Julian Anslinger / Jaroslava Huber / Michael Haslgrübler / Anita Thaler (Hrsg.)

Verantwortungsvolle Einbindung von KI-Assistenzsystemen am Arbeitsplatz

Ein Handbuch für Arbeitnehmende und ihre Vetretungen

Mit einem Gastbeitrag von Susanne Haslinger

GEFÖRDERT DURCH Digifonds

AK WIEN

IMPRESSUM

Medieninhaber und Hersteller:
IFZ – Interdisziplinäres Forschungszentrum für Technik, Arbeit und Kultur
ZVR: 193954517
Schlögelgasse 2, 8010 Graz, Österreich, www.ifz.at

Herausgeber*innen und Autor*innen
Mag. Dr. Julian Anslinger, IFZ
Jaroslava Huber, MSc, Pro2Future
DI Michael Haslgrübler, Pro2Future
MMag.a Dr.in Anita Thaler, IFZ

Autorin Gastbeitrag
Mag.a Susanne Haslinger, PRO-GE

Lektorat
Dr. Rosemarie Stern

Layout und Design
Mag.(FH) Marion Dorner

Barrierefreiheit
Dipl.-Ing. Gerhard Nussbaum, Kompetenznetzwerk Informationstechnologie zur Förderung der Integration von Menschen mit Behinderungen (KI-I)

Finanzierung
Das Projekt „Verantwortungsvolle Einbindung von KI-Assistenzsystemen am Arbeitsplatz“ (VEKIAA) und dieses daraus resultierende Handbuch wurden gefördert aus Mitteln des Digitalisierungsfonds Arbeit 4.0 der AK Wien.

Haftung
Alle Angaben in dieser Publikation erfolgen trotz sorgfältiger Bearbeitung ohne Gewähr. Eine Haftung der Herausgeber*innen oder Autor*innen ist ausgeschlossen. Rechtsausführungen stellen die unverbindliche Meinung der Autor*innen dar und können der Rechtsprechung der unabhängigen Gerichte keinesfalls vorgreifen.

Copyright
CC BY-SA 4.0

Zitiervorschlag
Anslinger, Julian, Jaroslava Huber, Michael Haslgrübler, und Anita Thaler, Hrsg. 2022. Verantwortungsvolle Einbindung von KI-Assistenzsystemen am Arbeitsplatz. Ein Handbuch für Arbeitnehmende und ihre Vertretungen. 10.17605/OSF.IO/98B4H.

Verfügbarkeit
Dieses Buch ist als digitale Version im Open Science Framework (OSF) unter folgendem Link abrufbar: *https://osf.io/98b4h/?view_only=e6579c2e52c944fe87c92d6c46235ecd.* Eine Printversion gibt es nicht.

Barrierefreiheit
Dieses Buch orientiert sich an den Richtlinien WCAG 2.1 des WAI / W3C für Barrierefreiheit. Die PDF-Datei ist zudem PDF/UA-konform.

Geschlechtersensible Sprache
Zur verbesserten Interpretierbarkeit durch Vorlese-Anwendungen für Menschen mit Sehbehinderungen zielen wir darauf ab, geschlechtsneutrale Formulierungen (z. B. „Vorgesetzte“ oder „Belegschaft“) zu verwenden. Wenn nicht möglich, greifen wir auf das Gender-Sternchen (z. B. „Arbeiter*innen“) zurück, um Menschen mit nicht-binären Geschlechtsidentitäten sichtbar einzuschließen. Bezeichnungen wie „der Betriebsrat“ und „der Arbeitgeber“ weisen auf das Organ bzw. die Funktion hin. Das heißt beispielsweise, auch wenn der Begriff „Betriebsrat“ im sogenannten generischen Maskulinum formuliert ist, kann es sich bei den Mitgliedern des Betriebsrats um Betriebsrät*innen jeden Geschlechts handeln.

GEFÖRDERT DURCH Digifonds

Inhalt

4 Vorwort

7 ABSCHNITT A: Einführung in das Handbuch „Verantwortungsvolle Einbindung von KI-Assistenzsystemen am Arbeitsplatz“

12 Verantwortungsvolle Künstliche Intelligenz: Ein kurzer Überblick
14 KI-Assistenzsysteme – was ist das überhaupt?
18 Wie KI für Unternehmen und den gesamten Planeten von Nutzen sein kann
21 Warum man bei (KI-)Algorithmen am Arbeitsplatz genauer hinsehen sollte

26 ABSCHNITT B: Technologiegestaltung: Wie sollten KI-Assistenzsysteme (nicht) gestaltet werden

27 Technologieentwicklung mit Nutzer*innen – warum denn das?
32 Assistenzsysteme zur Förderung psychischer Gesundheit am Arbeitsplatz?
36 Wie KI mit Daten lernt und welche Vorurteile dabei erzeugt werden
42 Erklärbare Künstliche Intelligenz

47 ABSCHNITT C: Mitbestimmung: Die KI-Systeme kommen – was können Sie tun?

48 KI im Betrieb: Wer soll wie mitreden?
54 Mitbestimmung bei Betriebsvereinbarungen
61 Gesetzliche Rahmenbedingungen für den Einsatz von KI-Assistenzsystemen am Arbeitsplatz
74 Fragen, die Sie bei der (Mit-)Gestaltung von KI-Assistenzsystemen am Arbeitsplatz stellen sollten

80 ABSCHNITT D: Darf`s ein bisserl mehr sein? Hintergrundwissen über die KI für Detailverliebte

81 Überblick über Infrastruktur, Entwicklungsumgebungen und Verbrauchsmodelle von KI-Systemen
84 Wie KI-Assistenzsysteme ihre Umwelt wahrnehmen: Sensoren
89 KI aus systemischer Sicht und ihre Risiken

93 Autor*innenverzeichnis
94 Datenschutzerklärung zum Übernehmen

Vorwort

Was genau ist eigentlich Künstliche Intelligenz? Was sind KI-Assistenzsysteme? Welche Vorteile und Gefahren bringen diese Technologien für Arbeitnehmende mit sich?

Mit diesem Handbuch möchten wir Ihnen auf all diese Fragen verständliche Antworten geben und Ihnen einen leichten Einstieg in das Themengebiet Künstliche Intelligenz am Arbeitsplatz ermöglichen. Damit Sie jedoch nicht nur wissen, was KI ist, sondern KI am Arbeitsplatz erfolgreich mitgestalten können, finden Sie in diesem Handbuch darüber hinaus eine Vielzahl an Tipps, Methoden und Listen, die Sie bei der Gestaltung von KI-Assistenzsystemen am Arbeitsplatz unterstützen.

Die Informationen, die Sie in diesem Buch finden, wurden in dem Projekt VEKIAA (Verantwortungsvolle Einbindung von KI-Assistenzsystemen am Arbeitsplatz) zusammengetragen, das in den Jahren 2020–2022 durch den Digitalisierungsfonds Arbeit 4.0 der Arbeiterkammer Wien gefördert wurde. Das VEKIAA-Projektteam, bestehend aus dem *IFZ - Interdisziplinäres Forschungszentrum für Technik, Arbeit und Kultur* sowie dem Forschungsunternehmen *Pro2Future*, verfolgte hierzu einen partizipativen Forschungsansatz für Technologieentwicklung. Mit diesem Ansatz wurden relevante Interessenvertretungen der Arbeitswelt aktiv in den Forschungsprozess miteingebunden, um sicherzustellen, dass die tatsächlichen Umstände rund um die Implementierung von KI-Assistenzsystemen an Arbeitsplätzen berücksichtigt werden und die abgeleiteten Methoden und Werkzeuge möglichst praxisnah und nützlich sind.

Wir möchten die Gelegenheit an dieser Stelle nutzen, um uns bei allen Arbeitnehmer*innen, Betriebsrät*innen, Gewerkschaftsverteter*innen und Expert*innen zu bedanken, die uns in zahlreichen E-Mails, Telefonaten, Diskussionsrunden und Workshops ihre Zeit und ihr Wissen zur Verfügung gestellt haben. Stellvertretend möchten wir folgenden Personen besonderen Dank aussprechen:

- Prof.(FH) Mag. Dr. Reinhard Altenburger, IMC FH Krems
- Dr. Bernhard Dieber, JOANNEUM RESEARCH
- Univ.-Prof.in Dr.in Elke Gruber, Universität Graz, Arbeiterbetriebsrat
- Franz Jantscher, voestalpine Stahl Donawitz, Zentralbetriebsrat
- Werner Müller, Eaton Industries Austria GmbH
- DIin Anita Stallinger, EFKON GmbH
- Mag.a Magdalena Wicher, Institut für Höhere Studien und Betriebsratsvorsitzende
- Ing. Manfred Zierler, Infineon Technologies Austria AG

Bevor es richtig losgeht und wir verständlich erklären, was KI überhaupt ist und wie Sie sicherstellen können, dass KI möglichst positive Auswirkungen auf Ihren Arbeitsplatz hat, möchten wir Menschen und Institutionen aus Österreich zu Wort kommen lassen. Wir haben wichtige Interessenvertretungen danach gefragt, welche Chancen sie im Einsatz von Künstlicher Intelligenz für den Wirtschaftsstandort Österreich beziehungsweise für Arbeitnehmende an österreichischen Arbeitsplätzen sehen. Das waren ihre Antworten:[1]

1 Disclaimer: Die Befragten hatten vorab keinen Einblick in konkrete Inhalte dieses Handbuchs.

© Daniel Novotny

FRIDOLIN HERKOMMER, MSC

Leiter des Büros für Digitale Agenden in der AK Wien, Digitalisierungsfonds Arbeit 4.0 und bis Oktober 2021 Mitglied des Österreichischen Rats für Robotik und KI

„Entscheidend ist, welches Ziel KI-Technologie verfolgt, wie sie das tut und inwieweit Betroffene in Technikentwicklung und Einsatz eingebunden sind. Ziel muss eine Qualitätssteigerung sein, das gilt für Produkte, Arbeit und Kundenzufriedenheit gleichermaßen. Dazu braucht es ehrliche Einbindung und soziale Innovationen, welche die Veränderung der Arbeitsabläufe und Qualifikationsanforderungen begleiten. Grundsätzlich gilt, nur was zielführend, zweckmäßig und auch datensparsam ist, kann und soll umgesetzt werden und die Erfüllung dieser Anforderungen ist unter breiter Einbindung, insbesondere der Beschäftigten, sicherzustellen. Unter diesen Voraussetzungen kann KI wettbewerbsentscheidend sein, sowohl, was Arbeitsplatzzufriedenheit als auch, was Produkte und Dienstleistungen betrifft.“

© Cajetan Perwein

LEONORE GEWESSLER, BA

Bundesministerin für Klimaschutz, Umwelt, Energie, Mobilität, Innovation und Technologie

„Künstliche Intelligenz kann Unternehmen entscheidende Wettbewerbsvorteile verschaffen und zugleich wichtige gesellschaftliche und ökologisch notwendige Transformationsprozesse in den verschiedensten Bereichen unterstützen. KI-Anwendungen begegnen uns bereits jetzt in der Arbeitswelt als ‚intelligentes Werkzeug‘ und werden das in Zukunft vermehrt tun. Diese intelligenten Systeme haben das Potential, unseren Arbeitsalltag effizienter, angenehmer und in manchen Bereichen auch sicherer zu gestalten. Bei diesen Entwicklungen ist es aber wesentlich, den Nutzen für den Menschen und das Wohl des Menschen in den Mittelpunkt zu stellen und die Arbeitnehmer:innen frühzeitig einzubinden.“

© Dragan Tatic

UNIV.-PROF. MAG. DR. MARTIN KOCHER

Bundesminister für Arbeit sowie für Digitalisierung und Wirtschaftsstandort

„Systeme künstlicher Intelligenz sind schon jetzt integraler Bestandteil von Organisationen, oft ohne, dass wir es merken. Sie können eine große Hilfe für Arbeitnehmerinnen und Arbeitnehmer darstellen, nicht zuletzt bei der Inklusion von Menschen mit gesundheitlichen Einschränkungen. Ich bin froh, dass es gerade im Bereich der Inklusion viele Initiativen in Österreich gibt. Wie immer bieten neue Technologien Chancen und Risiken. Wichtig aus meiner Sicht ist es, bei ihrer Anwendung die großen Chancen zu nutzen und dabei auf die Transparenz von Algorithmen und deren Kontrolle zu achten, um mögliche Risiken oder unerwünschte Nebenwirkungen zu minimieren.“

© Mirjam Reither

KORINNA SCHUMANN
Vizepräsidentin und Bundesfrauenvorsitzende Österreichischer Gewerkschaftsbund

„Die praktische Anwendung von KI und Algorithmen in der Arbeitswelt ist bisher noch auf einzelne Bereiche begrenzt, etwa beim Wiedereinstieg in den Arbeitsmarkt. KI hat als Entscheidungssystem Potenzial, kann aber die Gefahr in sich tragen, gerade Diskriminierungen von Frauen, älteren Menschen oder Menschen mit Migrationshintergrund fortzuschreiben. Technologie muss immer nach menschlichen Werten und Bedürfnissen gestaltet werden. Die Herausforderung besteht darin, dass technologische Entwicklungen Solidarität und Mitbestimmung unterstützen und nicht erodieren."

© Daniel Novotny

DR.IN EVA ANGERLER
Gewerkschaft GPA, Abteilung Arbeit und Technik

„KI-Anwendungen werten große Datenmengen aus und machen Prognosen und Entscheidungsvorschläge, die in 99 % der Fälle befolgt werden (Automation Bias). Solche Systeme haben daher massive Auswirkungen auf die Betroffenen. Die Risiken liegen in der technischen Überwachung und algorithmischen Steuerung. Sollen KI-Anwendungen zur Unterstützung und zur Arbeitserleichterung für die ArbeitnehmerInnen eingesetzt werden, muss KI erklärbar und verstehbar werden. Als Voraussetzung dafür braucht es bei der Einführung von KI in der Arbeitswelt verpflichtende Mindeststandards, unabhängige Kontrollmechanismen und starke Mitbestimmung."

PRO-GE
Produktionsgewerkschaft

„Mit Hilfe von KI können wir Arbeitsplätze noch sicherer und vor allem gesünder gestalten. Da BetriebsrätInnen immer öfter mit dem Einsatz von neuen Technologien konfrontiert sind, ist es notwendig, ihnen die entsprechenden Werkzeuge bereitzustellen. Dieses Handbuch beinhaltet die wichtigsten Informationen, um bei der Umgestaltung zu KI-Arbeitsplätzen mitzureden und mitzuwirken."

Einführung in das Handbuch „Verantwortungsvolle Einbindung von KI-Assistenzsystemen am Arbeitsplatz“

Einführung in das Handbuch „Verantwortungsvolle Einbindung von KI-Assistenzsystemen am Arbeitsplatz“

DIESES KAPITEL IST FÜR SIE VON INTERESSE, WENN …

… Sie erfahren möchten, warum KI-Assistenzsysteme ihren Weg an den Arbeitsplatz finden.

… Sie wissen möchten, wie dieses Handbuch aufgebaut ist.

… Sie lernen möchten, wie Sie dieses Handbuch am besten nutzen.

KI-Assistenzsysteme – mehr als nur ein Schlagwort?

Die Digitalisierung der Arbeitswelt schreitet immer weiter fort. In den 1970er-Jahren zu Beginn der dritten industriellen Revolution war es der Mikroprozessor, der Arbeitsplätze von Grund auf zu verändern begann. Arbeitnehmenden standen fortan Computer und Industrieroboter zur Seite, die sie bei der Arbeit unterstützen sollten und gleichzeitig völlig neue Kompetenzen erforderlich machten. Zwei Dekaden später, in den 1990er-Jahren, erhielt das Internet zunehmend Einzug in die Arbeitswelt und erhöhte Vernetzung und Kommunikationsgeschwindigkeit in einem kaum zu ahnenden Ausmaß – die vierte industrielle Revolution war eingeläutet. Die im deutschsprachigen Raum mit einer Anspielung auf die Nummerierung für Softwaresysteme auch „Industrie 4.0“ genannte Umwälzung der Arbeitswelt ist gekennzeichnet durch eine nie dagewesene Automatisierung und Verflechtung von Technologien. Heute bestimmen Schlagworte wie „Big Data“, das „Internet der Dinge“, „Cloud Computing“, „Maschinelles Lernen“, „Blockchains“, „Metaverse“ sowie „Künstliche Intelligenz“ den medialen und wissenschaftlichen Diskurs über die Digitalisierung der Arbeitswelt. Und während die Welt diskutiert, sind Arbeitnehmende längst mittendrin.

KÜNSTLICHE INTELLIGENZ IN KÜRZE

Wenn auch in aller Munde, ist der Begriff „Künstliche Intelligenz“ auf Anhieb nicht für alle Menschen verständlich. Dies ist dem Umstand geschuldet, dass es sich bei der „Künstlichen Intelligenz“ um ein Modewort handelt, das viele unterschiedliche Ansätze der Informatik unter einem Begriff vereint. Viele dieser Ansätzen haben mit dem Intelligenzbegriff, den wir mit Menschen oder Tieren assoziieren, tatsächlich nur wenig zu tun. In Lehrbüchern werden KI-Technologien, etwas sperrig, als Systeme beschrieben, die ihre Umwelt erfassen und Handlungen setzen, die die Chancen maximieren, gesetzte Ziele zu erreichen[1].

Ein Bilderkennungsalgorithmus hat beispielsweise auf der Basis verschiedener Fotos (*Erfassung der Umwelt*) gelernt, Gegenstände mit einer gewissen Wahrscheinlichkeit ▶

1 Z. B. Poole, Mackworth, und Goebel, 1998.

▶ richtig (*Chancenmaximierung*) zu erkennen (*Zielerreichung*). Das „Lernen“ erfolgt hierbei entgegen gängiger Annahmen üblicherweise nicht vollautomatisch, sondern mit erheblicher menschlicher Unterstützung, unter anderem durch vorherige Bild-Kategorisierung und Programmierung der Unterscheidungsmerkmale.

Andere KI-Algorithmen wiederum funktionieren zwar nach den gleichen Prinzipien der Umwelterfassung, Chancenmaximierung und Zielerreichung, aber auf andere Art und Weise. Mehr hierzu gibt es im Abschnitt A in Kapitel KI-Assistenzsysteme – was ist das überhaupt? zu lesen.

Unter anderem setzen Unternehmen zunehmend auf KI-basierte Assistenzsysteme, um die es auch in diesem Buch gehen soll. Es gibt viele verschiedene Arten von KI-Assistenzsystemen. Im Folgenden ein paar Beispiele:

Sogenannte Augmented-Reality-Displays (s. Abbildung 1) – also am Kopf zu befestigende Durchsichtbildschirme – können beispielsweise Arbeiter*innen in Produktionsstätten nützliche Zusatzinformationen zum Gesehenen anzeigen. Exoskelette (s. Abbildung 2) wiederum können dabei helfen, körperliche Arbeiten wie das Heben von schweren Kisten durchzuführen. Andere KI-basierte Assistenzsysteme können Informationen sammeln, zusammenstellen und bei Entscheidungen helfen.

Abbildung 1: Augmented-Reality-Display der Marke Vuzix Smart Glasses. Foto von *Axrxax,* lizensiert unter *CC BY-SA 4.0*

Abbildung 2: Exoskelett der Marke ExoHeaver. Foto von *Vladislav Davidov li*zensiert unter *CC BY-SA 4.0*

Der Nutzen solcher Systeme im Hinblick auf die Optimierung von Arbeitsprozessen und die Verbesserung der Produktqualität ist augenscheinlich. Aber auch für die Arbeitnehmenden selbst gibt es viele Vorteile. Es gibt KI-gesteuerte Assistenzsysteme, die dabei helfen, Unfallgefahren zu minimieren. Andere wiederum tragen dazu bei, die betriebliche Inklusion zu verbessern, indem sie für Arbeitnehmende mit Hörbehinderungen für den Arbeitsprozess relevante akustische Informationen in visuelles und haptisches Feedback umwandeln. Wieder andere KI-Assistenzsysteme reduzieren die Informationslast und dadurch den notwendigen kognitiven Verarbeitungsaufwand, indem sie große Mengen an Informationen zu einer besser verdaulichen Auswahl reduzieren.

Diesen und anderen Vorteilen KI-basierter Assistenzsysteme stehen jedoch auch mögliche Nachteile gegenüber, die es zu berücksichtigen gilt. Wegen der großen Datensammelwut von KI-Systemen sind nie dagewesene Formen von Leistungsüberwachung und Einschränkungen der Privatsphäre möglich, mit unmittelbaren Konsequenzen für das psychische Wohlbefinden und die Gesundheit von Arbeitnehmenden[2]. Darüber

2 Z. B. Gimpel u. a., 2019.

hinaus sind von KI-Systemen getroffene Entscheidungen oft wenig transparent und nachvollziehbar, sowie manchmal nicht einmal anfechtbar. Auch gibt es Beispiele von KI-Systemen, die in der Gesellschaft vorhandene diskriminierende Strukturen fortschreiben und ganze Menschengruppen benachteiligen.

Was braucht's? Technologische Kompetenzen

> Während Arbeitgeber naturgemäß vor allem direkte ökonomische Vorteile von KI-Assistenzsystemen im Blick haben, sind es oft die Arbeitnehmenden und ihre Vertretungen (Arbeitnehmer*innenvertretungen, Gleichstellungsbeauftragte, Beauftragte für Belange von Menschen mit Behinderungen, Datenschutzbeauftragte, etc.), die die Bürde tragen, sich im Spannungsfeld zwischen positiven und negativen Auswirkungen von KI-Assistenzsystemen zurechtzufinden.

Arbeitnehmer*innen(-vertretungen) sind gezwungen, die im Betrieb geplanten und eingesetzten Systeme zu verstehen, um durch ihre Werkzeuge der Mitbestimmung Änderungen erwirken zu können. Dabei reicht es jedoch nicht, allein die Funktionsweise der KI-Systeme nachzuvollziehen. Vielmehr benötigt es Wissen, das über das hinausgeht, was in den Schulungen über die Systeme vermittelt wird. Der Sozialphilosoph und Gewerkschaftsforscher Oskar Negt spricht hier von „technologischen Kompetenzen“[3]. Diese erlauben es, die Auswirkungen der neuartigen Technologien auf Gesellschaft und Umwelt (und den Betrieb) kritisch zu reflektieren und die Technologie als kontrollierbar zu begreifen.[4] Zugegebenermaßen ist das keine leichte Aufgabe, die durch die hohe Innovationsgeschwindigkeit, die steigende Komplexität der Technologien und rechtlichen Rahmenbedingungen, und vor allem durch die fehlende Priorisierung in Unternehmen zusätzlich erschwert wird.

3 Negt, 1998.

4 S. auch Schröter, 2019.

Was bieten wir? Aufgaben und Inhalte dieses Handbuchs

Mit diesem Handbuch möchten wir an dieser Stelle ansetzen und Abhilfe schaffen. Wir möchten ganzheitliche technologische Kompetenzen mit Blick auf Künstliche Intelligenz und KI-Assistenzsysteme fördern und Arbeitnehmende und ihre Vertretungen dazu befähigen, die im Unternehmen geplanten oder bereits eingesetzten KI-basierten Technologien besser einschätzen und deren Einsatz mitbestimmen zu können.

Das Handbuch ist in vier Abschnitte gegliedert. Im Abschnitt A Einführung in das Handbuch, in dem sich auch dieses Kapitel befindet, geben wir einen einführenden Überblick über die Technologie der KI-Assistenzsysteme und ihre möglichen Vorteile und potentiellen Nachteile, insbesondere am Arbeitsplatz. Im Abschnitt B Technologiegestaltung: Wie sollten KI-Assistenzsysteme (nicht) gestaltet werden legen wir unter anderem dar, warum Nutzer*innen bei der Gestaltung eines KI-Assistenzsystems einbezogen werden sollten, warum KI-Assistenzsysteme zu Diskriminierung führen können und wie man KI-Algorithmen erklärbar macht. In Abschnitt C Mitbestimmung: Die KI-Systeme kommen – was können Sie tun? zeigen wir Mittel und Wege auf, wie betriebliche Mitbestimmung bei der Entwicklung und Implementierung von KI am Arbeitsplatz gelingen kann. Sie finden dort Informationen über die Gestaltung von Betriebsvereinbarungen, rechtliche Rahmenbedingungen bis hin zu einem How-To der Einbindung von Interessenvertretungen. Auch finden Sie dort eine Liste an Fragen, die Sie bei der (Mit-)Gestaltung von KI-Assistenzsystemen am Arbeitsplatz stellen sollten. Im letzten Abschnitt D Darf's ein bisserl mehr sein? Hintergrundwissen über KI für Detailverliebte finden Sie einen technischeren Einblick in die Funktionsweise und Risiken von KI-Assistenzsystemen.

Wie soll dieses Handbuch genutzt werden?

Um die Frage aus der Überschrift direkt zu beantworten: Genau so, wie Sie es brauchen! Wir haben uns bemüht, die Abschnitte und Kapitel des Handbuchs modular aufzubereiten. Das heißt, wenn Sie sich für ein bestimmtes Thema interessieren, blättern Sie auf die im Inhaltsverzeichnis angeführte Seite und lesen Sie los. Möchten Sie Detailinformationen zu den angeführten Informationen, können Sie die zitierte Literatur konsultieren, die in den Fußnoten in kurzer Form und im Literaturverzeichnis des jeweiligen Kapitels in ausführlicher Form angeführt ist.

Zum Abschluss möchten wir Ihnen noch ein paar Symbole erklären, die Sie dabei unterstützen, sogleich zu den wesentlichen Informationen zu gelangen.

Dieses Symbol finden Sie am Anfang jedes Kapitels. Daneben finden Sie eine kurze Beschreibung der wesentlichen Kapitelinhalte.

Wenn Sie dieses Symbol gefunden haben, befinden Sie sich mitten in einem Kapitel und haben die Chance, vertiefende Informationen in einer Infobox nachzulesen.

Dieses Symbol befindet sich an vielen Kapitelenden. Hier finden Sie unsere Empfehlungen für Texte, Webseiten, Videos, Podcasts, etc. zum jeweiligen Thema.

Auch dieses Symbol findet sich am Ende mancher Kapitel. Es verweist auf Ansprechpersonen, die Sie bei Interesse am Thema kontaktieren können.

Dieses Symbol verweist auf die Literaturverzeichnisse, die sich am Ende der Kapitel befinden.

Wir wünschen Ihnen viel Spaß beim Lesen und viel Erfolg beim Mitgestalten von KI-basierten Systemen an Ihrem Arbeitsplatz!

Julian Anslinger, Jaroslava Huber, Michael Haslgrübler & Anita Thaler

KONTAKT GEFÄLLIG?

Mag. Dr. Julian Anslinger, IFZ
Leiter des Projekts „Verantwortungsvolle Einbindung von KI-Assistenzsystemen am Arbeitsplatz (VEKIAA)“
Julian.Anslinger@ifz.at

LITERATURVERZEICHNIS

Gimpel, Henner, Julia Lanzl, Christian Regal, Nils Urbach, Sascha Wischniewski, Patricia Tegtmeier, Mathias Kreilos, u. a. 2019. *Gesund digital arbeiten?! Eine Studie zu digitalem Stress in Deutschland.* Augsburg: Projektgruppe Wirtschaftsinformatik des Fraunhofer FIT. *https://doi.org/10.24406/fit-n-562039.*

Negt, Oskar. 1998. *Lernen in einer Welt gesellschaftlicher Umbrüche.* Herausgegeben von Heinrich Dieckmann und Bernd Schachtsiek. Lernkonzepte im Wandel. Die Zukunft der Bildung. Klett-Cotta: Stuttgart.

Poole, David, Alan Mackworth, und Randy Goebel. 1998. *Computational Intelligence: A Logical Approach.* Oxford University Press: Oxford.

Schröter, Welf. 2019. Der mitbestimmte Algorithmus. Arbeitsweltliche Kriterien zur sozialen Gestaltung von Algorithmen und algorithmischen Entscheidungssystemen. In *Der mitbestimmte Algorithmus. Gestaltungskompetenz für den Wandel der Arbeit,* herausgegeben von Welf Schröter, 101–50. Mössingen-Talheim: Talheimer Verlag.

Verantwortungsvolle Künstliche Intelligenz: Ein kurzer Überblick

Ob unter dem Stichwort verantwortungsvolle, faire, vertrauensvolle oder ethische Künstliche Intelligenz, die Forderung nach KI-Systemen, die keine Nachteile für Personen und Umwelt nach sich ziehen, ist in aller Munde. Sie wird medial diskutiert und von politisch, wissenschaftlich sowie ökonomisch motivierten Akteuren, wie Staaten, Universitäten oder großen Firmen, in Leitlinien gefasst. Bis dato lassen sich viele hunderte verschiedene Leitlinien zu ethischer KI finden. Bei diesen handelt es sich um keine verpflichtenden Regeln, sondern lediglich um richtungsweisende Ideen, die jedoch einen relevanten Einfluss auf Entscheidungsprozesse in Entwicklungs- und Gestaltungsprozessen Künstlicher Intelligenz ausüben können[1].

Interessanterweise weisen Ethik-Leitlinien keinen einheitlichen Tenor im Hinblick darauf auf, was eine ethische KI überhaupt ausmacht oder wie man diese realisieren könnte. Je nach Interessenlage der jeweiligen Autor*innen werden verschiedene Aspekte ethischer KI in den Vordergrund gestellt und diskutiert[2]. Was sind nun also die relevantesten Aspekte im Diskurs über ethische KI?

Eine Forschungsgruppe der ETH Zürich rund um Anna Jobin befasste sich mit dieser Frage und legte im renommierten Forschungsjournal „Nature Machine Intelligence“ eine Inhaltsanalyse von 84 verschiedenen KI-Leitlinien vor[3]. Insgesamt identifizierten sie 11 wiederkehrende ethische Grundsätze und Prinzipien. Davon waren fünf in mehr als der Hälfte der untersuchten Leitlinien vertreten. Die Grundsätze, aufgelistet nach Häufigkeit des Auftretens in absteigender Reihenfolge, waren:

- **Transparenz:** ein KI-System sollte verständlich und nachvollziehbar sein.
- **Gerechtigkeit und Fairness:** ein KI-System sollte niemanden benachteiligen.
- **Schadlosigkeit:** ein KI-System sollte niemandem Schaden zufügen.
- **Verantwortung:** es sollte geklärt sein, wer die Verantwortung für die Auswirkungen eines KI-Systems trägt.
- **Datenschutz:** das Recht auf den Schutz personenbezogener Daten sollte aufrechterhalten bleiben.
- **Nutzen:** eine KI sollte zum menschlichen Wohl, Frieden und Glück beitragen.
- **Freiheit und Autonomie:** eine KI sollte menschliche Freiheit und Autonomie nicht schwächen.
- **Vertrauen:** Vertrauen in KI sollte gestärkt und hinterfragt werden.
- **Nachhaltigkeit:** eine KI sollte zum Umweltschutz und der Gleichstellung verschiedener sozialer Gruppen beitragen.
- **Würde:** eine KI sollte die Würde des Menschen nicht antasten.
- **Solidarität:** eine KI sollte weniger geschützte soziale Gruppen unterstützen.

Das wohl bekannteste und einflussreichste Dokument über ethische KI wurde von der Europäischen Kommission beauftragt und beeinflusste u. a. maßgeblich den Verordnungsvorschlag der Europäischen Kom-

1 Jobin, Ienca, und Vayena, 2019.

2 Gießler und Haas, 2020.

3 Jobin, Ienca, und Vayena, 2019.

mission zur Regulierung Künstlicher Intelligenz[4]. Diese „Ethik-Leitlinien für eine vertrauenswürdige KI“ der Hochrangigen Expertengruppe für Künstliche Intelligenz (HEG-KI) umfassen neben ethischen Aspekten zwei weitere Komponenten. Alle drei Komponenten sollen während des gesamten Lebenszyklus eines KI-Systems erfüllt sein. Die HEG-KI fordert:[5]

- 1. Künstliche Intelligenz sollte **rechtmäßig** sein und somit geltendes Recht und alle gesetzlichen Bestimmungen einhalten.
- 2. Künstliche Intelligenz sollte **ethisch** sein und somit die Einhaltung ethischer Grundsätze und Werte garantieren.
- 3. Künstliche Intelligenz sollte **robust** sein, und somit sowohl in technischer als auch sozialer Hinsicht die Wahrscheinlichkeit für unabsichtliche Fehler geringhalten.

In diesem Handbuch gehen wir – mit explizitem Fokus auf Künstliche Intelligenz am Arbeitsplatz – auf alle drei genannten Komponenten vertrauenswürdiger KI und viele der obig genannten Grundsätze ethischer KI ein. Beispielsweise finden Sie in Abschnitt B In den Kapiteln Wie KI mit Daten lernt und welche Vorurteile dabei erzeugt werden und Erklärbare Künstliche Intelligenz Beispiele und Lösungsmöglichkeiten für diskriminierende und intransparente Künstliche Intelligenz. In Abschnitt C im Kapitel Gesetzliche Rahmenbedingungen für den Einsatz von KI-Assistenzsystemen am Arbeitsplatz finden Sie Informationen dazu, welche Gesetze es bei der Implementierung von KI an österreichischen Arbeitsplätzen zu befolgen gilt. Darüber hinaus finden Sie dort die Kapitel KI im Betrieb: Wer soll wie mitreden? und Mitbestimmung bei Betriebsvereinbarungen, in denen wir erklären, wie KI-Systeme am Arbeitsplatz robuster gestaltet werden können.

SIE HABEN NOCH NICHT GENUG?

Sehen Sie diese SWR-Dokumentation auf YouTube an:

- *Von Chatbots bis zu Waffensystemen – Fluch und Segen der Künstlichen Intelligenz*[6]

LITERATURVERZEICHNIS

EU. 2021. *Proposal for a Regulation of the European parliament and of the council laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts,* EUR-Lex – 52021PC0206 – EN §. *https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1623335154975&uri=CELEX%3A52021PC0206*.

Europäische Kommission. 2019. *Ethik-Leitlinien für eine vertrauenswürdige KI. https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60425*.

Gießler, Sebastian, und Leonard Haas. 2020. *Ethics between business lingo and politics: Why bother? https://inventory.algorithmwatch.org/about*.

