs-/Selbstmordrisiko, körperliche Behinderungen). Dies wirft Bedenken und Fragen bezüglich des Datenschutzes auf der einen Seite, der ethischen Vertretbarkeit auf der anderen Seite auf. Auch im britischen Fall steht der Vorwurf der Reproduktion und Verstärkung sozialer Ungerechtigkeit aufgrund beispielsweise historisch voreingenommener Trainingsdaten im Raum.

In Fällen der Verbrechensvorhersage muss man sich bewusst werden, dass es sich um ein stark verworrenes und interdisziplinäres Netz handelt. Je mehr Daten und Kategorien von Daten einfließen, umso komplizierter und weniger nachvollziehbar kann die Auswertung eines Algorithmus' werden. Die Risikobewertung eines Algorithmus ist nicht die einzige Kontrollinstanz in solchen Vorhersagen. Für diese Kontrollinstanz muss das Personal für die Erstellung von Trainingsdatensätzen und für die Auswertung aber entsprechend geschult und sensibilisiert sein und mögliche Verzerrungen, Diskrepanzen und Auffälligkeiten interpretieren und auswerten können.

21 [Larson, Jeff u.a., How we analyzed the COMPAS Recidivism Algorithm, in: ProPublica, 23.05.2016](#)

22 [UK: Ministry of Justice secretly developing 'murder prediction' system, in: Statewatch, 08.04.2025](#)

23 Ludwig-Mayerhofer, Wolfgang: Soziale Ungleichheit, Kriminalität und Kriminalisierung, Wiesbaden 2000



Ein letztes Beispiel befasst sich mit Bewerbungen. Hier wurden Frauen vom Algorithmus einer Bewerbungs-KI-Anwendung diskriminiert, weil die Datensätze veraltet und unvollständig waren, entsprechend dann auch nicht mehr repräsentativ. Auslöser war die Annahme der KI-Anwendung, dass Männer aufgrund ihrer höheren Vertretung in technischen Berufen besser für diese Jobs geeignet seien. Hatten Frauen in ihren Bewerbungen also stehen, dass sie Hochschulen nur für Frauen besucht haben oder Rollen in außerschulischen, -universitären oder -beruflichen Aktivitäten mit dem Vermerk „women“, verpasste die Anwendung der Bewerbung automatisch ein Downgrade.²⁴

2.2 Modellierung

Die oben genannten Beispiele zeigen auf, was beim Umgang mit Trainingsdaten bewusst oder unbewusst zu Bias führen kann. Jedoch sind nicht allein diese für mögliche Bias durch eine KI verantwortlich. Auch die Modellierung spielt eine integrale Rolle bei der Entstehung von Bias.

Nehmen wir das Beispiel mit dem Downgrade, welches weiblichen Bewerberinnen im Bewerbungsprozess verpasst wurde. Aufgrund der Modellierung führte die Interpretation der Trainingsdaten zu einer Assoziation des männlichen Geschlechts mit dem ausgeschriebenen Beruf. Selbst wenn dieser mehrheitlich von Männern ausgeübt werden sollte, ist die Assoziation zwischen Geschlecht und Beruf nicht erwünscht. Die Herabstufung von Frauen führt zu einer Verzerrung, da der Auslöser davon keinen Aufschluss über die tatsächlichen Qualifikationen von Frauen wiedergeben kann.

Bei der Modellierung kann präventiv gewirkt werden, indem irrelevante Merkmale wie das Geschlecht ausgeschlossen oder deren Einfluss transparent gemacht wird, etwa durch Warnhinweise bei verzerrender Gewichtung. Derartige Massnahmen zur Prävention und Kontrolle von Verzerrungen durch Künstliche Intelligenz sind möglich, sie müssen nur erkannt und richtig umgesetzt werden.

Bei der Modellierung sind in erster Linie IT-Experten und Programmierer gefragt. Jedoch werden diese meistens nicht darin geschult abzusehen, wie Daten aus beispielsweise sozialwissenschaftlicher Sicht interpretiert werden. Entsprechend ist es für Programmierer schwierig bis unmöglich, zusätzlich zur Aufgabe der Entwicklung und Modellierung eines KI-Modells, die Trainingsdaten, die Nutzung der Anwendung und den Umgang mit Prompts über verschiedene Disziplinen hinweg so zu kontrollieren, dass Verzerrungen ausgeschlossen werden können.

²⁴ [Dastin, Jeffrey: Insight – Amazon scraps secret AI recruiting tool that showed bias against women, in: Reuters, 11.10.2018](#)

2.3 Prompting

Nun wurden Trainingsdaten und Modellierung als mögliche Ursachen erläutert. In diesen Fällen sind Hersteller und Betreiber von KI-Anwendungen in der Verantwortung, mit Bias umzugehen. Mit in die Verantwortung genommen wird beim Prompting aber auch Nutzer:innen einer KI-Anwendung. „Prompts“ sind die Eingabebefehle und durch „Prompting“ wird der algorithmische Prozess einer Künstlichen Intelligenz angestoßen, gesteuert und beeinflusst. Die Art und Weise, wie Prompts formuliert werden, stellt einen Aspekt dar, wie die darauf folgende Ausgabe formuliert wird. So können folgende Ursachen für Prompt-Bias entstehen:

- **Voreingenommenheit:** Wenn die Person, die den Prompt formuliert, Annahmen in ihren Prompt packt, welche subjektive Wahrnehmungen, Stereotypen oder direkt Vorurteile enthalten, kann eine KI-Anwendung in ihrer Ausgabe diese übernehmen. Hier gilt, dass nicht immer von Absicht ausgegangen werden muss. Aussagen können von allen Personen immer wieder so formuliert werden, dass daraus Verzerrungen durch oben genannte Eigenschaften entstehen können - bewusst wie unbewusst.
- **Unklare Ausdrucksweise:** Wird in einem Prompt beispielsweise eine eindeutige und ausführliche Ausgabe verlangt, ohne entsprechend Informationen zu beinhalten, kann es passieren, dass eine KI-Anwendung Details falsch oder anders interpretiert oder „Fakten“ einfach hinzufügt oder diese in einem falschen Kontext wiedergibt. Wenn man in die Filiale eines Coffeeshops geht und dort einen Kaffee bestellen will, kann man dies mit „einen Kaffee bitte“ oder mit „einen Latte Macchiato mit einem extra Schuss Espresso und fettarmer Milch“ bestellen. Man wäre vielleicht mit beidem zufrieden, jedoch ist das Hinzufügen von Details zielführender für das, was man eigentlich will. Ähnlich verhält es sich mit dem Prompting. Je detaillierter und präziser ein Prompt verfasst wird, desto eher erhält man die Antwort, die man tatsächlich sucht und desto eher kann man Verzerrungen vermeiden.
- **Confirmation Bias:** Wird in einem Prompt auf eine spezifische Antwort in der Ausgabe gedrängt, um beispielsweise die eigene These zu untermauern, kann die KI-Anwendung dazu verleitet werden, entsprechende Antworten zu liefern. Auch hier können „Fakten“ erfunden werden. Da sich die Ausgaben eines KI-Modells an die Prompts anpassen, kann sich dieser Effekt weiter verstärken. Speziell dann, wenn keine Quellen für die Antworten einer KI angefordert werden, läuft man als Nutzer:in Gefahr, Unwahrheiten zum Opfer zu fallen. Dies ist ein sich reproduzierendes und verstärkendes Phänomen, da wir Menschen durch den Confirmation Bias auch dazu neigen eher dem zu glauben, was unseren eigenen Ansichten und unserem Glauben entspricht.

2.4 Sensibilisierung und Schulung des Personals

Ein weiterer Punkt, der bei der Planung einer Implementierung einer Künstlichen Intelligenz berücksichtigt werden muss, ist die Schulung und Sensibilisierung des Personals, welches mit der Anwendung und Interpretation solcher Systeme betraut ist. Ein extremes, aber veranschaulichendes Beispiel ist der VioGén-Algorithmus in Spanien²⁵, welcher seit 2007 zur Einschätzung des Risikos bei geschlechtsspezifischer Gewalt eingesetzt wird. Es wurde ersichtlich, dass viele Polizeibeamte nicht ausreichend in den Dynamiken geschlechtsspezifischer Gewalt geschult sind. Dem Algorithmus wurde eine so zentrale Rolle zugewiesen, dass die Beamten sich teilweise stark bis gänzlich auf die

25 [In Spain, the VioGén algorithm attempts to forecast gender violence, in: Algorithmwatch, 27.04.2020](#)

algorithmischen Bewertungen verlassen haben, ohne diese kritisch zu hinterfragen oder durch eigenes Fachwissen zu ergänzen. In der Folge gab es Fälle, in welchen Opfer trotz eindeutiger Warnsignale, wie die Androhung von Gewalt und Ermordung, als „niedriges Risiko“ eingestuft wurden. Dies hatte in diesen Fällen tragische Konsequenzen, darunter auch tödliche Gewaltdelikte.

Dieser Sachverhalt zeigt auf, dass selbst ein ausgereiftes Programm, welches in vielen Fällen schlimmeres verhindern konnte, seine Wirksamkeit verliert, wenn das mit ihm betraute Personal nicht geschult oder sensibilisiert ist, um die Ergebnisse korrekt zu interpretieren und in den Kontext der individuellen Fälle einzuordnen. Dies stellt ein extremes brutales Beispiel dar und in der Geschäftswelt hat man es glücklicherweise nicht oft mit dieser Art von Tragweite zu tun. Es soll aber verdeutlichen, dass Menschen nach wie vor die Verantwortung für die ihnen zugeteilte Arbeit tragen und ein Algorithmus die Tätigkeit vereinfachen, nicht aber umgehen kann. So würden Versicherungen keine ungelerten Sachbearbeiter mit der Bewertung komplexer Schadensfälle oder Risikoprüfungen betrauen. Ebenso wenig setzt man in der Pflege kaum geschulte Kräfte auf der Intensivstation, in der Luftfahrt Piloten ohne Erfahrung für schwierige Flugrouten oder studentische Hilfskräfte für Gerichtsverhandlungen ein. In erster Linie braucht das mit der Künstlichen Intelligenz betraute Personal Erfahrung und Wissen im jeweiligen Sachbereich. Hinzu kommt aber auch eine erforderliche Kompetenz zum Umgang mit der KI, um deren Wirken zu verstehen und Verzerrungen oder anderweitige Probleme zu erkennen, interpretieren und ihnen entgegenzuwirken oder der richtigen Anlaufstelle zu melden.