Jobin, Anna, Marcello Ienca, und Effy Vayena. 2019. *The global landscape of AI ethics guidelines. Nature Machine Intelligence 1,* Nr. 9: 389–99. *https://doi.org/10.1038/s42256-019-0088-2*.

SWR. 2021. *Von Chatbots bis zu Waffensystemen – Fluch und Segen der Künstlichen Intelligenz. https://www.youtube.com/watch?v=oNk6ESLpxKI*.

4 Der Verordnungsvorschlag wird zurzeit überarbeitet und findet vermutlich frühestens im Jahr 2024 seinen Weg in die europäische Gesetzgebung. EU, 2021.

5 Europäische Kommission, 2019.

6 SWR, 2021.

KI-Assistenzsysteme – was ist das überhaupt?

DIESES KAPITEL IST FÜR SIE VON INTERESSE, WENN …

… Sie wissen wollen, was Künstliche Intelligenz ist.

… Sie verstehen wollen, was KI-Assistenzsysteme sind.

Was ist Künstliche Intelligenz?

Bereits in 40er-Jahren des letzten Jahrhunderts wurde über Künstliche Intelligenz nachgedacht, beispielsweise in Alan Turings Computational Theory[1]. Dort postulierte Turing mit dem Turing-Test ein Werkzeug, mit dem man feststellen können sollte, ob Maschinen wie Menschen denken können. Er definierte, dass Maschinen denken können, wenn eine beobachtende Person glaubt, dass das beobachtbare Verhalten der Maschine der eines Menschen entspricht.

Heute gibt es eine Vielzahl an Definitionen Künstlicher Intelligenz. Viele der Definitionen verstehen Künstliche Intelligenz als das maschinelle Nachahmen menschlicher Fähigkeiten (s. Infobox). Um welche Fähigkeiten es sich genau handelt und ab wann eine Ähnlichkeit zum Menschen überhaupt gegeben ist, bleibt jedoch in vielen Fällen vage. Alan Turings Operationalisierung denkender Maschinen war somit sogar greifbarer als heutige allgemeine Definitionen Künstlicher Intelligenz.

DEFINITIONEN VON KÜNSTLICHER INTELLIGENZ

Definition (Gabler Wirtschaftslexikon[2]): Erforschung intelligenten Problemlösungsverhaltens sowie die Erstellung intelligenter Computersysteme. Künstliche Intelligenz (KI) beschäftigt sich mit Methoden, die es einem Computer ermöglichen, solche Aufgaben zu lösen, die, wenn sie vom Menschen gelöst werden, Intelligenz erfordern.

Definition (Deutsche Wikipedia[3]): ist ein Teilgebiet der Informatik, das sich mit der Automatisierung intelligenten Verhaltens und dem maschinellen Lernen befasst.

Definition (Bundesregierung Österreich[4]): Als Künstliche Intelligenz werden Computersysteme bezeichnet, die intelligentes Verhalten zeigen, d. h. die in der Lage sind, Aufgaben auszuführen, die in der Vergangenheit menschliche Kognition und menschliche Entscheidungsfähigkeit erfordert haben.

Definition (Arbeiterkammer Wien[5]): Von Künstlicher Intelligenz wird meist dann gesprochen, wenn IT-Anwendungen im Betrieb selbst lernen und die dafür notwendigen Daten erheben und Zusammenhänge analysieren. ▶

1 Turing, 1950.
2 Siepermann, 2021.
3 Wikipedia, 2021.
4 Bundesministerium für Klimaschutz, Umwelt, Energie, Mobilität, Innovation und Technologie, 2021.
5 Angerler, Chlestil, und Heiling, 2021.

▶ **Definition (John McCarthy[6]):** It is the science and engineering of making intelligent machines, especially intelligent computer programs.
Übersetzung: Es ist die Forschung und Entwicklung vom Erzeugen intelligenter Maschinen, insbesondere intelligenter Computerprogramme.

Definition (PWC[7]): AI, shorthand for artificial intelligence, defines technologies emerging today that can understand, learn, and then act based on that information.
Übersetzung: Unter AI versteht man entstehende Technologien, die verstehen, lernen und dann auf Basis von Informationen handeln können.

Definition (Wikipedia[8]): Artificial intelligence (AI) is intelligence demonstrated by machines, as opposed to the natural intelligence displayed by humans or animals.
Übersetzung: Künstliche Intelligenz (KI) ist Intelligenz, die von Maschinen demonstriert wird, im Gegensatz zu natürlicher Intelligenz bei Mensch oder Tier.

Definition (Europäisches Parlament[9]): AI is the ability of a machine to display human-like capabilities such as reasoning, learning, planning and creativity.
Übersetzung: Künstliche Intelligenz ist die Fähigkeit einer Maschine, menschliche Fähigkeiten wie logisches Denken, Lernen, Planen und Kreativität zu imitieren.

Was KI-Expert*innen unter KI verstehen: Starke und schwache KI

Unter starker KI versteht man eine Künstliche Intelligenz, die sich, ähnlich wie wir Menschen, hypothetisch allen Problemstelllungen widmen kann, die ihr begegnen. Sie kann denken und hat ein Bewusstsein. Starke KI ist bis dato und zumindest für die kommenden Jahrzehnte jedoch ein rein hypothetisches Konstrukt, das im Bereich der Science Fiction anzusiedeln ist.

Wenn KI-Expert*innen heute über KI-Systeme sprechen, beziehen sie sich zumeist auf die sogenannte schwache KI. Schwache KI besteht aus komplexen Algorithmen, die ausschließlich Lösungen für klar definierte Problemstellungen finden können. Programmierer*innen bedienen sich hierbei vieler unterschiedlicher Technologien, die in den vergangenen 70 Jahren erforscht wurden.

KI-Methoden

Die EU-KI-Direktive[10] definiert folgende Techniken und Konzepte Künstlicher Intelligenz:

1. Konzepte des Maschinellen Lernens, mit beaufsichtigtem, unbeaufsichtigtem und bestärkendem Lernen unter Verwendung einer breiten Palette von Methoden, einschließlich des tiefen Lernens (Deep Learning)

2. Logik- und wissensgestützte Konzepte, einschließlich Wissensrepräsentation, induktiver (logischer) Programmierung, Wissensgrundlagen, Inferenz- und Deduktionsmaschinen, (symbolischer) Schlussfolgerungs- und Expertensysteme

3. Statistische Ansätze, Bayessche Schätz-, Such- und Optimierungsmethoden

Maschinelles Lernen (engl. Machine Learning; ML) ist vermutlich die verbreitetste Form von Künstlicher Intelligenz. Vor allem in Unternehmen spielt sie eine große Rolle. Beim Maschinellen Lernen werden Zusammenhänge, Konzepte, Regeln, etc. auf Basis von vorhandenen Daten automatisch erlernt. Zum Beispiel: Einem Computerprogramm werden oft genug unterschiedliche Bilder von einem Hund oder einer Katze gezeigt, bis es erlernt, diese beiden voneinander zu unterscheiden. Danach kann es neue Bilder von Hund und Katze völlig autonom betrachten und entscheiden, ob eine Katze oder ein Hund dargestellt wird (s. auch Abschnitt B, Kapitel Wie KI mit Daten lernt und welche Vorurteile dabei erzeugt werden).

Ein weiterer wichtiger Ansatz Künstlicher Intelligenz sind sogenannte Expertensysteme, die aus einer ausgiebigen Wissensbasis schöpfen, sich im Gegensatz zu Machine-Learning-Algorithmen jedoch nicht selbstständig optimieren (s. auch Infobox Symbolische und subsymbolische Wissensverarbeitung).

6 McCarthy, 1998.
7 PWC, 2017.
8 Wikipedia, 2021.
9 European Parliament, 2021.
10 Europäisches Parlament, 2021.

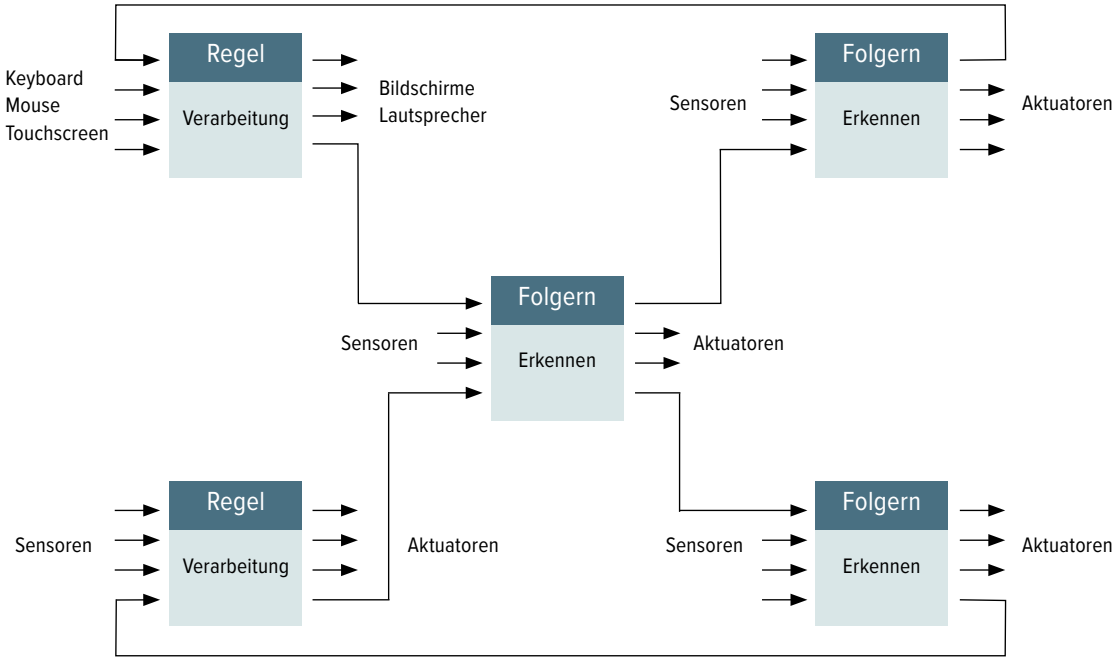


Abbildung 1: Vernetzte KI-Assistenzsysteme verwenden auch Daten von anderen Systemen, um ihre Schlussfolgerungen zu treffen und auf die Umwelt zu wirken. Diese Auswirkungen können wiederum von weiteren Systemen erfasst und verarbeitet werden. Auch Wechselwirkungsschleifen sind möglich.

SYMBOLISCHE UND SUBSYMBOLISCHE WISSENSVERARBEITUNG

KI-Forschung unterscheidet zwischen symbolischer und subsymbolischer Wissensverarbeitung[11]. Eine symbolische KI kommt auf einem nachvollziehbaren Weg zu Ergebnissen, indem Symbole (Wörter, Buchstaben, Ziffern o. Ä.) nach vorprogrammierten Regeln kombiniert werden. Ein Beispiel dafür wäre die klassische Logik (Symbol 1: Alle Menschen sind sterblich; Symbol 2: Sokrates ist ein Mensch; Schluss: Sokrates ist sterblich). Expertensysteme gelten gemeinhin als symbolische KI.

Eine subsymbolische KI hingegen kommt nicht durch die Kombination von Symbolen und Regeln zu einem Ergebnis, sondern gewinnt Informationen durch die Anwendung von mathematischen Formeln und optimiert diese so lange, bis sie das gewünschte Ergebnis erhält. In diesem Fall sprechen wir vom Erfahrungslernen, wozu auch das Maschinelle Lernen zählt.

Während im Fall von symbolischer KI alle Schritte lesbar, die Repräsentationen erkennbar und der Prozess der Beweisführung logisch verstanden und nachvollzogen werden kann, sind die kausalen Zusammenhänge bei der subsymbolischen KI nicht nachvollziehbar.

An manchen Einsatzorten Künstlicher Intelligenz nimmt man die fehlende Nachvollziehbarkeit subsymbolischer KI wegen ihrer herausragenden Fähigkeiten ohne Zögern in Kauf (z. B. bei Go- und Poker-Spielen[12]). Bei KI-basierten Entscheidungen, die hohe Risiken mit sich bringen, wie bei selbstfahrenden Autos oder medizinischen Diagnosen, stellt die fehlende Nachvollziehbarkeit subsymbolischer KI-Systeme jedoch eine Herausforderung dar, die es zu umschiffen gilt.

11 Bhatia, 2017.

12 Dix, 2018.

Was sind nun KI-Assistenzsysteme?

KI-Assistenzsysteme sind Systeme, die den Menschen bei der Verrichtung von Tätigkeiten unterstützen und bei denen Künstliche Intelligenz als zentrale Technologie eingesetzt wird. Üblicherweise nutzen KI-Assistenzsysteme Verfahren Maschinellen Lernens. KI-Assistenzsysteme werden anhand von Daten trainiert, Sensorinformationen zu interpretieren und eine Schlussfolgerung aus diesen zu treffen. Die Schlussfolgerung wird in Folge an sogenannte Aktuatoren, also KI-Umweltschnittstellen, weitergegeben. Ein KI-Assistenzsystem könnte beispielsweise mit Fotos (Daten) trainiert werden, Kamerabilder (Sensorinput) zu interpretieren, um eine Personenerkennung (Schlussfolgerung) durchzuführen und in Folge ein Türschloss (Aktuator) zu entriegeln.

Oft sind KI-Assistenzsysteme darüber hinaus eng mit anderen digitalen Systemen (klassisch und KI-basiert) vernetzt (s. Abbildung 1). Das Türschloss im obigen Beispiel könnte die Information über seine Entriegelung beispielsweise in einer Sicherheitsdatenbank protokollieren. Darüber hinaus könnte es Informationen an die Gebäudeverwaltung weitergeben, die basierend auf den jeweiligen Raumbelegungen das Gebäudeklima mittels KI automatisiert steuert.

SIE HABEN NOCH NICHT GENUG?

In dem folgenden auf YouTube erhältlichem Vortrag von Informatiker Univ.-Prof. Dr. Alois Ferscha wird erklärt, wie Forschung denkende Systeme (im Sinne der KI) umsetzt, mit welchen Sensoren und Algorithmen man Assistenzfunktionalität entwickeln kann und wie die Zukunft mit solchen Systemen aussehen könnte.

- *Cognitive Industrial Systems*[13]

LITERATURVERZEICHNIS

Angerler, Eva, Martina Chlestil, und Michael Heiling. 2021. *Künstliche Intelligenz in der Arbeitswelt. https://wien.arbeiterkammer.at/interessenvertretung/arbeitdigital/policypapers/Policy_Paper_KI_in_der_Arbeitswelt.pdf*.

Bhatia, Richa. 2017. *Understanding the difference between symbolic AI and non symbolic AI. https://analyticsindiamag.com/understanding-difference-symbolic-ai-non-symbolic-ai/*.

Bundesministerium für Klimaschutz, Umwelt, Energie, Mobilität, Innovation und Technologie. 2021. *Strategie der Bundesregierung für Künstliche Intelligenz. https://www.bmdw.gv.at/dam/jcr:25d7d1f6-b32c-4a6b-ae39-ffa3825a73ff/2021-AIM_AT_2030_UA-bf.pdf*.

Dix, Jürgen. 2018. Von Holmes zu Watson: Zu welchem Ende studiert man KI? *Jahrbuch 2017*, 75-79. *https://publikationsserver.tu-braunschweig.de/servlets/MCRFileNodeServlet/dbbs_derivate_00044622/Jahrbuch_2017_12_Dix.pdf*.

Europäisches Parlament. 2021. *Verordnung des Europäischen Parlaments und des Rates zur Festlegung harmonisierter Vorschriften für Künstliche Intelligenz (Gesetz über Künstliche Intelligenz) und zur Änderung bestimmter Rechtsakte der Union. https://eur-lex.europa.eu/legal-content/DE/TXT/?uri=CELEX:52021PC0206*.

European Parliament. 2021. *What is artificial intelligence and how is it used? https://www.europarl.europa.eu/news/en/headlines/society/20200827STO85804/what-is-artificial-intelligence-and-how-is-it-used*.

Institute of Pervasive Computing, JKU Linz. 2021. *Cognitive Industrial Systems: Vortrag von Univ.-Prof. Dr. Alois Ferscha. https://www.youtube.com/watch?v=YSwqAXiSuW4*.

McCarthy, John. 1998. *What is artificial intelligence? http://jmc.stanford.edu/articles/whatisai/whatisai.pdf*

PWC. 2017. *Bot.Me: A revolutionary partnership. How AI is pushing man and machine closer together. https://www.pwc.in/assets/pdfs/consulting/digital-enablement-advisory1/pwc-botme-booklet.pdf*.

Siepermann, Markus. 2021. Definition: Was ist „Künstliche Intelligenz (KI)“? *Gabler Wirtschaftslexikon. https://wirtschaftslexikon.gabler.de/definition/kuenstliche-intelligenz-ki-40285*.

Turing, Alan. 1950. Mind. *Computing machinery and intelligence* 59, Nr. 236: 433–60.

Wikipedia. 2021. *Künstliche Intelligenz. https://de.wikipedia.org/wiki/K%C3%BCnstliche_Intelligenz*.

Wikipedia. 2021. *Artificial intelligence. https://en.wikipedia.org/wiki/Artificial_intelligence*.

13 Institute of Pervasive Computing, JKU Linz, 2021.

Wie KI für Unternehmen und den gesamten Planeten von Nutzen sein kann

DIESES KAPITEL IST FÜR SIE VON INTERESSE, WENN …

… Sie in Kürze wissen wollen, warum KI für Unternehmen nützlich sein kann.

… Sie erfahren möchten, wie KI dabei helfen kann, Nachhaltigkeitsziele für unseren Planeten zu erreichen.

KI für Unternehmen

Durch Digitalisierungsprozesse, verstärkte Protokollierung und die zunehmende Vernetzung verschiedener Unternehmenseinheiten und digitaler Systeme werden enorme Mengen an Daten verfügbar gemacht. Diese auszuwerten, kann sehr gewinnbringend für ein Unternehmen sein. Die Verarbeitung solch großer Datenmengen ist jedoch nicht einfach. Denn sowohl das menschliche Gehirn als auch einfache, regelbasierte Algorithmen („Wenn A, dann B“) stoßen hier an ihre natürlichen Grenzen[1]. Eine Lösung bieten selbstlernende KI-Technologien. Diese eignen sich unter anderem bestens dafür, große Datenbanken zu analysieren, in ihnen selbstständig Muster zu erkennen, daraus Schlüsse zu ziehen und Prognosen zu erstellen.[2]

Ein KI-System kann beispielsweise in einem Vertrieb eingesetzt werden, um aus bereits getätigten Verkäufen Prognosen für Produkte und Kund*innen zu erstellen. Das System kann die Wahrscheinlichkeit von bestimmten Ereignissen vorhersagen, also zum Beispiel die Kaufwahrscheinlichkeit von Kund*in A für ein Produkt B berechnen. Das Ergebnis kann die Erfahrung und Intuition der Vertriebler*innen ergänzen, mit empirischen Kennzahlen untermauern und letztendlich die Qualität von Entscheidungen verbessern.[3]

Künstliche Intelligenz bietet also vor allem die Möglichkeit, die Effizienz und Genauigkeit von Arbeiten zu erhöhen. Bereits heute setzen Unternehmen auf eine Vielzahl von KI-Algorithmen. KI-Systeme unterstützen beispielsweise Manager*innen beim Recruiting, Devisenhändler*innen beim Trading, ITler*innen beim Abschätzen von Ausfallrisiken und Produktionsmitarbeiter*innen beim Finden von Lackschäden.

Doch Vorsicht! Oft wird das Narrativ bedient, der Einsatz von KI-Systemen werde langfristig dazu führen, Arbeitnehmer*innen von mühsamen Routinearbeiten zu entlasten und ihnen mehr Raum für spannende und kreative Tätigkeiten einzuräumen, die Maschinen nicht übernehmen können. Dass KI-Systeme im Gegenteil von manchen Unternehmen dazu eingesetzt werden, den Umsatz auf Kosten von Arbeitnehmer*innen zu erhöhen, können Sie im Kapitel Warum man bei (KI-) Algorithmen am Arbeitsplatz genauer hinsehen sollte nachlesen.

1 Bremmer, 2017.

2 Wuttke, 7. September 2020.

3 Brynjolfsson und Mitchell, 22. Dezember 2017.

Abbildung 1: Die 17 Ziele der Vereinten Nationen für nachhaltige Entwicklung. Für mehr Informationen sehen Sie unter: THE 17 GOALS | Sustainable Development (un.org). Übersetzung und Grafik von FairStyria.

KI für den Planeten

Die Agenda 2030, der Aktionsplan der Vereinten Nationen[4], umfasst 17 verschiedene Ziele für nachhaltige Entwicklung (s. Abbildung 1). Diese auch Sustainable Development Goals (SDGs) genannten Ziele sollen dabei helfen, nachhaltigen Frieden, Wohlstand und den Schutz unseres Planeten zu sichern.

Auch KI-Algorithmen sollen dabei helfen können, die Nachhaltigkeitsziele zu erreichen. Es wird geschätzt, dass mit der Unterstützung von KI-Technologien 134 von 169 Teilzielen erreicht werden könnten.[5]

- **KI und gesellschaftliche Ziele:** KI-Systeme könnten beispielsweise bei der Organisation der bedarfsorientierten Verteilung von Lebensmitteln, Wasser und Dienstleistungen helfen. Auch kann KI dafür eingesetzt werden, hochwertige Bildung bereitzustellen und saubere Energie zu liefern.
- **KI und wirtschaftliche Ziele:** Neben Produktivitätssteigerung könnte KI dabei helfen, Ursachen für Ungleichheit und Konflikte ans Licht zu bringen und diese zu minimieren. Ricardo Vinuesa und Kolleg*innen[6] schlagen beispielsweise vor, Simulationen an virtuellen Gesellschaften durchzuführen, um besser antizipieren zu können, wie sich verschiedene Maßnahmen auswirken.
- **KI und Ziele für die Umwelt:** KI-basierte Analyse und Simulation der Auswirkungen des Klimawandels können dabei helfen, rechtzeitig notwendige Maßnahmen zu setzen. Beispielsweise kann Wüstenbildung durch die Verarbeitung von Satellitenbildern frühzeitig erkannt und verhindert werden.

Doch Vorsicht! Wie jede andere Technologie auch, bietet KI nicht nur das Potential, gesetzte Ziele zu erreichen, sondern kann auch zum Verhindern von Zielen führen. Beispielsweise sollte der gesteigerte Energieverbrauch durch den Einsatz von KI-Systemen nicht außer Acht gelassen werden.

4 United Nations, 2022.
5 Vinuesa u. a., 2020.
6 Vinuesa u. a., 2020.

SIE HABEN NOCH NICHT GENUG?

Im Strategiepaper der Bundesregierung zur Künstlichen Intelligenz kann man nachlesen, wie obige Ziele in Österreichs verankert werden:

- *Strategie der Bundesregierung für Künstliche Intelligenz – AIM AT 2030*[7]

In diesem Video-Beitrag von und über das Deutsche Forschungszentrum für Künstliche Intelligenz (DFKI) wird erklärt, wie mit KI-Forschung Nachhaltigkeitsziele adressiert werden:

- *DFKI4planet – Künstliche Intelligenz für Umwelt und Nachhaltigkeit*[8]

LITERATURVERZEICHNIS

BMK. 2021. *Strategie der Bundesregierung für Künstliche Intelligenz – AIM AT 2030. https://www.bmk.gv.at/themen/innovation/publikationen/ikt/ai/strategie-bundesregierung.html.*

Bremmer, Manfred. 2017. *Robotic Process Automation: Kollege Roboter, übernehmen Sie! https://www.computerwoche.de/a/kollege-roboter-uebernehmen-sie,3331269.*

Brynjolfsson, Erik, und Tom Mitchell. 2017. What Can Machine Learning Do? Workforce Implications. *Science* 358, Nr. 6370: 1530–34. *https://doi.org/10.1126/science.aap8062.*

German Research Center for Artificial Intelligence. 2021. *DFKI4planet – Künstliche Intelligenz für Umwelt und Nachhaltigkeit. https://www.youtube.com/watch?v=LiLSZ95NEnE.*

BMK. 2021. *Strategie der Bundesregierung für Künstliche Intelligenz* – AIM AT 2030. *https://www.bmk.gv.at/dam/jcr:93f327ac-b69c-4ac7-a9aa-30eee51cc221/AIM_AT_2030_UA.pdf.*

United Nations. 2022. *THE 17 GOALS | Sustainable Development. https://sdgs.un.org/goals.*

Vinuesa, Ricardo, Hossein Azizpour, Iolanda Leite, Madeline Balaam, Virginia Dignum, Sami Domisch, Anna Felländer, Simone Daniela Langhans, Max Tegmark, und Francesco Fuso Nerini. 2020. The Role of Artificial Intelligence in Achieving the Sustainable Development Goals. *Nature communications* 11, Nr. 1: 1–10.

Wuttke, Laurenz. 2020. *Künstliche Intelligenz im Vertrieb: Vorteile, Nutzen und Anwendungsmöglichkeiten.* datasolut GmbH. *https://datasolut.com/kuenstliche-intelligenz-im-vertrieb/.*

7 BMK, 2021.

8 German Research Center for Artificial Intelligence, 2021.

Warum man bei (KI-)Algorithmen am Arbeitsplatz genauer hinsehen sollte

DIESES KAPITEL IST FÜR SIE VON INTERESSE, WENN …

… Sie wissen möchten, warum Technologie niemals neutral ist.

… Sie lernen möchten, wie Algorithmen eingesetzt werden, um Arbeitnehmende zu kontrollieren.

… Sie erfahren möchten, warum auch österreichische Betriebe nicht vor problematischen KI-Systemen gefeit sind.

Technologie: Von Steinen, Machtverhältnissen und Künstlicher Intelligenz

„Technology is neither good nor bad; nor is it neutral"[1]. Dieser Satz geht auf den Technikphilosophen Melvin Kranzberg zurück und drückt aus, dass Technologien für sich selbst keine moralische Subjektposition zugeschrieben werden kann. In anderen Worten: Technologien können keine Schuld tragen oder für etwas verantwortlich gemacht werden: „Technology is neither good nor bad." Gleichzeitig seien Technologien jedoch genauso (wenig) unparteiisch, objektiv oder neutral wie die Menschen, die sie schaffen: „nor is it neutral". Beispielsweise ist ein Stein, der als Technologie begriffen werden kann, sobald er mit einer Handlungsintention aufgehoben wird, bereits durch seine Härte nicht neutral[2]: Bei einem erfolgreichen Wurf auf eine andere Person verursacht er Schmerzen; beim Hausbau wiederum kann er zur Stabilität beitragen.

Der Technikphilosoph Langdon Winner kommt in seinem berühmten Aufsatz „Do Artifacts Have Politics?"[3] zu dem Schluss, dass Technologien unweigerlich politisch seien, weil Machtverhältnisse in Technologien eingeschrieben würden. Auch wenn dieser Aufsatz schon lange vor dem Aufkommen von KI-Assistenzsystemen verfasst wurde, hat er auch heute nichts von seiner Aktualität eingebüßt. Denn genau diese Machtverhältnisse, die im schlechtesten Fall zum Ausschluss oder der Kontrolle von Menschen führen können, erfordern es, bei Künstlicher Intelligenz genauer hinzusehen und die Wechselwirkung mit einer differenzierten Umwelt zu berücksichtigen.[4] Wie jede andere Technologie birgt Künstliche Intelligenz neben Licht- auch Schattenseiten. Auf der einen Seite hilft Künstliche Intelligenz, Arbeitsprozesse effizienter zu gestalten, auf der anderen Seite häufen sich die Berichte von KI-Systemen, die Menschengruppen gezielt ausschließen und diskriminieren. Zum einen tragen KI-Systeme dazu bei, Sicherheit in der IT, auf Flughäfen oder beispielsweise im Verkehr zu erhöhen, zum anderen werden sie von Staaten wie China zur gezielten Überwachung von Bürger*innen und zur Unterdrückung demokratischer Bestrebungen eingesetzt. Nicht nur werden KI-Systeme

1 Kranzberg, 1995.

2 Strate, 2012.

3 Winner, 1980.

4 S. auch Riesenecker-Caba und Astleithner, 2021.

immer besser im Erkennen von Tumoren in CT-Scans, sie werden auch immer wirkungsvoller im Töten von Menschen als Teil autonomer Waffensysteme.

Algorithmenbasierte Kontrolle am Arbeitsplatz

Doch bereits auch am Arbeitsplatz, bei der Einbindung von KI-Assistenzsystemen, ist es notwendig, positive und negative Konsequenzen möglichst früh zu antizipieren und diese durch konkrete Maßnahmen zu fördern oder ihnen präventiv entgegenzuwirken (s. Abschnitt C Mitbestimmung: Die KI-Systeme kommen – was können Sie tun?). Im Kapitel Wie KI für Unternehmen und den gesamten Planeten von Nutzen sein kann haben wir die Potentiale von KI am Arbeitsplatz diskutiert. Im Folgenden möchten wir auf potentielle Gefahren eingehen.

Wissenschaftlerinnen des MIT und der Stanford University haben in einem Artikel mit dem Titel „Algorithms at Work: The New Contested Terrain of Control“ analysiert[5], wie Betriebe Algorithmen einsetzen, um ihre Kontrolle über Arbeitnehmer*innen zu erhöhen. Ihre Analysen lehnten sie an ein in den 1970ern publiziertes Modell über betriebliche Methoden zur Maximierung des Wertes zu leistender Arbeit an. Dieses stellt drei zentrale Kontrollmechanismen von Arbeitgebern dar: Richtungsvorgabe, Bewertung und Disziplinierung (engl. direction, evaluation, and discipline)[6]. Diese Kontrollmechanismen finden sich auch im Kontext von Algorithmen wieder. Im Folgenden sind einige Negativbeispiele algorithmischer Kontrollmechanismen im Betrieb angeführt:

- Kontrolle via algorithmischer Richtungsvorgabe[7]
  - In einem Telekommunikationsunternehmen waren die Angestellten dazu angehalten, ihr Verhalten bei Verkaufsgesprächen auf Basis der Empfehlungen eines intransparenten Algorithmus zu treffen. Diese Empfehlungen widersprachen jedoch häufig der Expertise der Angestellten. Da ihre Vergütung provisionsbasiert war, entschieden die Angestellten sich, nur so zu tun, als ob sie der Empfehlung folgten. Dies führte zu einem innerbetrieblichen Konflikt, der in der Entlassung der jeweiligen Mitarbeiter*innen resultierte.[8] [9]
  - Die Personenbeförderungsunternehmen Uber und Lyft nutzen einen Algorithmus, der Fahrer*innen die nächste Fahrt zuweist, noch bevor die letzte geendet hat. Diese Funktion aktiviert sich automatisch bei jedem neuen Login in das Programm und muss händisch pausiert werden. Dadurch fühlen sich die Fahrer*innen getrieben und machtlos.[10]
  - Die Job-Plattform Upwork nutzt einen Chatbot, der die registrierten Freiberufler*innen an ihr Einverständnis erinnert, nicht außerhalb der Plattform zu arbeiten, sobald diese im Chatkontakt mit potentiellen Klient*innen Wörter wie „Skype“, „Telefon“ oder „E-Mail“ nutzen.[11]

  → *Arbeitnehmende, die solchen algorithmischen Richtungsvorgaben unterliegen, kommen nicht umhin, sich machtlos und manipuliert zu fühlen, da sie die Kontrolle über ihr eigenes Verhalten zu verlieren scheinen.*

- Kontrolle via algorithmischer Bewertung[12]
  - In einem großen Unternehmen für Lagerabwicklung erfasste ein mobiles Gerät die Verhaltensweisen der Angestellten, wie die Dauer einzelner Tätigkeiten oder etwaige Verspätungen. Aus den Daten wurde eine Punktzahl errechnet, die bei einem zu niedrigen Wert einen Alarm bei den Führungspersonen auslöste und diese zur Handlung aufforderte.[13] [14]
  - Ein anderes Unternehmen berechnete Punkte für Auftragnehmende, basierend auf Bewertungen durch Führungspersonen, Kolleg*innen und Kund*innen sowie gemessenen Fähigkeiten, Persönlichkeitseigenschaften und ihrer objektiven Einhaltung von Budgets und Fristen. Personen, die höhere Punktzahlen aufwiesen, hatten einen besseren Zugang zu weiteren Projekten.[15] [16]

5 Kellogg, Valentine, und Christin, 2020.
6 Edwards, 1979.
7 Kellogg, Valentine, und Christin, 2020.
8 Pachidi u. a., 2014.
9 S. auch Christl, 2021.
10 Leicht-Deobald u. a., 2019.
11 Jarrahi u. a., 2020.
12 Kellogg, Valentine, und Christin, 2020.
13 McClelland.
14 S. auch Christl, 2021.
15 Lix und Valentine, 2019. Zitiert von Kellogg, Valentine, und Christin, 2020.
16 S. auch Christl, 2021.

→ *Arbeitnehmende, die solchen algorithmischen Bewertungen unterliegen, erleben konstante Überwachung, was zumeist zum gefühlten Verlust der Privatsphäre und erheblichem Stress führt* (siehe Kapitel Assistenzsysteme zur Förderung psychischer Gesundheit am Arbeitsplatz? in Abschnitt B).

- Kontrolle via algorithmischer Disziplinierung und Belohnung[17]
  - Uber[18]-Fahrer*innen, deren durchschnittliche Bewertung durch ihre Kund*innen unter den regionalen Durchschnitt fällt (ca. 4,6 von 5 Sternen), erfahren eine plötzliche temporäre oder permanente Suspendierung[19]. Dies ist zusätzlich problematisch, weil Bewertungen auf öffentlichen Plattformen anfällig für rassistische und sexistische Einstellungen der Bewertenden sind[20].
  - Amazon Mechanical Turk, eine Online-Plattform für bezahltes Crowdsourcing[21], nutzt eine Belohnungsstruktur, die das Verhalten der Freelancer so lenkt, dass sie möglichst wenig kosten: Während die meisten Aufträge nur wenige Dollar pro Stunde einbringen, gibt es vereinzelte Aufträge, die 10–20 Dollar ausmachen und den Anreiz erhöhen, möglichst lange am Stück auf der Plattform zu arbeiten.[22]
  - Viele Unternehmen (u. a. Nike, Google, Microsoft, Amazon und Samsung) haben damit begonnen, an Computerspiele angelehnte Computerprogramme und Apps zu implementieren, um individuelle und teambezogene Arbeitserfolge (z. B. Verkäufe) abzubilden und vergleichbar zu machen[23]. Solche auch „Gamification" genannten Maßnahmen zielen darauf ab, mittels Ausschüttung von Dopamin besonders motivationsfördernd und verhaltensverstärkend zu sein.[24]

→ *Arbeitnehmende, die solchen algorithmischen Disziplinierungs- und Belohnungsmaßnahmen ausgesetzt sind, empfinden Kontrollverlust und Stress, u. a., da die Wirkungsmechanismen oft wenig nachvollziehbar sind* (siehe Kapitel Assistenzsysteme zur Förderung psychischer Gesundheit am Arbeitsplatz in Abschnitt B). *Darüber hinaus fördern die genannten Maßnahmen prekäre Arbeitsverhältnisse und innerbetriebliche Konkurrenz.*

Algorithmen in österreichischen Betrieben

Sie haben es sich sicher bereits gedacht: Bei vielen der oben genannten Beispiele handelt es sich um Fälle aus den USA. In Österreich bzw. Europa sind manche der angeführten Maßnahmen bereits aufgrund der Gesetzeslage kaum vorstellbar. Trotzdem sollte man sie nicht außer Acht lassen.

> Es besteht nämlich die unmittelbare Gefahr, dass solche algorithmischen Kontrollmaßnahmen durch die Hintertür Einzug in österreichische Betriebe erhalten.[25]

Zwar werden importierte KI-Systeme üblicherweise so angepasst, dass sie den hiesigen Gesetzen unterliegen. Einerseits schützt jedoch auch die aktuelle Gesetzeslage nicht immer vor negativen Konsequenzen für Arbeitnehmende[26] und andererseits werden Gesetze auch nicht immer befolgt. Ein gutes Beispiel hierfür ist Software, die bereits von vielen österreichischen Betrieben eingesetzt wird: Microsoft 365.

Microsoft 365 ermöglicht seinen Nutzer*innen, über das Tool „Microsoft Viva Insights" (formals Microsoft MyAnalytics) Einblick in die eigene Arbeitsweise zu erhalten. Wöchentliche Berichte geben Einsicht in Kennzahlen wie die Dauer von Besprechungen oder die benötigte Zeit für das Schreiben von E-Mails.[27] Auch wird man darauf hingewiesen, wenn man es wagt, E-Mails während Besprechungen zu formulieren.[28] In Europa ist

17 Kellogg, Valentine, und Christin, 2020.

18 Uber ist eine US-amerikanische Firma, die über ein Online-Portal Vermittlungen für Personenbeförderungen anbietet. Die gesetzlichen Rahmenbedingungen variieren je nach Land.

19 Rosenblat u. a., 2017.

20 Z. B. Kauff u. a., 2021.

21 Beim sog. Crowdsourcing werden Aufgaben an (viele) Personen außerhalb eines Unternehmens vergeben – oftmals Internet-User*innen – und zu geringeren Kosten und ohne entsprechenden Arbeitnehmer*innenschutz umgesetzt.

22 Lehdonvirta, 2018.

23 Kim, 2018.

24 Pacewicz, 2015.

25 S. auch Christl, 2021.

26 Ein europäisches KI-Gesetz ist erst im Entstehen: EU, 2021.

27 madehmer und olprod, 2022.

28 Angerler, Heiling, und Chlestil, 2021.

die Verarbeitung solcher personenbezogener Daten von der Datenschutzgrundverordnung (DSGVO) geschützt (s. auch Abschnitt C, Kapitel Gesetzliche Rahmenbedingungen für den Einsatz von KI-Assistenzsystemen). Vorgesetzte können die jeweiligen Informationen nur in anonymisierter, aggregierter Form über ganze Gruppen hinweg abrufen. Trotzdem ist hier Vorsicht geboten, denn in der Praxis könnten Vorgesetzte einfach um die „freiwillige" Vorlage der wöchentlichen Viva-Insights-Berichte bitten – und im Herrschaftsverhältnis von Arbeitgebenden und Arbeitnehmenden ist Freiwilligkeit doch ein relatives Konzept.

Dass man auch in Österreich genauer hinsehen sollte, wenn Algorithmen am Arbeitsplatz implementiert werden, zeigt eine von der Arbeiterkammer geförderte Studie von Wolfie Christl mit dem Titel „Digitale Überwachung und Kontrolle am Arbeitsplatz"[29], aus der wir zwei Beispiele über eingesetzte Smartphone-Apps mitgebracht haben:

- Betriebsrät*innen eines Unternehmens für Montage und Wartung im Anlagenbau berichteten, dass die Beschäftigten eine Smartphone-App nutzen müssen, die alle Bewegungen, Zeiträume und Arbeitsschritte im Außendienst dokumentiere und darüber hinaus KI-gesteuerte Vorgaben für kommende Arbeitsschritte mache. Die für die Arbeitsschritte benötigte Zeit werde in wöchentlichen Gesprächen mit Vorgesetzten minutengenau diskutiert. Entsprechend führe die Smartphone-App bei den Beschäftigten zur Reduktion von Autonomie, Selbstbestimmung und Sinnstiftung, Beschleunigung und Verdichtung von Arbeit und Kontrollerleben.[30]

- Ein/e ehemalige Betriebsrät*in eines österreichischen Essenszustelldiensts durch Fahrradbot*innen berichtete von automatisierter Leistungssteuerung und Sanktionierung per Smartphone-App. Die Leistung der Fahrradbot*innen (reguläre Angestellte, freie Dienstnehmer*innen sowie selbstständige Subunternehmer*innen), d. h. die Schnelligkeit und Häufigkeit von Zustellungen, würde anhand von mittels Smartphone gesammelten Daten bestimmt und in einem Gesamtranking abgebildet. Die bestgereihten Bot*innen bekämen als Erstes die Chance, Schichten für die Folgewoche auszuwählen und somit die Höhe des potentiellen Verdienstes zu beeinflussen.[31]

SIE HABEN NOCH NICHT GENUG?

Für weitere spannende Beispiele und einen ausgezeichneten Überblick über algorithmische Kontrolle von Arbeitstätigkeiten empfehlen wir die Studie

- *Digitale Überwachung und Kontrolle am Arbeitsplatz. Von der Ausweitung betrieblicher Datenerfassung zum algorithmischen Management?* von Wolfie Christl[32]

sowie den englischsprachigen Literaturüberblick

- *Artificial Intelligence at Work: An Overview of the Literature* von Didem Özkiziltan und Anke Hassel[33]

Wenn Sie sich darüber hinaus weiter dafür interessieren, warum Technologien nicht neutral sind und welche gesellschaftlichen Auswirkungen KI hat, sehen Sie dieses YouTube-Video an: Ein Wissenschaftsgespräch mit

- *Helga Nowotny über ihr Buch „In AI We Trust"*[34]

LITERATURVERZEICHNIS

Angerler, Eva, Michael Heiling, und Martina Chlestil. 2021. Es geht nur gemeinsam. Wie künstliche Intelligenz im Betrieb Beschäftigten (nicht) hilft. *Arbeit&Wirtschaft Blog. https://awblog.at/kuenstliche-intelligenz-im-betrieb/.*

Bruno Kreisky Forum für internationalen Dialog. 2022. Helga Nowotny: IN AI WE TRUST. *https://www.youtube.com/watch?v=-ZiCvYmQTfU.*

Christl, Wolfie. 2021. *Digitale Überwachung und Kontrolle am Arbeitsplatz. Von der Ausweitung betrieblicher Datenerfassung zum algorithmischen Management? https://crackedlabs.org/dl/CrackedLabs_Christl_UeberwachungKontrolleArbeitsplatz.pdf.*

Edwards, Richard. 1979. *Contested Terrain: The Transformation of the Workplace in the Twentieth Century.* New York: Basic Books.

29 Christl, 2021.
30 Christl, 2021.
31 Christl, 2021.
32 Christl, 2021.
33 Özkiziltan und Hassel, 2021.
34 Bruno Kreisky Forum für internationalen Dialog, 2022.

EU. 2021. *Proposal for a Regulation of the European parliament and of the council laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts,* EUR-Lex – 52021PC0206 – EN §. *https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1623335154975&uri=CELEX%3A52021PC0206*.

Jarrahi, Mohammad Hossein, Will Sutherland, Sarah Beth Nelson, und Steve Sawyer. 2020. Platformic Management, Boundary Resources for Gig Work, and Worker Autonomy. *Computer Supported Cooperative Work (CSCW)* 29, Nr. 1: 153–89. *https://doi.org/10.1007/s10606-019-09368-7*.

Kauff, Mathias, Julian Anslinger, Oliver Christ, Moritz Niemann, Michaela Geierhos, und Lars Huster. 2021. Ethnic and gender-based prejudice towards medical doctors? The relationship between physicians' ethnicity, gender, and ratings on a physician rating website. *The Journal of Social Psychology,* 1–9.

Kellogg, Katherine C., Melissa A. Valentine, und Angéle Christin. 2020. Algorithms at Work: The New Contested Terrain of Control. *Academy of Management Annals* 14, Nr. 1: 366–410. *https://doi.org/10.5465/annals.2018.0174*.

Kim, Tae Wan. 2018. Gamification of Labor and the Charge of Exploitation. *Journal of Business Ethics* 152, Nr. 1: 27–39. *https://doi.org/10.1007/s10551-016-3304-6*.

Kranzberg, Melvin. 1995. Technology and History: "Kranzberg's Laws". *Bulletin of Science, Technology & Society* 15, Nr. 1: 5–13. *https://doi.org/10.1177/027046769501500104*.

Lehdonvirta, Vili. 2018. Flexibility in the Gig Economy: Managing Time on Three Online Piecework Platforms. *New Technology, Work and Employment* 33, Nr. 1: 13–29. *https://doi.org/10.1111/ntwe.12102*.

Leicht-Deobald, Ulrich, Thorsten Busch, Christoph Schank, Antoinette Weibel, Simon Schafheitle, Isabelle Wildhaber, und Gabriel Kasper. 2019. The Challenges of Algorithm-Based HR Decision-Making for Personal Integrity. *Journal of Business Ethics* 160, Nr. 2: 377–92. *https://doi.org/10.1007/s10551-019-04204-w*.

Lix, Katharina, und Melissa A. Valentine. 2019. *Kharma scores and team learning in software development gigs.* Unpublished Manuscript.

madehmer, und olprod. 2022. *Week in the life metrics in Workplace Analytics. https://docs.microsoft.com/de-de/viva/insights/use/explore-metrics-week-in-the-life*.

McClelland, Mac. "I Was a Warehouse Wage Slave": Reporter Mac McClelland on Life Inside the Online Shipping Machine. Democracy Now! *http://www.democracynow.org/2013/8/7/i_was_a_warehouse_wage_slave*.

Özkiziltan, Didem, und Anke Hassel. 2021. Artificial Intelligence at Work: An Overview of the Literature. *SSRN Electronic Journal. https://doi.org/10.2139/ssrn.3796746*.

Pacewicz, Krzysztof. 2015. The dopamine loop and its discontents. Analysis of "Gamification by design" as biopolitical power/knowledge. In *Gamification. Critical Approaches,* herausgegeben von Jarosław Kopeć und Krzysztof Pacewicz, 27–39. *https://depot.ceon.pl/bitstream/handle/123456789/8013/Gamification-Critical-Approaches.pdf?sequence=1&isAllowed=y#page=27*.

Pachidi, Stella, Hans Berends, Samer Faraj, Marleen Huysman, und Inge van de Weerd. 2014. What Happens When Analytics Lands in the Organization? Studying Epistemologies in Clash. *Academy of Management Proceedings* 2014, Nr. 1: 15590. *https://doi.org/10.5465/ambpp.2014.15590abstract*.

Riesenecker-Caba, Thomas, und Franz Astleithner. 2021. *Verarbeitung personenbezogener Beschäftigtendaten und Grenzen betrieblicher Mitbestimmung in einer digitalisierten Arbeitswelt*. Forba. *https://www.forba.at/wp-content/uploads/2021/06/Verarbeitung-persbez-Daten-und-MitbestimmungFORBA-Bericht2021_DigiFonds.pdf*.

Rosenblat, Alex, Karen E.C. Levy, Solon Barocas, und Tim Hwang. 2017. Discriminating Tastes: Uber's Customer Ratings as Vehicles for Workplace Discrimination. *Policy & Internet* 9, Nr. 3: 256–79. *https://doi.org/10.1002/poi3.153*.

Strate, Lance. 2012. If It's Neutral, It's Not Technology. *Educational Technology* 52, Nr. 1: 6–9.

Winner, Langdon. 1980. Do Artifacts Have Politics? *Daedalus* 109, Nr. 1: 121–36.

Technologiegestaltung: Wie sollten KI-Assistenzsysteme (nicht) gestaltet werden

Technologieentwicklung mit Nutzer*innen – warum denn das?

DIESES KAPITEL IST FÜR SIE VON INTERESSE, WENN …

… Sie schon immer wissen wollten, warum Technologien nach ihrer Markteinführung von manchen Nutzer*innen abgelehnt oder anders als geplant verwendet werden.

… Sie denken, dass Technik neutral ist und Geschlecht oder soziale Herkunft keine Rolle spielen, weder bei der Entwicklung noch bei der Nutzung von Technologien.

… Sie wissen, dass Technologien dazu beitragen können, gesellschaftliche Ungleichheiten zu verstärken und Sie dies bei der Entwicklung von KI-Assistenzsystemen berücksichtigen wollen.

… Sie Fakten brauchen, um die Einbindung von Arbeitnehmer*innenvertretungen in die Implementierung eines KI-Assistenzsystems im Betrieb zu argumentieren.

Einbezug von Nutzer*innen und Technikfolgenabschätzung

Rückmeldungen über eine Technologie erst nach der Markteinführung einzuholen, kann Technikentwicklungen verteuern. So kann widerständiges Nutzungsverhalten aufgrund unausgereifter Produkte oder nicht berücksichtigter Anwendungsmöglichkeiten dazu führen, dass die Technologieentwicklung wieder aufgenommen werden muss. Beispielsweise wurde die Textnachrichtfunktion („SMS“) von Mobiltelefonen zunächst unterschätzt und erst aufgrund von entsprechendem Nutzer*innenverhalten weiter ausgebaut. Dies führte zu einer Fülle von text- und später auch bildbasierten Nachrichtenfunktionen heutiger Smartphones.

Dass auch die ursprüngliche Funktion von Mobiltelefonen, das Telefonieren, an Bedeutung verloren hat, hätten aber Nutzer*innen der 1990er-Jahre noch nicht wissen können. Hier zeigt sich, dass neben Rückmeldungen über die Technologienutzung auch Umweltfaktoren, wie zum Beispiel entsprechende Tarifkonditionen (günstige SMS vs. teureres Telefonieren in den Anfangsjahren der Mobiltelefonie) zur Fortentwicklung von Technologien beitragen. Solche gesellschaftlichen, aber auch ethischen und ökologischen Auswirkungen von Technik werden im Forschungsfeld der Technikfolgenabschätzung behandelt. Technikfolgenabschätzung zielt darauf ab, Potentiale und Risiken von Technologien zu beurteilen.

Fehleranfällige Flur-Test-Methode

Die mangelnde Einbindung von Nutzer*innen kann darüber hinaus zu fehleranfälligen Technologieentwicklungen führen. Zum Beispiel führten auf Stimmen reagierende automatische Garagentore, die nur Stimmen in einem bestimmten, tieferen Frequenzbereich erkannten, dazu, dass Nutzer*innen mit höheren Stimmen – vielfach Frauen – trotz korrektem Codewort vor sprichwörtlich verschlossenen Türen standen. Was war hier passiert? Der Sensor bzw. der damit gekoppelte Öffnungsmechanismus für das Garagentor wurde nicht mit einer Bandbreite an Stimmfrequenzen trainiert.

Solche Technologieentwicklungen vorbei an tatsächlichen Anwender*innen haben ihre Ursprünge in der sogenannten Ich-Methode („I-Methodology"). Dabei gehen Technikentwickler*innen zum einen von sich selbst und ihren Lebensumständen aus und konstruieren Lösungen für Probleme, die für sie relevant sind. Ergänzt wird diese Methode in der Technikforschung von Rückmeldungen, die von Kolleg*innen am Flur eingeholt werden („Hallway-Testing Method"[1]).

DIE „ICH-METHODE" IN KÜRZE

Ein Beispiel für die „Ich-Methode"[2] aus der Berufswelt ist die Einführung von Computersystemen in Callcentern. Beispielsweise konnte eine deutsche Studie aufzeigen, dass die in Callcentern eingesetzten Programme hauptsächlich an Sachaufgaben, wie dem Erfassen von Daten, orientiert waren. Menschen, die bei einer Hotline anrufen, beispielsweise, um eine Reise zu buchen oder ein Problem mit ihrem Mobiltelefon zu lösen, wollen jedoch in erster Linie ihr Anliegen bearbeiten und nicht damit aufgehalten werden, durch das System geforderte Daten durchzugeben. „Emotionsarbeit" (engl. Emotional Labour) wie das Beruhigen und Abholen der anrufenden Menschen durch die Callcenter-Agents wurde in den Computerprogrammen jedoch nicht angemessen berücksichtigt. Die Autor*innen der Studie kamen deshalb damals zum Schluss: „Es fehlt den EntwicklerInnen offensichtlich ein ausreichendes Wissen über die Erfordernisse der Arbeitssituation".[3]

Das Konzept der Ich-Methode geht davon aus, dass Technik-Entwickler*innen nicht ausreichend mit den Nutzer*innen ihrer Technologien kommunizieren, sondern hauptsächlich versuchen, sich selbst in die Rolle der Nutzenden hineinzuversetzen. Das Problem dabei ist, dass damit oft unbewusst von den Anforderungen und Bedürfnissen der eigenen Lebenswelten auf die der Nutzenden geschlossen wird[4]. Vor dem Hintergrund, dass die Technologiebranche weiterhin wenig divers ist und Teams überwiegend aus jungen, männlichen, gut gebildeten Technologieentwickler*innen bestehen, wirkt sich die Ich-Methode negativ auf die Kompatibilität von Technologien mit denjenigen aus, die diesen Kategorien nicht entsprechen.

Mangelnde Diversität in der Technikbranche

Die Technik- und insbesondere Informations- und Kommunikationstechnologie-Branche weist eine große Homogenität auf. Sowohl in Ausbildungsberufen, Höheren Technischen Lehranstalten (HTL) als auch in technischen Studiengängen an Fachhochschulen und Universitäten lernen und lehren überwiegend männliche Personen. An Österreichs größter technischer Universität, der TU Wien, studieren rund 85% männliche Studierende Computerwissenschaften[5]; im Bereich der Künstlichen Intelligenz sind weltweit nur rund 12 Prozent der Forscher*innen weiblich[6].

Zudem sind Arbeiterkinder und Menschen mit nichtdeutscher Erstsprache in höheren Bildungsinstitutionen in Österreich generell unterrepräsentiert. So schließen 57% der jungen Erwachsenen aus Akademiker*innenhaushalten ein Hochschulstudium ab, während dies nur 7% aus Elternhäusern mit Pflichtschulabschluss gelingt[7]. Zudem wählen Personen ohne Migrationshintergrund zu 49% ein Studium, während nur rund 20% Personen aus der ersten und zweiten Generation diesen Bildungsweg einschlagen[8]. Das bedeutet, dass es in der Gruppe der akademisch ausgebildeten Techniker*innen weniger Frauen, weniger Menschen mit Migrationshintergrund und weniger Menschen aus Arbeiter*innenhaushalten gibt als in der übrigen Gesamtbevölkerung.

Folglich sind Technikentwickler*innen eine homogenere Gruppe und weniger divers als die Nutzer*innen, für die sie technologische Innovationen produzieren und implementieren.

1 Burtscher, 2019.
2 Oudshoorn, Rommes, und Stienstra, 2004.
3 Maaß, Theissing, und Zallmann, 2001.
4 Oudshoorn, Rommes, und Stienstra 2004.
5 Burtscher, 2019.
6 UNESCO, 2019.
7 Statistik Austria, 2018.
8 Die Prozentangaben errechnen sich durch die Gegenüberstellung von Studienanfänger*innen zum Anteil der Bevölkerung im selben Alter, man spricht hier von der „Hochschulzugangsquote", siehe Unger und Kolleg*innen, 2019, S. 91 ff.

Darüber hinaus fallen manche Menschen, und damit auch Technikanwender*innen, in mehrere der angegebenen Kategorien (z. B. weibliche Arbeiter*innenkinder). Es hat sich gezeigt, dass es an solchen „sozialen Kreuzungen“ gehäuft zu Ausschlüssen und Diskriminierung kommt, sowohl gesamtgesellschaftlich als auch in Hinblick auf die Technikentwicklung. Beispielsweise wiesen Gesichtserkennungssoftwares Frauen mit dunklerer Hautfarbe öfters fälschlicherweise als Männer aus als Frauen mit hellerer Hautfarbe[9]. Bei der Technikentwicklung ist es somit nicht nur wichtig, einzelne soziale Kategorien zu berücksichtigen, sondern auch, die wechselseitigen Einflüsse der jeweiligen Kategorien miteinzubeziehen. Dieser Ansatz wird in der sozialwissenschaftlichen Forschung als Intersektionalitätsanalyse bezeichnet (siehe Infokasten).

INTERSEKTIONALITÄT

Die auf Kimberlé Crenshaw zurückgehende Metapher der Straßenkreuzung (engl. intersection) nimmt ineinandergreifende Strukturen von Ungleichheit in den Blick und macht mehrfach diskriminierte Gruppen sichtbar. Als sogenannte (Ausschluss-)Kategorien gelten soziale und ethnische Herkunft, Geschlecht, Alter, sexuelle Orientierung, etc., die nicht selten als Mehrfachidentitäten zusammenfallen und somit strukturelle Diskriminierungserfahrungen noch potenzieren können (mehr dazu unter: *https://www.politische-medienkompetenz.de/debatte/ki-und-intersektionalitaet/*).

In der Studie “Gender Shades”[10] wurde beispielsweise demonstriert, dass drei große marktübliche Gesichtserkennungsprogramme Frauen mit dunklerem Hautton am öftesten unkorrekt klassifizierten (im Vergleich zu Männern, gleich welchen Hauttyps bzw. im Vergleich zu Frauen mit hellerem Hautton), und diese somit in bis zu 34,7% der Fälle fälschlicherweise nicht als weiblich identifizierten.

Zusammenfassend führt mangelnde Diversität in der Technik dazu, dass unter anderem vermittelt über das unbewusste Anwenden der Ich-Methode und die Einholung von Rückmeldungen „am Flur“ häufiger privilegierte, männliche, akademische Sichtweisen in Technologieentwicklungen einfließen und diversere Perspektiven unberücksichtigt bleiben.

Soziale, ökologische und gesellschaftliche Aspekte

Lange Zeit galten soziale, ökologische und gesellschaftliche Aspekte von Technik als nicht notwendige Zusatzfächer in der Technikausbildung. Mit großen technologiebasierten Krisen, wie zum Beispiel Atomreaktorunfällen, aber auch gesellschaftlichen Debatten über Gentechnik oder die BSE-Krise[11], wurde die sogenannte Technikfolgenabschätzung auch von Medien und Politik als relevant erkannt, und daraus Forderungen nach einer entsprechenden Anreicherung von Technikbildung abgeleitet. (Diplom-)Ingenieur*innen sollten nicht mehr nur das Rüstzeug für technologische Entwicklungen vermittelt bekommen, sondern auch ihre Verantwortung innerhalb der Gesellschaft reflektieren.

Leider sind die entsprechenden wenigen Lehrveranstaltungen in österreichischen Technikstudien oftmals frei wählbare Fächer. Viele Technikstudierende bekommen in ihrem Studium lediglich einen ersten kleinen Einblick und kaum Werkzeuge für die konkrete Umsetzung für den Einbezug sozialer, ökologischer und gesellschaftlicher Aspekte. Ein Beispiel für ein solches Werkzeug finden Sie im Infokasten, wo wir mit dem GERD-Modell ein Tool vorstellen, mit dem informationstechnische Fragestellungen mit soziotechnischen Perspektiven erweitert werden können.

9 Buolamwini und Gebru, 2018.

10 Buolamwini und Gebru, 2018.

11 Bovine Spongiforme Enzephalopathie, umgangssprachlich „Rinderwahn“ genannt (siehe *https://www.verbrauchergesundheit.gv.at/tiere/krankheiten/bse.html*).

MIT GERD ZU EINER BESSEREN TECHNIKENTWICKLUNG

Ein Ansatz zur Herstellung von Gerechtigkeit und Fairness entstammt aus einer Forschungsrichtung, die IT- mit Gender- und Diversitätsforschung verknüpft. Mit der Unterstützung von Fachexpert*innen entwickelten Wissenschaftler*innen das sogenannte GERD-Modell (Gender Extended Research and Development-Modell[12]). Das Modell kann dabei helfen, Forschungs- und Entwicklungsprozesse in der Informatik um eine soziotechnische Perspektive zu erweitern, also einen Blickwinkel, der die gesellschaftlich-kulturellen Einflüsse auf und die Auswirkungen von technischen Systemen berücksichtigt. Obwohl der Begriff „Gender“ bereits im Namen des Modells steckt, zielt es gleichermaßen auf die nähere Betrachtung weiterer Merkmale und Aspekte ab, die für die Herstellung oder Aufrechterhaltung von Gerechtigkeit und Fairness von zentraler Bedeutung sein können (z. B. Alter, Ethnizität oder Macht- und Hierarchieverhältnisse). Zentrales Element des Modells ist die angeleitete Bewusstseinsschaffung über die soziotechnischen Aspekte des jeweiligen informationstechnischen Betrachtungsgegenstands (z. B. eines KI-Algorithmus). Das GERD-Modell schlägt hierfür konkrete Reflexionsaspekte mit Reflexionsfragen vor.

Das Besondere an dem Modell ist, dass es sich in der Anwendung nahe an den Kernprozessen und Teilaufgaben der Informatikforschung und -entwicklung bewegt. Es bietet eine prozessorientierte Hilfestellung, die flexibel zu beliebigen Zeitpunkten während der Entwicklung in Anspruch genommen werden kann. Für einen möglichst gewinnbringenden Einsatz des Modells möchten wir die Prozessbegleitung durch eine Person mit entsprechender Gender- und Diversity-Perspektive empfehlen.

Entlang der einzelnen GERD-Kategorien können folgende Fragen zum Denken anregen:

1. Relevanz
   a. Welche gesellschaftlichen Bereiche wurden bislang bevorzugt behandelt? Welche Bereiche wurden vernachlässigt?
   b. Wie wird mit gesellschaftlich marginalisierten Themen umgegangen?

2. Nutzen
   a. Wem soll die Forschung nutzen?
   b. Welche Bevölkerungsgruppen werden mitgedacht? Schließen die Technologien evtl. bestimmte Bevölkerungsgruppen aufgrund sozialer Kategorien wie z. B. Alter oder körperlicher Befähigung aus? Wenn ja, welche und warum?

3. Wissen
   a. Wer gilt als Expert*in?
   b. Gibt es zu dem Forschungsfeld relevante Arbeiten aus den Gender- und Diversity Studies, Disability Studies/Behindertenpädagogik, Gesundheits- und Pflegewissenschaften o. Ä.?

4. Werte
   a. Welche Wertvorstellungen und Grundeinstellungen unterliegen Forschung und Entwicklung? Z. B., was Themenbereiche betrifft wie:
      - Privacy, Security
      - Hierarchien, Verteilung von Arbeit
      - Gleichberechtigung und Teilhabe
      - Mitgestaltungsmöglichkeiten der Nutzenden
      - Nachhaltigkeit, ökologische Verantwortung
   b. Für welche Personen ist das technische System durchschaubar? Wie verständlich ist es? Wer kann in technische Prozesse eingreifen?

5. Machtverhältnisse
   a. Für welche Tätigkeiten oder Anwendungsbereiche wird ein Forschungs- oder Entwicklungsvorhaben geplant?
   b. Wer produziert die Technik und unter welchen Bedingungen?
   c. Wer kann sich die Technik leisten?

6. Menschenbild
   a. Welche und wessen Lebenswelt, welche Alltagszusammenhänge werden aufgegriffen?
   b. Welche Vorannahmen über Geschlecht fließen in das Forschungsfeld mit ein? Wie bestimmen diese das Forschungsfeld? Wie sind sie begründet?

12 Draude, Maaß, und Wajda, 2014.

7. Arbeitskultur
   a. Werden soziale Faktoren und ihr Einfluss auf die Arbeitskultur als wichtig erachtet?
   b. Lassen sich besondere persönliche Situationen, wie z. B. Elternschaft, Pflege Angehöriger, besondere körperliche oder seelische Befähigungen, mit der Arbeit und den Arbeitszeiten in Einklang bringen?

8. Sprache
   a. Wird geschlechtergerechte und nicht-rassistische Sprache benutzt?
   b. Wird mit Stereotypen gearbeitet oder wird versucht, Vielfalt herzustellen und herkömmliche (Geschlechter-)Erzählungen zu durchbrechen?

SIE HABEN NOCH NICHT GENUG?

In diesem Video erklärt die Informatikerin Joy Buolamwini, warum es bei Gesichtserkennungs-Algorithmen zu Diskriminierung von Frauen mit dunklerem Hautton kommt und wie sie dagegen vorgeht:

- *How I'm fighting bias in algorithms*[13]

Wenn Sie sich darüber hinaus näher für den intersektionalen Ansatz interessieren, empfehlen wir den folgenden Audiobeitrag:

- *Podcast über Intersektionalität und deren Bedeutung für die Bildungsarbeit*[14]

ANSPRECHPERSON

MMag.a Dr.in Anita Thaler, IFZ
Expertin für gender- und diversitysensible Technologiegestaltung
Anita.Thaler@ifz.at

LITERATURVERZEICHNIS

Buolamwini, Joy. How I'm fighting bias in algorithms. TEDxBeaconStreet, 2016. *https://www.ted.com/talks/joy_buolamwini_how_i_m_fighting_bias_in_algorithms/transcript*.

Buolamwini, Joy, und Timnit Gebru. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. *In Proceedings of the 1st Conference on Fairness, Accountability and Transparency,* FAT 2018, 23-24 February 2018, New York, NY, USA, herausgegeben von Sorelle A. Friedler und Christo Wilson, 81:77–91. Proceedings of Machine Learning Research. PMLR. *http://proceedings.mlr.press/v81/buolamwini18a.html*.

Burtscher, Sabrina. 2019. *Literature Review: Gender Research in Human Computer Interaction.* GENDER EQUALITY IN ENGINEERING THROUGH COMMUNICATION AND COMMITMENT (GEECCO). *http://www.geecco-project.eu/fileadmin/t/geecco/geecco/material/GEECCO_WP6_Literatur_Review_Gender_research_in_HCI.pdf*.

Draude, Claude, Susanne Maaß, und Kamila Wajda. 2014. *Gender-/Diversity-Aspekte in der Informatikforschung: Das GERD-Modell.* 67–77.

Gunda Werner Institut, Heinrich Böll Stiftung. 2019. Feminismus Backstage #002: Intersektionalität. *https://soundcloud.com/boellstiftung/feminismus-backstage-002-intersektionalitat*.

Maaß, Susanne, Florian Theissing, und Margita Zallmann. 2001. Computereinsatz und Arbeitsgestaltung in Call-Centern. In *Mensch und Computer 2001. https://doi.org/10.1007/978-3-322-80108-1_8*.

Oudshoorn, Nelly, Els Rommes, und Marcelle Stienstra. 2004. Configuring the User as Everybody: Gender and Design Cultures in Information and Communication Technologies. *Science, Technology & Human Values* 29 (Dezember). *https://doi.org/10.1177/0162243903259190*.

Statistik Austria. 2018. *Vererbung von Bildungschancen.* Dezember 2018. Statistics Brief. *http://www.statistik.at/wcm/idc/idcplg?IdcService=GET_PDF_FILE&dDocName=119813*.

Unger, Martin, Binder, David, Dibiasi, Anna, Engleder, Judith, Schubert, Nina, Terzieva, Berta, Thaler, Bianca, Zaussinger, Sarah, und Zucha, Vlasta. 2019. *Studierenden-Sozialerhebung 2019*. Projektbericht. Wien: IHS. Zugegriffen 31. März 2022. *http://www.sozialerhebung.at/images/Berichte/Studierenden-Sozialerhebung_2019_Kernbericht.pdf*.

United Nations Educational, Scientific and Cultural Organization, UNESCO. 2019. *I'd blush if I could: closing gender divides in digital skills through education. https://unesdoc.unesco.org/ark:/48223/pf0000367416.page=1*.

13 Buolamwini, (TEDxBeaconStreet, 2016).

14 Gunda Werner Institut, 2019.

Assistenzsysteme zur Förderung psychischer Gesundheit am Arbeitsplatz?

DAS FOLGENDE KAPITEL IST FÜR SIE VON INTERESSE, WENN …

… Sie mehr über digitalen Stress und psychische Gesundheit am Arbeitsplatz erfahren wollen.

… Sie sich schon immer gefragt haben, wie psychische Belastungen am Arbeitsplatz reduziert werden können.

… Sie wissen wollen, welche gesetzlichen Verpflichtungen Arbeitgeber zum Erhalt der psychischen Gesundheit von Arbeitnehmer*innen haben.

… Sie Assistenzsysteme zur Stressprävention im Betrieb einsetzen wollen.

Informations- und Kommunikationstechnologien sind von vielen Arbeitsplätzen nicht mehr wegzudenken. Die negativen Auswirkungen sind unter dem Schlagwort „digitaler Stress“ bekannt. Dazu zählen das Gefühl der ständigen Erreichbarkeit, die Entgrenzung zwischen Arbeits- und Privatleben und der subjektiv erlebte Druck, durch immer mehr und schneller bereitgestellte Informationen auch selbst immer schneller arbeiten zu müssen. Doch Technologien müssen nicht automatisch Stress verursachen. Wenn Technologien gezielt eingesetzt werden und Arbeitnehmende gut mit ihnen umgehen können, erleben sie weniger Verunsicherung und damit weniger digitalen Stress[1]. Die kontinuierliche Evaluierung von psychischen Belastungen am Arbeitsplatz gehört seit 2013 zu den Pflichten von Arbeitgebern.