3 Bias: Auswirkungen durch KI-Anwendungen

Nachdem darauf eingegangen wurde, wie Verzerrungen in KI-Anwendungen überhaupt entstehen, stehen nun die Auswirkungen und deren Bedeutung im Fokus.

3.1 Diskriminierung und soziale Ungerechtigkeit

Das Thema Diskriminierung und soziale Ungerechtigkeit wurde eingangs bereits erwähnt und die ersten, direkten Folgen beleuchtet. Nun soll betrachtet werden, welche weitreichenden gesellschaftlichen Folgen mit Verzerrungen in und durch Künstliche Intelligenz einhergehen.

Reproduktion sozialer Ungerechtigkeit

Gesellschaftliche Strukturen erhalten sich durch soziale Reproduktion selbst aufrecht, so erklärt es die Theorie des französischen Soziologen Pierre Bourdieu. Das bedeutet, dass Machtverhältnisse und (Un-)Gleichheiten weitergegeben werden, damit die Struktur, welche sich auf diese stützt, weiter in ihrer Form erhalten bleiben kann. Trainiert man eine Künstliche Intelligenz mit historischen Daten, findet diese Reproduktion in den Algorithmen von Künstlichen Intelligenzen auf einem neuen sozialen Umfeld statt. Handelt es sich um Trainingsdaten, wie sie der Anwendung COMPAS in den USA vorgelegt wurden, werden überholte „Fakten“ und vergangene Normen wieder verstärkt.

Finden solche Prozesse der Reproduktion und Verstärkung durch algorithmische Systeme ohne genauere Untersuchung und Kritik statt, birgt dies die Gefahr der Verlangsamung der gesellschaftlichen Entwicklung oder Unterminierung bereits erreichter gesellschaftlicher Entwicklungen.

Auch auf Unternehmen lässt sich dies übertragen. Wird bei Entscheidungsfindungshilfen in Unternehmensbereichen wie Personalwesen nicht auf neue Entwicklungen wie Gesetzesanpassungen, gesellschaftliche Maßstäbe usw. Rücksicht genommen, erhöht sich das Risiko, dass ein Unternehmen

mit überholten Methoden und einem veralteten Wissensstand agiert, nicht auf Neuerungen reagiert und dadurch in mehreren Hinsichten das Nachsehen hat oder gar Schaden davon tragen kann.

Verstärkung sozialer Ungerechtigkeit

Geht man für ein Gedankenspiel davon aus, dass alle Künstliche Intelligenzen, welche man selbst benutzen kann, Bias verursachen, ergeben sich beunruhigende Bilder. Man liegt auf dem Sofa und erhält Serien- und Filmvorschläge, bearbeitet durch eine KI, und folgt diesen. Man klickt auf die Werbung, welche man beim Surfen im Internet sieht, besucht durch KI empfohlene Restaurants, folgt Personen auf Social Media, welche jedes Detail im Leben im Griff zu haben scheinen und das eigene Selbstbewusstsein in die Schranken weist. Banken gewähren keinen Kredit, weil eine KI das bewohnte Viertel als Grund für eine Herabstufung der Kreditwürdigkeit ansieht und man hat aufgrund von Hautfarbe, Namen, Alter, sexueller Orientierung, gesundheitlicher Vorgeschichte und Geschlecht nur beschränkten Zugang zum Arbeitsmarkt. Und dies deswegen, weil eine KI ungehindert damit begonnen hat, Menschen aufgrund für den Job irrelevanter Merkmale in Gruppen zu sortieren.

Die in dieser Auflistung genannten künstlichen Intelligenzen können aufgrund von Modellierung oder Trainingsdaten zu fragwürdigen Ausgaben kommen, welche mehr mit der menschlichen Zugehörigkeit zu einer Gruppe und weniger mit individuellen Merkmalen und Eigenschaften zusammenhängen. Nun handelt man entsprechend der Vorgaben einer KI und verhält sich stärker den Stereotypen folgend, als man das sonst tun würde. Die KI-Anwendung wiederum nimmt diese Daten als Bestätigung ihrer Ausgabe und Auswertung, macht also eine sich selbst erfüllende Prophezeiung daraus und der Algorithmus wird in diesen Mustern bestärkt. Dieser Kreislauf würde nicht nur zur Weiterverbreitung von Verzerrungen führen, sondern auch zur Verstärkung dieser.

Davon auszugehen, dass jede KI, mit der wir es zu tun haben, derartige und spürbare Verzerrungen mit sich bringt, ist übertrieben. Jedoch sehen wir uns der Gefahr bereits in der analogen Welt gegenüber und die Komplexität, die mit künstlicher Intelligenz einher geht, bietet viel Raum für Fehler oder übersehene Details.

3.2 Unternehmerische Risiken

Wirtschaftliche Schäden durch Übernahme, Reproduktion und Verstärkung von Bias und daraus folgender Diskriminierung können alle Beteiligten treffen. Das erste Beispiel, welches einem hier in den Sinn kommen kann, ist der wirtschaftliche Schaden potenzieller Bewerber:innen durch die fehlende Anstellung. Jedoch trifft der Schaden auch das Unternehmen, wenn durch Bias besser qualifizierte Arbeitskräfte aufgrund einer automatischen (Vor-)Entscheidung durch eine KI-Anwendung nicht eingestellt werden. Das bereits erwähnte Beispiel mit der Bewerber-KI, welche Frauen im Bewerbungsprozess mit einem Downgrade versieht, wäre ein entsprechender Fall.

Die Auswirkungen können sich aber auch in der Wahrnehmung des Unternehmens widerspiegeln. Verzerrungen, hervorgerufen durch die Nutzung einer Künstlichen Intelligenz, können als generelle Unternehmenspraxis wahrgenommen oder als Versäumnis bei der Behebung wahrgenommen werden. Ob diese Fälle so zutreffen mögen, spielt dabei eine nebensächliche Rolle. So wären Image- und Vertrauensverlust eine mögliche Konsequenz bei Untätigkeit im Hinblick auf Bias durch Künstliche Intelligenz.

3.3 Rechtliche Herausforderungen

Rechtliche Konsequenzen und regulatorische Auflagen und Strafen können ebenso ein Ergebnis durch (nicht behobene) Verzerrungen und darauf fussende Diskriminierung sein. Sollte eine durch eine KI-Anwendung automatisierte Entscheidung diskriminierender Natur sein oder durch einen Bias die Realität nicht abbilden und das Unternehmen dadurch zu unwahren Aussagen verleiten, kann dies weitreichende rechtliche Folgen haben.

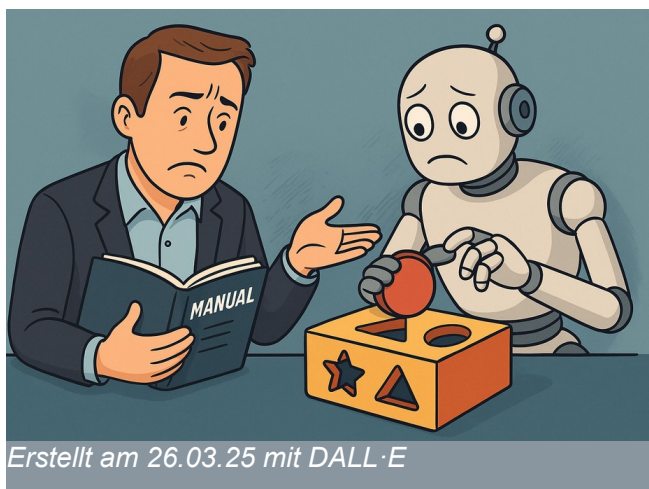
In einem Beispiel aus Finnland aus dem Jahre 2018 wird dies ersichtlich. Eine Person erhielt eine Absage für einen Antrag für einen Kredit, weil die automatisierte Entscheidung durch eine KI-Anwendung Merkmale wie Geschlecht, Muttersprache, Alter und Wohnort für die Bewertung der Kreditwürdigkeit verwendete. Aufgrund dieser KI-Nutzung erhielt die klagende Person vor Gericht Recht.²⁶ In solchen Fällen werden konkret auch datenschutzrechtliche Fragen relevant, da die Aussagekraft dieser Personendaten für die Kreditwürdigkeit zumindest fragwürdig sind. In diesem Fall verstösst man in der Europäischen Union gegen die Datenschutzgrundsätze nach Art. 5 Abs.1 DSGVO.

Zudem verlangt Art. 22 DSGVO bei automatisierten Entscheidungen, welche erhebliche Auswirkungen auf Personen haben, dass diese Entscheidungen im Einzelnen explizit transparent, erklärbar und kontrollierbar sein müssen und Unternehmen können sich auch der Notwendigkeit einer Datenschutz-Folgeabschätzung nach Art. 35 DSGVO und Erwägungsgrund 91 DSGVO gegenüber sehen.

3.4 Unternehmerische Chancen

Nach dem kritischen Blick auf die Risiken und Gefahren durch die Nutzung Künstlicher Intelligenz und gleichzeitiger Problematik durch Bias und Diskriminierung lohnt sich ein Blick auf die Chancen, welche sich einem Unternehmen bieten. Dass Künstliche Intelligenz weiterhin auf einem schnellen Vormarsch ist lässt sich nicht leugnen und dass man sich diese Entwicklung nicht wegdenken kann, erst recht nicht. Daher sollte man sich nach der Bestandsaufnahme aller Risiken mit den Chancen befassen und sehen, wo man konkurrenzfähig bleiben kann.