Evaluierung psychischer Belastungen am Arbeitsplatz

Seit der Novelle des Arbeitnehmerschutzgesetzes von 2013[2] ist, neben der physischen, auch die psychische Gesundheit von Mitarbeiter*innen in die Aufmerksamkeit von österreichischen Betrieben gerückt. Damit geht auch die Prävention arbeitsbedingter psychischer Belastungen einher, die zu Fehlbeanspruchungen wie Stress oder psychischer Ermüdung führen.

Laut Gesetz sind alle Arbeitgeber verpflichtet, einen Arbeitsplatzevaluierungsprozess zu implementieren. Dieser besteht aus der Erhebung und Beurteilung von Gefahren, abgeleiteten Maßnahmen und Kontrolle derer Wirksamkeit und Dokumentierung.

1 Gimpel, Lanzl, Regal, u. a., 2019.

2 ASchG-Novelle, BGBl. Nr. I 118/2012, abrufbar unter: *https://www.ris.bka.gv.at/eli/bgbl/I/2012/118/20121228*.

Folgende Gefahren sind Beispiele aus der Arbeitswelt, die zu arbeitsbedingten psychischen Belastungen führen können:

- Anhaltender Zeit- und Termindruck
- Parallele Arbeitsanforderungen (Multitasking)
- Informationsmangel oder -überflutung
- Knappe Personalressourcen
- Entgrenzung von Arbeit und Freizeit (z. B. im Homeoffice)
- Angst vor Arbeitsplatzverlust
- Fehlende Spielräume und Beteiligung in der Arbeit
- Fehlende soziale Kontakte in der Arbeit

Nach erfolgter Arbeitsplatzevaluierung werden gemeinsam mit den betroffenen Beschäftigten geeignete Maßnahmen festgelegt, um zum Beispiel Arbeitsabläufe, -anforderungen, -organisation (inklusive Gestaltung der Arbeitszeit), soziale Beziehungen sowie die Arbeitsumgebung zu verbessern.

Eine erfolgreiche Arbeitsplatzevaluierung hat Prozesscharakter und analysiert laufend relevante Daten, um psychische Belastungen am Arbeitsplatz konsequent zu reduzieren.

GESUNDHEIT UND SICHERHEIT AM ARBEITSPLATZ IN EUROPA

Innerhalb der EU regelt die Rahmenrichtlinie 89/391/EWG die Verbesserung der Sicherheit und des Gesundheitsschutzes von Arbeitnehmer*innen. Die Rahmenrichtlinie umfasst Bestimmungen zur Gestaltung von Arbeitsplätzen, der Auswahl von Arbeitsmitteln und Arbeits- und Fertigungsverfahren. Die Richtlinie zielt vor allem auf eine Erleichterung bei eintöniger Arbeit und maschinenbestimmtem Arbeitsrhythmus sowie auf die Gefahrenverhütung ab. Des Weiteren steht die kohärente Verknüpfung von Technik, Arbeitsorganisation und -bedingungen, sozialen Beziehungen und der Umwelt im Mittelpunkt. ▶

▶ Theoretisch sind drei Arten von Interventionen möglich:

- Primäre Interventionen zielen auf Prävention ab und versuchen psychosoziale Risiken bereits an der Quelle zu vermeiden, z. B. durch sinnvolle Gestaltung von Schichtplänen und Verbesserungen der Kommunikation und des physischen Arbeitsumfelds.
- Sekundäre Interventionen setzen daran an, die Fähigkeit von Arbeitnehmer*innen zur Wahrnehmung und Bewältigung von psychosozialen Risiken und Gefahren zu verbessern, z. B. Schulungen in allgemeinen Bewältigungskompetenzen.
- Tertiäre Interventionen beinhalten die Behandlung von Symptomen und sind häufig Teil von Rehabilitationsprozessen, z. B. Behandlungen psychischer Erkrankungen wie Burnout, Depression oder Angst. Arbeitsstätten können durch Arbeitsunfähigkeitsmanagement die Rückkehr an den Arbeitsplatz unterstützen.

Die Wirksamkeit der Interventionen hängt davon ab, wie sie organisiert werden. So erhöhen

1. Mitarbeiter*innenbeteiligung
2. Evaluierungen und das Zuschneiden der Maßnahmen auf den jeweiligen Arbeitskontext und
3. Ethik, insbesondere Maßnahmen zur Gewährleistung der Vertraulichkeit von Informationen

Wirkung und Effizienz der Interventionsmaßnahmen.[3]

Digitale Assistenzsysteme zwischen Stressprävention und Überwachung

Tragbare Gesundheitsmonitoring-Systeme (sog. „Wearables“, z. B. Fitnesstracker, Smartwatches) könnten dazu eingesetzt werden, um Arbeitnehmer*innen anhand von Herzfrequenzdaten und Bewegungstracking rückzumelden, wann ein idealer Zeitpunkt für eine Pause, sich zu bewegen oder etwas zu trinken wäre.

Da sich zunehmend mehr Menschen über Gesundheitsapps ihrer Smartphones oder diverser Fitnesstracker über ihr Bewegungsverhalten und andere Gesundheitswerte informieren, sind Krankenkassen daran interessiert, diese Technologien für die Gesundheits-

3 Europäische Kommission, 2017.

prävention einzusetzen. Marktgängige „Wearables“ wurden deshalb bereits dahingehend untersucht, für welche Anwendungsbereiche welche Technologien am besten zu nutzen wären[4].

Eine Metauntersuchung von verschiedenen wissenschaftlichen Studien zur elektronischen Überwachung am Arbeitsplatz ergab, dass diese negative Auswirkungen auf das subjektive Empfinden (z. B. von Stress, Beanspruchung, Kontrolle und Zufriedenheit) haben kann[5].

Also aufgepasst! Technologien zur Stressreduktion und Burnout-Vorbeugung können genau gegenteilige Effekte erzielen, weil die Systeme als Überwachungstechnologien erlebt werden!

Digitale Assistenzsysteme zur Stressprävention richtig eingesetzt

Welche konkreten Empfehlungen können nun aus diesen wissenschaftlichen Erkenntnissen abgeleitet werden, um die Wirksamkeit der Assistenzsysteme zu steigern und besser auf die Bedürfnisse der Arbeitnehmer*innen abzustimmen?

- Vor der Einführung eines Assistenzsystems sollten Ziele und Wirkungsweise offen kommuniziert und grundsätzliche Informationen zur Technologie bereitgestellt werden. Das steigert das Vertrauen der Belegschaft.
- Auch führt eine hohe wahrgenommene Nützlichkeit bzw. ein erlebter persönlicher Vorteil der Technologie (wie z. B. Feedback zur Gesundheitsvorsorge) zu einer größeren Akzeptanz.
- Mitarbeiter*innen sollen und müssen darüber hinaus nicht nur genau informiert werden, welche Daten erhoben und wie diese weiterverarbeitet werden, sondern haben auch ein Anrecht auf ihre persönlichen Daten[6] (Abschnitt C, Kapitel Gesetzliche Rahmenbedingungen für den Einsatz von KI-Assistenzsystemen am Arbeitsplatz).
- Abschließend steigt die Zufriedenheit mit Technologien, wenn diese partizipativ gestaltet und implementiert werden (siehe dazu Abschnitt C, Kapitel KI im Betrieb: Wer soll wie mitreden?). Insbesondere in Bezug auf die Datenerfassung gilt: Arbeitszufriedenheit und -motivation leiden, wenn die Privatsphäre verletzt bzw. auch, wenn persönliche Daten erhoben werden. Das heißt, wenn Beschäftigte in partizipativen Prozessen rückmelden können, welche Aspekte ihrer Tätigkeit sinnvollerweise erfasst werden sollen und welche nicht, dann können diese Wünsche in der Ausgestaltung der Technologie berücksichtigt und somit die Akzeptanz der Belegschaft gesteigert werden[7].

SIE HABEN NOCH NICHT GENUG?

Ausführlichere Erklärungen und Beispiele für arbeitsbedingte psychische Belastungen finden Sie in diesem

- *Merkblatt der Arbeitsinspektion*[8]

Weiterführende Informationen sowie praktische Materialien zur Arbeitsplatzevaluierung finden Sie auf dieser

- *Webseite der Arbeitsinspektion*[9]

In diesem Dokument finden Sie weitere Informationen über ein Programm der österreichischen Allgemeinen Unfallversicherungsanstalt (AUVA), das arbeitsbedingte psychische und physische Belastungen (z. B. des Bewegungs- und Stützapparats) analysiert und darauf aufbauend gezielte Maßnahmen für Verbesserungen der Arbeit oder deren Ausführungsbedingungen vorschlägt – ganz ohne technologische Überwachungssysteme:

- *AUVAfit Verbesserung der Arbeitsqualität*[10]

ANSPRECHPERSONEN

Expert*innen der Sektion Arbeitsrecht und Zentral-Arbeitsinspektorat des Bundesministeriums für Arbeit
https://www.arbeitsinspektion.gv.at/Gesundheit_im_Betrieb/psychische_Belastungen/Arbeitsplatzevaluierung_psychischer_Belastungen.html

4 Ellegast, Schellewald, Weber, Weber, und Hartmann, 2017.
5 Backhaus, 2018.
6 Backhaus, 2018.
7 Backhaus, 2018.
8 Bundesministerium für Arbeit, 2013.
9 Bundesministerium für Arbeit, 2022.
10 Allgemeine Unfallversicherungsanstalt, 2015.

LITERATURVERZEICHNIS

Allgemeine Unfallversicherungsanstalt, AUVA. 2015. *AUVAfit – Verbesserung der Arbeitsqualität. https://www.auva.at/cdscontent/load?contentid=10008.543844&version=1491389244*.

Backhaus, Nils. 2018. *Review zur Wirkung elektronischer Überwachung am Arbeitsplatz und Gestaltung kontextsensitiver Assistenzsysteme. https://doi.org/10.21934/BAUA:BERICHT20180726*.

Bundesministerium für Arbeit, Soziales und Konsumentenschutz. 2022. *Arbeitsplatzevaluierung psychischer Belastungen nach dem Arbeitnehmerinnenschutzgesetz (ASCHG). Merkblatt. https://www.arbeitsinspektion.gv.at/Zentrale_Dokumente/Gesundheit_im_Betrieb/psychische_Belastungen/merkblatt_arbeitsplatzevaluierung_psychischer_belastungen.pdf*.

Bundesministerium für Arbeit, Soziales und Konsumentenschutz. 2022. *Arbeitsplatzevaluierung psychischer Belastungen.* Arbeitsinspektion. *https://www.arbeitsinspektion.gv.at/Gesundheit_im_Betrieb/psychische_Belastungen/Arbeitsplatzevaluierung_psychischer_Belastungen.html*.

Ellegast, Rolf, Schellewald, Vera, Weber, Britta, Weber, Anika, und Hartmann, Ulrich. 2017. Kategorisierung von Wearables zur Erfassung der körperlichen Aktivität am Arbeitsplatz. In *Soziotechnische Gestaltung des digitalen Wandels – kreativ, innovativ, sinnhaft.* Frühjahrskongress 2017 in Brugg. Hrsg. von Gesellschaft für Arbeitswissenschaft e.V., Dortmund. *https://gfa2017.gesellschaft-fuer-arbeitswissenschaft.de/inhalt/A.5.4.pdf*.

Gimpel, Henner. 2019. Gesund digital arbeiten?! *https://doi.org/10.24406/FIT-N-562039*.

Europäische Kommission. 2019. *Förderung der psychischen Gesundheit am Arbeitsplatz: Leitfaden für die Umsetzung eines umfassenden Konzepts.* Publications Office. *https://doi.org/doi/10.2767/423581*.

Wie KI mit Daten lernt und welche Vorurteile dabei erzeugt werden

DAS FOLGENDE KAPITEL IST FÜR SIE VON INTERESSE, WENN …

… Sie grundsätzlich verstehen wollen, wie Künstliche Intelligenz lernt.

… Sie sich gefragt haben, warum es so wichtig ist, KI-Systeme mit qualitativ hochwertigen Daten zu füttern.

… Sie wissen möchten, was der Begriff „Bias“ im Zusammenhang mit KI bedeutet.

… Sie konkrete Beispiele für die Auswirkungen von KI-Bias erfahren möchten.

… Sie nach Lösungen suchen, um Bias zu verhindern.

Beim Maschinellen Lernen werden Zusammenhänge, Konzepte und Regeln auf Basis von eingegebenen Daten automatisch erlernt. Dies geschieht, indem das maschinelle Lernprogramm versucht, Verallgemeinerungen und Schlüsse aus den vorhandenen Daten abzuleiten. Diesen Vorgang nennt man auch „Training“.

Aus Daten werden KI-Entscheidungen

Es gibt unterschiedliche Arten von Trainings. Eine häufig verwendete Form ist das Klassifizieren: Aus Mustern bzw. Eigenschaften in den Daten wird gelernt, ob etwas einer Klasse bzw. einem Typ entspricht. Nehmen wir beispielsweise an, in einer Obstsortieranlage wird ein Farbsensor verwendet, um zu erkennen, welches Obst genau am Fließband liegt. Hierbei muss der KI-Algorithmus im Trainingsprozess eine Beziehung zwischen den Eingabedaten (Sensorwert) und dem zu klassifizierenden Typ feststellen (siehe Abbildung 1).

Wirklichkeit	Sensorwert		Klassifizierter Typ
	grün	KI?	Apfel
	rot		Kirsche
	gelb		Banane

Abbildung 1: Beim Klassifizieren, einer speziellen Form des Maschinellen Lernens, muss eine KI selbstständig anhand von Trainingsdaten (Beispielen, wie hier abgebildet) Regeln finden, um auf Basis von Eingabedaten (Sensorwert) den Klassifizierungstyp zu bestimmen.

Auf Basis der Beispiele in den Trainingsdaten könnte eine KI nun folgende einfache Regeln aufstellen:

- Wenn das Objekt grün ist, ist es ein Apfel.
- Wenn das Objekt rot ist, ist es eine Kirsche.
- Wenn das Objekt gelb ist, ist es eine Banane.

Solange nur grüne Äpfel, rote Kirschen und gelbe Bananen auf dem Fließband anrollen, wird die KI gute Dienste leisten. Jedoch würde die KI einen roten Apfel als Kirsche klassifizieren oder eine grüne Banane als Apfel.

Unsere Zufriedenheit mit den Ergebnissen hängt also davon ab, ob wir die KI mit ausreichend Beispielen (Daten) trainiert haben und ob diese Daten auch repräsentativ für die Wirklichkeit sind (engl. Ground Truth).

Wie kann eine KI mit solchen Ambivalenzen (siehe Abbildung 2) nun umgehen?

Wirklichkeit	Sensorwert		Klassifizierter Typ
	grün	KI?	Apfel
	rot		Kirsche
	gelb		Banane
	rot		Apfel
	grün		Apfel
	rot		Kirsche

Abbildung 2: In Trainingsdaten gibt es oft Ambivalenz. Hier soll die KI auf Basis der Sensorwerte die Klassifizierung vornehmen, aber der Farbwert rot ist sowohl Kirschen als auch Äpfeln zugordnet.

Wenn wir unsere KI mit neuen Beispielen trainieren, die der Wirklichkeit besser entsprechen, kann der Algorithmus keine widerspruchsfreien Regeln mehr aufstellen (nicht jedes rote Obst ist eine Kirsche). Der Algorithmus basiert seine Entscheidung daher auf einer Schätzung. Solche auch „Faustregeln“ oder „Heuristiken“ genannten Schätzungen ermöglichen es, Entscheidungen bei ambivalentem und unvollständigem Wissen oder auch in begrenzter Zeit zu treffen[1]. Da die in Abbildung 2 dargestellten Trainingsdaten mehr rote Kirschen als rote Äpfel enthalten, würde die Schätzung in unserem Fall lauten: Rotes Obst ist wahrscheinlich eine Kirsche.

Wenn Sie nun denken, „das ist aber keine wirklich gute Schätzung“, haben Sie natürlich völlig recht. Rote Äpfel machen immerhin ein Drittel allen roten Obstes im Trainingsdatensatz aus. Die Wahrscheinlichkeit für eine Fehlentscheidung wäre also ziemlich hoch. In echten Anwendungsfällen Künstlicher Intelligenz würde ein solcher Algorithmus seine Entscheidung wahrscheinlich auf einer nuancierteren Farbpalette und eventuell weiteren Sensordaten aufbauen. Ein gewisse, wenn auch kleinere Wahrscheinlichkeit für eine Fehlentscheidung bliebe jedoch aufgrund der statistischen Schätzmethode auch dann vorhanden.

In Kürze: Die von KI-Systemen genutzten Schätzheuristiken ermöglichen zwar die Entscheidungsfindung bei ambivalenten Datensätzen, bringen jedoch auch immer eine Wahrscheinlichkeit für Fehler mit sich. Diese Fehler nennt man Bias.

Bias hat nicht nur Auswirkungen auf die Qualität der Entscheidungsfindung, sondern beispielsweise auch auf die mit dem jeweiligen KI-System arbeitenden Menschen. Im obigen Beispiel könnte ein großer Bias bei den in der Sortieranlage arbeitenden Menschen beispielsweise mehr Stress aufgrund eines höheren Bedarfs an Nachsortierung auslösen.

INFOBOX: PRAXISBEISPIEL

Das Sortieren von Obst ist ein Anschauungsbeispiel mit realem Bezug. Bei Pro²Future wird gerade an einem ähnlich gelagerten Projekt (ReAIcle) gearbeitet, wo Mensch und KI zusammenarbeiten, um die Sortierung von Müll zu verbessern. In einem anderen Projekt (REWAI) wird KI dafür verwendet, um Produktionsprozesse so zu optimieren, dass weniger Müll und Energie anfällt und diese Prozesse so transparent gemacht werden, dass der Mensch die Gründe für die Veränderung gut nachvollziehen kann und diese auch freigibt.

1 Gigerenzer und Todd, 1999.

Beispiele für Bias und deren Auswirkungen

Zuvor haben wir am Beispiel einer Obstsortierungsanlage datenbasierte Bias erläutert. Im Folgenden sind Fallbeispiele für Sie zusammengestellt, die neben daten- auch algorithmen- und kulturbasierte Bias umfassen[2].

1. Datenbasierte Bias: Ein wichtiger datenbasierter Bias ist der sogenannte Sample-Bias, der entsteht, wenn die einem Algorithmus zugrundeliegenden Daten nicht die (relevante) Wirklichkeit widerspiegeln. Bekannt geworden sind beispielsweise Bilderkennungsalgorithmen, die Probleme damit hatten, dunkelhäutige Frauen korrekt als solche zu identifizieren, während die Gesichter von Männern oder weißen Menschen häufiger richtig klassifiziert wurden[3]. Ein anderer Bilderkennungsalgorithmus wiederum identifizierte Männer, die in Küchen abgebildet waren, fälschlicherweise als Frauen[4]. In beiden Fällen ist die schlechte Datenlage verantwortlich. Im ersten Beispiel enthielt der zum Trainieren verwendete Datensatz hauptsächlich Fotos von weißen Personen und Männern; im zweiten Beispiel waren im zugrundeliegenden Datensatz kochende Männer weniger repräsentiert als kochende Frauen.

2. Algorithmenbasierte Bias: Algorithmenbasierte Bias können unter anderem dann auftreten, wenn Nutzer*innen nicht ausreichend in die Erstellung der Software miteingebunden werden. Beispielsweise ist es relevant darauf zu achten, wie man die Ergebnisse von KI-Algorithmen so darstellt, dass diese nutzbringend und richtig interpretiert werden können. Forschungsarbeiten konnten demonstrieren, dass Visualisierungen einfacher und schneller wahrgenommen werden können als Datentabellen. Es hat sich jedoch gezeigt, dass Visualisierungen verhindern können, dass sich Menschen tiefergehend mit dem Betrachtungsgegenstand auseinandersetzen. In einer Studie von der Universität Michigan und Microsoft zeigte sich zum Beispiel, dass selbst Datenwissenschaftler*innen Visualisierungen von KI-Ergebnissen zur Interpretation heranzogen, auch wenn diese für sie nicht vollständig nachvollziehbar waren[5]. Darüber hinaus erhöhten die teilweise nicht nachvollziehbaren Visualisierungen trotzdem das Vertrauen der Expert*innen in den Machine-Learning-Algorithmus. Dass Menschen von Computern dargestellte Ergebnisse fälschlicherweise als besonders objektiv und korrekt wahrnehmen, ist ein Phänomen, das als „Automation Bias“ beschrieben wird[6]. Zusammenfassend muss auch bei der Programmierung von Algorithmen genauestens berücksichtigt und bestenfalls untersucht werden, wie sich die Algorithmen und deren Ergebnisdarstellung auswirken könnten.

3. Kulturbasierte Bias: Man stelle sich einen KI-Algorithmus vor, dessen Daten so ausgewählt sind, dass sie die Wirklichkeit perfekt repräsentieren und dass der Algorithmus darüber hinaus ohne jegliche Fehlentscheidungen programmiert wurde. Eine perfekte nicht-gebiaste KI, oder? Nein, nicht zwangsläufig. In einigen Fällen von gebiasten KI-Systemen kann es passieren, dass der Bias nicht aus Daten oder Algorithmen resultiert, sondern in der Gesellschaft besteht und in Daten oder Algorithmen reproduziert wurde. Ein Beispiel hierfür ist ein von Amazon eingesetzter KI-Algorithmus, der von 2014 bis 2015 genutzt wurde, um die Lebensläufe von Bewerber*innen zu bewerten[7]. Zum Trainieren des Algorithmus wurden historische aufgezeichnete Lebenslaufdaten und deren Erfolgsraten früherer Bewerber*innen genutzt. Um diskriminierende Effekte von vornherein auszuschließen, wurde der Algorithmus so eingestellt, dass Geschlecht als Feature nicht berücksichtigt werden sollte. Das Machine-Learning-Programm fand jedoch einen indirekten und unvorhergesehenen Weg, männliche und weibliche Bewerber*innen zu unterscheiden. Und zwar stellte es fest, dass erfolgreiche Bewerbungen für Jobs im Technik-Bereich zumeist von Personen eingereicht wurden, die keine mit Weiblichkeit verknüpften Attribute aufwiesen. So führte beispielsweise die Erwähnung des Wortes „Frau“ (wie z. B. in „Frauensportverein“) automatisch zu einer schlechteren Bewertung der jeweiligen Lebensläufe. Die KI setz-

2 Für weitere Arten von Bias s. Mehrabi u. a., 2019.
3 Z. B. Buolamwini und Gebru, 2018; Zou und Schiebinger, 2018.
4 Zhao u. a., 2017.
5 Kaur u. a., 2020.
6 Mehrabi u. a., 2019.
7 Dastin, 2018.

te somit ein insbesondere in der westlichen Gesellschaft verankertes Phänomen[8], den Umstand, dass vergleichsweise wenige Frauen in Technikberufen tätig sind, fort. Sich bewerbende Frauen bekamen somit eventuell gar nicht die Chance, ihre Technikkompetenz unter Beweis zu stellen und wurden von vornherein ausgeschlossen. In anderen Worten: Sie wurden diskriminiert.

Ein weiteres Beispiel für die Fortschreibung von kulturell bestehenden Ungleichheiten unterschiedlicher Gruppen ist ein vom Arbeitsmarktservice (AMS) programmierter Algorithmus. Der vom AMS als „Arbeitsmarktchancen-Assistenz-System“ (AMAS) bezeichnete Algorithmus soll zur Bewertung von Jobchancen dienen und die Einschätzung der Förderwürdigkeit von Arbeitssuchenden durch Arbeitsvermittler*innen unterstützen. Personen mit, durch den Algorithmus berechneten, hohen und niedrigen Jobchancen sollen hierbei weniger gefördert werden als Personen mit mittleren Jobchancen. Der Algorithmus basiert zwar nicht auf Verfahren, die es zuließen, ihn als Künstliche Intelligenz zu bezeichnen, er eignet sich jedoch als anschauliches Beispiel für kulturbasierte Bias von Algorithmen, auch wegen seines potentiell immensen Einflusses. Eine von der AK OÖ beauftragte und durch das Institut für Technikfolgenabschätzung durchgeführte Studie untersuchte AMAS[9]. Die Studie konnte zeigen, dass AMAS Personen aufgrund von Personenmerkmalen wie Geschlecht, Alter, Migrationshintergrund und Behinderung unterschiedlich in ihren Jobchancen bewertet. Die Bewertung basiert auf historisch aufgezeichneten Daten, die zeigen, dass Personen mit bestimmten Merkmalen oder Merkmalskombinationen eher einen Job fanden als andere. Beispielsweise schätzt AMAS die Jobchancen von Frauen mit dem Merkmal Kinderbetreuungspflichten schlechter ein als diejenigen von Männern mit Kinderbetreuungspflichten. Dieser Unterschied in den Jobchancen mag sich aufgrund verschiedener Gegebenheiten in der Gesellschaft (z. B. aufgrund von diskriminierenden Einstellungspraktiken[10]) tatsächlich so zeigen, sollte sich aber eigentlich nicht negativ auf die Förderungsempfehlungen von Arbeitssuchenden auswirken. Eine ähnliche Herabstufung im AMAS-Ratingsystem findet sich für Frauen mit Migrationshintergrund. Statt diese Personengruppen besonders zu fördern, um sie besser auf dem Arbeitsmarkt zu integrieren und eine langfristige Gleichstellung zu erreichen, entscheidet sich das AMS für die Diskriminierung dieser Gruppen und eine Fortschreibung von gesellschaftlich bestehenden Ungleichheiten.

Bias verhindern

Wie können Bias und weitere negative Auswirkungen von KI-Systemen in Unternehmen nun verhindert werden? Hier eine kurze und naturgemäß unvollständige Übersicht wichtiger Maßnahmen:

- Die Entwicklung eines KI-Systems sollte nie abgeschlossen sein. Eventuelle Bias und weitere Auswirkungen müssen von Beginn an mitbedacht und im Laufe des Prozesses berücksichtigt werden. Das beginnt mit dem Unternehmensinteresse und der Frage danach, welche KI-Systeme und Daten tatsächlich benötigt werden, geht über den Prozess der Technologieentwicklung (d. h. Entwicklung, Training und Testung des Algorithmus, der Implementierung) bis hin zur Implementierung und einem langfristigen Monitoring.

- Dem Training zugrundeliegende Daten sollten möglichst die relevante Wirklichkeit widerspiegeln. Ändern sich die Umstände oder wird die Relevanz von bisher unberücksichtigten Variablen deutlich, sollte ein Algorithmus neu trainiert werden.

- Es empfiehlt sich, zu überprüfen, ob das System zu anderen, unerwünschten Ergebnissen kommt, wenn man bestimmte vom System erfasste Merkmale verändert (z. B. das Geschlecht oder die Hautfarbe).

- Besteht die Gefahr des Fortschreibens gesellschaftlich bestehender Ungleichheiten, sollten algorithmische Korrekturmaßnahmen getroffenen werden. Beispielsweise könnten Daten mit unterschiedlichen Gewichtungen versehen werden, um bestehende Bias in den Trainingsdaten zu korrigieren.

8 Mellström, 2009.

9 Allhutter u. a., 2020.

- KI-Systeme sollten möglichst transparent und verständlich gestaltet sein, um eventuelle Bias bereits während des Betriebes erkennen oder im Nachhinein nachvollziehen zu können (s. Abschnitt B, Kapitel Erklärbare Künstliche Intelligenz). Das Annotieren der Trainingsdaten sollte einem genau festgelegten Standard folgen und wenn möglich parallel durch mehrere Menschen durchgeführt werden, sodass die Übereinstimmung der annotierenden Personen berechenbar ist.

- Ethische sowie gender- und diversitätsbezogene Fragestellungen sollten bereits während der KI-Programmierung berücksichtigt werden. Gute Tipps für eine Vorgehensweise in verschiedenen wissenschaftlichen Feldern finden sich zum Beispiel auf: *https://genderedinnovations.stanford.edu.*

- „I-Methodology" (d. h. die Einschreibung von fast ausschließlich eigenen Erfahrungen durch Technik-Entwickler*innen) sollte verhindert werden. Hierfür lohnt sich, neben User-centered-Design, eine inter- und transdisziplinäre Arbeitsweise in möglichst gemischten Teams. Diversität, Inter- und Transdisziplinarität erhöhen die Robustheit von Technologien und wissenschaftlichen Erkenntnissen.

- Die Bedürfnisse und Charakteristiken aller potentiell Betroffener sollten stets im Auge behalten werden. Hierfür lohnt es sich, Betroffene gleich von Anfang an einen gemeinsamen Tisch zu holen und zum Beispiel in Workshops Interessen und Bedürfnisse abzuwägen. Dieser Prozess ist sehr gut im WeBuildAI-Framework von Min Kyung Lee[11] beschrieben.

- In Betrieben sollten Arbeitnehmendenvertretungen stets im direkten Dialog mit der Unternehmens-IT stehen, um den Bedürfnissen der Arbeitnehmenden deutliches Gehör zu verschaffen (s. Abschnitt C, Kapitel KI im Betrieb: Wer soll wie mitreden?).

Eine ausführliche Auflistung zu berücksichtigender Aspekte bei der Implementierung von KI finden Sie im Abschnitt C im Kapitel Fragen, die Sie bei der (Mit-)Gestaltung von KI-Assistenzsystemen am Arbeitsplatz stellen sollten.

SIE HABEN NOCH NICHT GENUG?

Im Folgenden englischsprachigen YouTube-Video finden Sie weitere spannende und gut erklärte Beispiele für Bias bei Künstlicher Intelligenz. Möchten Sie das Video lieber auf Deutsch ansehen, nutzen Sie doch die KI-erzeugten (und hoffentlich bias-freien) YouTube-Untertitel:

- *Algorithmic Bias and Fairness: Crash Course AI #18*[12]

In diesem Podcast wird über Bias, Künstliche Intelligenz und den Unterschied zu Unconscious Bias diskutiert:

- *Unconscious Bias – Warum Künstliche Intelligenz diskriminieren und Vorurteile verschärfen kann. Ein Gespräch mit Prof. Katharina Hölzle und Rea Eldem*[13]

Dieses YouTube-Video, eine Vorlesungseinheit des MIT über Deep Learning, empfehlen wir für technisch Interessierte:

- *Vortrag von Ava Soleimany (MIT): "AI Bias and Fairness"*[14]

ANSPRECHPERSONEN

DI Michael Haslgrübler, Pro²Future
Experte für Angewandtes Maschinelles Lernen
michael.haslgruebler@pro2future.at

Mag. Dr. Julian Anslinger, IFZ
Experte für die Auswirkungen von KI-Bias und Kognitiven Bias
julian.anslinger@ifz.at

LITERATURVERZEICHNIS

Allhutter, Doris, Astrid Mager, Florian Cech, Fabian Fischer, und Gabriel Grill. 2020. *Der AMS-Algorithmus: Eine Soziotechnische Analyse des Arbeitsmarktchancen-Assistenz-Systems (AMAS).* ITA-Projektbericht. *https://epub.oeaw.ac.at/0xc1aa5576_0x003bfdf3.pdf.*

10 S. bspw. Hipp, 2020.
11 Lee u. a., 2019.
12 CrashCourse, 2019.
13 Neuland, 2021.
14 Soleimany, 2021.

Buolamwini, Joy, und Timnit Gebru. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, 77–91. PMLR. *https://proceedings.mlr.press/v81/buolamwini18a.html*.

CrashCourse. 2019. *Algorithmic Bias and Fairness: Crash Course AI #18. https://www.youtube.com/watch?v=gV0_raKR2UQ*.

Dastin, Jeffrey. 2018. *Amazon scraps secret AI recruiting tool that showed bias against women.* Reuters. *https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G*.

Gigerenzer, Gerd, und Peter M. Todd. 1999. *Simple heuristics that make us smart*. USA: Oxford University Press.

Hipp, Lena. 2020. Do Hiring Practices Penalize Women and Benefit Men for Having Children? Experimental Evidence from Germany. *European Sociological Review* 36, Nr. 2: 250–64. *https://doi.org/10.1093/esr/jcz056*.

Kaur, Harmanpreet, Harsha Nori, Samuel Jenkins, Rich Caruana, Hanna Wallach, und Jennifer Wortman Vaughan. 2020. Interpreting Interpretability: Understanding Data Scientists' Use of Interpretability Tools for Machine Learning. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems,* 1–14. Honolulu HI USA: ACM. *https://doi.org/10.1145/3313831.3376219*.

Lee, Min Kyung, Daniel Kusbit, Anson Kahng, Ji Tae Kim, Xinran Yuan, Allissa Chan, Daniel See, u. a. 2019. WeBuildAI: Participatory Framework for Algorithmic Governance. *Proceedings of the ACM on Human-Computer Interaction* 3, Nr. CSCW 7. Nov. 2019: 181:1-181:35. *https://doi.org/10.1145/3359283*.

Mehrabi, Ninareh, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, und Aram Galstyan. 2019. A Survey on Bias and Fairness in Machine Learning. *arXiv.* 1908.09635 [cs]. *http://arxiv.org/abs/1908.09635*.

Mellström, Ulf. 2009. The Intersection of Gender, Race and Cultural Boundaries, or Why is Computer Science in Malaysia Dominated by Women? *Social Studies of Science* 39 2009: 885–907. *https://doi.org/10.1177/0306312709334636*.

Neuland. 2021. Unconscious Bias – Warum Künstliche Intelligenz diskriminieren und Vorurteile verschärfen kann. Gespräch mit Prof. Katharina Hölzle und Rea Eldem. *Hasso Plattner Institut. Wissenspodcast*. *https://podcasts.apple.com/at/podcast/unconscious-bias-warum-k%C3%BCnstliche-intelligenz-diskriminieren/id1448992875?i=1000541329741*.

Soleimany, Ava. 2021. MIT Introduction to Deep Learning 6.S191. Lecture 8: AI Bias and Fairness. *https://youtu.be/wmyVO-Dy_WD8*.

Zhao, Jieyu, Tianlu Wang, Mark Yatskar, Vicente Ordonez, und Kai-Wei Chang. 2017. Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints. *arXiv.* 1707.09457 [cs.AI]. *http://arxiv.org/abs/1707.09457*.