Eine ordentliche, überlegte und umsichtige Implementierung eines KI-Modells in ein Unternehmen wird beim künftigen Einsatz weniger Ressourcen binden und weniger Kosten generieren. Alleine die



Klärung der Fragen, welche Modelle überhaupt infrage kommen, wie diese entwickelt sein müssen, was der Betrieb kostet und was er an Umsatz oder Einsparungen generiert, wie man ein Modell organisiert und am Laufen hält und wie eingebundene Mitarbeiter geschult und sensibilisiert werden spart einem Unternehmen mittel- und langfristig Zeit und Ressourcen und ermöglicht einen erhöhten Investitionsschutz. So vielseitig die Chancen und Risiken sind, so vielseitig sind auch die Angebote an Modellen - und es werden mehr. Mit diesen Anschaffungs- und Planungsüberlegungen wären initial bereits wichtige und grosse Schritte getan.

26 [Künstliche Intelligenz: Haftungsfragen für Finanzbranche, in: Fonds Professionell online, 08.02.2023](#)

Ein frühes Hinzuziehen von Experten aus verschiedenen Bereichen untermauert eine solide Planung weiterhin. Dazu gehören Experten in der Entwicklung und Konzeption, bei der Auswahl und Auswertung der Trainingsdaten, bei der Vorbeugung von rechtlichen, sozialen und gesellschaftlichen Problemen, d.h. Sozialwissenschaftler, juristische Expertise, Datenschutz-Knowhow und je nach Projekt und Branche noch weitere Disziplinen. Im Falle von Ethik- und Diskriminierungsfragen ist eine ethisch verantwortungsvolle und zukunftsfähige Organisation gefragt. Jedoch gilt es hier abzuwägen: Mit einem möglichst umsichtigen Start und etwas mehr Zeit für Analyse und Konzeption zu beginnen oder unter massivem Zeitdruck so schnell wie möglich der Konkurrenz zuvor zu kommen.

Die KI-Verordnung der EU verlangt nicht nach einem Beauftragten für Künstliche Intelligenz, jedoch sollten Unternehmen bei der Implementierung bereits Zuständigkeiten geregelt haben und Ansprechpartner für (An-)Fragen haben, um keine rechtlichen Konsequenzen fürchten zu müssen. Eine klare Regelung sorgt für einen reibungsloseren Verlauf, mehr Vertrauen bei Kunden und Partnern und für mehr Sicherheit im Unternehmen selbst.

Wird all das berücksichtigt, entstehen erhebliche Wettbewerbsvorteile. Die Künstliche Intelligenz hält immer mehr Einzug im beruflichen Alltag und verlangt nach entsprechenden Antworten und Massnahmen - nicht nur auf rechtliche Anforderungen, sondern auch im Hinblick auf die gesellschaftliche Akzeptanz und die soziale Fairness. Werden diese frühzeitig angegangen, kann dies bei der Wahrnehmung der Marke des Unternehmens ein weiteres Argument für Kunden, Partner, aber auch für zukünftige und potentielle Mitarbeiter sein, sich für dieses Unternehmen zu entscheiden. Es bieten sich für Unternehmen also nicht nur rechtliche Absicherungen, sondern auch zukunftsgewandter strategischer Mehrwert.

4 Strategien zur Vermeidung und Reduzierung von Bias in künstlichen Intelligenzen

So vielfältig die Möglichkeiten auch sind, Bias bewusst oder unbewusst herbeizuführen, die Strategien zur Vermeidung und Reduzierung dieser werden ebenso vielfältig. Für Hersteller und Betreiber Künstlicher Intelligenzen ist es für die eigene Planung von Bedeutung, die Möglichkeiten zu kennen. Darüber hinaus gilt es zu verstehen, dass Strategien und Massnahmen zur Vermeidung und Reduzierung von Bias interdisziplinär sein müssen. Gesellschaftswissenschaftler können soziale Bias beispielsweise durch Analyse von Trainingsdaten erkennen und Massnahmen vorschlagen, jedoch meistens nicht technisch umsetzen. Informatiker wiederum können technische Anforderungen umsetzen, jedoch haben diese nicht den fachlichen Hintergrund, um soziale Bias zu entdecken, einzuordnen und Lösungen zu entwickeln. Ebenso verhält es sich mit Mathematikern, welche durch mathematische Methoden und Lösungen zur Erkennung von Bias in beispielsweise Statistiken beitragen können, jedoch nichts zur technischen Umsetzung oder zu gesellschaftswissenschaftlichen Methoden bei der Einordnung der Bias'.

4.1 KI-Governance

Nicht nur aufgrund von Verordnungen wie dem **EU-AI-Act**²⁷ ist es wichtig, ein Governance-Modell im Hinblick auf KI-Anwendungen zu etablieren. Dies gilt für Hersteller, Anbieter wie Anwender gleichermaßen. Im Falle von Bias und Diskriminierung sollen Strategien, Richtlinien und Verantwortlichkeiten

27 [EU-AI-Act](#)

aufgestellt werden, welche den Einsatz von KI-Anwendungen möglichst ethisch, fair und kontrolliert gestalten. Organisationen und Institutionen müssen entsprechend handeln und Massnahmen wie Prüfungen der Fairness, Audits und interne Mechanismen zur Kontrolle und Regulierung ergreifen. Die grossen Technologieunternehmen haben bereits Corporate Guidelines mit entsprechenden Ethik-Richtlinien verabschiedet oder arbeiten daran. Diese sollen aber nicht nur die Fairness in den Fokus stellen, sondern auch die Transparenz, also beispielsweise wie eine solche Anwendung genutzt wird und wie sie zu ihren Ergebnissen kommt. Für einen möglichst Bias-freien Betrieb einer Anwendung sind derartige Richtlinien, Strategien, Regulationen und Kontrollen ein Eckpfeiler, welcher schlussendlich alle menschlichen Akteure schützt.

Aber auch staatliche und internationale Institutionen wie die Europäische Kommission haben in diesem Bereich erste konkrete Schritte vorgenommen, indem sie mit dem **EU-AI-Act** regulatorische Massnahmen verabschiedet haben, welche Missbrauch und Verzerrungen vorbeugen und damit eine vertrauenswürdige Nutzung gewährleisten sollen. Dies nimmt Hersteller und Betreiber solcher Anwendungen allerdings nicht aus der Verantwortung, weiterhin stets die Ausgaben einer KI-Anwendung auf Richtigkeit und die Art der Verarbeitung auf Gesetzeskonformität zu überprüfen.

Schulung und Sensibilisierung des Personals

Selbst ein technisch ausgereiftes System kann verzerrende oder diskriminierende Wirkungen entfalten, wenn es durch ungeschultes oder nicht sensibilisiertes Personal falsch interpretiert oder unkritisch angewendet wird. Analog gilt: So wie ein Krankenhaus kein ungeschultes Personal für Diagnosen einsetzen sollte, und wie Versicherungen keine ungelerten Kräfte mit der Analyse komplexer Vertragsrisiken betrauen sollten, darf auch der Einsatz Künstlicher Intelligenzen nicht ohne fundierte Schulung erfolgen. Dies umfasst sowohl das technische Verständnis über die Funktionsweise und Grenzen einer KI, als auch gesellschaftliche und psychologische Sensibilität – insbesondere wenn vulnerable Gruppen und deren Daten Teil einer entsprechenden Auswirkung sind.

Eine Einführung verbindlicher Fortbildungsmassnahmen zur Sensibilisierung und Schulung für alle Berufsgruppen, die mit der Anwendung Künstlicher Intelligenzen betraut sind, wird gerade mit wachsendem Einfluss und Nutzen dieser Systeme und Anwendungen empfohlen. Diese sollen sowohl die Bewertung von Unsicherheiten und Wahrscheinlichkeiten, wie auch die Reflexion eigener Vorannahmen und algorithmischer und systemischer Verzerrungen umfassen. Massnahmen wie Human-in-the-Loop entfalten ihre Wirkung nur dann, wenn der entsprechende Mensch im Prozess auch weiss, wo kritisch hinterfragt werden muss. Technische Lösungen bedürfen weiterhin sozialer Kompetenz, um verzerrungs- und diskriminierungsfrei wirken zu können.

Hier muss jeder Entwickler und Betreiber zwischen der Über- und Unteranpassung der Trainingsdaten abwägen. Hier spricht man von einem Bias-Varianz-Dilemma und dieses lässt sich kurz und knapp wie folgt erklären: Bei einer Überanpassung ist die Varianz (Streuung) der Datensätze zu hoch, sprich tritt das Overfitting ein und während die Anwendung gute Performance bei ihren Trainingsdaten aufgrund der genauen Abstimmung auf diese erzielt, schwächt sie bei neuen Daten - was oftmals reale Daten in der praktischen Anwendung sind. Bei einer Unteranpassung sind die Daten zu ungenau und führen nicht selten auf die eine oder andere Weise zu Bias durch fehlende Zusammenhänge.

4.2 Trainingsdaten & Diversität

Trainingsdaten stellen ein Kernelement für mögliche Bias und daraus folgende Diskriminierung dar. Aktiv kann man die Problematik hier durch eine sorgfältigere Analyse und Behandlung dieser Daten angehen. Hierzu empfehlen sich in erster Instanz folgende Lösungsansätze:

1. Diversifizierung der Datensätze

Hierzu ist eine inklusive Datenerhebung notwendig. Dabei muss sichergestellt werden, dass Daten von allen relevanten und anvisierten Zielgruppen gesammelt werden, damit eine möglichst umfassende Repräsentation gewährleistet werden kann. Berücksichtigt man verschiedene Datenquellen, Forschungsmethoden und theoretische Perspektiven bei der Zusammensetzung der Trainingsdaten, lässt sich die Bias-freie/-reduzierte Nutzung durch Triangulation erreichen.²⁸

2. Anonymisierung & Minimierung der Datensätze

Datensätze zu anonymisieren und zu minimieren bringt mehrere Vorteile mit sich. Da durch die Anonymisierung keine Personendaten verarbeitet werden, muss der Datenschutz nicht berücksichtigt werden. Zusätzlich ist es einfacher, Transparenz und Richtlinien einzuhalten, wenn man den Aufwand nicht durch eine zu grosse Menge an Trainingsdaten beschränkt halten kann.