Zou, James, und Londa Schiebinger. 2018 AI Can Be Sexist and Racist — It's Time to Make It Fair. *Nature* 559, Nr. 7714, Juli 2018: 324–26. *https://doi.org/10.1038/d41586-018-05707-8*.

Erklärbare Künstliche Intelligenz

DAS FOLGENDE KAPITEL IST FÜR SIE VON INTERESSE, WENN …

… Sie wissen möchten, warum erhöhte Nachvollziehbarkeit von Künstlicher Intelligenz wichtig ist.

… Sie erfahren möchten, was der Unterschied zwischen Black- und White-Box-Modellen ist.

… Sie lernen möchten, welche Möglichkeiten es gibt, die Nachvollziehbarkeit zu erhöhen.

… Sie sich fragen, warum nachvollziehbare KI-Systeme weniger populär sind, als man meint und welche Gefahren sie mit sich bringen.

Bei der Arbeit mit KI-Algorithmen empfiehlt es sich, einen besonderen Wert darauf zu legen, nachvollziehbar zu machen, wie der Algorithmus zu seinem Ergebnis kommt.[1] Eine gute Nachvollziehbarkeit eines KI-Algorithmus kann dabei unterstützen, eventuelle Bias in den Daten frühzeitig zu entdecken (s. Abschnitt B im Kapitel Wie KI mit Daten lernt und welche Vorurteile dabei erzeugt werden). Auch können etwaige Fehlschlüsse, wie das Klassifizieren jedes roten Obstes als Kirsche (ebda.), schnell aufgedeckt werden.

KI-Modelle

White-Box-Modelle

KI-Algorithmen mit der größten Nachvollziehbarkeit sind sogenannte White-Box-Modelle.[2] Typische Modelle sind regelbasierte Modelle, Entscheidungsbäume, Nächste-Nachbarn-Klassifikation, lineare Modelle sowie Bayessche Modelle (s. Abbildung 1). Diese Modelle haben den Nachteil, dass sie für viele Anwendungsfälle weit weniger präzise sind als Modelle, die auf einem Machine-Learning-Ansatz basieren.

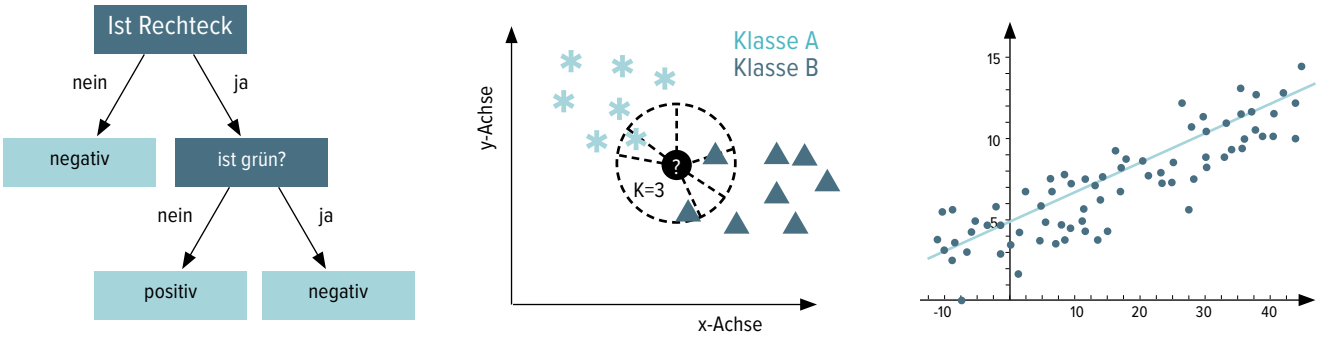


Abbildung 1: Beispiele für (transparente) White-Box-KIs, Entscheidungsbaum (links); Nächste-Nachbarn-Klassifikation (Mitte); Linear-Klassifikation (rechts)

1 Barredo Arrieta u. a., 2020.

2 Loyola-Gonzalez, 2019.

Black-Box-Modelle

Im Unterschied zu White-Box-Modellen sind die Entscheidungen von Machine-Learning-Algorithmen jedoch weitaus weniger transparent (für einen Vergleich der Algorithmusarten s. Abschnitt A, Kapitel KI-Assistenzsysteme – was ist das überhaupt?) Dies ist insbesondere für Deep-Learning-Verfahren der Fall. Diese spezielle Form des Maschinellen Lernens beruht auf einem vereinfachten Modell des menschlichen Gehirns als selbstständig lernendes System (neuronales Netz) und arbeitet mit Millionen von Parametern. Die Entscheidungsfindung eines Deep-Learning-Algorithmus ist von außen kaum nachvollziehbar. Wegen der fehlenden Einsicht und Interpretierbarkeit werden Deep-Learning-Algorithmen auch häufig als „Black Box“ bezeichnet. Aber auch bei der Entwicklung von Black-Box-Modellen kann die Nachvollziehbarkeit erhöht werden.

Verbesserte Nachvollziehbarkeit

Sichtbarmachung der Entscheidungsfindung

Ein wichtiges Instrument zur Erhöhung der Nachvollziehbarkeit ist die Erklärung des Entscheidungsfindungsprozesses, d. h., bei der Ausgabe des Algorithmus darzustellen, warum ein KI-Algorithmus zu seinem Ergebnis gekommen ist. Sehen wir uns beispielhaft eine Ausgabe eines Bilderkennungsalgorithmus an (s. Abbildung 2).

Abbildung 2: Beispiel für eine wenig nachvollziehbare Annotation eines Bilderkennungsalgorithmus

Sie wundern sich vielleicht, warum der Algorithmus das Bild mit den Worten „Pferd“ und „Hund“ versehen hat. Die Entscheidung des Algorithmus ist auf den ersten Blick wenig nachvollziehbar. Eine nachvollziehbarere Ausgabe können Sie Abbildung 3 entnehmen.

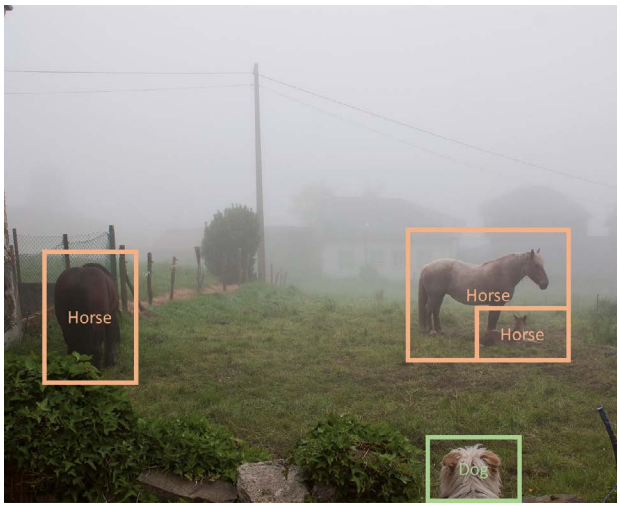


Abbildung 3: Beispiel für eine besser nachvollziehbare Annotation eines Bilderkennungsalgorithmus

In Abbildung 3 wurden alle Tierarten in unterschiedlich farbige Kästchen gefasst und beschriftet. Der Hund unten rechts ist schnell zu erkennen. Eine ähnliche Art der Darstellung bietet die farbliche und semi-transparente Markierung aller betroffenen Bildbereiche wie in Abbildung 4.

Abbildung 4: Beispiel für eine (inkorrekte) semi-transparente Annotation eines Bildbereichs

In Abbildung 4 erkennen Sie darüber hinaus, warum die Nachvollziehbarkeit von KI-Algorithmen wichtig ist. Entscheidungsfehler wie der Umstand, dass der markierte Bereich keine Katze enthält, sind schnell erkennbar. Der KI-Algorithmus könnte nun korrigiert bzw. das Ergebnis durch eine menschliche Aufsichtsperson überstimmt werden. Weitere Maßnahmen zur Erhöhung der Sichtbarmachung umfassen die Offenlegung aller in einen Algorithmus eingegangen Variablen, Visualisierungen, das Zeigen von analogen Beispielen sowie zusätzliche textuelle Erklärungen.

Sichtbarmachung von Limitationen

Eine weitere Empfehlung zur Erhöhung der Transparenz ist die Sichtbarmachung von Limitationen eines KI-Systems[3]. Die KI-Nutzenden könnten beispielsweise darüber aufgeklärt werden, dass die in den KI-Algorithmus eingegangenen Daten mit Einschränkungen versehen sind. Bei einem Bilderkennungsalgorithmus zur Erkennung von Tierarten könnte man beispielsweise darauf hinweisen, dass sich in dem zum Training verwendeten Datensatz ausschließlich Bilder von Tieren aus Europa befinden.

Entsprechend empfiehlt sich, bereits beim Training eines Algorithmus nur auf Datensätze zurückzugreifen, die mit sogenannten Metainformationen versehen sind[4]. Mit Metainformationen ausgestattete Datensätze sind insbesondere dann nützlich, wenn man datenbasierte Bias verhindern möchte (s. Abschnitt B, Kapitel Wie KI mit Daten lernt und welche Vorurteile dabei erzeugt werden). Metainformationen können beispielsweise Informationen über die demographische Zusammensetzung von Personendaten (d. h. Alter, Geschlecht, Ethnizität, etc.), Erhebungsort und Erhebungszeitraum beinhalten.

Technische Transparenz

Eine weitere Maßnahme, die Nachvollziehbarkeit von KI-Systemen zu erhöhen, ist technische Transparenz[5]. Technische Transparenz umfasst die Offenlegung des Quellcodes, der zugrundeliegenden Dokumentation sowie, wenn datenschutzmäßig vertretbar, der in den Algorithmus eingegangenen Daten. Eine derartige Offenlegung ermöglicht die Untersuchung der KI durch unabhängige Expert*innen.

Darüber hinaus ist dieser Ansatz eine Antwort auf die sogenannte Reproduzierbarkeits-Krise. Dabei handelt es sich um den Umstand, dass Forschungs- und Entwicklungsergebnisse nicht repliziert werden können; eine Situation, von der die KI-Forschung ähnlich betroffen ist wie in der Vergangenheit die Psychologie oder Medizin[6]. Volle technische Transparenz würde es ermöglichen festzustellen, unter welchen Bedingungen Ergebnisse von Algorithmen repliziert werden können.

Abschließend würde technische Transparenz den technologischen Fortschritt erheblich beschleunigen, da nicht jede*r Programmierer*in ihre eigene Problemlösungen finden müsste, sondern, wie bei Open-Source-Software üblich, bestehende Lösungen durch Kollaboration stetig verbessert würden.

Verringerte Nachvollziehbarkeit

Nachvollziehbare Algorithmen sind trotz ihrer Vorteile weniger populär, als man annehmen möchte. Dies hat verschiedene Gründe. Einerseits bedeutet eine nachvollziehbare Gestaltung eines Algorithmus auch einen höheren ökonomischen Mehraufwand, sowohl durch einen erhöhten Programmieraufwand als auch durch eine größere Rechenleistung. Darüber hinaus setzen Firmen bewusst Black-Box-Modelle ein, um ihr geistiges Eigentum zu schützen. Hinzu kommt, dass mögliche Alternativen aus der Familie der White-Box-Modelle gar nicht erst in Betracht gezogen werden, weil bei Programmierer*innen die fälschliche Überzeugung besteht, dass unbekannte Datenmuster ausschließlich mit Black-Box-Modellen gefunden werden könnten.[7] Glücklicherweise gibt es Bestrebungen der EU, die Nachvollziehbarkeit von Algorithmen gesetzlich zu steuern.[8]

Nachvollziehbarkeit, aber richtig

Forschungsarbeiten haben gezeigt, dass die Erhöhung der Transparenz durch die Offenlegung der Eingangsvariablen einen negativen Einfluss auf die Ergebnisinterpretation haben kann. In einer Studie von Forough

3 Department of Health and Social Care, 2021.
4 Bender und Friedman, 2018.
5 Z. B. Council of Europe (CEPEJ), 2018.
6 Hutson, 2018.
7 Rudin, 2019.
8 EU, 2021.

Poursabzi-Sangdeh und Kolleg*innen stellte sich heraus, dass Personen, denen alle in die KI eingehenden Variablen vorgelegt wurden, vermutlich wegen Informationsüberfrachtung (engl. Information Overload), weniger dazu imstande waren, gravierende Entscheidungsfehler der KI zu bemerken und zu korrigieren[9].

Selbst die nachvollziehbare Gestaltung über Visualisierungsmethoden hat sich als problematisch herausgestellt. Visualisierungen können zwar einfacher und schneller wahrgenommen werden, was Information Overload reduzieren kann; gleichzeitig können sie jedoch verhindern, dass sich Menschen tiefergehend mit dem Betrachtungsgegenstand auseinandersetzen. In einer Studie der Universität Michigan und Microsoft zeigte sich zum Beispiel, dass selbst Datenwissenschaftler*innen Visualisierungen von KI-Ergebnissen zur Interpretation heranzogen, auch wenn diese für sie nicht vollständig nachvollziehbar waren.[10] Darüber hinaus erhöhten die teilweise nicht nachvollziehbaren Visualisierungen trotzdem das Vertrauen der Expert*innen in den Machine-Learning-Algorithmus. Dass Menschen von Computern dargestellte Ergebnisse fälschlicherweise als besonders objektiv und korrekt wahrnehmen, ist ein Phänomen, das als „Automation Bias“ beschrieben wird[11].

Was bedeutet das für das Gestalten von nachvollziehbaren KI-Systemen? Zusammengefasst müssen die Offenlegung und Darstellung von zusätzlichen Informationen fallweise abgewogen werden. Mehr Nachvollziehbarkeit ist gut, Informationsüberfrachtung nicht. Auch empfiehlt es sich, je nach Informationsdarstellung und Komplexität der KI-Anwendung, die Auswirkungen der Darstellung auf die Ergebnisinterpretation genau zu untersuchen, und so unerwünschte Effekte von vornherein auszuschließen.

SIE HABEN NOCH NICHT GENUG?

Lesen, sehen und hören Sie, wie Industrie und Forschung im Publikumsgespräch des Report Verlag über transparente KI diskutiert: Text:

- *KI – transparent und vertrauenswürdig*[12]

YouTube:

- *Publikumstalk „KI transparent und vertrauenswürdig in der Praxis“*[13]

Wir empfehlen darüber hinaus folgendes Whitepaper von TÜV AUSTRIA und Johannes Kepler Universität, Linz:

- *Towards Certification of Machine Learning Applications*

LITERATURVERZEICHNIS

Barredo Arrieta, Alejandro, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, u. a. 2020. Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion* 58: 82–115. *https://doi.org/10.1016/j.inffus.2019.12.012*.

Bender, Emily M., und Batya Friedman. 2018. Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. *Transactions of the Association for Computational Linguistics* 6: 587–604. *https://doi.org/10.1162/tacl_a_00041*.

Council of Europe (CEPEJ). 2018. *European ethical charter on the use of Artificial Intelligence in judicial systems and their environment*, Nr. December: 1–77.

Department of Health and Social Care, UK. *A Guide to Good Practice for Digital and Data-Driven Health Technologies*. GOV.UK, 2021. *https://www.gov.uk/government/publications/code-of-conduct-for-data-driven-health-and-care-technology/initial-code-of-conduct-for-data-driven-health-and-care-technology*.

9 Poursabzi-Sangdeh u. a., 2021.

10 Kaur u. a., 2020.

11 Osoba und Welser IV, 2017.

12 Legat, 2022.

13 Report Verlag, 2021.

EU. *Proposal for a Regulation of the European parliament and of the council laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts*. 2021. EUR-Lex – 52021PC0206 – EN § 2021. *https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1623335154975&uri=CELEX%3A52021PC0206*.

Hutson, Matthew. 2018. Artificial intelligence faces reproducibility crisis. *Science* 359, Nr. 6377: 725–26. *https://doi.org/10.1126/science.359.6377.725*.

Kaur, Harmanpreet, Harsha Nori, Samuel Jenkins, Rich Caruana, Hanna Wallach, und Jennifer Wortman Vaughan. 2020. Interpreting Interpretability: Understanding Data Scientists' Use of Interpretability Tools for Machine Learning. In *CHI Conference on Human Factors in Computing Systems Proceedings*, 1–14. *https://doi.org/10.1145/3313831.3376219*.

Legat, Karin. 2021. *KI – transparent und vertrauenswürdig*. www.report.at. *https://www.report.at/mehr/report-talk/19735-ki-transparent-und-vertrauenswuerdig*.

Loyola-Gonzalez, Octavio. 2019. Black-box vs. white-box: Understanding their advantages and weaknesses from a practical point of view. *IEEE access: practical innovations, open solutions* 7: 154096–113.

Osoba, Osonde, und William Welser IV. 2017. *An Intelligence in Our Image. The Risks of Bias and Errors in Artificial Intelligence.* RAND Corporation.

Poursabzi-Sangdeh, Forough, Daniel G. Goldstein, Jake M. Hofman, Jennifer Wortman Vaughan, und Hanna Wallach. 2021. Manipulating and Measuring Model Interpretability. *arXiv.* 1802.07810 [cs]. *http://arxiv.org/abs/1802.07810*.

Report Verlag. 2021. Publikumstalk „KI transparent und vertrauenswürdig in der Praxis". *https://www.youtube.com/watch?v=XpJrNOSph2k*.

Rudin, Cynthia. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence 1,* Nr. 5: 206–15.

Winter, Philip Matthias, Sebastian Eder, Johannes Weissenböck, Christoph Schwald, Thomas Doms, Tom Vogt, Sepp Hochreiter, und Bernhard Nessler. 2021. *White Paper – Trusted Artificial Intelligence: Towards Certification of Machine Learning Applications. https://www.tuv.at/wp-content/uploads/2022/03/Whitepaper_Trusted-AI_TUeV-AUSTRIA_JKU.pdf*.

Mitbestimmung: Die KI-Systeme kommen – was können Sie tun?

KI im Betrieb: Wer soll wie mitreden?

DAS FOLGENDE KAPITEL IST FÜR SIE VON INTERESSE, WENN …

… Sie ein KI-Assistenzsystem im Betrieb implementieren wollen.

… Sie sich schon immer gefragt haben, wen Sie mitreden lassen sollten, um möglichst große Akzeptanz bei der Einführung eines KI-Assistenzsystems zu erreichen.

… Sie sich dafür interessieren, wie Sie mehr Diversity in Beteiligungsprozessen erlangen.

… Sie wissen wollen, wann und wie unterschiedlichste Beteiligte in den Prozess der KI-Implementierung am besten einzubeziehen sind.

Um Assistenzsysteme am Arbeitsplatz passgenau und möglichst breit akzeptiert implementieren zu können, empfiehlt sich der Einsatz partizipativer Methoden. Damit können neben Arbeitgebern und Technologieentwickler*innen auch betroffene Arbeitnehmer*innen bzw. deren Interessenvertretungen am Implementierungsprozess von KI-Assistenzsystemen teilhaben.

Generell gibt es vier kritische Zeitpunkte der Entwicklung und Implementierung, die für eine faire und verantwortungsvolle KI im Betrieb zu beachten sind. Erstens, wenn die KI-Methode selbst entwickelt wird, zweitens, wenn das KI-System mit bestimmten Daten trainiert wird, drittens, wenn vortrainierte KI-Systeme im Unternehmen implementiert werden und viertens, wenn die mit dem KI-System gewonnenen Daten abgespeichert und aufbewahrt werden.[1] Bei der Einführung eines KI-Assistenzsystems am Arbeitsplatz ist es wichtig zu wissen, in welchem Stadium sich das KI-System befindet und welchen Handlungsspielraum es für die jeweiligen Interessenvertretungen (im Folgenden „Stakeholder“) gibt.

VIER KRITISCHE ZEITPUNKTE FÜR INTERVENTIONEN IN RICHTUNG FAIRE KI (Z. B. ZUR BIAS-PRÄVENTION):

1. Während der Entwicklung der KI-Methode (z. B. neuronales Netzwerk) und des Algorithmus
2. Während des Trainings des KI-Systems (d. h. Auswahl und Annotierung der Daten)
3. Während der Implementierung vortrainierter KI-Systeme (z. B. bei der Einbindung vortrainierter Systeme in Unternehmen)
4. Abspeicherung/Aufbewahrung der Daten (z. B. Wer hat Zugriff? Wie geschützt sind die Daten?)

Abbildung 1: Vier kritische Zeitpunkte für Interventionen in Richtung fairer KI[2]

1 Thaler et al., 2021.

2 Thaler et al., 2021.

Auch wenn KI-Systeme bereits entwickelt (Zeitpunkt 1) und mit externen Daten vortrainiert (Zeitpunkt 2) für den Betrieb angekauft wurden, sollten die Algorithmen und Trainingsdaten kritisch durchleuchtet werden (s. auch Kapitel Fragen, die Sie bei der (Mit-)Gestaltung von KI-Assistenzsystemen am Arbeitsplatz stellen sollten). Spätestens für die erfolgreiche Implementierung von KI-Assistenzsystemen (Zeitpunkt 3) und den Umgang gewonnener Daten (Zeitpunkt 4) ist die partizipative Einbindung von betroffenen Arbeitnehmenden bzw. deren Interessenvertretungen anzuraten.

Personen, die vom potentiellen Anwendungsfall betroffen sind oder einen Einfluss auf diesen haben (genannt „Stakeholder“), sollten in einem ersten Schritt identifiziert (zum Beispiel mit einem sog. Stakeholder Mapping) und dann mit einem partizipativen Verfahren an der Implementierung beteiligt werden.

Stakeholder Mapping

Um zu bestimmen, wer sinnvollerweise in einen Partizipationsprozess eingebunden werden soll, empfiehlt sich ein sogenanntes Stakeholder Mapping. Dabei spielt die Betroffenheit der potentiell Beteiligten durch das KI-Assistenzsystem eine wesentliche Rolle, wenn es darum geht, wie die Stakeholder konkret eingebunden werden sollen. Arbeitnehmer*innen, an deren Arbeitsplatz ein KI-Assistenzsystem eingesetzt werden soll, sind unmittelbar betroffen, während zum Beispiel Techniker*innen, die das neue KI-System technisch betreuen sollen oder Datenschutzbeauftragte, die das Einhalten von gesetzlichen Vorgaben bei der Speicherung und dem Abrufen von Daten des KI-Systems im Auge haben, mittelbar Beteiligte sind. Expertise bringen alle Stakeholder mit, denn auch eine unmittelbare Betroffenheit kann als spezifische Kompetenz und die eingebrachte Perspektive als wertvoller Beitrag im gemeinsamen Partizipationsprozess der Technologieentwicklung gesehen werden.

Aller Anfang ist leicht – eine erste Liste

Ein Stakeholder Mapping beginnt zunächst mit der Erstellung einer ersten Liste an potentiellen Beteiligten. Diese sollte möglichst divers sein (auch, was Alter, Geschlecht, Bildung, etc. angeht). Daher empfiehlt es sich, die Liste in Tabellenform anzulegen und relevante Diversity-Kriterien genauso anzuführen wie den Bezug der Person zum Thema – d. h., wie ein einzuführendes KI-Assistenzsystem diese betrifft (z. B. betroffene Arbeitnehmer*innen, Betriebsrät*innen).

Hierbei ist es wichtig, erstens bewusst Menschen zu suchen, die von der eigenen Sichtweise abweichen oder auch solche, die oft weniger mit Künstlicher Intelligenz bzw. Technologie assoziiert werden. Zweitens sollten Menschen einbezogen werden, die von dem KI-Anwendungsfall direkt und indirekt betroffen sein können.

Das zahlenmäßige Verhältnis der in dieser ersten Liste repräsentierten Personengruppen muss sich dabei nicht zwangsläufig an der betrieblichen oder gesellschaftlichen Repräsentanz orientieren, sondern sollte eher darauf abzielen, eine möglichst große Bandbreite an relevanten Personengruppen abzudecken.

Wer wird zum Mitreden eingeladen?

Im nächsten Schritt werden die gesammelten Personen aus der oben beschriebenen Liste (bzw. Tabelle) hinsichtlich ihres Einflusses auf das KI-Assistenzsystem und ihres Interesses in der Sache eingeschätzt (das eigentliche „Stakeholder Mapping“). Dazu können entweder die Namen der Beteiligten selbst, Initialen oder Nummern in einem entsprechenden Koordinatensystem verortet werden.

Abbildung 2: Visualisierung des Stakeholder Mappings[3]

3 Thaler et al., 2021.

Die Einschätzung des Einflusses und Interesses wird am besten in einem kleinen Team von Menschen getroffen, die die Beteiligten kennen bzw. deren Motivation beurteilen können. Es geht hier um keine wissenschaftlich genaue Analyse, sondern den Prozess der Teilhabe, strukturiert zu planen und somit zu vermeiden, dass wichtige Stimmen nicht gehört werden.

Das Interesse der Beteiligten („Stakeholder") ist ein Gradmesser dafür, wie motiviert und kontinuierlich jemand an einem Entwicklungsprozess teilnehmen wird. Dieses Interesse hängt wiederum von Ressourcen (z. B. Bekommen Arbeitnehmer*innen die benötigte Zeit als Arbeitszeit angerechnet?) und dem Willen der Veranstalter*innen ab, die Einbindung mehr oder weniger inklusiv zu gestalten (z. B. Wie viel Wissen wird vorausgesetzt? In welcher Sprache werden die partizipativen Methoden angewandt?)

Der Einfluss der Stakeholder orientiert sich an ihrer Position im Unternehmen und inwiefern sie auch die Interessen anderer mitvertreten (z. B. Repräsentant*innen von Belegschaftsgruppen wie Betriebsrät*innen oder Gleichstellungs- und Inklusionsbeauftragte) oder über die Organisationshierarchie in einer Multiplikationsfunktion tätig sind (z. B. Manager*innen, Betriebsärzt*innen, Personalentwickler*innen). Beteiligte im Unternehmen, die direkt betroffen sind (IT-Abteilung, Datenschutzbeauftragte, etc.), sollten jedenfalls eingebunden werden.

Mitreden ist nicht gleich mitreden – die vier Stufen der Partizipation

Durch die Positionierung der Stakeholder anhand der beiden Dimensionen *Interesse* und *Einfluss* zeigt sich, wie diese am besten im Prozess einzubinden sind. Daraus ergeben sich vier Stufen der Partizipation, d. h. der Teilhabe:

1. Personen mit geringem Interesse und geringem Einfluss werden – wie alle Mitarbeiter*innen – in regelmäßigen Abständen über den Partizipationsprozess und den Stand der Implementierung der Technologieentwicklung informiert. Das kann über Mitarbeiter*innenversammlungen, schriftliche direkte Nachrichten (SMS, E-Mails, etc.) oder als Kommunikation an die gesamte Belegschaft (kurze Erklärvideos, Unternehmenszeitungen, Newsletter, Aushänge, etc.) erfolgen. Wichtig ist hierbei immer, konkrete Ansprechpersonen zu nennen, bei denen weitere Informationen eingeholt werden können, sowie auf den gesamten Prozess (Wer macht was wann und warum?) zu verweisen, d. h., nachvollziehbar und transparent zu kommunizieren.

2. Stakeholder, die sehr viel Einfluss im Betrieb haben, aber wenig Interesse besitzen, an einem längerfristigen Prozess teilzunehmen, sollten in regelmäßigen Abständen konsultiert und deren Rat eingeholt werden; dies kann in strukturierten Gesprächen stattfinden (z. B. Expert*inneninterview).

3. Personen, die wiederum wenig Einfluss im Unternehmen, dafür aber großes Interesse an der partizipativen Technikgestaltung haben, sollten unbedingt aktiv eingebunden werden (z. B. mittels Umfragen).

4. Die letzte Gruppe an Stakeholdern, deren Interesse und Einfluss relativ groß ist, ist die Gruppe, aus der die konkreten Teilnehmer*innen für den eigentlichen Partizipationsprozess ausgewählt werden sollen. Diese sind ideal für „das Mitreden", weil sie durch ihren Einfluss einen Multiplikator*inneneffekt in der Organisation haben. Darüber hinaus trägt ihr Interesse am Thema generell und an der partizipativen Technikgestaltungsserie konkret dazu bei, dass sie kontinuierlich aktiv mitreden werden.

Mitreden, aber wie? – Der Partizipationsprozess

Neue Technologien, die am Arbeitsplatz eingesetzt werden, verursachen manchmal Verunsicherung bei Arbeitnehmer*innen, erst recht die nicht unmittelbar durchschaubaren KI-Assistenzsysteme (s. Abschnitt B, Kapitel Erklärbare Künstliche Intelligenz). Manche Unsicherheiten sind begründet, gerade, wenn es um den Umgang mit sensiblen Daten geht (s. Abschnitt C, Kapitel Gesetzliche Rahmenbedingungen für den Einsatz von KI-Assistenzsystemen am Arbeitsplatz), und das im betrieblichen Umfeld, in dem hierarchische Abhängigkeitsverhältnisse gegeben sind (s. Abschnitt B, Kapitel Warum man bei (KI-)Algorithmen am Arbeitsplatz genauer hinsehen sollte). Manche Ängste können jedoch durch eine gute Kommunikation mit den Betroffenen rechtzeitig abgebaut und die Einführung von KI-Systemen auch durch Einbeziehung von verschiedensten Beteiligten („Stakeholdern") verantwortungsvoll erfolgen.

Wir haben bereits im vorigen Abschnitt erläutert, dass, je nach Interesse und Einfluss, Stakeholder unter-

schiedlich beteiligt werden können. Methoden reichen von der regelmäßigen Information (über Newsletter, Aushänge, Videos, etc.) in allen für die Belegschaft relevanten Sprachen über Interviews mit einzelnen ausgewählten Expert*innen bis zu großen Umfragen im Betrieb.

Für die sehr interessierten und relativ einflussreichen Beteiligten empfiehlt sich die Planung eines Prozesses, der das regelmäßige Abhalten von Besprechungen und Workshops umfasst. Entsprechende Veranstaltungen können Präsentationen (von Wissen, Technologievorhaben, etc.) und Interaktionen (Diskussionen, Kleingruppenarbeiten, etc.) beinhalten. Diese sollten jedoch gut moderiert, d. h. auch entsprechend vor- und nachbereitet werden. Insbesondere benötigt es für solche Partizipationsprozesse klar kommunizierte und kommunizierende Verantwortliche, entsprechende Ressourcen, einen vorgegebenen Zeitablauf sowie Zielformulierung – d. h. nachvollziehbare Rahmenbedingungen: Wer soll wobei wie mitreden?

Rahmenbedingungen der Partizipation

Damit alle, die beim Implementieren eines KI-Assistenzsystems mitreden – aber auch alle anderen Kolleg*innen im Betrieb – wissen, worum es beim Partizipationsprozess geht, sollte dieser von Beginn an transparent und nachvollziehbar sein.

Folgende Aspekte sollten kommuniziert werden:

- Ziel und Absicht: Was wollen wir erreichen?
- Verantwortliche und deren Aufgaben: Wer tut was?
- Spielregeln des Prozesses: Wie gehen wir miteinander um?
- Spielraum: Was geschieht mit den Ergebnissen?
- Ablauf des Beteiligungsprozesses: Wann gibt es Zwischen- und Endergebnisse?
- Ressourcenaufwand: Wer investiert wie viel Zeit und bringt welche Kompetenzen mit?

Ein längerfristiger Partizipationsprozess benötigt eine Moderation, die möglichst wertfrei und ergebnisoffen durch den Verlauf begleitet. Diese muss auch dafür sorgen, dass alle Beteiligten diskriminierungsfrei und gleichgestellt mitreden können. Es empfiehlt sich, eine externe Person mit Moderationserfahrung dafür zu beauftragen.

Methoden der Partizipation

Wenn ein KI-Assistenzsystem am Arbeitsplatz eingeführt werden soll, können die Arbeitnehmenden also in unterschiedlichem Ausmaß beteiligt werden. Wir haben in diesem Kapitel bereits erläutert, dass die Beteiligung stark vom individuellen Interesse und vom Einfluss im Betrieb abhängt, d. h., das sogenannte Stakeholder Mapping mündet in die vier Stufen der Partizipation, d. h. es legt fest, wer wie intensiv eingebunden wird. Im Folgenden sehen wir uns nun die aktive Beteiligung und Teilhabe von Arbeitnehmenden bzw. deren Vertretungen genauer an. Mit welchen Methoden kann ein Partizipationsprozess, vom ersten Kennenlernen bis zur Abschlusspräsentation der Ergebnisse, angereichert werden? Wie kann sichergestellt werden, dass alle Stimmen gleich gehört und alle Perspektiven eingebracht werden können?

Hier finden Sie eine Auswahl an beliebten Methoden, die sowohl für einen längerfristigen Prozess einer Gruppe als auch für singuläre Partizipations-Ereignisse (z. B. auch zur Konfliktlösung) geeignet sind:

Dialog	Die Beteiligten erarbeiten in 1–3 Stunden eine gemeinsame Position zu einem Gegenstand, indem sie nacheinander (mithilfe eines Objekts, das als „Sprechstab“ fungiert) zum vorgegebenen Thema sprechen.
Konsensuskonferenz	Gut vorbereitete dreitägige Konferenz, in der Expert*innen Kurzreferate halten, alle Beteiligten sich austauschen und diskutieren und am Ende ein gemeinsames Konsenspapier erstellen.
Mediation	Klar strukturiertes – zeitlich flexibles – Verfahren zur Konfliktlösung mithilfe von externen, unparteiischen Mediator*innen.
Planspiel	Mehrstündige Simulation der komplexen Realsituation mithilfe vereinfachter Rahmenbedingungen und zugewiesenen Rollen. Durch Veränderungen von Spielbedingungen und sich daraus ergebenden Konsequenzen können Schlüsse für die Realität gezogen werden.
Runder Tisch	Durch eine neutrale Person moderierte Diskussion, in der alle gleichberechtigten Interessensgruppen durch gleich viele Personen vertreten sind, um eine gemeinsame Lösung zu erarbeiten.
World Café	An mehreren Tischen werden von fixen Gastgeber*innen Themen für 20–30 min moderiert. Die Diskussionsbeteiligten bringen sich verbal und/oder schriftlich (Papier als „Tischtuch“) ein, ehe sie zum nächsten Tisch weiterwechseln und dort wieder 20–30 min diskutieren.
Zukunftswerkstatt	Eintägige Kreativmethode, die von einer Kritik- über eine Utopie- zur Realisierungsphase verläuft und durch Anbieten verschiedenster Materialien die Fantasie der Beteiligten beflügeln soll.

Tabelle 1: Liste an verschiedenen Methoden für partizipative Prozesse

Der Vorteil einer derart partizipativ durchgeführten, gemeinsamen Wissensproduktion und Technologieentwicklung (sogenannte „Co-Creation“) liegt darin, dass von Beginn an Akzeptanz für die Technologie auf- und Ängste abgebaut werden können.

Je nach Prozess ist eventuell eine Datenschutzerklärung vonnöten. Eine Vorlage können Sie unter dem Kapitel Datenschutzerklärung zum Übernehmen entnehmen.

SIE HABEN NOCH NICHT GENUG?

Das folgende digitale Buch des Bundesministeriums für Kunst, Kultur, Öffentlichen Dienst und Sport thematisiert, wie digitale Technologien dazu eingesetzt werden können, um Mitbestimmung zu befördern:

- *Grünbuch: Partizipation im digitalen Zeitalter*[4]

Die Partizipations-Initiative des Bundesministeriums für Klimaschutz, Umwelt, Energie, Mobilität, Innovation und Technologie bietet eine umfassende Darstellung sämtlicher oben erwähnter und weiterer partizipativer Methoden – wir empfehlen diese praktische Plattform unbedingt:

- *partizipation.at*

Konkrete Workshopkonzepte zur Beförderung von fairer und vertrauenswürdiger Künstlicher Intelligenz in Österreich finden Sie beim Projekt dAIalog.at:

- *How to make it fAIr – Methoden partizipativer Technikgestaltung für das Anwendungsfeld der Künstlichen Intelligenz*[5]

ANSPRECHPERSON

MMag.a Dr.in Anita Thaler, IFZ
Expertin für technikbezogene Partizipationsprozesse
Anita.Thaler@ifz.at

LITERATURVERZEICHNIS

Anita Thaler, Vogel, Susanna, Anslinger, Julian, Frauenberger, Christopher, Bentegeac, Philippe, Fiser, Scarlet, Krischkowsky; Alina, und Lindlbauer, Andreas, 2021. *How to make it fAIr – Methoden partizipativer Technikgestaltung für das Anwendungsfeld der Künstlichen Intelligenz.* dAIalog.at. *https://doi.org/10.13140/RG.2.2.20812.95360*.

BMKÖS. 2020. *Grünbuch: Partizipation im digitalen Zeitalter. https://www.oeffentlicherdienst.gv.at/verwaltungsinnovation/oeffentlichkeitsbeteiligung/201103_Partizipation_Gruenbuch_A4_BF_1.pdf?7t15d4*.

4 BMKÖS, 2020.

5 Thaler et al., 2021.

Mitbestimmung bei Betriebsvereinbarungen

Relevante Punkte im Arbeitsverfassungsgesetz

Wie kann Mitbestimmung bei der Einführung und Gestaltung von KI-Assistenzsystemen am Arbeitsplatz gelingen? Ein Weg führt über Betriebsvereinbarungen. In Österreich hat nämlich der Betriebsrat, der bereits ab fünf stimmberechtigten Arbeitnehmer*innen einzurichten ist,[1] mit Betriebsvereinbarungen (BV) eines der stärksten Mitbestimmungsrechte im Betrieb. Je nach Tatbestand *können* oder *müssen* Betriebsvereinbarungen abgeschlossen werden. Im letzteren Fall spricht man von sogenannten „Erlaubnistatbeständen". Diese sind in den §§ 96 und 96a des Arbeitsverfassungsgesetzes (ArbVG) geregelt[2]. Alle anderen Angelegenheiten, wegen derer Betriebsvereinbarungen geschlossen werden *können*, sind in §97 ArbVG angeführt. Für Betriebsvereinbarungen zur Einführung und Gestaltung von KI-Assistenzsystemen sind insbesondere die in Tabelle 1 angeführten Paragraphen relevant, die die Anwendungsbereiche vieler Systeme abdecken.

§ 96 Abs 1 Z 3 ArbVG *(BV erforderlich)*	„die Einführung von Kontrollmaßnahmen und technischen Systemen zur Kontrolle der Arbeitnehmer, sofern diese Maßnahmen (Systeme) die Menschenwürde berühren;"
§ 96a Abs 1 Z 1 ArbVG *(BV erforderlich)*	„Die Einführung von Systemen zur automationsunterstützten Ermittlung, Verarbeitung und Übermittlung von personenbezogenen Daten des Arbeitnehmers, die über die Ermittlung von allgemeinen Angaben zur Person und fachlichen Voraussetzungen hinausgehen. Eine Zustimmung ist nicht erforderlich, soweit die tatsächliche oder vorgesehene Verwendung dieser Daten über die Erfüllung von Verpflichtungen nicht hinausgeht, die sich aus Gesetz, Normen der kollektiven Rechtsgestaltung oder Arbeitsvertrag ergeben;"
§ 96a Abs 1 Z 2 ArbVG *(BV erforderlich)*	„die Einführung von Systemen zur Beurteilung von Arbeitnehmern des Betriebes, sofern mit diesen Daten erhoben werden, die nicht durch die betriebliche Verwendung gerechtfertigt sind."
§ 97 Abs 1 Z 6 ArbVG	„Maßnahmen zur zweckentsprechenden Benützung von Betriebseinrichtungen und Betriebsmitteln;"
§ 97 Abs 1 Z 9 ArbVG	„Maßnahmen zur menschengerechten Arbeitsgestaltung;"

Tabelle 1: Überblick über die für die Einführung von KI-Assistenzsystemen relevanten Paragraphen im Arbeitsverfassungsgesetz

Mit einer Betriebsvereinbarung wird ein Interessensausgleich zwischen der Arbeitnehmer*innenschaft, vertreten durch den Betriebsrat, und dem Arbeitgeber geschaffen.

1 Das Fehlen eines Betriebsrates ist jedoch an keinerlei Rechtsfolgen geknüpft. Wichtig ist, dass der Arbeitgeber die Wahl eines Betriebsrates in keiner Weise verhindern darf: *https://www.usp.gv.at/mitarbeiter/betriebsrat.html*

2 Bundesministerium für Digitalisierung und Wirtschaftsstandort, 2022.

Analyse der Zweck-Mittel-Relation

Was gilt es dabei in Hinblick auf KI-Assistenzsysteme nun zu beachten? Aus Sicht des Betriebsrats ist es ratsam, sich zu Beginn der Aushandlung mit der Ergründung einer angemessenen **Zweck-Mittel-Relation** zu befassen[3]. Das heißt, es muss geklärt werden, welchen Zweck der Arbeitgeber mit der Einführung des KI-Assistenzsystems verfolgt und ob dieser Zweck legitim und rechtlich zulässig ist. Anschließend muss geprüft werden, ob sich das KI-Assistenzsystem überhaupt eignet, den Zweck zu erfüllen und ob es nicht andere, gelindere Mittel gäbe, das gesetzte Ziel zu verfolgen. Sind diese Fragen hinreichend geklärt, lohnt sich ein Detailcheck des geplanten KI-Assistenzsystems, zum Beispiel unter Zuhilfenahme der in Kapitel Fragen, die Sie bei der (Mit-)Gestaltung von KI-Assistenzsystemen am Arbeitsplatz stellen sollten angeführten Checkliste dieses Abschnitts. Eine Übersicht dieser Schritte findet sich in Abbildung 1.

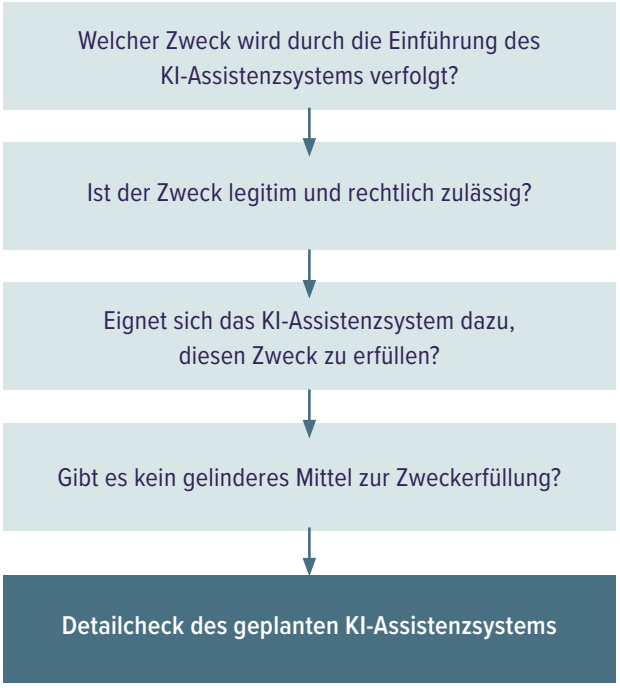


Abbildung 1: Prozess der Ergründung der adäquaten Zweck-Mittel-Relation

Besonders empfehlenswert: Rahmenbetriebsvereinbarungen

Das Aushandeln von Betriebsvereinbarungen für jedes geplante technische System ist mit erheblichem Aufwand verbunden. Vor dem Hintergrund der Häufigkeit und Schnelligkeit, mit der neue technische Systeme in Betrieben eingeführt werden sollen und dem großen Aufgabenbereich von Betriebsrät*innen kann das detailgenaue Überprüfen der einzuführenden Systeme schon einmal auf der Strecke bleiben – so die Rückmeldung einiger im Rahmen von VEKIAA befragten Arbeitnehmer*innenvertretungen. Vor diesem Hintergrund möchten wir empfehlen, Rahmenbetriebsvereinbarungen abzuschließen, die die Einführung und/oder Gestaltung von Künstlicher Intelligenz bzw. KI-Assistenzsystemen am Arbeitsplatz für alle zukünftigen einzuführenden und bestehenden Systeme regeln.

Folgende Aspekte halten wir beim Vereinbaren von Rahmenbetriebsvereinbarungen für besonders relevant:

1. **Festlegung von Beschaffungs- oder Zulassungskriterien**
   Eine Rahmenbetriebsvereinbarung kann erstens dazu dienen, betriebsinterne Beschaffungs- und/oder Zulassungskriterien für KI-Assistenzsysteme zu formulieren, wie es sie beispielsweise auch oft im Hinblick auf die Barrierefreiheit anderer Technologien gibt. Entsprechende Kriterien können Punkte umfassen wie zum Beispiel KI-Transparenz, hohe Qualität der Trainingsdaten oder eine obligatorische Technikfolgenabschätzung (s. auch Abschnitt C, Kapitel Fragen, die Sie bei der (Mit-)Gestaltung von KI-Assistenzsystemen am Arbeitsplatz stellen sollten). Es empfiehlt sich, die Anforderungen auf den Betrieb und die möglichen KI-Assistenzsysteme abzustimmen und diese möglichst präzise zu formulieren.

3 Das Finden einer richtigen Zweck-Mittel-Relation wird in Bezug auf datenschutzrelevante Fragestellungen von Clara Fritsch und Susanne Haslinger im empfehlenswerten Buch „Beschäftigtendatenschutz. Handbuch für die betriebliche Praxis“ beschrieben. Fritsch und Haslinger, 2020.

Das Festlegen solcher Beschaffungskriterien wirkt sich nicht nur auf die implementierten Systeme aus, sondern erzeugt auch einen gewissen Marktdruck, an den sich Anbieter von KI-Systemen anzupassen haben. Entsprechend könnten Anbieter von KI-Assistenzsystemen sogar dazu aufgefordert werden, bestimmte Informationen über die jeweiligen Systeme zur Verfügung zu stellen. Beispielsweise könnte festgelegt werden, dass jeder Anbieter die ausgefüllte, von der EU gestaltete ALTAI-Checkliste[4] zur Verfügung zu stellen hat, um nachvollziehen zu können, wie vertrauenswürdig das KI-System ist.

2. **Informationspflichten des Betriebs**
   In einer Rahmenbetriebsvereinbarung kann festgelegt werden, wann und worüber der Betrieb den Betriebsrat (und evtl. die Arbeitnehmer*innenschaft) zu informieren hat. Beispielsweise könnte geregelt werden, dass der Betrieb den Betriebsrat mit einer festgelegten Vorlaufzeit über einen geplanten Einkauf eines KI-Assistenzsystems zu informieren hat. Der Betrieb sollte dem Betriebsrat hierbei sogleich Informationen über den Zweck des Systems, die vorauszusehenden Auswirkungen und eine Systembeschreibung zur Verfügung stellen. Auch könnte der Betrieb dazu aufgefordert werden, Fragen zu beantworten, wie sie in unserer Checkliste angeführt sind (z. B., welche Datenkategorien verwendet werden). Achtung: Sobald es um personenbezogene Daten geht, muss der Betriebsrat eingebunden werden (s. auch Abschnitt C, Kapitel Gesetzliche Rahmenbedingungen für den Einsatz von KI-Assistenzsystemen am Arbeitsplatz).

3. **Recht auf Ansprechpersonen und System-Demonstration**
   Auch kann festgelegt werden, dass der Betrieb dem Betriebsrat relevante Ansprechpersonen zur Verfügung zu stellen hat. Beispielsweise könnte geregelt werden, dass Betriebsrät*innen jederzeit mit der IT-Abteilung des Betriebs in Dialog treten können, um System-Demonstrationen anzufordern und Detailfragen zu klären.

4. **Etablierung einer Testphase und Ausarbeitung eines Rückbaukonzepts**
   Das Vereinbaren eines Testzeitraumes für neue KI-Assistenzsysteme ist eine vertrauensbildende Maßnahme, die es Betriebsrat und Arbeitgeber ermöglicht, die Einführung der Systeme in einem festgelegten Zeitraum anzupassen oder sogar rückgängig zu machen. Dies könnte beispielsweise dann geschehen, wenn das KI-Assistenzsystem zuvor festgelegten Kriterien nicht entspricht oder unvorhergesehene Aspekte die Bedürfnisse der Belegschaft einschränken. Für den Betrieb empfiehlt es sich, Rückbaukonzepte auszuarbeiten, sodass jederzeit zu der ursprünglichen Verfahrensweise zurückgekehrt werden kann.

Auch für zukünftige Updates eines KI-Assistenzsystems sollte ein Rückbaukonzept vereinbart werden. Denn in der Praxis zeigt sich, dass diese oft am Betriebsrat vorbei geschehen,[5] und ungewünschte Änderungen mit sich bringen können.

Selbst eine Betriebsvereinbarung kann einer Testphase unterzogen und als modifizierbar vereinbart werden. Mehr dazu findet sich in der Infobox zu lebenden Betriebsvereinbarungen.

LEBENDE BETRIEBSVEREINBARUNGEN

Einen innovativen Weg der Betriebsvereinbarungen gingen Leitung und Betriebsrat der GASAG-Gruppe in Deutschland. Zum Regeln einer Plattform für internes Crowdsourcing riefen sie eine „Lebende Konzernbetriebsvereinbarung“ ins Leben.

- „Die ‚Lebende Konzernbetriebsvereinbarung‘ für plattformbasiertes internes Crowdsourcing eröffnet der Arbeitskultur und dem IC neue Handlungsfelder. Konzernleitung und Konzernbetriebsrat verstehen die Vereinbarung als Einstieg in einen nach vorne offenen, ganzheitlichen Prozess der vorausschauenden Arbeitsgestaltung. Beide Partner verpflichten sich auf ein kooperatives Veränderungsmanagement auf gleicher Augenhöhe, das sowohl ergebnisorientiert wie fachlich und methodisch agil organisiert wird.”[6]

- „Die ‚Lebende Konzernbetriebsvereinbarung‘ will auf der Basis eines agilen, kooperativen Changemanage- ▶

4 High-Level Expert Group on AI, 2020.
5 S. z. B. Christl, 2021.
6 Otte und Schröter, 2018.

▶ ments wechselseitiges, konzernöffentliches Lernen und den Erwerb von Gestaltungskompetenz befördern. Dazu finden in vereinbarten regelmäßigen Zeitabständen gemeinsame Überprüfungen, gemeinsame Auswertungen und gemeinsame Updates der Konzernbetriebsvereinbarung wie auch gemeinsam strukturierte Evaluierungen ihrer praktischen Umsetzung statt. Beide Partner passen die Konzernbetriebsvereinbarung den erworbenen Erfahrungen und Lernschritten jeweils gemeinsam an."[7]

Auch in Bezug auf KI-Assistenzsysteme ist eine solche lebende Betriebsvereinbarung durchaus denkbar. Arbeitgeber und Betriebsrat können sich über einen längeren Zeitraum ein Bild von der eingeführten Technologie und ihren Folgen für den Betrieb machen, lernen dazu und passen die Betriebsvereinbarung in einem ergebnisoffenen Dialogprozess an. Sogar eine formalisierte regelmäßige Evaluierung oder Auditierung ist denkbar.

Eine Beschreibung des Prozesses und die Betriebsvereinbarung des obigen Beispiels kann hier im Wortlaut nachgelesen werden: *http://www.blog-zukunft-der-arbeit.de/wp-content/uploads/2018/07/Lebende_KBV_Otte_Schroeter.pdf*

5. **Regelmäßige Evaluierung /Auditierung**
   Unabhängig von einer Testphase kann eine regelmäßige Evaluierung/Auditierung bestimmt werden. Hierbei sollte vereinbart werden, wie oft diese stattfindet, was durch wen evaluiert/auditiert wird und welche Maßnahmen bei negativen Ergebnissen getroffen werden.

6. **Informations-, Kontroll- und Zugangsrechte des Betriebsrats**
   Eine Rahmenbetriebsvereinbarung kann dem Betriebsrat auch konkrete Informations-, Kontroll- und Zugangsrechte zur Seite stellen. Es könnte dem Betriebsrat beispielsweise freigestellt werden, Einsicht in die genaue Projektplanung und die Dokumentation des KI-Assistenzsystems zu nehmen. Auch wäre das Zurverfügungstellen von Leseberechtigungen für die vom KI-Assistenzsystem verarbeiteten Daten denkbar.

7. **Recht auf Qualifizierung und das Hinzuziehen von Expert*innen**
   In der Arbeit mit Betriebsrät*innen haben wir die Rückmeldung erhalten, wie wichtig es eigentlich wäre, KI-Expertise im Betriebsrat weiter auszubauen. Hierfür fehlten jedoch leider häufig die Ressourcen. In einer Rahmenbetriebsvereinbarung könnte der Aufbau von KI-Expertise mit einem Schulungsrecht des Betriebsrats mit Kostenübernahme durch den Arbeitgeber festgelegt werden.[8]

8. **Recht auf Technikfolgenabschätzung oder Zertifizierung**
   Auch das vom Betrieb finanzierte Hinzuziehen von KI-Expert*innen zur Technikfolgenabschätzung oder Zertifizierung des KI-Assistenzsystems wäre durchaus denkbar in eine Rahmenbetriebsvereinbarung einzubauen – übrigens eine Maßnahme, die mit der Einführung des Betriebsrätemodernisierungsgesetzes in Deutschland im Jahr 2021 gesetzlich festgelegt wurde. Hier gilt die Hinzuziehung einer sachverständigen Person bei der Einführung oder Nutzung eines KI-Systems sogar als erforderlich (§ 80 Abs 3 BetrVG)[9]!

9. **Institutionalisierung der Einbindung von Arbeitnehmer*innen, deren Vertretungen und weiteren Interessenvertretungen**
   In Kapitel KI im Betrieb: Wer soll wie mitreden? (Abschnitt C) beschreiben wir, wie bei der Gestaltung und Einführung von KI-Assistenzsystemen die Einbindung von Arbeitnehmenden und ihren Vertretungen von Anfang an gelingen kann. In einer Rahmenbetriebsvereinbarung gibt es die Möglichkeit, diese Form der Mitbestimmung für jede Einbindung oder Gestaltung von KI-Assistenzsystemen zu erwirken. Konkret bedeutet dies, dass sich der Betrieb dazu verpflichten würde, vor dem Einkauf oder der Gestaltung eines KI-Assistenzsystems alle relevanten Interessenvertretungen an einen Tisch zu holen, das System und seine Auswirkungen zu diskutieren sowie eventuelle Modifizierungen auszuarbeiten. Dies kann unter anderem umfassen: den Betriebsrat, das Management, die IT-Technik, Nutzer*innen, Datenschutzbeauftragte, Beauftragte für Menschen mit Behinderungen, Gleichstellungsbeauftragte sowie externe Expert*innen oder Gremien. Wenn möglich, würden wir sogar empfehlen,

7 Otte und Schröter, 2018.

8 S. auch GPA-djp, Abt. Arbeit & Technik, 2017.

9 Bundesamt für Justiz, 2022.

eine offene, aber festgelegte Gruppe einzurichten, die sich regelmäßig mit den im Betrieb auftretenden KI-bezogenen Fragestellungen befasst.

10. **Konkretisierung der Umsetzung von gesetzlichen Regelungen**
    Weiters kann eine Rahmenbetriebsvereinbarung dazu dienen, Gesetze in konkrete betriebliche Regelungen auszugestalten. Dies gilt beispielsweise für die Rechte der Arbeitnehmer*innen nach Art 15-22 DSGVO[10], d. h., das Recht auf Auskunft, Berichtigung, Löschung, Verarbeitungseinschränkung, Widerspruch, etc. (s. auch Abschnitt C, Kapitel Gesetzliche Rahmenbedingungen für den Einsatz von KI-Assistenzsystemen am Arbeitsplatz). Aber auch der geplante Verordnungsvorschlag der Europäischen Kommission zur Festlegung harmonisierter Vorschriften für Künstliche Intelligenz[11] könnte hier berücksichtigt werden, sobald er schlagend wird.

Keine Frage, alle oben beschriebenen Aspekte in einer Rahmenbetriebsvereinbarung zu berücksichtigen, verlangt Verhandlungsgeschick und ausreichend Ressourcen. Aber der Aufwand lohnt sich. Wie ein von uns befragter Betriebsrat ausdrückte: „Das Vereinbaren einer Rahmenbetriebsvereinbarung spart Zeit, Nerven und erleichtert das Leben aller Beteiligten."

Was passieren kann, wenn Mitbestimmungsverfahren nicht durchgeführt werden oder scheitern, können Sie in der Infobox zu Algoaktivismus nachlesen.

ALGOAKTIVISMUS

Nicht immer können KI-Assistenzsysteme bzw. Algorithmen, die an einem Arbeitsplatz eingeführt werden, aktiv mitgestaltet werden. In manchen Fällen stellt sich der Betrieb quer und unterbindet jegliche Mitbestimmung, die darüber hinausgeht, was das Gesetz verlangt. In anderen Fällen fehlen dem Betriebsrat die Ressourcen, um sich neben dem Alltagsgeschäft damit auseinanderzusetzen, welche Auswirkungen ein Algorithmus auf die Arbeitnehmenden hat. In wieder anderen Fällen und kleineren Firmen gibt es vielleicht gar keine Arbeitnehmendenvertretung, die sich dem Themengebiet annehmen kann. Doch auch in diesen scheinbar ausweglosen Situationen etablierten Arbeitnehmende Verhaltensweisen, die es ihnen erlauben, sich den negativen Auswirkungen von KI-Assistenzsystemen zu entziehen. ▶

▶ Von Wissenschaftlerinnen wurde dieses jüngst entdeckte Phänomen als „Algoaktivismus" bezeichnet. Katherine Kellog vom MIT und Kolleginnen aus Stanford beschreiben in ihrem Artikel „Algorithms at Work: The New Contested Terrain of Control"[12] verschiedene Formen von Taktiken des Widerstandes, die Arbeitnehmer*innen gegen am Arbeitsplatz eingeführte Algorithmen einnehmen.

Im Folgenden möchten wir einige dieser Taktiken auflisten und kurz beschreiben:

- **Kooperationsunterlassung durch Ignoranz**: Die Empfehlungen eines Algorithmus werden ignoriert. Dies kann insbesondere dann passieren, wenn die Hinweise nicht mit der Expertise der Arbeitnehmenden übereinstimmen oder wenig nachvollziehbar sind.

- **Kooperationsunterlassung durch Abschaltung**: Das System wird abgeschaltet, wenn es unerwünschte Auswirkungen hat. Ein Beispiel dafür liefern Fahrer*innen des Personenbeförderungsunternehmens Uber, die beobachtet wurden, wie sie sich regelmäßig aus dem System ausloggten, um die Wahrscheinlichkeit für lange Fahrten zu reduzieren oder wie sie das Programm außerhalb von Wohngebieten abschalteten, um Taxifahrten für Barbesucher*innen zu umgehen.

- **Kooperationsunterlassung durch Manipulation:** Auch wurden Fälle beobachtet, in denen Arbeitnehmende die Funktionsweise eines intransparenten Algorithmus durch Probieren herausfanden, um ihn dann durch eine Anpassung des Inputs zu einem gewünschten Ergebnis zu zwingen.

- **Kollektives Handeln:** Ein Beispiel für kollektives Handeln (engl. collective action) stellt die sogenannte #slaveroo-Bewegung deutscher, englischer, irischer und niederländischer Essens-Lieferant*innen dar, die sich aufgrund von prekären Arbeitssituationen gegen ihre Arbeitgeber (Deliveroo, Uber Eats, etc.) auflehnten. Neben scheinbar willkürlichen Lohnkürzungen standen vor allem durch das algorithmenbasierte Management erfolgte Dehumanisierung und Diskriminierung im Mittelpunkt.[13] ▶

10 Europäisches Parlament und Europäischer Rat, 2016.
11 EU, 2021.
12 Kellogg, Valentine, und Christin, 2020.
13 Klikauer und Young, 2022.

- **Besetzung des Diskurses über KI-Ethik:** Der Diskurs über verantwortungsvolle KI wird maßgeblich von Wissenschaftler*innen, Betrieben und Staaten(-Verbünden) geführt. Doch immer mehr unmittelbar betroffene Arbeitnehmer*innen schalten sich in die Diskussionen ein und bestimmen mit, was als verantwortungsvolle, faire und transparente KI am Arbeitsplatz geführt wird. Hier scheinen betriebsübergreifende Netzwerke besonders erfolgreich.

SIE HABEN NOCH NICHT GENUG?

Wenn Sie einen vertiefenden Einblick in Datenschutz in der betrieblichen Praxis erhalten möchten, empfehlen wir das ausgezeichnete Buch:

- *Beschäftigtendatenschutz. Handbuch für die betriebliche Praxis*, herausgegeben von Susanne Haslinger, Andreas Krisch und Thomas Riesenecker-Caba[14]

Eine Muster-Rahmenbetriebsvereinbarung der GPA-djp finden Sie hier:

- *Rahmenbetriebsvereinbarung über die Verwendung personenbezogener Beschäftigtendaten*[15]

ANSPRECHPERSONEN

Abteilung Recht der Produktionsgewerkschaft PRO-GE
Beratung von Betriebsrät*innen (Produktion) zu Datenschutz am Arbeitsplatz
recht@proge.at

Dr.in Eva Angerler, GPA-djp, Abteilung Arbeit und Technik
Beratung von Betriebsrät*innen (Angestellte Privatwirtschaft) zu Datenschutz am Arbeitsplatz
Eva.Angerler@gpa.at

Fridolin Herkommer
Leiter des Büros für Digitale Agenden in der AK Wien, Digitalisierungsfonds Arbeit 4.0 und bis Oktober 2021 Mitglied des Österreichischen Rats für Robotik und KI
Fridolin.Herkommer@akwien.at

14 Haslinger, Krisch und Riesenecker-Caba, 2020.

15 GPA-djp, 2017.

LITERATURVERZEICHNIS

Bundesamt für Justiz. 2022. *Betriebsverfassungsgesetz*. *https://www.gesetze-im-internet.de/betrvg/__80.html*.

Bundesministerium für Digitalisierung und Wirtschaftsstandort. 2022. *Bundesrecht konsolidiert: Gesamte Rechtsvorschrift für Arbeitsverfassungsgesetz*. *https://www.ris.bka.gv.at/GeltendeFassung.wxe?Abfrage=Bundesnormen&Gesetzesnummer=10008329*.

Christl, Wolfie. 2021. *Digitale Überwachung und Kontrolle am Arbeitsplatz. Von der Ausweitung betrieblicher Datenerfassung zum algorithmischen Management? https://crackedlabs.org/dl/CrackedLabs_Christl_UeberwachungKontrolleArbeitsplatz.pdf*.

EU. 2021. *Proposal for a Regulation of the European parliament and of the council laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts.* EUR-Lex – 52021PC0206 – EN §. *https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1623335154975&uri=CELEX%3A52021PC0206*.

Europäisches Parlament und Europäischer Rat. 2016. *Verordnung zum Schutz natürlicher Personen bei der Verarbeitung personenbezogener Daten, zum freien Datenverkehr und zur Aufhebung der Richtlinie 95/46/EG (Datenschutz-Grundverordnung)*, Verordnung (EU) 2016/679 §. *https://eur-lex.europa.eu/legal-content/DE/TXT/HTML/?uri=CELEX:02016R0679-20160504&from=EN*.

Fritsch, Clara, und Susanne Haslinger. 2020. DSGVO im Betrieb? Die Betriebsvereinbarung bringt's! In *Beschäftigtendatenschutz. Handbuch für die betriebliche Praxis,* herausgegeben von Susanne Haslinger, Andreas Krisch, und Thomas Riesenecker-Caba, 2. Aufl., 154–71.

GPA-djp, Abt. Arbeit & Technik. 2017. *Rahmenbetriebsvereinbarung über die Verwendung personenbezogener Beschäftigtendaten. https://bildung.gpa-djp.at/files/2017/10/GPA-djp-Muster-Betriebsvereinbarung-Datenschutz-DSGVO.pdf*.

Haslinger, Susanne, Andreas Krisch, und Thomas Riesenecker-Caba, Hrsg. 2020. *Beschäftigtendatenschutz. Handbuch für die betriebliche Praxis*. 2. Aufl.

High-Level Expert Group on AI. 2020. *Assessment List for Trustworthy Artificial Intelligence (ALTAI) for Self-Assessment. https://doi.org/10.2759/002360*.

Kellogg, Katherine C., Melissa A. Valentine, und Angéle Christin. 2020. Algorithms at Work: The New Contested Terrain of Control. *Academy of Management Annals* 14, Nr. 1: 366–410. *https://doi.org/10.5465/annals.2018.0174*.

Klikauer, Thomas, und Meg Young. 2022. Slaveroo @ Deliveroo – Platform Labor. *countercurrents.org* (Blog). *https://countercurrents.org/2022/03/slaveroo-deliveroo-platform-labor/*.

Otte, Andreas, und Welf Schröter. 2018. *Lebende Konzernbetriebsvereinbarung als soziale Innovation. Internes Crowdsourcing in der GASAG-Gruppe. Bedeutung – Bewertung – Wortlaut. http://www.blog-zukunft-der-arbeit.de/wp-content/uploads/2018/07/Lebende_KBV_Otte_Schroeter.pdf*.

Gesetzliche Rahmenbedingungen für den Einsatz von KI-Assistenzsystemen am Arbeitsplatz

Ein Gastbeitrag von Susanne Haslinger

DAS FOLGENDE KAPITEL IST FÜR SIE VON INTERESSE, WENN …

… Sie wissen wollen, welche rechtlichen Rahmenbedingungen beim Einsatz von KI in der Arbeitswelt zu berücksichtigen sind.

… Sie Näheres darüber wissen wollen, warum Maschinen niemals neutral sind und wer für allfällige Diskriminierung zur Verantwortung gezogen werden kann.

… es Sie interessiert, ob und wie unser aktuelles Datenschutz- und Antidiskriminierungsrecht auf KI angewendet werden kann.

Algorithmische Entscheidungsfindung ist im Alltag mittlerweile sehr präsent geworden – sie begegnet uns bei Internet-Suchmaschinen, personalisierter Werbung, automatisierten Kreditvergaben etc. Neben einer Vielzahl an Alltagsvereinfachungen, die die sog. „Künstliche Intelligenz“ verspricht, nehmen wir einen zunehmenden Verlust von Autonomie und unserer informationellen Selbstbestimmung wahr[1]. Künstliche Intelligenz greift in unseren Alltag ein, ihre Ergebnisse sind für uns in aller Regel nicht nachvollziehbar oder durchschaubar, ihre Entscheidungen können von uns nicht beeinflusst und auch nicht beeinsprucht werden. Wenig überraschend ist gerade die Arbeitswelt ein besonders heikler Bereich für den Einsatz algorithmenbasierter Entscheidungs- und/oder Überwachungssysteme. Kurzum: *„Die potenziellen Auswirkungen auf menschenwürdige Arbeitsbedingungen sind beträchtlich. Während die algorithmische Entscheidungsfindung im Prinzip auch zur Verbesserung der Arbeitsqualität beitragen kann, haben Fallstudien bereits eine Reihe beunruhigender Auswirkungen aufgezeigt, die von ständiger Mikroüberwachung bis zu automatisierter Diskriminierung reichen“*[2].