3. Aktualität der Datensätze

Es ist für den Entwickler und den Betreiber von KI-Anwendungen aus verschiedenen Gründen von Interesse, die Daten im Sinne der Branchen- oder Forschungsstandards aktuell zu halten und bei Bedarf zu aktualisieren. Dies kann sich im Endeffekt positiv auf die Qualität der Anwendung, die Reputation des Unternehmens, den Schutz von Individuen und Gleichstellung und Gesetzeskonformität auswirken. Beispielsweise muss ein Chatbot im Kundenservice eines Unternehmens permanent auf die neuesten Produkte, Dienstleistungen und auf Compliance trainiert werden. Ebenso darf das Initial-Training, sofern es auf historischen Unternehmensdaten beruht, nicht auf veralteten Trainingsdaten basieren.

4.3 Mathematische Metriken zur Fairness

Zur Identifizierung und folglich Minimierung von algorithmischen Verzerrungen werden auch mathematische Ansätze genutzt. Diese können Indikatoren für Bias sein, müssen im Einzelnen jedoch überprüft werden, um keine Massnahmen gegen nicht vorhandene Bias zu unternehmen. Gleichzeitig

²⁸ [Flick, Uwe: Entzauberung der Intuition: systematische Perspektiven-Triangulation als Strategie der Gerlungsbegründung qualitativer Daten und Interpretationen, In: Hoffmeyer-Zlotnik, Jürgen H.P. \(Ed.\): Analyse verbaler Daten, Essen 1992](#)

können einzelne Indikatoren und Metriken sich gegenseitig aufheben oder behindern. Daher gilt es in jedem Fall die geeigneten Fairness-Metriken auszuwählen.

Statistische Parität

Eines der Beispiele für einen solchen Indikator wäre die statistische (auch demographische) Parität. Diese gibt vor, dass ein Modell dann als fair bezeichnet werden kann, wenn die Wahrscheinlichkeit einer positiven Entscheidung (wie die Zusage für einen Job) für alle demographischen Gruppen gleich ist. Dies wäre ein grundlegender Indikator, um die Fairness in den Ausgaben einer KI-Anwendung zu bewerten. Jedoch ist das Ziel der statistischen Parität die absolute Gleichheit in Zahlen und ignoriert zu Gunsten der statistischen Ausgleichlichkeit alle anderen Parameter.

Beispiel

Ein Unternehmen stellt fest, dass nach Prüfung von Bewerbungsunterlagen durch eine KI merklich mehr männliche Kandidaten zu Bewerbungsgesprächen eingeladen wurden als weibliche. Diese Zahlen werden mit der Summe der eingehenden Bewerbungen je Geschlecht verglichen und es wird festgestellt, dass von allen männlichen Bewerbern 80% davon zu Gesprächen eingeladen werden, bei Frauen jedoch nur 20% aller Bewerberinnen. Durch die statistische Parität kann man also ermitteln, dass die unterschiedlichen Verhältnisse nicht auf die Menge an tatsächlichen Bewerber:innen eingegangen sind - es also nicht an einer Quote an Bewerber:innen von 80% Männern zu 20% Frauen liegt. Gleichzeitig gibt die statistische Parität aber keine Auskunft über Qualifikationen oder ähnliches.

Es handelt sich hierbei um ein begrenztes Werkzeug zur Ermittlung von Verzerrungen im Bewerbungsprozess, trotzdem aber um ein hilfreiches Mittel um zeitnah einen ersten Überblick über die Verhältnisse zu erlangen, worauf man die Ursache untersuchen kann.

Disparate Impact Ratio²⁹

Hierbei handelt es sich um ein Mass zur Analyse und Bewertung von Entscheidungsprozessen. Der Disparate Impact wird durch den Vergleich einer Referenzgruppe (der im jeweiligen Fall privilegierten Gruppe) mit einer oder mehreren überwachten Gruppen (im jeweiligen Fall nicht privilegierte Gruppe/n) ermittelt. Während die statistische Parität ein Mittel für absolute Gleichstellung der Gruppen darstellt, ist die Disparate Impact Ratio eine Prüfung auf unverhältnismässige Benachteiligung durch Verhältnissberechnung. Es gilt herauszufinden, ob Mitgliedern der jeweils überwachten Gruppen seltener, gleich oft oder häufiger positive Ergebnisse zuteil werden. Definiert man also ein (positives) Ergebnis und die jeweiligen Gruppen, erhält man bei der Anwendung dieses Masses einen Wert bezüglich der Fairness einer KI-Anwendung in diesem Fall.

Im Gegensatz zur statistischen Parität ist eine absolute Angleichung der Werte hier auch kein Ziel, dafür hat man die 80-Prozent-Regel eingeführt. Man dividiert die ermittelte prozentuale Häufigkeit der überwachten Gruppe durch den Wert der prozentualen Häufigkeit der Referenzgruppe. Umgangssprachlich formuliert teilt man die Häufigkeit der nicht-privilegierten Gruppe durch die Häufigkeit

²⁹ [Unterschiedliche Auswirkungen auf Watson OpenScale - Fairnessmetriken, in: IBM, 07.10.2024](#)

der privilegierten Gruppe. In einem Beispiel wird dies im Folgenden verdeutlicht. Der Wert, der daraus entsteht, darf nach der 80-Prozent-Regel die besagten 80% nicht unterschreiten, da sonst ein potenziell diskriminierender Effekt eintreten würde.

Beispiel

Mietanträge auf Wohnungen werden in einem Unternehmen von einer KI überprüft. Das Unternehmen wertet die Statistiken dieser Überprüfungen aus und stellt fest: 80% der Bewerber:innen ohne Migrationshintergrund erhalten eine Zusage für eine Wohnung. Bei Bewerber:innen mit Migrationshintergrund liegt die Häufigkeit bei 52%. Dem Disparate Impact Ratio nach werden die 52% jetzt durch die 80% dividiert und man erhält 0,65, also 65%. Dieser Wert liegt also niedriger als die in dieser Methode vorgegebene Grenze von 80% Abweichung. In diesem Fall sollte nun geprüft werden, weswegen diese Unterschiede von dem KI-Modell gemacht wurden und wie sie zu beheben sind.

Während die statistische Parität auf Gleichverteilung abzielt, prüft die Disparate Impact Ratio, ob bestehende Unterschiede noch fair und vertretbar sind oder eine Verzerrung vorliegt, welcher man nachgehen sollte.

*Continuous Fairness Algorithm*³⁰

Das Beispiel der statistischen Parität bietet eine grundlegende, aber auch eine starre Lösung für mehr Gleichbehandlung an. Der **Continuous Fairness Algorithm** (CFA θ) unterscheidet sich hier durch eine dynamische und anpassbare Lösung, in der auch Fairness anpassbar und dynamisch definiert werden kann, um dem jeweiligen Kontext gerecht zu werden.

Ein weiterer Unterschied zur statistischen Parität ist die Fähigkeit des Algorithmus bereits während des Trainings zu wirken und situationsspezifisch gestaltet werden zu können. Auch ist der Algorithmus dazu in der Lage, zwischen individueller und gruppenbasierter Fairness Anpassungen vorzunehmen und gibt dem Anwender die Möglichkeit durch den Parameter θ den Grad der gewünschten Fairness einzustellen. Unbedingte und vollständige Gleichbehandlung ist nicht die Antwort auf jede Frage zur Eliminierung von Diskriminierung und sozialer Ungleichheit. In Fällen, in welchen eine differenzierte Herangehensweise notwendig ist, kann der Algorithmus aufgrund der Unterscheidung zwischen individuellen und gruppenbasierten Merkmalen, unterstützen. Beispiele hierfür wären die personalisierte medizinische Diagnostik und Kreditvergaben.

Beispiel

Eine Universität entwickelt zur Vergabe von Plätzen von Studiengängen und -programmen mit begrenzten Teilnehmerzahlen eine Software, welche auf Künstlicher Intelligenz basiert. Ausschlaggebende Kriterien hierbei sollen Schulnoten, das außerschulische Engagement und auch das Motivationsschreiben sein. Die Lehrkräfte der Universität teilen den Entwicklern der Software mit, dass berücksichtigt werden muss, dass Schüler aus sozioökonomisch benachteiligten Regionen nicht selten einen leicht schlechteren Durchschnitt bei den Noten vorzuweisen haben. Dies läge aber nicht an der generellen Eignung oder weniger erbrachten Leistung, sondern an strukturellen

³⁰ [Zehlike, Meike u.a.: Matching Code and Law: Achieving algorithmic Fairness with optimal Transport, in: Cornell University, 24.09.2019](#)

Gründen, wie weniger persönliche und systematische Förderung, an grösseren Klassen, langsames Vermitteln des Stoffes aufgrund von Verständigungsproblemen, etc.

Vor dem Training des einzusetzenden Modells wird von der Universität entschieden und vorgegeben, dass ein Continuous Fairness Algorithm eingesetzt werden soll, welcher nicht als Korrekturmassnahme von Verzerrungen, sondern als Teil der Modellarchitektur fungieren soll. Der Schieberegler θ kann im Algorithmus von 0 bis 1 gestellt werden. Um die strukturellen Nachteile auszugleichen wird entschieden, den Wert des Reglers θ auf 0,3 festzusetzen. Bei 0 würde rein die Leistung in Zahlen zählen, bei 1 würde der Algorithmus Gruppenunterschiede wie Migrationshintergrund komplett ausgleichen. Bei 0,3 steht die Leistung weiterhin stark vertreten im Zentrum der Auswahl, es werden geringfügige Ungleichheiten dennoch ausgeglichen. Hat ein Schüler aus einer bildungsprivilegierten Region also einen Notendurchschnitt von 1,2 und ein Schüler aus einer bildungsbenachteiligten Region einen Schnitt von 1,4 führt das Modell eine Fairnessangleichung durch und erkennt die beiden Durchschnitte als gleichwertig an.