In (unter anderem) folgenden Bereichen lässt sich immer öfter der Einsatz algorithmenbasierter Technologie in der Arbeitswelt beobachten[3]:

- Recruitingsoftware bei der Bewerbung und Begründung des Dienstverhältnisses, inkl. Persönlichkeitstests
- Automatische Verwarnungen bis hin zur Beendigung des Dienstverhältnisses, Entscheidungsgrundlage für Beendigungen
- Smart Factory, intelligente Produktionsmaschinen, Überwachung als „Nebeneffekt“ von Betriebsdatensammlung

1 Vgl. u.a. WP249 11ff.

2 Prassl-Adams, 195.

3 Diese Aufzählung ist keinesfalls abschließend. Für eine ausführliche Aufstellung und Analyse siehe z. B. Christl 2021.

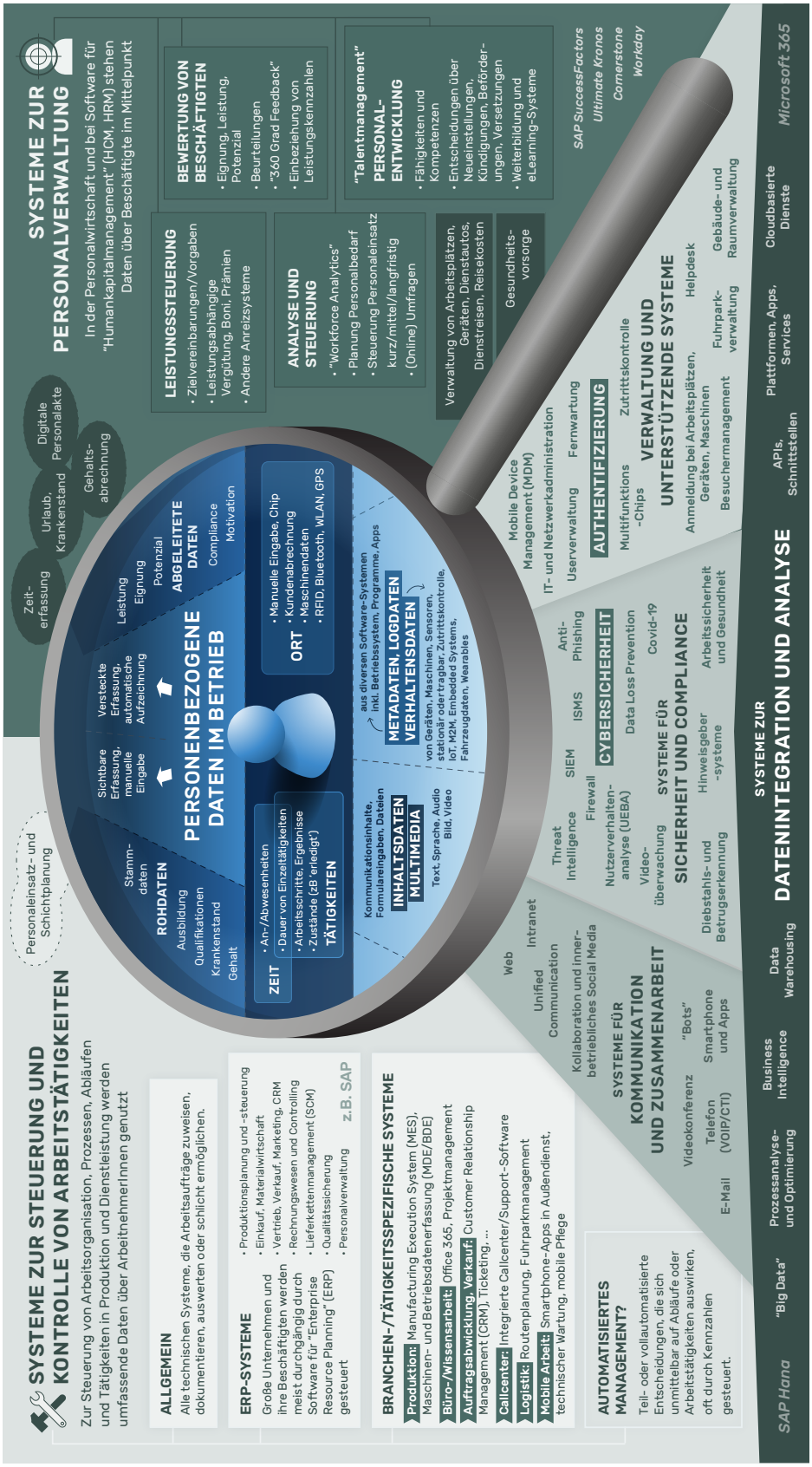
- Überwachung mobiler Arbeit (Montage, Pflege); Auftragsdatenerfassung und -verarbeitung, Auftragszeiten etc.
- Automatische Zuweisung des/der nächsten Auftrags/Kund*in/Gesprächs anhand der Analyse und Kategorisierung von Kund*innen- und/oder Mitarbeiter*innendaten
- Vorgabe von Arbeitsschritten bei Wartungen, Reparaturen, Standardabläufen etc.
- Kontrolle und Überwachung im Rahmen von Office-Software
- HR (Personalmanagement, Lohnverrechnung, Personalentwicklung und -beurteilung), Dienst- und Arbeitszeiteinteilung
- Unterstützende Tools (Smart Devices) wie Datenbrillen, Wearables etc.
- Verhaltenssteuerung durch Anreizsysteme und/oder Druck (Ampelsysteme, Gamification, Punktevergabe etc.)
- Eigene Systeme wie z. B. Plattformarbeit[4]

Neben klassischen datenschutzrechtlichen Problemen und Mitbestimmungsfragen (insb. Kontrolle und Überwachung der Arbeitnehmer*innen) ergeben sich daraus für das Arbeitsverhältnis eine Reihe neuer Problemstellungen, die konkrete, schwerwiegende rechtliche Folgen nach sich ziehen können:

- **Entscheidungen**, die nicht mehr vom Menschen (allein) getroffen werden, sondern an algorithmusbasierte Entscheidungssysteme ausgelagert werden oder die sich in ihrer Entscheidungsgrundlage zumindest zentral auf deren Output stützen (während die Entscheidung selbst nach wie vor von einer natürlichen Person getroffen wird), z. B. Selektion von Bewerbungen, Personalbeurteilung, Beförderungen etc.
- **Zufallsfunde**, die arbeitsrechtliche Auswirkungen mit sich bringen und im Rahmen der exzessiven Datenerhebung und Analyse durch Maschinelles Lernen, z. B. in der Produktion, bei mobilen Endgeräten etc. Wir kennen derlei aus „herkömmlichen“ (wenn auch digitalen) Überwachungsmaßnahmen, was nicht zuletzt den Grund für die diesbezügliche Betriebsvereinbarungspflicht darstellt, sobald das System in seiner potentiellen Überwachungsintensität die Menschenwürde der Arbeitnehmer*innen berührt.
- Ableitungen aus **maschinell entdecktem „abweichenden“ Verhalten**, aus dem Rückschlüsse getroffen werden (aus der Analyse der verfassten E-Mails ergibt sich, dass eine Mitarbeiterin mutmaßlich in nächster Zeit das Dienstverhältnis kündigen wird). Die Maschine kann „Auffälligkeiten“ orten, die dem Menschen ob der zu durchforstenden Datenmenge unbekannt bleiben. Hier gilt es m. E., einen gesellschaftlichen Konsens zu schaffen, ob und wieweit wir das insbesondere in sensiblen Bereichen zulassen wollen.
- **Unsachliche Diskriminierung** von Arbeitnehmer*innen durch das algorithmusbasierte System, sei es aufgrund einer bereits diskriminierenden Programmierung und „Anfütterung“ des Systems oder einer nachfolgenden unerwünschten und nicht vorhergesehenen Pfadeinschlagung durch ein selbstlernendes System.

Einen guten, aber erschreckenden Überblick über Datensammlung und -verarbeitung durch den Arbeitgeber mit dem Potential einer Mitarbeiter*innenkontrolle und -überwachung bietet Wolfie Christl in seiner Studie „Digitale Überwachung und Kontrolle am Arbeitsplatz. Von der Ausweitung betrieblicher Datenerfassung zum algorithmischen Management?“. Die Grafik auf der nachfolgenden Seite beschränkt sich freilich nicht auf algorithmische Entscheidungs- und Überwachungssysteme, angesichts der Rasanz, mit der ihr Einsatz im Steigen begriffen ist, gibt sie jedoch einen guten Einblick in mögliche zukünftige Szenarien.

4 Vgl. dazu bspw. Prassl-Adams, 195ff.



© 2021 Cracked Labs | Pascale Osterwalder, Wolfie Christl | CC BY-SA

Abbildung 1: Landkarte betrieblicher Datenpraktiken und Systeme. Cracked Labs 2021. Wolfie Christl und Pascale Osterwalder. *CC BY-SA 4.0*.

Eine Vielzahl an algorithmusbasierten Entscheidungssystemen in der Arbeitswelt basiert auf sog. **„Wissenssystemen“** und ist nicht selbstlernend. Die im Wesentlichen auf einer „Wenn-dann“-Systematik basierenden Entscheidungen können (auch bei einer gewissen Komplexität) noch nachvollzogen werden. Vorgegeben wird dem System eine Handlungsanleitung, Ergebnisse und Ableitungen können (und sollen) vom System natürlich angepasst werden, nicht aber die Regeln selbst.

Eine neue Dimension bzw. Fragestellung bringen **selbstlernende Systeme** mit sich, die (gewollt) den ursprünglichen Pfad verlassen, indem sie ihn „intelligent“ weiterentwickeln. Eine maschinelle Lernfähigkeit entwickeln diese Systeme aus der Vielzahl an Daten, die sie sammeln und analysieren (Big Data). Im Arbeitsleben sollen sie die Prozesse beschleunigen, erleichtern oder gar erst ermöglichen – die konkreten Auswirkungen auf die Arbeitnehmer*innen werden jedoch selten beleuchtet und oft auch lediglich als „Kollateralschaden“ wahrgenommen.

Aus **datenschutzrechtlicher Sicht** liegt zunächst bereits eine der Gefahren in der umfangreichen Sammlung an Daten, die die Durchforstung der gigantischen Datenmengen, die für das Maschinelle Lernen notwendig sind, mit sich bringt, und das dadurch entstehende Überwachungspotential. Besonders prekär ist dies hinsichtlich der Daten, die einen Rückschluss auf die Persönlichkeit des/der Arbeitnehmer*in zulassen, einerseits, und jener, die eine minutiöse Nachvollziehbarkeit jeden einzelnen Schrittes im Tagesverlauf ermöglichen, andererseits.

Automatisierte Entscheidungsfindung führt selbst beim „intelligentesten“ System nicht selten (aufgrund der statistischen Verallgemeinerung) zu einer ungewollten und mitunter auch unentdeckten Diskriminierung (s. Abschnitt B, Kapitel Wie KI mit Daten lernt und welche Vorurteile dabei erzeugt werden). Mitunter haben derartige Entscheidungen gravierende Auswirkungen (z. B. personalrechtliche Konsequenzen), sie unterliegen also erhöhten Anforderungen hinsichtlich **Nachvollziehbarkeit** (Transparenz; s. Abschnitt B, Kapitel Erklärbare Künstliche Intelligenz) und **Rückführbarkeit** auf eine durch einen Menschen getroffene Entscheidung.

Bei der rechtlichen Betrachtung des Einsatzes von KI und insbesondere Fragen der **Zurechnung von Verantwortung** begeben wir uns an vielen Stellen in **komplettes Neuland**. Der Beitrag in diesem Handbuch wirft entsprechend auch zahlreiche Fragen auf, deren Lösung es noch zu finden gilt. Dabei ist insbesondere ein gesellschaftlicher Konsens entlang unserer Moralvorstellungen relevant, den es zu finden und allfälligen Regulatorien zugrunde zu legen gilt. Für die praktische betriebliche Handhabe sei an dieser Stelle auf die wichtige **frühzeitige Einbeziehung aller im Betrieb Beteiligten** verwiesen und auf das essenzielle Instrument der **Betriebsvereinbarung**, mit dem (wenn gut gemacht) typische und absehbare Risiken abgefangen werden können (siehe Abschnitt C, Kapitel Mitbestimmung bei Betriebsvereinbarungen).

Welche datenschutzrechtlichen Grundlagen sind zu beachten?

Der Einsatz von KI ist datenschutzrechtlich äußerst sensibel und verlangt eine besondere Transparenz und Information der Betroffenen hinsichtlich der verarbeiteten Daten, der Prozesse, mittels derer dies geschieht und der allfälligen Ergebnisse, der Ableitungen daraus und der Auswirkungen auf die Betroffenen.

Grundsätze der Datenverarbeitung

Ebenso wie bei Anwendung herkömmlicher Datenverarbeitungssysteme sind bei Anwendung von algorithmusbasierten Systemen die Grundsätze der Datenverarbeitung der Datenschutzgrundverordnung (DSGVO) einzuhalten.

Art 5 der DSGVO behandelt diese Grundsätze der Datenverarbeitung und gliedert sie in sechs Punkte:

- Rechtmäßigkeit, Verarbeitung nach Treu und Glauben, Transparenz (lit a)
- Zweckbindung (lit b)
- Datenminimierung (lit c)
- Richtigkeit und Recht auf Löschung („Aktualität“) (lit d)
- Speicherbegrenzung (lit e)
- Integrität und Vertraulichkeit (lit f)

Grundfrage jeglicher Verarbeitung von Daten ist, ob sie überhaupt zulässig (= rechtmäßig) ist – erst dann kommt überhaupt das entsprechende Schutzregime der DSGVO zu tragen. Die DSGVO folgt hier dem Konzept, dass **eine Verarbeitung personenbezogener Daten grundsätzlich unzulässig** ist, es sei denn, es liegt ein **Rechtfertigungsgrund** dafür vor. Zusammenfassend lässt sich festhalten, dass eine Datenverarbeitung dann rechtmäßig ist, wenn eine konkrete Rechtsgrundlage zur Verarbeitung der Daten vorliegt.

Dies kann:

- die Einwilligung der betroffenen Personen
- eine vertragliche oder gesetzliche Verpflichtung oder
- die Wahrung bestimmter als berechtigt angesehener Interessen sein

Der **Zweck**, für den die jeweiligen Daten erhoben werden, muss **ausdrücklich** *vorab* festgelegt werden und er muss **eindeutig** sein (vage, offene oder gar widersprüchliche Formulierungen genügen diesen Anforderungen nicht).

- Es muss sich um einen legitimen (rechtmäßigen und nicht überschießenden) Zweck handeln.
- Die Daten dürfen nur verarbeitet werden, wenn der Zweck der Verarbeitung nicht in zumutbarer Weise durch andere Mittel erreicht werden kann (Zweck-Mittel-Relation).
- Eine Weiterverwendung der Daten, die nicht mit dem ursprünglichen Erhebungs-Zweck vereinbar ist, ist im Wesentlichen (mit Ausnahmen z. B. für Archive oder wissenschaftliche Zwecke, bestimmte berechtigte Interessen) unzulässig.

Gerade an den Zweckbindungserfordernissen scheitern jene KI-Systeme, deren „Optimierungs-Ziel“ zunächst noch nicht feststeht, sondern die zunächst eine Vielzahl an Daten erheben, diese „zufallsanalysieren“ (Big Data) und erst aus diesen Findings Ableitungen hinsichtlich z. B. möglicher Effizienzpotentiale in der Produktion treffen. Hier empfehle ich ein Ausweichen auf Systeme, die keine personenbezogenen Daten erheben oder diese verlässlich z. B. mittels Avataren pseudonymisieren.

Details zu den Grundsätzen der Datenverarbeitung können bspw. in Haslinger, Susanne / Krisch, Andreas / Riesenecker-Caba, Thomas (Hrsg).: „Beschäftigtendatenschutz – Handbuch für die betriebliche Praxis“ nachgelesen werden. An dieser Stelle wollen wir uns mit den spezifischen rechtlichen Anforderungen an den Einsatz von KI auseinandersetzen.

Spezifische datenschutzrechtliche Regelungen i.Z.m. vollautomatisierten Entscheidungsfindungssystemen

Gemäß Artikel 22 DSGVO hat die betroffene Person das Recht, keiner ausschließlich auf einer vollständig automatisierten Verarbeitung — einschließlich Profiling — beruhenden Entscheidung oder Maßnahme unterworfen zu werden, die ihr gegenüber **rechtliche Wirkung entfaltet** oder sie **in ähnlicher Weise erheblich beeinträchtigt**. Eine solche rechtliche Wirkung kann sich bspw. auf die Begründung oder Beendigung eines (Arbeits-)Vertrags beziehen, oder in Rechte wie die Vereinigungsfreiheit oder das Recht, rechtliche Schritte einzuleiten, eingreifen[5]. ErwGr 71 nennt bspw. die vielfach zitierte automatisierte Entscheidung über eine Kreditvergabe, aber auch die automatische Selektion von Bewerbungen auf einen Job ohne jegliches menschliche Eingreifen.

Eine **erhebliche Beeinträchtigung in „ähnlicher Weise“** bedeutet unter anderem die Beeinträchtigung des Verhaltens oder der Entscheidung des/der Betroffenen, einen Ausschluss z. B. von Gesundheitsdiensten, einem Arbeitsplatz, Bildung etc. oder eine Diskriminierung des/der Betroffenen[6]. Im Arbeitsleben fallen darunter jedenfalls der Zugang zu Beförderungen; Personalbeurteilungen, die Auswirkungen auf den weiteren Verlauf des Arbeitsverhältnisses oder Prämienzahlungen haben, der „*eingreifende Charakter des Profiling-Prozesses*“ z. B. aufgrund seines Überwachungspotentials hinsichtlich der Arbeitsleistung, die „*Ausnutzung von Schwachstellen der betroffenen Person*“[7] oder dem/der Betroffenen vermittelte Ergebnisse, die zu hochriskantem Verhalten führen können.

5 WP251, 23.

6 WP251, 23f.

7 WP251, 24.

Dieses in Art 22 DSGVO verbriefte Recht, einem derartigen vollautomatisierten Entscheidungsprozesses nicht unterworfen zu werden, bedeutet nicht, dass der/die Betroffene aktiv tätig werden und sich dagegen wehren muss, sondern gilt unabhängig davon[8].

INFOBOX: PROFILING

... ist jegliche Form automatisierter Verarbeitung personenbezogener Daten unter Bewertung der persönlichen Aspekte in Bezug auf eine natürliche Person, insbesondere zur Analyse oder Prognose von Aspekten bezüglich Arbeitsleistung, wirtschaftlicher Lage, Gesundheit, persönlicher Vorlieben oder Interessen, Zuverlässigkeit oder Verhalten, Aufenthaltsort oder Ortswechsel der betroffenen Person, soweit dies rechtliche Wirkung für die betroffene Person entfaltet oder sie in ähnlicher Weise erheblich beeinträchtigt.

Von dieser Grundregel gibt es natürlich auch **Ausnahmen** – die allerdings erhöhte Anforderungen an die Maßnahmen zur Wahrung der Rechte und Freiheiten sowie der berechtigten Interessen des/der Betroffenen mit sich bringen:

Eine rein algorithmenbasierte Entscheidungsfindung ist dann erlaubt, wenn die Entscheidung für den Abschluss oder die Erfüllung eines Vertrags zwischen der betroffenen Person und dem*der Verantwortlichen erforderlich ist, wenn eine derartige Verarbeitung nach dem Unionsrecht oder dem nationalen Recht ausdrücklich (!) für zulässig erklärt wird oder die betroffene Person eingewilligt hat (Art 22 Abs 2 DSGVO).

- Eine **„freiwillige" Einwilligung** in die Datenverarbeitung im Beschäftigungskontext ist schon bei herkömmlichen Datenanwendungen eine heikle Sache, dies gilt umso mehr bei vollautomatisierten Entscheidungen. Auch die Art-29-Gruppe (G29) sieht dies als problematische Rechtsgrundlage an: *„Angesichts der Abhängigkeit, die sich aus dem Verhältnis Arbeitgeber/Arbeitnehmer ergibt, ist es unwahrscheinlich, dass die betroffene Person ihrem Arbeitgeber die Einwilligung in die Datenverarbeitung verweigern kann, ohne Angst zu haben oder wirklich Gefahr zu laufen, dass diese Weigerung zu Nachteilen führt."*[9]
- Auch die **Notwendigkeit zur Vertragserfüllung** steht im Arbeitsverhältnis auf hölzernen Beinen und kann wohl eher ausgeschlossen werden. Diesfalls ist der Arbeitgeber in der Beweispflicht, dass die vollautomatisierte Entscheidungsfindung nicht nur zur Vertragserfüllung erforderlich ist, sondern auch kein gelinderes Mittel zumutbar ist. Denkbar ist eine Anwendung im vorvertraglichen Bereich, wenn auf Stellenanzeigen mehrere hunderte oder gar tausende Bewerbungen eingehen[10].

Für **sensible Daten** („besondere Kategorien personenbezogener Daten" gem. Art 9 DSGVO), das sind Daten, aus denen die rassische und ethnische Herkunft, politische Meinungen, religiöse oder weltanschauliche Überzeugungen oder die Gewerkschaftszugehörigkeit hervorgehen, sowie die genetischen Daten, biometrische Daten, Gesundheitsdaten, Daten zum Sexualleben oder der sexuellen Orientierung, gelten gem. Art 22 Abs 4 DSGVO noch einmal strengere Anforderungen.

Rechte der Betroffenen

Ist eine rein algorithmusbasierte Entscheidung im Ausnahmefall zulässig, verlangen Art 22 Abs 3 und ErwGr 71 entsprechende Garantien:

- Der/die Verantwortliche muss **spezifische, leicht zugängliche und aussagekräftige Informationen** an die betroffene Person zur Verfügung stellen (in besonderer Ausformung von Art 12 ff DSGVO). Im Konkreten bedeutet das Informationen über:
  - die erfassten Daten und deren Quelle
  - die automatisierte Entscheidungsfindung und die involvierte Logik (Wie soll der Prozess zu welchem Ergebnis führen? In der Regel beinhaltet dies aber nicht die Offenlegung des Algorithmus selbst) sowie
  - die Tragweite und die angestrebte Auswirkung der Datenverarbeitung für die betroffene Person (mit Untermauerung durch „echte, greifbare Beispiele"[11])

8 WP251, 21.
9 WP259, 7.
10 Vgl. WP251, 25.
11 WP251, 29.

- Ein Anspruch auf **direktes Eingreifen einer Person** auf Seiten des/der Verantwortlichen, der/die „zur Änderung der Entscheidung befugt und befähigt" ist, was auch eine gründliche Prüfung aller relevanten Daten impliziert[12].

- Ein Anspruch auf **Darlegung des eigenen Standpunkts** und auf **Erläuterung** der nach einer entsprechenden Bewertung getroffenen **Entscheidung**[13] sowie

- das **Recht auf Anfechtung der Entscheidung.**

- **Kinder** sollten **keinesfalls** derartigen automatischen Entscheidungsfindungen unterworfen werden.

Die DSGVO verpflichtet ferner den/die Verantwortliche*n zur Verwendung **geeigneter mathematischer oder statistischer Verfahren**, um eine Diskriminierung hintanzuhalten und zum Ergreifen **geeigneter technischer und organisatorischer Maßnahmen**, die auf die spezifische Situation der automatisierten Datenverarbeitung eingehen: Das heißt, *„dass Faktoren, die zu unrichtigen personenbezogenen Daten führen, korrigiert werden und das Risiko von Fehlern minimiert wird"*[14].

Aufgrund des **„Black-Box"-Problems** (siehe unten) ist fraglich, ob neuronale Netzwerke, bei denen das Ergebnis nicht mehr zur Entscheidungskette rückgeführt werden kann, überhaupt DSGVO-konform eingesetzt werden können. An dieser Stelle kann hier nicht weiter darauf eingegangen werden, auch weil anzunehmen ist, dass derlei Fragen mutmaßlich durch den derzeit noch in Diskussion befindlichen Rechtsrahmen für KI der EU beantwortet werden[15].

Datenschutz-Folgenabschätzung (DSFA)

Art 35 DSGVO verlangt die Durchführung einer sog. Datenschutz-Folgenabschätzung, wenn die Verwendung neuer Technologien, **aufgrund der Art, des Umfangs, der Umstände** und **der Zwecke der Verarbeitung** voraussichtlich ein **hohes Risiko für die Rechte und Freiheiten natürlicher Personen** zur Folge hat. Dieses Risiko gilt gemeinhin bei algorithmenbasierten Systemen und solchen, die auf einer Vielzahl gesammelter Daten (Big Data) operieren als verwirklicht, Gleiches gilt für diskriminierungsgeneigte Systeme[16].

Die DSFA hat zumindest folgende Inhalte zum Gegenstand[17]:

- eine Bewertung der **Notwendigkeit und Verhältnismäßigkeit** der Verarbeitungsvorgänge in Bezug auf den Zweck

- eine **Risikoabschätzung** in Hinblick auf die Rechte und Freiheiten der Betroffenen

- die geplanten **Abhilfemaßnahmen** (technischer und organisatorischer Art), die getroffen werden, um sicherzustellen, dass die identifizierten Risiken nicht eintreten

Die Artikel-29-Datenschutzgruppe empfiehlt zudem im Rahmen der DSFA auch, *„den Umfang der menschlichen Beteiligung an der Entscheidungsfindung und die Phase, in der sie erfolgt"*, zu ermitteln und aufzuzeichnen"[18].

Auf nationaler Ebene schließt die **Verordnung der Datenschutzbehörde über Verarbeitungsvorgänge, für die eine Datenschutz-Folgenabschätzung durchzuführen ist (DSFA-V)**, in § 2 Abs 2 vorletzter Satz die Verpflichtung zur Durchführung einer DSFA dann aus, wenn eine **Betriebsvereinbarung** abgeschlossen wurde oder die Personalvertretung der Datenanwendung zugestimmt hat. Offenbar wollte der VO-Geber damit die Relevanz von Betriebsvereinbarungen hervorstreichen, die Sinnhaftigkeit der Bestimmung wird jedoch auf jene Fälle zu reduzieren sein, in denen die Betriebsvereinbarung dieselben Prüfschritte durchläuft, die ansonsten im Zuge einer DSFA untersucht werden. Andernfalls empfiehlt es sich als Verantwortliche*r allein schon aus Haftungsgründen, auch trotz Vorliegens einer Betriebsvereinbarung (die zwar ein ähnliches, aber doch anderes Ziel verfolgt), eine umfassende und gründliche DSFA derart riskanter und potentiell eingriffsintensiver Systeme vorzunehmen.

---

12 WP251, 30.

13 Art 22 Abs 3 DSGVO.

14 DSGVO ErwGr 71.

15 COM(2021) 206 final.

16 Vgl. dazu ausführlich Greif/Kullman, 69f.

17 Art 35 Abs 7 DSGVO.

18 WP251, 22.

Betriebsverfassungsrechtliche Regelungen

Das Arbeitsverfassungsgesetz (ArbVG) und insb. die Betriebsvereinbarungs-Tatbestände der §§ 96 und 96a ArbVG sind „technikneutral“ ausgestaltet und daher vollumfänglich geeignet, zum Zeitpunkt der Entstehung der relevanten §§ nicht absehbare technische Entwicklungen „einzufangen“ (siehe zur Gestaltung derartiger Betriebsvereinbarungen Kapitel Mitbestimmung bei Betriebsvereinbarungen dieses Abschnitts).

Gleiches gilt m. E. auch für die Informationsrechte bzgl. Datenanwendung des § 91 Abs 2 ArbVG, dennoch ist eine Klarstellung, wie sie z. B. jüngst im spanischen „Rider-Gesetz“ (hinsichtlich der Regelung plattformbasierter Zustelldienste) erfolgte, sicher sinnvoll: Dort heißt es u. a., dass der Arbeitnehmer*innen-Vertretung Informationen über „*die Parameter, Regeln und Anweisungen, auf denen die Algorithmen oder Systeme der künstlichen Intelligenz beruhen, die sich auf die Entscheidungsfindung auswirken und Auswirkungen auf die Arbeitsbedingungen und den Zugang zu und die Aufrechterhaltung von Arbeitsplätzen haben können, einschließlich der Erstellung von Profilen*“[19], bereitzustellen sind.

Diskriminierungsschutz

Im Unterschied zu „herkömmlichen“ Datenanwendungen stellt sich i. Z. m. mit KI besonders die Frage des Diskriminierungsschutzes. Derzeit gibt es keine spezifische Regulierung (weder bei Programmierung noch bei Verwendung von KI), wir sind somit im Arbeitskontext auf althergebrachtes, sprich das herkömmliche Gleichbehandlungsrecht angewiesen.

Hinter jeder KI stehen Menschen an unterschiedlichen Schnittstellen, von der Programmierung bis zur Verwendung allfälliger Ergebnisse, doch ist die Zurechnung und Verantwortung bis dato nicht geklärt. Regelmäßig wird davon ausgegangen, dass Systeme nur dann diskriminieren, wenn der Mensch sie „falsch programmiert“ oder mit „falschen Daten“ gefüttert hat. Tatsächlich basieren KI-Systeme – vereinfacht gesagt – stets auf Kategorisierungen anhand bestimmter Merkmale. Werden also beispielsweise Arbeitnehmer*innen anhand eines oder mehrerer bestimmter Merkmale einer Kategorie zugeordnet und hat dies im weiteren Verlauf Folgen für sie oder andere (von dieser Kategorie ausgeschlossene Personen), liegt zunächst einmal eine **Ungleichbehandlung** vor (zum Proxy-Problem siehe Infobox). Im Bereich der rechtlich unzulässigen **Diskriminierung** finden wir uns dann, wenn diese Ungleichbehandlung **ungerechtfertigterweise** erfolgt (d. h. Gleiches ungleich oder Ungleiches gleich behandelt wird).

Folgende drei Fallgruppen scheinen hier dienlich:

- **Der Faktor Mensch (scheinbar neutraler Algorithmus):**
  „Automation Bias“ – der Glaube, dass der Computer bessere Entscheidungen treffen kann als der Mensch. Der Automation Bias führt auch dazu, dass man dem System diskriminierungsfreies „Verhalten“ zutraut, während dies beim Menschen nicht der Fall ist. Das wiederum führt nicht selten zu einem höheren Vertrauen in vom Algorithmus vorgeschlagene Lösungen und Zögern beim Eingreifen. Gerade bei einem Diskriminierungsvorwurf kann dies aber von entscheidender Bedeutung sein.

- **Die Schnittstelle Mensch-Maschine (Input Bias/ Output Bias etc.):**
  Tatsächlich steht am Anfang jeder KI der Mensch mit allen Fehlern, Wertungen und Fehlentscheidungen. Weder Anlerndaten noch erlernte Ergebnisse sind dabei neutral, sie spiegeln insbesondere auch gesellschaftliche Wertvorstellungen:
  - Trainingsdaten: unmittelbarer oder mittelbarer **Input Bias**, z. B. Amazon Recruiting Software: Sie bevorzugt Männer, da sie mit den erfolgreichen (hauptsächlich männlichen) Bewerbungen der Vergangenheit trainiert wurde (s. Kapitel Wie KI mit Daten lernt und welche Vorurteile dabei erzeugt werden in Abschnitt B).
  - Diskriminierung durch Technik, z. B. Sensoren oder Zugang zum System
  - Diskriminierung durch die ursprünglichen Entscheidungsbäume, zu erreichende Benchmarks (z. B. bei Personalbeurteilungssystemen) etc.

Wer überwacht das (noch immer) einwandfreie Funktionieren des Systems? So kennen wir bspw. das Phänomen, dass ursprünglich diskriminierungsfreie Systeme sich durch das analysierte Umfeld ein diskriminierendes Verhalten „antrainieren“ (siehe das berühmte Beispiel des Microsoft-Chatbots Tay, die binnen kurzer Zeit ras-

19 Prassl-Adams, 199.

sistisch und ausfällig wurde, weil das Internet ihr dies vorgelebt hatte) (**Output Bias**). Wer haftet für derartigen „diskriminierenden Output", wenn er nachweislich nicht auf diskriminierenden Input zurückzuführen ist?[20]

Potentielle Maßnahmen: Regelmäßige Analyse des Systems und Bewertung der verarbeiteten Datensätze, um Verzerrungen und allfällige Diskriminierungen aufzuspüren; aktives Entwickeln von Maßnahmen, um verzerrende Zusammenhänge zu entschärfen; regelmäßige Überprüfung der Richtigkeit und der Relevanz der Entscheidungen[21].

Hindernisse: Größe und Komplexität der Datenmengen und des Quellcodes[22]

- Die Black Box (autonome, nicht einmal mehr kontrollierbare Maschine):
  - Wie kann Diskriminierung beim Deep Learning überhaupt noch festgestellt werden? Das Black-Box-Problem beginnt oft schon damit, dass nicht einmal der Sachverhalt feststellbar ist! Selbst wenn das gelingt, ist die Zurechnung einer verantwortlichen Person noch lange nicht geklärt.
  - Kann man einer Maschine möglichst diskriminierungsfreie Entscheidung antrainieren? Parameter wie Gerechtigkeit, sachliche/unsachliche Differenzierungen sind in aller Regel normative Vorstellungen und damit aus Sicht der KI „kontrafaktische Größe"[23].

Potentielle Maßnahmen: Externe Instanz zur Kontrolle von Algorithmen/algorithmenbasierten Entscheidungen[24]. Für weitere Maßnahmen siehe Kapitel Erklärbare Künstliche Intelligenz in Abschnitt B.

DAS PROXY-PROBLEM

Zentral ist in allen drei Fallgruppen das Zusammenspiel von **Korrelation** und **Kausalität** (Correlation-proxies challenge)[25]: Algorithmenbasierte Systeme funktionieren im Wesentlichen auf Kategorisierung von Daten anhand ihrer Häufigkeit. Eine statistische Häufung trifft jedoch noch keinerlei Aussage über die Kausalität (Was ist die Ursache für dieses Ergebnis?). Wenig überraschend führen vorschnelle Rückschlüsse auf eine (in Wahrheit gar nicht gegebene) Kausalität schnell zur Diskriminierung oder verstärken eine bereits vorhandene Diskriminierung sogar noch. Gerade einer selbstlernenden Maschine kann ich nun aber nicht die Augen verbinden, indem ich z. B. bei den Anlerndaten das Geschlecht weglasse. Über verschiedene Proxies (z. B. bestimmte öfter von einem Geschlecht verwendete Schlagworte in Bewerbungsschreiben, Teilzeitarbeit etc.) wird das System regelmäßig eine statistische Häufung bestimmter Merkmale feststellen, die danach positiv oder negativ bewertet werden. Die Diskriminierung, die dadurch vermieden werden soll, setzt sich damit regelmäßig verdeckt fort.

Die Diskriminierung aufgrund der statistischen Häufung von Proxies ist als geradezu klassische mittelbare Diskriminierung zu werten.

Die Proxy-Debatte ist zwar unter dem Diskriminierungs-Gesichtspunkt entstanden, m. E. finden wir jedoch das exakt selbe Phänomen auch im Zusammenhang mit datenschutzrechtlichen Fragen: Insbesondere seit Inkrafttreten der DSGVO wird deutlich, dass die Erhebung oder gar Analyse bestimmter Daten ein No-Go im Arbeitsverhältnis sein muss, ebenso die „Kollateral-Überwachung" der Arbeitnehmer*innen bei der Verfolgung vermeintlich gänzlich anderer Ziele. In der Folge haben sich zahlreiche Stellvertreter-Variablen entwickelt (so wird nicht die Arbeitsleistung überwacht, sondern nur die optimale Programmauslastung etc.), bei denen es sich nicht selten um relativ ungeschickte „Tarnmanöver" handelt.