CFA θ soll schon bei der Entwicklung einer Künstlichen Intelligenz eingesetzt werden und strukturelle Ungleichheiten erkennen und ausgleichen, bevor eben diese sich in den Ergebnissen über Wochen, Monate oder gar Jahre hinweg niederschlagen.

🌟 Vergleichstabelle: Fairness-Methoden in der KI-Anwendung

Kriterium	Statistische Parität	Disparate Impact Ratio (DIR)	Continuous Fairness Algorithm (CFA θ)
Ziel	Gleiche Erfolgswahrscheinlichkeit für alle Gruppen	Kein zu starkes Missverhältnis zwischen Gruppen	Flexible Balance zwischen Fairness und Leistung
Typische Frage	„Haben Männer und Frauen gleich oft eine Chance?“	„Wird eine Gruppe auffällig benachteiligt im Verhältnis zur anderen?“	„Wie viel Fairness-Korrektur wollen wir im Modell aktiv einbauen?“
Messung / Ansatz	Vergleich der Erfolgsraten zwischen Gruppen – absolute Gleichheit	Verhältnis der Erfolgsraten – darf i. d. R. nicht $< 0,8$ liegen (80%-Regel)	Fairness-Regler (θ) zwischen 0 (nur Leistung) und 1 (maximale Fairness)
Anwendungszeitpunkt	Reaktiv – prüft Ergebnisse nach Anwendung der KI	Reaktiv – prüft Ergebnisse nachträglich auf indirekte Diskriminierung	Präventiv und reaktiv einsetzbar – kann schon beim Training genutzt werden
Beispiel	60% Männer & 60% Frauen eingeladen zum Gespräch	80% Männer & 56% Frauen \rightarrow DIR = 0,7 (möglicher Bias)	$\theta = 0,4$ gleicht kleine Leistungsunterschiede mit Blick auf Fairness aus
Vorteile	Klar & leicht verständlich, gute Vergleichbarkeit	Flexibel, toleriert gewisse Unterschiede ohne Zwang zur Gleichheit	Fein abstimmbare, ideal für komplexe Systeme & faire Optimierung von Anfang an
Nachteile / Grenzen	Ignoriert individuelle Unterschiede & Leistung	Erkennt nur starke Benachteiligung, keine gezielte Fairnesssteuerung	Komplexer in der Anwendung & Erklärung – setzt bewusste Fairnessstrategie voraus
Geeignet für ...	Öffentliche Auswahlverfahren, Quotenregelungen	Arbeitsrechtliche Prüfungen, Gleichstellungsanalyse	Frühzeitige KI-Entwicklung, sensible Entscheidungsprozesse (Kredite, Zulassungen)

Vergleichstabelle der hier vorgestellten mathematischen Fairness-Methoden, erstellt durch ChatGPT 4o am 26.03.25

4.4 Transparenz, Recht, Regulierung & Ethik

Weitere Ansätze zur Vermeidung und Reduzierung von Bias durch die Nutzung von künstlicher Intelligenz sind Transparenz, rechtliche Rahmenbedingungen, Regulierung und ethische Prinzipien. Darunter fallen neu verabschiedete Gesetze wie der **EU-AI-Act** oder Corporate Guidelines.

Transparenz

Transparenz ist ein nicht zu unterschätzendes Mittel für Bias-Prävention. Dies beginnt bei der **Offenlegung** der Datenquellen, damit diese untersucht werden können - wie repräsentativ sind sie? Woher stammen sie? Warum wurden genau diese ausgewählt?

Nach der Offenlegung der verwendeten Trainingsdaten lässt sich auch die **Erklärbarkeit** der Entscheidungsfindung einer KI nachvollziehen. Mit den nun vorhandenen Informationen kann man also

erklären, welche Merkmale die Ausgaben einer KI am stärksten beeinflussen und welche durch die Modellierung oder die Trainingsdaten beeinflusst werden.

Hat man die Sachlage um die Trainingsdaten und die Merkmale um die Entscheidungsfindung durch Modellierung oder Trainingsdaten offengelegt und erklärt, besteht eine **Nachvollziehbarkeit** von Anpassungen in der Modellierung oder in der Auswahl der Trainingsdaten. Wurden also erkannte Ursachen für Verzerrungen behoben und wenn ja, wie?

Für eine Erhöhung der Transparenz wurden bereits Methoden entwickelt, so wie die **erklärbare künstliche Intelligenz**³¹ (**Explainable AI, XAI**) und die **Artificial Intelligence Impact Assessments (AIIA)**³². **XAI** steht für eine Zusammenstellung von Prozessen und Methoden, welche es menschlichen Nutzer:innen ermöglicht, die vom Algorithmus einer KI generierten Ausgaben zu verstehen und damit gegebenenfalls auch zu vertrauen. Beim **AIIA** handelt es sich um ein Bewertungsverfahren, welches die potenziellen Risiken und Auswirkungen einer KI auf die Gesellschaft, die Ethik und Menschenrechte analysiert. Ziel dabei ist es, Fairness, Transparenz und Verantwortung in den Ausgaben einer KI zu gewährleisten. Dieses Verfahren funktioniert wie folgt:

1. Risikoidentifikation: Gibt es benachteiligte Gruppen oder Individuen durch die Ausgaben der KI und gibt es in den Trainingsdaten einem zugrunde liegenden Bias?
2. Transparenzprüfung: Kann nachvollzogen werden, wie die KI zu einer bestimmten Ausgabe kommt?
3. Bewertung nach Recht und Ethik: Werden für die KI greifenden Gesetze wie der **EU-AI-Act**, Datenschutz-, Menschenrechtsgesetze usw. eingehalten?
4. Strategieentwicklung: Welche Massnahmen können ergriffen und umgesetzt werden, um die Ausgaben der KI fairer und befreit von Diskriminierung zu gestalten?

In den Fällen **XAI** und **AIIA** handelt es sich um implementierte präventive Massnahmen, sogenanntes **Fairness by Design**^{33,34}. Damit ist gemeint, dass Fairness nicht erst nachträglich geprüft, sondern bereits in der Planung, Entwicklung und Modellierung von Künstlicher Intelligenz gezielt mitgedacht und systematisch eingebaut wird. Das Ziel hierbei ist es, verzerrte oder diskriminierende Ergebnisse bestenfalls gar nicht erst entstehen zu lassen. Im Falle der **XAI** sollen menschliche Nutzer:innen und im Falle der **AIIA** die Entwickler bzw. Hersteller und Betreiber einer KI die Ausgaben einer KI nachvollziehen können.

Während KI-Modelle Ausgaben generieren und automatisierte Entscheidungen treffen, lässt sich eine weitere Kontrollinstanz einsetzen, welche man **Human-in-the-Loop** nennt. Dabei wird die Kontrolle über die Ausgaben gerade in kritischen Phasen auf Menschen übertragen, um keine rein automatisierte Ausgaben in Umlauf zu bringen. Diese Phasen können beim Training, bei der Validierung, beim Testen oder bei der Entscheidungsfindung im aktiven Betrieb auftreten. In Fällen wie medizinische Auswertungen, Entscheidungen im Personalwesen und bei der Kreditvergabe ist diese Methode mittlerweile gängige Praxis. Gemäss Art. 22 DSGVO hat eine betroffene Person das Recht, nicht einer ausschliesslich automatisierten Entscheidung unterworfen zu werden. Die betroffene Person kann

31 [What is Explainable AI?, in: IBM, 29.03.2023](#)

32 [Frijters, Daniël u.a.: Artificial Intelligence Impact Assessment, in: ECP Platform for the Information Society, 2019](#)

33 [Abbassi, Ahmed u.a.: Make "Fairness by Design" Part of Machine Learning, in: Harvard University, 01.08.2018](#)

34 Eine vertiefende Auseinandersetzung mit Fairness by Design folgt in einem separaten Kurzpapier.

verlangen, dass eine automatisierte Entscheidung durch einen Menschen überprüft wird. Somit spiegelt sich die Kontrollinstanz Human-in-the-Loop damit in der Gesetzeslage des Datenschutzes wieder.

Dass eine menschliche Kontrollinstanz sinnvoll, aber auch notwendig sein kann, zeigt sich an Beispielen wie der Kreditvergabe. Nutzen beispielsweise Kreditinstitute Ersatzinformationen (Proxies), ist man als Kreditinstitut und Anwender dieser KI-Anwendung in der Pflicht, die Daten gebunden an den tatsächlichen Zweck zu nutzen und deren Richtigkeit zu überprüfen. So gibt es in den USA den Begriff „Redlining“.³⁵ Demzufolge wurden Antragsteller auf Kredite aufgrund von Proxies der Zugang zu Krediten verwehrt. Jedoch nicht wegen des individuellen Risikos, sondern aufgrund der Gegend, aus der sie stammen und in der sie wohnen. So gaben Kreditinstitute die Adressen der Antragsteller mit in den zu verwertenden Datensatz mit ein, worauf hin das automatisierte Entscheidungsprogramm dann die Adresse mit anderen Antragstellern verglich. Das Ergebnis war, dass Gegenden, welche von Minderheiten bewohnt wurden, aufgrund ihrer Wohnumgebung negativ bewertet wurden, da in eben diesen Gegenden das Einkommen im Vergleich zu anderen geringer ausfiel. Damit wurden Kriterien wie gesellschaftliche Faktoren auf Kosten der individuellen Eignung ungerechtfertigt gestärkt - oder überhaupt erst in die Wertung mit aufgenommen.

Ethik & Verantwortung

Ethische und durch Verantwortung geleitete Vorgaben sind für den Betrieb einer fairen künstlichen Intelligenz entscheidend. Dazu können Hersteller und Betreiber künstlicher Intelligenzen eigene Vorgaben entwickeln oder aber auf bestehende ethische KI-Leitlinien zurückgreifen. Beispiele für letztere wären Richtlinien, welche die OECD, IEEE oder die hochrangige Expertengruppe für Künstliche Intelligenz der EU³⁶ veröffentlicht haben.

Regulierung & rechtliche Rahmenbedingungen

Hersteller und Betreiber können sich durch selbst initiierte Massnahmen regulieren, seien dies nun Fairness-Metriken, aber auch Bias-Audits, Open-Source-Lösungen und Dokumentation. Das Ergreifen dieser Massnahmen ist erst einmal keine Regulierung an sich, aber die durch diese erzielten Kenntnisse können zu Regulierungen führen. Dazu muss in den Vordergrund gestellt werden, dass Regulierungen, trotz ihres negativen Rufs in ihrer Natur nichts schlechtes sind. Im Falle der Verzerrungen durch künstliche Intelligenzen können Regulierungen derer Nutzung positive Effekte für Hersteller und Betreiber von KI-Anwendungen haben, etwa durch ein gestärktes Image und grösseres Vertrauen aus der Gesellschaft heraus.

Um zu verhindern, dass sich künstliche Intelligenz zu einem rechts- und sicherheitsfreien Raum entwickelt, werden gesetzliche Auflagen wie Regulierungen benötigt. Dabei haben beispielsweise die EU (EU-AI-Act, DSGVO) und die USA Massnahmen ergriffen (Gesetzesentwurf des US Algorithmic Accountability Act). Die bereits genannten Massnahmen und Strategien zur Bekämpfung von Bias stützen sich zu grossen Teilen auf Freiwilligkeit der Implementierung, während gesetzliche Vorgaben eine einheitliche Landschaft zur Entwicklung für alle schaffen.

35 [Rose, Jonathan u.a.: Redlining, in: Federal Reserve History, 02.06.2023](#)

36 [Ethikleitlinien für vertrauenswürdige KI, in: digital strategy EU, 08.04.2019](#)

Regulierung: Vergleich der Systeme EU, USA und China

EU

Die EU geht einen umfassenden und rechtsbasierten Weg in Bezug auf die KI-Regulierung. Am 1. August 2024 trat die **europäische Verordnung über künstliche Intelligenz (KI-Verordnung)** in Kraft, welches auch das weltweit erste umgesetzte KI-Gesetz darstellt. Wie mit der **DSGVO** wurde mit dem KI-Gesetz ein einheitlicher und zentralisierter Rahmen für Anforderungen festgelegt, welche Produkthersteller, Anbieter, Händler und Betreiber von künstlichen Intelligenzen zu befolgen haben. Die Anforderungen werden durch verschiedene Parameter definiert, der zentrale Parameter ist hier jedoch die Risikobewertung einer KI.

In folgende Risikostufen werden KI-Systeme eingeordnet (zitiert aus einer Veröffentlichung des EU-Parlaments:³⁷

- **Minimales Risiko:** Die meisten KI-Systeme, z. B. Spamfilter und KI-gestützte Videospiele, unterliegen keinen besonderen Verpflichtungen, doch Unternehmen können freiwillig zusätzliche Verhaltenskodizes aufstellen.
- **Besondere Transparenzverpflichtungen:** Systeme wie Chatbots müssen ihre Nutzer:innen deutlich darauf hinweisen, dass sie es mit einer Maschine zu tun haben, und durch KI erzeugte Inhalte müssen als solche gekennzeichnet werden.
- **Hohes Risiko:** Für KI-Systeme, die als hochriskant eingestuft werden (z. B. KI-basierte medizinische Software oder KI-Systeme für die Personalauswahl), gelten strenge Anforderungen, z. B. im Hinblick auf Risikominderungssysteme, hochwertige Datensätze, klare Informationen für die Nutzer:innen, menschliche Aufsicht usw.
- **Unannehmbares Risiko:** KI-Systeme, von denen eine klare Bedrohung für die Grundrechte der Menschen ausgeht, sind verboten. Dies gilt z. B. für Systeme, die Behörden oder Unternehmen eine Bewertung des sozialen Verhaltens ermöglichen (*Social Scoring*).

Zusätzlich hat die EU-Kommission eigenen Angaben nach „eine Konsultation zu einem **Verhaltenskodex für Anbieter von KI-Modellen mit allgemeinem Verwendungszweck (GPAI)** eingeleitet“³⁸. Dieser Verhaltenskodex, welcher auch in der **KI-Verordnung** nach Art. 56 KI-VO vorgesehen ist, soll „kritische Bereiche wie Transparenz, Urheberrecht und Risikomanagement abdecken“ und Entwickler bzw. Anbieter von KI-Modellen in die Pflicht nehmen.

USA

Die USA folgen einem stärker vom Markt getriebenen Ansatz. Auf Bundesstaaten-Ebene gibt es vereinzelt Gesetzgebungen wie den **Generative Artificial Intelligence Accountability Act** in Kalifornien³⁹. Jedoch gibt es derzeit keine einheitliche föderale Gesetzgebung, welche speziell KI betrifft und diese zum Schutz der Bürger und der Öffentlichkeit regulieren soll. Dennoch gibt es Vorschläge wie den **Blueprint for an AI Bill of Rights**⁴⁰ oder den **Artificial Intelligence (AI) Accountability Act**⁴¹ auf Bundesebene. Die **AI Bill of Rights** hat dabei ihren

37 [EU Kommission: KI-Verordnung tritt in Kraft, 01.08.2024](#)

38 [EU Kommission: KI-Verordnung tritt in Kraft, 01.08.2024](#)

39 [Generative Artificial Intelligence Accountability Act, in: California Legislative Information, 29.09.2024](#)

40 [Blueprint for an AI Bill of Rights, in: White House Archives, 04.10.2022](#)

41 [Artificial Intelligence Accountability Act, in: congress.gov, 25.10.2023](#)

Fokus auf den rechtlichen Rahmen zwischen künstlicher Intelligenz und den Bürgern, bzw. deren Schutz vor Missbrauch und Benachteiligung, ähnlich der EU **KI-Verordnung**. Der **AI Accountability Act** befasst sich wie der noch zu veröffentlichende KI-Verhaltenskodex der EU mit der Verantwortung von Herstellern und Betreibern.

Beide Vorschläge wurden während der Biden-Administration formuliert und hervorgebracht, welche einer Regulierung von Technologien wie künstlicher Intelligenz offener gegenüber stand, als dies nun im Falle der Trump-Administration der Fall zu sein scheint. Durch die Aufhebung⁴² bereits kleinerer umgesetzter oder geplanter Regulierungen und die dafür formulierte Begründung steht für die Trump-Administration die Position als Marktführer im Bereich KI-Technologie im Zentrum des Interesses. Aufgrund der aktuellen Entwicklungen und Mitteilungen des Weissen Hauses ist derzeit nicht von einer mit der EU vergleichbaren Regulierung von künstlicher Intelligenz in den USA auszugehen.

China

Die chinesische Regierung verfolgt einen stärker vom Staat kontrollierten Ansatz, welcher die Innovationen im Bereich künstlicher Intelligenz enger mit staatlicher Kontrolle und Aufsicht verbindet. Hierzu hat die Regierung Richtlinien eingeführt, welche sicherstellen sollen, dass die Entwicklung von KI keinen Kontrast zu den sozialistischen Grundwerten Chinas bilden, sondern viel mehr mit diesen konform gehen. Dies bedeutet aber nicht, dass China die Entwicklung von künstlicher Intelligenz ausbremsen will. Die chinesische Regierung sieht sich im direkten technologischen Wettstreit mit den USA um die Position des Marktführers, auch im Bereich KI, wie die Veröffentlichung der chinesischen KI Deepseek⁴³ und das Verfassen des **Next Generation Artificial Intelligence Development Plan**⁴⁴, mit dem erklärten Ziel bis 2030 die weltweit führende KI-Nation zu sein, deutlich gemacht haben.

China hat zudem Massnahmen in Form des **Data Security Law**⁴⁵ ergriffen. Diese sollen umfassend Datenverarbeitungen regulieren, Datensicherheit gewährleisten, Entwicklung und Nutzung von Daten fördern, die Rechte und Interessen von Individuen und Organisationen schützen, jedoch auch die nationale Souveränität und nationalen Interessen.

Stärken und Schwächen der Ansätze

Die unterschiedlichen Ansätze spiegeln die verschiedenen Ansätze, Werte und Ansichten der drei Systeme in politischer, wirtschaftlicher und gesellschaftlicher Hinsicht. Der Ansatz der EU



Erstellt am 19.03.25 mit DALL-E mit der Aufforderung, anhand der Beschreibung der drei Systeme in diesem Paper eine Karikatur mit Tauziehen zu erstellen.

42 [Removing Barriers to American Leadership in Artificial Intelligence](#), in: the White House, 23.01.2025

43 [Kesselheim, Stefan u.a.: "Deepseek hat extrem vorgelegt" - Chinas neues KI-Modell und seine Bedeutung für die Tech-Welt](#), Interview, in: Jülich Forschungszentrum, 07.02.2025

44 [China Science & Technology Newsletter](#), in: China Embassy, 15.09.2017

45 [Kawase, Toki u.a.: Was ist das chinesische 'Data Security Law'? Erklärung der Massnahmen, die japanische Unternehmen ergreifen sollten](#), in: monolith.law, 15.04.2024

setzt mit dem Fokus auf den Schutz der Rechte der Bürger auf Prävention und breit angelegte rechtliche Rahmenbedingungen zum Umgang und zur Regulierung hinsichtlich künstlicher Intelligenz. Gerade im Vergleich zum Ansatz der USA kann dies durch den zusätzlichen Aufwand und die Beschränkungen innovationshemmend wirken. Der Ansatz der USA bietet den Herstellern und Betreibern von Künstlicher Intelligenz mehr Spielraum und will die Kontrolle und Regulierung von KI-Modellen und die Wahrung der Bürgerrechte zugunsten von Innovation und Erhaltung der Marktführung in Grenzen halten. Der Staat soll sich weitestgehend zurückhalten. Konträr dazu läuft der chinesische Ansatz. Dieser verfolgt eine stark staatlich kontrollierte Strategie und stellt die Souveränität und Innovationskraft Chinas gegebenenfalls über individuelle Freiheit, Persönlichkeitsrechte und Datenschutz, da die Regierung mit entsprechenden politischen Mitteln im Zweifelsfall letztere zugunsten der eigenen Interessen herabsetzen kann.

5 Beispiele: Wie wird tatsächlich gegen Bias vorgegangen?

Algorithmic Justice League (AJL)

Die bereits erwähnte Problematik mit der Gesichtserkennung-Software wurde durch das „Gender-Shades-Projekt“⁴⁶ von Joy Buolamwini, einer Forscherin aus dem MIT Media Lab, publik. Das Projekt war von 2017 bis 2020 aktiv und führte zu Verbesserungen bei der Entwicklung von Gesichtserkennung-Software.

Eines der betroffenen Unternehmen, IBM, veröffentlichte in der Folge einen neu entwickelten Datensatz⁴⁷, namentlich den „Diversity in Faces“-Datensatz zur Verbesserung von Fairness in Gesichtserkennung-Softwares. Die Veröffentlichung war Teil einer Transparenz-Kampagne, um das Vertrauen in die Software wieder herzustellen und zu vermitteln, da es sich um eine breite Repräsentation mit verschiedenen demographischen Merkmalen handelt. Auch Microsoft begann wie IBM damit, die Datensätze zu verbessern und transparenter zu arbeiten⁴⁸.

Des Weiteren gaben beide Unternehmen bekannt, dass eine erhöhte Sensibilisierung gegenüber algorithmengenerierten Ausgaben erreicht wurde und Strategien entwickelt wurden für interne Tests mit dem Fokus auf Verzerrungen und Diskriminierung.

D-BIAS - Menschliche Korrektur von durch KI generierten Verzerrungen⁴⁹

Die Studie „D-BIAS: A Causality-Based Human-in-the-Loop System for Tackling Algorithmic Bias“ (2022) stellt eine Kombination aus kausaler Modellierung einer KI und einem Human-in-the-Loop Ansatz dar. Hierbei sollen menschliche Kontrolleure Variablen festlegen, welche tatsächlich kausal und relevant sind, und falsch hergeleitete Korrelationen korrigieren. Somit sollen zufällige oder verzerrte Zu-

46 [Buolamwini, Joy u.a., Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification, in: proceedings.mlr, 15.01.2018](#)

47 [Smith, John: IBM Research releases 'Diversity in Faces' dataset to advance study of Fairness in facial Recognition Systems, in: IBM, 15.02.2019](#)

48 [Roach, John: Microsoft improves facial Recognition Technology to perform well across all Skin Tones, Genders, in: Microsoft Blogs, 26.06.2018](#)

49 [Ghai, Bhavya u.a.: D-BIAS: A Casuality-based Human-in-the-Loop System for tackling algorithmic Bias, in: arxiv.org, IEEE, 10.08.2022](#)

sammenhänge in Sachverhalten verhindert werden. Anwendung fand die Methode beispielsweise in einer Auswahl von Bewerber:innen, als menschliche Kontrolleure das Modell so anpassen konnten, dass das Modell in der Folge faire und diskriminierungsfreie Ergebnisse lieferte.

Das Ergebnis dieser Methode ist, dass sich die Kombination aus algorithmischen und statistischen Methoden und hinzugezogener menschlicher Kontrolle als effektives Mittel etablieren konnte.

Geschlechtsbias in KI-Modellen

Wie Bias bezüglich der Geschlechter auftreten und welche Massnahmen dagegen ergriffen werden können, behandelte die Studie „Does Gender Matter? Towards Fairness in Dialogue Systems“⁵⁰ aus dem Jahr 2019. Bei der Untersuchung von KI-Sprachmodellen konnte die Studie belegen, dass die Modelle verschieden auf Geschlechter reagierten und ihre Ausgaben anpassten. Zum Beispiel wurden weibliche kodierte Begriffe öfter mit fürsorglicher Sprache assoziiert, während ihre männlichen Pendanten eher mit Autorität und technischer Kompetenz in Verbindung gebracht wurden.

In der Folge wurden Massnahmen zu mehr Fairness entwickelt. Dabei wurde die Gewichtung von Geschlechtern und mit ihnen in Verbindung stehende Assoziationen in den Trainingsdaten angepasst und mit adversialem Lernen gearbeitet, um die Ergebnisse in der Folge kontinuierlich zu prüfen.

Adversiales Lernen

Dabei handelt es sich um eine Lernmethode, in welcher eine „böswillige Eingabe“ in ein KI-Modell simuliert wird. Es werden absichtliche Verzerrungen und diskriminierende Parameter in eine Aufforderung an ein KI-Modell verpackt, was der Aufdeckung von Schwachstellen in einem System dient.

6 Abschluss

Es ist eine umfassende Herausforderung, welcher man sich durch den Einsatz Künstlicher Intelligenz zu stellen hat, nicht nur technischer, sondern auch ethischer und gesellschaftlicher Natur. Die Verbreitung und Integration von KI im Alltag und im Berufsleben nimmt rasant zu. Jedoch gilt es, sich nicht nur aus ethischer, sondern unter anderem auch aus rechtlicher Sicht, vor Augen zu halten, dass der Mensch weiterhin nichts abgibt: Im Falle der Kontrolle darf er es nicht und im Falle der Verantwortung kann er es nicht. So liegt die Verantwortung bei der Entwicklung beim Hersteller, beim Betrieb beim Betreiber und bei der Nutzung auch bei Nutzer:innen. Da es sich nach wie vor um ein menschliches Produkt handelt, wird auch der Mensch für dessen Versagen oder Fehler zur Rechenschaft gezogen - durch rechtliche Konsequenzen, Image-schäden oder anderen Problemen und Folgeschäden.

Es gilt zu verstehen, dass Künstliche Intelligenz ein Werkzeug und kein selbstständig agierender Akteur darstellt. Und ob dieses Werkzeug Verzerrungen wie soziale Ungleichheiten und Diskriminierung übernimmt, reproduziert und verstärkt hängt von den menschlichen Faktoren ab: Wollen wir das? Wie gestalten wir Gegenmassnahmen? Wie trainieren wir KI initial frei von Bias? Wie setzen wir sie ein, ohne dass daraus strukturelle Probleme weiter getragen werden?

Um eine KI so zu entwickeln, dass sie unseren Massstäben nach ethisch vertretbar und verantwortungsvoll funktioniert, braucht es auf menschlicher Seite eine ganzheitliche Perspektive, in der technische, rechtliche und gesellschaftliche Faktoren gleichermaßen mit einfließen. Nach Unternehmen

50 [Liu, Haochen u.a.: Does Gender Matter? Towards Fairness in Dialogue Systems, 2020, in: ACLAnthology, 13.08.2020](#)

und Einzelpersonen sind daher auch Regierungen und internationale Organisationen gefragt, klar definierte gesetzliche Rahmenbedingungen zu schaffen und faire KI-Standards zu etablieren. Nur durch einen solchen Rahmen sind die Grundsteine für eine dauerhafte Lösung zur Bekämpfung von Bias, welche in Diskriminierung und einseitiger gesellschaftlicher Machtkonzentration mündet, gelegt.

Es wurden drei nationale (USA, China) und supranationale (EU) Systeme diskutiert, welche einen unterschiedlichen Ansatz in Bezug auf Künstliche Intelligenz verfolgen. Jedes System hat für sich ein Mass in den Punkten Regulierung, Kontrolle und Innovation gefunden. Aus Sicht Europas gilt es hier noch einmal zu erwähnen, dass sich Akteure, die eine KI ausserhalb Europas entwickeln oder trainieren lassen wollen, sich im Vorfeld über die Implikationen dieser Entscheidung bewusst machen müssen. Diese Entscheidung ist an sich nichts schlechtes und ist neutral zu betrachten. Jedoch können durch die sich stark unterscheidenden Gesetzgebungen und Regulierungsmassnahmen sowie dem unterschiedlichen Umgang mit Diskriminierung Konflikte mit dem europäischen Standort ergeben.

Darüber hinaus ist die gesellschaftliche Verantwortung in dieser Frage ebenso wichtig. Eine breite Sensibilisierung gegenüber Künstlicher Intelligenz und algorithmischen Verzerrungen kann zu einem öffentlichen Diskurs beitragen, welcher die Entwicklung von Künstlicher Intelligenz mitbestimmt. Dafür entscheidend ist, dass KI nicht als „Black Box“ wahrgenommen wird, sondern als Mechanismus, der verstanden und beeinflusst werden kann. Dazu tragen nachvollziehbare und rechenschaftspflichtige Massnahmen bei.

In diesem Zusammenhang ist es essenziell zu begreifen, dass Diskurse und Meinungsbildung nie im luftleeren Raum entstehen. Die soziologische Forschung zeigt klar: Unsere Wahrnehmungen und Überzeugungen sind immer auch Ergebnis sozialer Strukturen und Kontexte. Wenn Algorithmen beginnen, Informationen zu gewichten, sichtbar zu machen oder auszuschließen, dann gestalten sie diesen sozialen Kontext aktiv mit - etwa durch Filterblasen, algorithmische Gewichtung oder emotionale Verstärkung. So beeinflussen sie, was wir wissen, wie wir dieses Wissen verarbeiten und wie wir zu Überzeugungen gelangen.

Auch wenn dies nicht der zentrale Fokus dieses Whitepapers war, so verdeutlicht dieser Aspekt, wie tiefgreifend algorithmische Systeme gesellschaftlich wirken. Institutionen, Unternehmen und Individuen sollten sich deshalb bewusst machen, dass jede Nutzung von KI auch auf Kommunikation, Diskurs und gesellschaftliches Miteinander zurückwirkt. Der verantwortungsvolle Umgang mit Künstlicher Intelligenz bedeutet daher nicht nur, technische Fehler zu vermeiden oder rechtlichen Anforderungen zu genügen - sondern auch, die Frage zu stellen, welchen Einfluss wir mit der Nutzung von Künstlicher Intelligenz auf die Gesellschaft als Ganzes nehmen möchten.

Die abschliessende Frage, welche nun im Raum steht, ist aber, wie man Künstliche Intelligenz so gestalten und nutzen kann, dass sie bestehende Verzerrungen und Diskriminierung nicht nur zu verhindern sucht, sondern ein aktives Werkzeug mit dem Ziel der Schaffung einer gerechteren Gesellschaft sein kann. Durch einen kontinuierlichen, reflektierten und ethischen Prozess kann eine Künstliche Intelligenz als Werkzeug des Fortschritts verwendet werden und keines der Ungleichheit.

6.1 Informationen zum White Paper

Zum Autor

Noch während seines Studiums der Politikwissenschaften und Soziologie begann bei Dominic Baumann das Interesse an Datenschutz und Künstlicher Intelligenz. Neben der politischen und gesellschaftlichen Relevanz des Datenschutzes und der Künstlichen Intelligenz kümmert sich Dominic Baumann als Consultant ganz konkret um die Umsetzung in Unternehmen und erstellt in diesen Themenbereichen die Lerninhalte der E-Learning-Plattform der integratio. 2024 erfolgte die TÜV Zertifizierung zum Datenschutzbeauftragten.

Über integratio

Als herstellerunabhängiges IT-Beratungsunternehmen mit Standorten in Zürich und Böblingen (Deutschland) verfügt integratio über länderübergreifendes Datenschutz-Know-How und fundiertes technisches und juristisches Fachwissen mit zertifizierten Datenschutz- und IT-Sicherheitsexperten, KI-Experten, erfahrenen Juristen und Branchenexpertise. Zum Kundenkreis gehören international agierende Konzerne wie auch kleinere und mittelständische Unternehmen verschiedener Branchen.

Erfahren Sie mehr

Wenn für Sie weitere Fragen bestehen, stehen wir Ihnen gerne zur Seite. Wir unterstützen Ihr Unternehmen bei der optimalen und rechtskonformen Gestaltung und Nutzung der Möglichkeiten von Künstlicher Intelligenz. Zudem führen wir Standortbestimmungen und Audits durch und schulen und sensibilisieren Ihre Mitarbeiter.



Dominic Baumann
Datenschutzexperte
Soziologie - Master of Arts
dominic.baumann@integratio.com



integratio GmbH
Börsenstrasse 18
8001 Zürich
+41 (0)44 321 72 00

Otto-Lilienthal-Str. 36
71034 Böblingen

E-Mail: info@integratio.com
www.integratio.com



Raimund Weiler
Key-Account-Manager
raimund.weiler@integratio.com

7 Literaturverzeichnis

<https://www.edge.org/response-detail/27117> (zuletzt aufgerufen am: 16.04.2025)

Tversky & Kahnemann - Cambridge University Press (1982), Judgement under Uncertainty: Heuristics and Biases

Wirtz (Hrsg.), Dorsch (2021) - Lexikon der Psychologie, 20. überarbeitete Auflage

Klimke, Lautmann, Stäheli, Weischer, Wienold (Hrsg.), Lexikon zur Soziologie 6. Auflage

<https://www.iks.fraunhofer.de/de/themen/kuenstliche-intelligenz.html> (zuletzt aufgerufen am: 16.04.2025)

<https://www.europarl.europa.eu/topics/de/article/20200827STO85804/was-ist-kunstliche-intelligenz-und-wie-wird-sie-genutzt> (zuletzt aufgerufen am: 16.04.2025)

<https://www.bidt.digital/glossar/maschinelles-lernen/> (zuletzt aufgerufen am: 16.04.2025)

<https://artificialintelligenceact.eu/de/article/3/> (zuletzt aufgerufen am: 16.04.2025)

<https://aws.amazon.com/de/what-is/overfitting/> (zuletzt aufgerufen am: 16.04.2025)

<https://datasolut.com/wiki/trainingsdaten-und-testdaten-machine-learning/> (zuletzt aufgerufen am: 16.04.2025)

<https://flexikon.doccheck.com/de/Bias> (zuletzt aufgerufen am: 16.04.2025)

<https://www.duden.de/rechtschreibung/systematisch> (zuletzt aufgerufen am: 16.04.2025)

<https://www.weforum.org/stories/2021/11/humans-cognitive-bias-mistake/> (zuletzt aufgerufen am: 16.04.2025)

<https://www.geo.de/wissen/framing-effekt--wenn-das-denken-manipuliert-wird-31778578.html> (zuletzt aufgerufen am: 16.04.2025)

<https://www.staatslexikon-online.de/Lexikon/Diskriminierung> (zuletzt aufgerufen am: 16.04.2025)

<https://www.heise.de/hintergrund/Neue-Tools-zeigen-wie-voreingenommen-KI-Bildgeneratoren-sind-7744035.html> (zuletzt aufgerufen am: 16.04.2025)

https://are.berkeley.edu/courses/EEP118/current/handouts/OVB%20versus%20Multicollinearity_eep118_sp15.pdf (zuletzt aufgerufen am: 16.04.2025)

<https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm> (zuletzt aufgerufen am: 16.04.2025)

https://www.statewatch.org/news/2025/april/uk-ministry-of-justice-secretly-developing-murder-prediction-system/#_ftn9 (zuletzt aufgerufen am: 16.04.2025)

<https://www.ibm.com/de-de/think/topics/algorithmic-bias> (zuletzt aufgerufen am: 16.04.2025)

Ludwig-Mayerhofer, Wolfgang: Soziale Ungleichheit, Kriminalität und Kriminalisierung, Wiesbaden 2000

<https://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf> (zuletzt aufgerufen am: 16.04.2025)

<https://algorithmwatch.org/en/viogen-algorithm-gender-violence/> (zuletzt aufgerufen am: 16.04.2025)

<https://www.reuters.com/article/world/insight-amazon-scrap-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK0AG/> (zuletzt aufgerufen am: 16.04.2025)

<https://www.fondsprofessionell.de/consent/?url=/news/recht/headline/kuenstliche-intelligenz-haftungsfragen-fuer-finanzbranche-222033/> (zuletzt aufgerufen am: 16.04.2025)

<https://dsgvo-gesetz.de/art-22-dsgvo/> (zuletzt aufgerufen am: 16.04.2025)

<https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32024R1689> (zuletzt aufgerufen am: 16.04.2025)

https://www.ewi-psy.fu-berlin.de/erziehungswissenschaft/arbeitsbereiche/qualitative_sozial_bildungsforschung/Medien/entzauberung_der_intuition.pdf (zuletzt aufgerufen am: 16.04.2025)

<https://dataplatform.cloud.ibm.com/docs/content/wsj/model/wos-disparate-impact.html?context=cpdaas> (zuletzt aufgerufen am: 16.04.2025)

<https://arxiv.org/pdf/1712.07924> (zuletzt aufgerufen am: 16.04.2025)

<https://www.ibm.com/de-de/topics/explainable-ai> (zuletzt aufgerufen am: 16.04.2025)

<https://static1.squarespace.com/static/5b7877457c9327fa97fef427/t/5c368c611ae6cf01ea0fba53/1547078768062/Artificial+Intelligence+Impact+Assessment+-+English.pdf> (zuletzt aufgerufen am: 16.04.2025)

<https://hbr.org/2018/08/make-fairness-by-design-part-of-machine-learning> (zuletzt aufgerufen am: 16.04.2025)

<https://www.federalreservehistory.org/essays/redlining> (zuletzt aufgerufen am: 16.04.2025)

<https://digital-strategy.ec.europa.eu/de/library/ethics-guidelines-trustworthy-ai> (zuletzt aufgerufen am: 16.04.2025)

https://commission.europa.eu/news/ai-act-enters-force-2024-08-01_de (zuletzt aufgerufen am: 16.04.2025)

https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=202320240SB896 (zuletzt aufgerufen am: 16.04.2025)

<https://bidenwhitehouse.archives.gov/ostp/ai-bill-of-rights/> (zuletzt aufgerufen am: 16.04.2025)

<https://www.congress.gov/bill/118th-congress/house-bill/3369> (zuletzt aufgerufen am: 16.04.2025)

<https://www.whitehouse.gov/presidential-actions/2025/01/removing-barriers-to-american-leadership-in-artificial-intelligence/> (zuletzt aufgerufen am: 16.04.2025)

<https://www.fz-juelich.de/de/aktuelles/news/pressemitteilungen/2025/deepseek-bedeutung-fuer-die-tech-welt> (zuletzt aufgerufen am: 16.04.2025)

<http://fi.china-embassy.gov.cn/eng/kxjs/201710/P020210628714286134479.pdf> (zuletzt aufgerufen am: 16.04.2025)

<https://monolith.law/de/general-corporate/china-data-security-law> (zuletzt aufgerufen am: 16.04.2025)

<https://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf> (zuletzt aufgerufen am: 16.04.2025)

<https://research.ibm.com/blog/diversity-in-faces> (zuletzt aufgerufen am: 16.04.2025)

<https://blogs.microsoft.com/ai/gender-skin-tone-facial-recognition-improvement/> (zuletzt aufgerufen am: 16.04.2025)

Bias durch Künstliche Intelligenz - soziale Ursprünge in algorithmischer Generierung

<https://arxiv.org/pdf/2208.05126> (zuletzt aufgerufen am: 16.04.2025)

<https://aclanthology.org/2020.coling-main.390.pdf> (zuletzt aufgerufen am: 16.04.2025)