Diskriminierungsverbot nach dem Gleichbehandlungsgesetz (GlBG)

Das Gleichbehandlungsgesetz (GlBG) beinhaltet ein umfassendes Diskriminierungsverbot[26], stellt dabei aber nur auf bestimmte, ausdrücklich genannte Diskriminierungsmerkmale ab. Demnach darf auf Grund des **Geschlechts**, insbesondere unter Bezugnahme auf den Familienstand oder den Umstand, ob jemand Kinder hat, der **ethnischen Zugehörigkeit**, der **Religion oder Weltanschauung,** des **Alters** oder der **sexuellen**

20 Beck et al., 17.
21 Vgl. z. B. WP251, 31.
22 Beck et al., 9.
23 Beck et al., 14.
24 Ebd.
25 Vgl. ausführlich Gerards/Xenidis, 44f.

Orientierung im Zusammenhang mit einem Arbeitsverhältnis niemand unmittelbar oder mittelbar diskriminiert werden. Ähnliche Diskriminierungsverbote bestehen auch aufgrund des **Behinderteneinstellungsgesetzes** (BEinstG).

Dieses Diskriminierungsverbot gilt insbesondere (aber nicht nur) bei der Begründung und Beendigung des Arbeitsverhältnisses, bei der Festsetzung des Entgelts und Gewährung von Sozialleistungen, beim beruflichen Aufstieg und bei den sonstigen Arbeitsbedingungen. Ein Verschulden für das Vorliegen einer Diskriminierung ist nicht erforderlich[27], was m. E. bedeutet, dass eine Zurechenbarkeit bereits bei **Verantwortung für den Einsatz des diskriminierenden Systems** gegeben ist (Kausalität).

UNMITTELBARE UND MITTELBARE DISKRIMINIERUNG (§ 5 UND § 19):

Unmittelbare Diskriminierung: Diese liegt dann vor, wenn eine Person auf Grund ihres Geschlechts, ihrer ethnischen Zugehörigkeit, ihrer Religion oder Weltanschauung, ihres Alters oder ihrer sexuellen Orientierung **in einer vergleichbaren Situation eine weniger günstige Behandlung erfährt,** als eine andere Person erfährt, erfahren hat oder erfahren würde (Abs 1).

Mittelbare Diskriminierung: Diese liegt vor, wenn **dem Anschein nach neutrale Vorschriften**, Kriterien oder Verfahren Personen, die einem Geschlecht oder einer ethnischen Gruppe angehören, oder Personen mit einer bestimmten Religion oder Weltanschauung, eines bestimmten Alters oder mit einer bestimmten sexuellen Orientierung in besonderer Weise gegenüber anderen Personen benachteiligen können, es sei denn, **die betreffenden Vorschriften, Kriterien oder Verfahren sind durch ein rechtmäßiges Ziel sachlich gerechtfertigt und die Mittel sind zur Erreichung dieses Zieles angemessen und erforderlich** (Abs 2).

Sachliche Rechtfertigung einer algorithmusbasierten Entscheidung bei einer mittelbaren Diskriminierung

Eine mittelbare Diskriminierung kann im Gegensatz zur unmittelbaren Diskriminierung dann gerechtfertigt sein, wenn mit dem Einsatz der betreffenden Kriterien und Prozesse ein rechtmäßiges Ziel verfolgt wird (d. h., das Ziel muss „**schützenswert** *sein und* **wichtig** *genug, um Vorrang vor dem Gleichbehandlungsgrundsatz zu haben*“[28]) und sie **zur Erreichung dieses Ziels geeignet, angemessen und erforderlich** sind. Je weiter wir bei Einsatz einer KI in den Bereich des Deep Learning kommen, desto schwieriger wird es, überhaupt nachzuweisen, um *welche* eingesetzten bspw. Proxies und sich daraus ergebende Prozesse es sich überhaupt handelt. Dies stellt mitunter eine höhere Hürde für den Arbeitgeber dar, da diesen letztlich eine Beweispflicht und nicht nur wie den/die Arbeitnehmer*in eine Pflicht zur Glaubhaftmachung (siehe sogleich) trifft. Die sachliche Rechtfertigung auf den Einsatz eines (selbstlernenden) algorithmischen Systems als solches zu beschränken[29] und nicht auf die Entscheidungsfindung selbst, scheint mir sowohl im Vergleich zu analogen Sachverhalten als auch angesichts der hohen Anforderungen der DSGVO zu kurz gegriffen.

Der Europäische Gerichtshof (EuGH) hat finanzielle Gründe/Kostenersparnis seitens des Unternehmens als sachliche Rechtfertigung ausdrücklich abgelehnt, in Frage kommt bspw. die Einhaltung gesetzlicher Verpflichtungen, wie Hygiene- oder Sicherheitsvorschriften[30]. Stets muss der Einsatz des konkreten Systems jedoch als Mittel zur Zielerreichung nicht nur geeignet, sondern in diesem Umfang auch erforderlich und keinesfalls überschießend sein (d. h., ein gelinderes Mittel steht mit zumutbarem Aufwand nicht zur Verfügung).

Rechtsdurchsetzung

Beruft sich eine betroffene Person auf einen Diskriminierungstatbestand, muss er/sie lediglich die **überwiegende Wahrscheinlichkeit** einer Diskriminierung **glaubhaft machen**[31]. Das bedeutet zwar eine Beweiserleichterung für die Betroffenen, bringt aber jedenfalls die Verpflichtung zur Vorlage von **Indizienbeweisen** mit sich, was bei zunehmend komplexen algorithmischen Systemen (insbesondere bei selbstlernenden Systemen) bereits eine unüberwindbare Hürde darstellen kann. Der Indizienbeweis verlangt, dass zumindest die ursächlichen Proxies benannt werden

26 §§ 3, 4, 17 und 18 GlBG.

27 Hopf/Mayr/Eichinger/Erler, § 5, Rz 8.

28 Hopf/Mayr/Eichinger/Erler, § 5, Rz 95

29 Vgl. z. B. Winkler, 35ff.

30 Hopf/Mayr/Eichinger/Erler, § 5, Rz 115

31 §§ 12 Abs 12 und 26 Abs 12 GlBG.

können oder im Worst Case sogar der Quellcode vorgelegt wird. Das wird nur im seltensten und augenscheinlichsten Fall gelingen, auch der „Rechengang" zwischen Proxy und Output wird nur selten in der notwendigen Nachvollziehbarkeit vorgelegt werden können[32].

Dem Arbeitgeber (oder sonstigen Verantwortlichen) obliegt es, das **Gegenteil zu beweisen**, nämlich, dass nach Abwägung aller Umstände ein anderes von diesem glaubhaft gemachtes (nicht diskriminierendes) Motiv/Kriterium ausschlaggebend für die unterschiedliche Behandlung war.

Mitunter schafft die **Offenlegungsverpflichtung nach Art 15 DSGVO** hier Abhilfe, wonach neben der Kategorie personenbezogener Daten, die verarbeitet werden, auch Auskunft über das Bestehen einer automatisierten Entscheidungsfindung und über die involvierte Logik sowie die Tragweite und die angestrebten Auswirkungen (siehe oben) erteilt werden muss. Gegenstand hitziger Debatten ist jedoch nach wie vor, inwieweit der Arbeitgeber dadurch zur Offenlegung des Algorithmus oder der aggregierten gesammelten Daten verpflichtet werden kann[33]. Der österreichische Gesetzgeber hat zudem im Rahmen einer zulässigen Öffnungsklausel die Rechte aus § 15 DSGVO dann ausgeschlossen, *„wenn durch die Erteilung dieser Auskunft ein Geschäfts- oder Betriebsgeheimnis des Verantwortlichen bzw. Dritter gefährdet würde"*[34].

Potentielle Maßnahmen: Anstelle eines Indizienbeweises zur Glaubhaftmachung vor Gericht muss der Anscheinsbeweis ausreichen.

Handlungs- und Regelungsbedarf für den Gesetzgeber
Trotz eines anwendbaren Rechtsrahmens gibt es angesichts der Komplexität der Problemlage noch großen Handlungsbedarf für den europäischen, aber auch den österreichischen Gesetzgeber. Neben den bereits genannten Maßnahmen zählen dazu vor allem:

- Erhöhte Dokumentationspflichten bereits bei der Programmierung

- Klare Zurechenbarkeit und Verantwortung: Denkbar ist z. B. eine erhöhte Nachbetreuungspflicht durch Hersteller einer entsprechenden Software, andernfalls sollen diese zur Verantwortung gezogen werden, wenn die KI nachfolgend einen Schaden anrichtet.[35]
- Erhöhter Rechtsschutz z. B. durch Verbandsklage: Eine individuelle Person hat selbst bei Beweiserleichterungen (Glaubhaftmachung, geforderte Übernahme des Ansatzes aus der Einkommensdiskriminierung etc.) wenig Möglichkeiten, systematisches Vorgehen des Systems aufzuzeigen (angesprochen werden z. B. wiederholt die Möglichkeiten von NGOs, mit anonymisierten z. B. Bewerbungen oder mithilfe von Avataren das Verhalten des Systems zu testen und nachvollziehbar aufzuzeigen), inkl. eines eigenständigen Auskunftsrechts von Antidiskriminierungs- und Beratungsstellen[36].

- Ausnahme bestimmter Bereiche von der Entscheidungsfindung durch KI, stattdessen ein Recht auf Einzelfallbeurteilung[37]

- Schaffung einer KI-Ombudsstelle[38] und Überprüfbarkeit der Entscheidungen durch eine unabhängige Instanz: Derzeit liefern Hinweise auf systematische Diskriminierung oft nur von Watchdog-Organisationen durchgeführte „Massen-Tests", ein Instrument, das Einzelpersonen nicht zur Verfügung steht. Es mehrt sich daher der Ruf nach einer verpflichtenden systematischen, regelmäßigen Überprüfung, quasi eines unabhängigen Audits des Algorithmus[39].

Am relevantesten scheint mir jedoch die gesellschaftliche Debatte über Risiken und Nutzen. Der „Marktfähigkeitsansatz", den die Europäische Kommission derzeit mit dem Entwurf zur KI-Richtlinie verfolgt[40], geht leider in die grundsätzlich entgegengesetzte Richtung und läuft damit manifest Gefahr, Arbeitnehmer*innenrechte gleichermaßen wie Konsument*innenrechte zu unterminieren. Wir müssen uns die Frage nach sozial wünschenswerter KI und ihrer Ausgestaltung stellen, andernfalls können wir auch die heiklen Haftungsfragen nicht klären.

---

32 Winkler, 38.
33 Vgl. ausführlich z. B. Greif/Kullmann: 68f.
34 § 4 Abs 6 DSG.
35 Beck et al., 17.
36 Vgl. Michot et al., 10.
37 Beck, 2019.
38 Siehe KI-Strategie 2030 der österreichischen Bundesregierung, 30.
39 Vgl. z. B. Adensamer/Klausner 10f oder Michot et al., 5.
40 COM(2021) 206 final.

SIE HABEN NOCH NICHT GENUG?

Einen umfassenden und sehr empfehlenswerten Überblick über herkömmliche und „intelligente" Überwachung in der Arbeitswelt liefert Wolfie Christl vom digitalen Labor Cracked Labs mit seiner Studie „Digitale Kontrolle und Überwachung am Arbeitsplatz. Von der Ausweitung betrieblicher Datenerfassung zum algorithmischen Management?"

- *https://crackedlabs.org/dl/CrackedLabs_Christl_UeberwachungKontrolleArbeitsplatz.pdf*

AlgorithmWatch ist eine gemeinnützige Forschungs- und Advocacy-Organisation mit dem Ziel, Systeme automatisierter Entscheidungsfindung (ADM) und deren Auswirkungen auf die Gesellschaft zu beobachten und zu analysieren.

- *https://algorithmwatch.org/de/*

ANSPRECHPERSONEN

Abteilung Recht der Produktionsgewerkschaft PRO-GE
Beratung von Betriebsrät*innen (Produktion) zu Datenschutz am Arbeitsplatz
recht@proge.at

Dr.in Eva Angerler, GPA-djp, Abteilung Arbeit und Technik
Beratung von Betriebsrät*innen (Angestellte Privatwirtschaft) zu Datenschutz am Arbeitsplatz
Eva.Angerler@gpa.at

Gleichbehandlungsanwaltschaft
Die Gleichbehandlungsanwaltschaft ist eine unabhängige staatliche Einrichtung zur Durchsetzung des Rechts auf Gleichbehandlung und Gleichstellung und zum Schutz vor Diskriminierung. Sie bietet für Menschen aus ganz Österreich vertrauliche und kostenlose Beratung und Unterstützung an.
https://www.gleichbehandlungsanwaltschaft.gv.at

Fridolin Herkommer
Leiter des Büros für Digitale Agenden in der AK Wien, Digitalisierungsfonds Arbeit 4.0 und bis Oktober 2021 Mitglied des Österreichischen Rats für Robotik und KI
Fridolin.Herkommer@akwien.at

LITERATURVERZEICHNIS

Adensamer, Angelika / Klausner, Lukas Daniel. *Algorithmen in der Entscheidungsfindung. Leitfaden zur Verantwortlichkeit und Rechenschaft.* Februar 2021 *https://wien.arbeiterkammer.at/interessenvertretung/arbeitdigital/DataPolitics/VerA_Leitfaden_Final.pdf.*

Artikel-29-Datenschutzgruppe. *Leitlinien in Bezug auf die Einwilligung gemäß Verordnung 2016/679* angenommen am 28. November 2017 zuletzt überarbeitet und angenommen am 10. April 2018. 17/DE WP259 rev.01.

Artikel-29-Datenschutzgruppe. Leitlinien zu automatisierten Entscheidungen im Einzelfall einschließlich Profiling für die Zwecke der Verordnung 2016/679 angenommen am 3. Oktober 2017 zuletzt überarbeitet und angenommen am 6. Februar 2018. 17/DE WP251 rev.01.

Artikel-29-Datenschutzgruppe. *Stellungnahme 2/2017 zur Datenverarbeitung am Arbeitsplatz* angenommen am 8. Juni 2017. 17/DE WP 249.

Beck, Susanne. *Wie diskriminierend ist künstliche Intelligenz? Ein Expertinnenbeitrag von Prof. Dr. Susanne Beck, Plattform Lernende Systeme. September 2019 https://www.wissenschaftsjahr.de/2019/neues-aus-der-wissenschaft/das-sagt-die-wissenschaft/wie-diskriminierend-ist-kuenstliche-intelligenz/.*

Beck, Susanne et al. Whitepaper. *Künstliche Intelligenz und Diskriminierung. Herausforderungen und Lösungsansätze.* AG IT-Sicherheit, Privacy, Recht und Ethik. Plattform für Künstliche Intelligenz. Juni 2019.

Bundesministerium für Klimaschutz, Umwelt, Energie, Mobilität, Innovation und Technologie (BMK). S*trategie der Bundesregierung für Künstliche Intelligenz.* Artificial Intelligence Mission Austria 2030 (AIM AT 2030). Februar 2021

Christl, Wolfie. *Digitale Kontrolle und Überwachung am Arbeitsplatz. Von der Ausweitung betrieblicher Datenerfassung zum algorithmischen Management?* Wien, 2021. *https://crackedlabs.org/dl/CrackedLabs_Christl_UeberwachungKontrolleArbeitsplatz.pdf*.

Gerards, Janneke / Xenidis, Raphaële. *Algorithmic discrimination in Europe: Challenges and opportunities for gender equality and non-discrimination law. A special report by the European network of legal experts in gender equality and non-discrimination*, European Commission, 2021.

Greif, Elisabeth und Kullmann, Miriam. Algortithmenbasiertes Personalrecruiting. Antidiskriminierungs- und Datenschutzaspekte. *ZAS* 2021/13.

Haslinger, Susanne / Krisch, Andreas / Riesenecker-Caba, Thomas (Hrsg). *Beschäftigtendatenschutz – Handbuch für die betriebliche Praxis.* Wien 2017.

Hopf/Mayr/Eichinger/Erler, *GlBG*[2] (2021) (Stand 1.1.2021, rdb.at).

Kolleck, Alma, und Carsten Orwat. 2020. *Mögliche Diskriminierung durch algorithmische Entscheidungssysteme und maschinelles Lernen – ein Überblick.* TAB Hintergrundpapier Nr. 24. Büro für Technikfolgen-Abschätzung beim Deutschen Bundestag.

Michot, Sarah / Mollen, Anne / Schiller, Anna Lena / Wulf, Jessica. *Algorithmen-basierte Diskriminierung. Warum Antidiskriminierungsgesetze jetzt angepasst werden müssen.* Digital Autonomy Hub, Policy Brief #5, Februar 2022.

Prassl-Adams, Jeremias. Der Algorithmus im Personalmanagement. *DRdA* 2022, 195. *https://www.drda.at/a/400_DRDA_1/Der-Algorithmus-im-Personalmanagement.*

Tinhofer, Andreas. Algorithmenbasierte Entscheidungen und Diskriminierung. *DRdA* 2022, 171. *https://rdb.manz.at/document/rdb.tso.LIrda202201a14*.

Winkler, Christoph. *Mittelbare Diskriminierung durch Algorithmen: Wo liegt die Gefahr?* Diplomarbeit zur Erlangung des akademischen Grades Magister der Rechtswissenschaften im Diplomstudium Rechtswissenschaften an der JKU Linz, Juli 2021.

Verordnung (EU) 2016/679 des Europäischen Parlaments und des Rates vom 27. April 2016 zum Schutz natürlicher Personen bei der Verarbeitung personenbezogener Daten, zum freien Datenverkehr und zur Aufhebung der Richtlinie 95/46/EG. Amtsblatt 2016 L119/1 (Datenschutz-Grundverordnung, DSGVO).

Vorschlag für eine Verordnung des Europäischen Parlaments und des Rates zur Festlegung harmonisierter Vorschriften für künstliche Intelligenz (Gesetz über künstliche Intelligenz) und zur Änderung bestimmter Rechtsakte der Union. COM(2021) 206 final.

Fragen, die Sie bei der (Mit-)Gestaltung von KI-Assistenzsystemen am Arbeitsplatz stellen sollten

In den vorangegangenen Kapiteln haben wir ethische und rechtliche Grundsätze sowie Prinzipien und Methoden vorgestellt, mit denen KI-Assistenzsysteme am Arbeitsplatz gestaltet werden können. Aus den vorgestellten Inhalten, der angeführten Literatur und Diskussionen mit Betriebsrät*innen haben wir eine Liste an Fragen erstellt, die Sie bei der (Mit-)Gestaltung von KI-Assistenzsystemen[1] am Arbeitsplatz stellen sollten.

Die Fragen können von Betriebsrät*innen dazu genutzt werden, mit der IT-Abteilung und allen weiteren relevanten Interessenvertretungen in Dialog zu treten (s. Kapitel KI im Betrieb: Wer soll wie mitreden? dieses Abschnitts). Auch könnte die Auflistung dazu genutzt werden, um konkrete Kriterien für die Einführung oder Gestaltung neuer KI-Assistenzsysteme abzuleiten und diese beispielsweise innerhalb einer Rahmenbetriebsvereinbarung festzulegen (s. Kapitel Mitbestimmung bei Betriebsvereinbarungen dieses Abschnitts). Grundsätzlich gilt: Je früher im Prozess von der KI-Entwicklung bis zum tatsächlichen Einsatz des KI-Assistenzsystems Sie beginnen, sich mit den folgenden Fragen zu beschäftigen, umso besser wird es Ihnen gelingen, Ihr KI-Assistenzsystem verantwortungsvoll zu gestalten.

Die folgenden Fragen stellen vor allen Dingen eine Anregung zum Nachdenken und Nachlesen dar. Die Frage danach, welche Aspekte für Ihr KI-Assistenzsystem relevant sind, müssen wir Ihnen überlassen.

Technikfolgenabschätzung

- Welche Auswirkungen wird die Einführung des KI-Assistenzsystems auf folgende Bereiche haben?
  - Betrieb
    - Arbeit und Arbeitsprozesse
    - Arbeitsgesundheit
    - Betriebsklima
    - Diversität an Menschen im Betrieb
    - Fähigkeiten von Arbeitnehmenden
  - Klima
  - Demokratie
  - Gesellschaft
- Gibt es eine schrittweise Einführung des KI-Assistenzsystems, die es ermöglicht, etwaige Problemstellungen bei gleichzeitiger Minimierung negativer Auswirkungen zu erkennen?

Zum Nachlesen empfehlen wir:

- Den Bericht *Künstliche Intelligenz und die Zukunft der digitalen Arbeitsgesellschaft: Konturen einer ganzheitlichen Technikfolgenabschätzung* von Thorben Albrecht und Christian Kellermann[2]
- Das Buch *Der mitbestimmte Algorithmus. Gestaltungskompetenz für den Wandel der Arbeit*, herausgegeben von Welf Schröter[3]

Fairness, Ethik und Bias

- Wurden bei der Gestaltung des KI-Assistenzsystems relevante Interessenvertretungen und Expertisen einbezogen (Betriebsrat, Datenschutzbeauftragte, Gleichstellungsbeauftragte, Inklusionsbeauftragte, etc.)?

1 Wie viele andere Inhalte dieses Buches lässt sich auch diese Liste auf KI-Systeme umlegen, die keine Assistenzsysteme sind.

2 Albrecht und Kellermann, 2020.

3 Schröter, 2019.

- Stimmt das KI-Assistenzsystem mit den ethischen Grundsätzen der Firma, des Betriebsrats, etc. überein?

- Ist das System inklusiv gestaltet und berücksichtigt es auch sogenannte vulnerable Gruppen bzw. Menschen in Minderheitenpositionen im Unternehmen (z. B. Menschen mit Beeinträchtigungen, Menschen, deren Erstsprache nicht Deutsch ist, Leiharbeiter*innen, Menschen mit Pflegeverantwortung)?

- Welche Maßnahmen wurden bei der Entwicklung und Einbindung des KI-Assistenzsystems ergriffen, um Bias und Diskriminierung zu verhindern?
  - Wurde das System mit Personen mit unterschiedlichen Bedürfnissen getestet?
  - Wurde getestet, ob das System zu anderen, unerwünschten Ergebnissen kommt, wenn man bestimmte vom System erfasste Merkmale verändert (z. B. das Geschlecht oder die Hautfarbe)?

- Welche Daten wurden/werden verwendet, um das System zu trainieren?

- Sind die Trainingsdaten von guter Qualität?
  - Spiegeln die Trainingsdaten die tatsächlichen Verhältnisse korrekt wider?
  - Wurden entsprechende Gewichtungen vorgenommen, falls die in den Trainingsdaten widergespiegelten tatsächlichen Verhältnisse nicht den gewünschten Verhältnissen entsprechen?

- Wird das sogenannte Human-In-The-Loop-Prinzip angewandt, wobei ein Mensch die Ergebnisse des KI-Assistenzsystems kontrolliert und diesem Feedback gibt?

- Wie wurden menschliche Fehlerneigungen im Umgang mit dem KI-Assistenzsystem überprüft (z. B., dass Menschen einem mittels KI berechneten Ergebnis zu sehr vertrauen, selbst dann, wenn sie es nicht nachvollziehen können)?

Zum Nachlesen empfehlen wir:

- Die folgenden Kapitel in diesem Buch:
  - Verantwortungsvolle Künstliche Intelligenz: Ein kurzer Überblick
  - Warum man bei (KI-)Algorithmen am Arbeitsplatz genauer hinsehen sollte
  - Technologieentwicklung mit Nutzer*innen – warum denn das?
  - Assistenzsysteme zur Förderung psychischer Gesundheit am Arbeitsplatz?
  - Wie KI mit Daten lernt und welche Vorurteile dabei erzeugt werden
  - KI aus systemischer Sicht und ihre Risiken

- Die Checkliste der High-Level Expert Group on AI: *ALTAI – The Assessment List on Trustworthy Artificial Intelligence*[4]

- Das Buch *Der mitbestimmte Algorithmus. Gestaltungskompetenz für den Wandel der Arbeit*, herausgegeben von Welf Schröter[5]

Transparenz

- Werden die Entscheidungen des KI-Assistenzsystems dokumentiert?

- Wie ist sichergestellt, dass Entscheidungen auch nachträglich nachvollzogen werden können (d. h. auch, welche Daten und Modelle zum Ergebnis geführt haben)?

- Werden die Entscheidungen des KI-Assistenzsystems den Nutzer*innen niederschwellig (z. B. multimedial) erklärt?

- Wird niederschwellig erklärt, wie das KI-Assistenzsystem zu der jeweiligen Entscheidung gekommen ist?

- Ist die Verwechslungsgefahr des KI-Assistenzsystems mit Menschen ausgeschlossen (z. B. in der Interaktion mit KI-gesteuerten Chatbots)?

4 High-Level Expert Group on AI, 2020.

5 Schröter, 2019.

Zum Nachlesen empfehlen wir:

- Das folgende Kapitel in diesem Buch:
  - Erklärbare Künstliche Intelligenz
- Die Checkliste der High-Level Expert Group on AI: *ALTAI – The Assessment List on Trustworthy Artificial Intelligence*[6]

Sicherheit allgemein

- Welche Maßnahmen gewährleisten die Integrität (d. h. Einheitlichkeit und Akkuratheit der Daten) und Robustheit (z. B. Resilienz gegen Hackangriffe oder technisches Versagen) des Systems über seinen gesamten Lebenszyklus hinweg?

- Wie lange sind Sicherheits-Updates für das System gewährleistet und wer stellt diese zur Verfügung?

- Wurde das KI-Assistenzsystem sicherheitszertifiziert (z. B. durch TÜV AUSTRIA)?

Zum Nachlesen empfehlen wir:

- Das folgende Kapitel in diesem Buch:
  - KI aus systemischer Sicht und ihre Risiken

- Das Whitepaper *Towards Certification of Machine Learning Applications* aus der Zusammenarbeit von TÜV AUSTRIA und Johannes Kepler Universität Linz[7]

Datenschutz

- Werden nur notwendige personenbezogene Daten erfasst?

- Welche Datenkategorien personenbezogener Daten werden konkret erfasst?
  - Ist die Erfassung dieser Datenkategorien personenbezogener Daten wirklich notwendig?
  - Welche dieser Datenkategorien können ggf. anonymisiert werden?

- Wie können Arbeitnehmer*innen einsehen, welche Daten über sie erhoben werden?

- Wer hat Zugriff auf die personenbezogenen Daten?
  - Wurde mittels IT oder offizieller Regelung festgelegt, wer die Daten auslesen und verändern darf?
  - Über welche technischen Schnittstellen ist ein Zugriff wie möglich?
  - Wie erfolgt die Protokollierung der Zugriffe, um nachvollziehen zu können, wann wer wie auf die personenbezogenen Daten zugegriffen hat?

- Welche Analysemöglichkeiten gibt es?
  - Ist sichergestellt, dass analysierte Gruppen groß genug sind, um Anonymität zu gewährleisten?

- Wie wird sichergestellt, dass die erfassten personenbezogenen Daten nicht missbräuchlich ausgewertet werden?
  - Wie wird sichergestellt, dass die erfassten personenbezogenen Daten auch in Zukunft mit bisher unbekannten Methoden nicht missbräuchlich ausgewertet werden können?

- Wie sind personenbezogene Daten vor unbefugten Zugriffen geschützt (z. B. durch Verschlüsselung, Speicherort, Firewall, etc.)?

- Wie wird sichergestellt, dass personenbezogene Daten auf den KI-Assistenzsystemen verbleiben und nicht vervielfältigt werden können?

- Wie wird sichergestellt, dass unterschiedliche Daten nicht miteinander in Verbindung gebracht werden können, um das Risiko für Identifikation und Tracking zu minimieren?

Zum Nachlesen empfehlen wir:

- Das folgende Kapitel in diesem Buch:
  - Gesetzliche Rahmenbedingungen für den Einsatz von KI-Assistenzsystemen am Arbeitsplatz

6 High-Level Expert Group on AI, 2020.

7 Winter u. a., 2021.

- Das Buch *Beschäftigtendatenschutz. Handbuch für die betriebliche Praxis*, herausgegeben von Susanne Haslinger, Andreas Krisch und Thomas Riesenecker-Caba[8]
- Die Studie *Digitale Überwachung und Kontrolle am Arbeitsplatz. Von der Ausweitung betrieblicher Datenerfassung zum algorithmischen Management?* von Wolfie Christl[9]

- Die Studie *Verarbeitung personenbezogener Beschäftigtendaten und Grenzen betrieblicher Mitbestimmung in einer digitalisierten Arbeitswelt* von Thomas Riesenecker-Caba und Franz Astleithner[10]

Verantwortung, Monitoring und Risk Management

- Wer trägt die Verantwortung für Fehler und unerwünschte Auswirkungen sowohl im juristischen Sinne als auch im innerbetrieblichen Kontext?
  - Wurden etwaige Entschädigungen vorab definiert?
  - Wurden etwaige Sanktionen der verantwortlichen Personen vorab definiert?

- Gilt das sogenannte Human-In-Command-Prinzip, nach dem Menschen ständig Aufsicht über die Operationen des KI-Assistenzsystems haben?

- Wie und durch wen wird das KI-Assistenzsystem und dessen Auswirkungen über den gesamten Lebenszyklus überwacht und kontrolliert?

- Wie oft finden Überprüfungen statt?

- Wurden Algorithmen implementiert, die die Qualität der KI-Ergebnisse laufend überprüfen?

- Wird der Umgang der Nutzer*innen mit dem KI-Assistenzsystem über den gesamten Lebenszyklus überwacht?

- Wie bzw. wem können Nutzer*innen Probleme mit dem KI-Assistenzsystem melden?

- Welche Maßnahmen werden ergriffen, um die Genauigkeit des KI-Assistenzsystems zu überprüfen?

Modifizierbarkeit

- Inwieweit ist das KI-Assistenzsystem an sich ändernde Umstände oder entdeckten Bias anpassbar?
  - Inwieweit wurde dies bei der Systemgestaltung berücksichtigt?

- Können Korrekturen am KI-Assistenzsystem vorgenommen werden?
  - Können beispielsweise neue Trainingsdaten eingelesen werden?

- Gibt es einen Plan dafür, das System, wenn notwendig, schnell stoppen zu können?

- Gibt es einen Plan dafür, das System, wenn notwendig, langfristig deaktivieren und entfernen zu können?

- Wie können Arbeitnehmende und ihre Vertretungen Modifizierungen anstoßen?
  - Gibt es dafür Ansprechpersonen und ein konkretes, einfaches und schnelles Prozedere?

- Gibt es regelmäßige Statusmeetings, bei denen über den aktuellen Stand der KI reflektiert wird?

Zum Nachlesen empfehlen wir:

- Die folgenden Kapitel in diesem Buch:
  - Wie KI mit Daten lernt und welche Vorurteile dabei erzeugt werden
  - KI aus systemischer Sicht und ihre Risiken

- Das Buchkapitel „Der mitbestimmte Algorithmus. Arbeitsweltliche Kriterien zur sozialen Gestaltung von Algorithmen und algorithmischen Entscheidungssystemen“ von Welf Schröter.[11]

8 Haslinger, Krisch, und Riesenecker-Caba, 2020.
9 Christl, 2021.
10 Riesenecker-Caba und Astleithner, 2021.
11 Schröter, 2019.

Information, Ansprechpersonen und Weiterbildung

- Wurden alle relevanten Interessenvertretungen und Expertisen einbezogen (Betriebsrat, Datenschutzbeauftragte, Gleichstellungsbeauftragte, Inklusionsbeauftragte, etc.) und über die geplante Einführung des KI-Assistenzsystems informiert?

- Gibt es für das KI-Assistenzsystem abgestellte Ansprechpersonen?
  - Werden alle Nutzer*innen und Personen, die von dem KI-Assistenzsystem betroffen sind, darüber informiert, wer diese Ansprechpersonen sind?

- Wie werden neue Nutzer*innen in der Anwendung des KI-Assistenzsystems geschult?

- Besteht eine ausreichende und nachvollziehbare Dokumentation etc. über das KI-Assistenzsystem?
  - Gibt es Anleitungen oder Erklärvideos für Nutzer*innen?
  - Wie können Nutzer*innen erfahren, wie verlässlich das KI-Assistenzsystem ist, welche Fähigkeiten, Limitationen und Risiken es birgt?
  - Werden die Ergebnisse des KI-Assistenzsystems den Nutzer*innen bei der Verwendung erklärt?

Zum Nachlesen empfehlen wir:

- Die folgenden Kapitel in diesem Buch:
  - KI im Betrieb: Wer soll wie mitreden?
  - Mitbestimmung bei Betriebsvereinbarungen

Wichtige gesetzliche Rahmenbedingungen

- Arbeitet das KI-Assistenzsystem im Einklang mit den Grundsätzen für die Verarbeitung personenbezogener Daten der DSGVO (Art. 5 lit a-f)?
  - Ist die Datenverarbeitung personenbezogener Daten rechtmäßig?
    - Z. B. besteht eine Einwilligung der betroffenen Personen, eine vertragliche oder gesetzliche Verpflichtung oder berechtigte Interessen zur Datenverarbeitung?
  - Ist der Zweck der Datenverarbeitung legitim und eindeutig?
    - Oder gäbe es z. B. einen anderen Weg, die Daten zu verarbeiten als die Nutzung des KI-Assistenzsystems?
  - Wird der Grundsatz der Datenminimierung eingehalten?
    - Oder sammelt das KI-Assistenzsystem mehr personenbezogene Daten als notwendig?
  - Wird das Recht auf Löschung und Aktualität eingehalten?
    - D. h. können die betroffenen Personen jederzeit ihre im KI-Assistenzsystem verankerten personenbezogenen Daten ändern oder löschen lassen?
  - Werden die personenbezogenen Daten im KI-Assistenzsystem nur so lange gespeichert wie für den definierten Zweck notwendig?
  - Sind Integrität und Vertraulichkeit der personenbezogenen Daten gewahrt?
    - D. h. sind die Daten vor unberechtigtem oder unbeabsichtigtem Zugriff oder Verlust geschützt?

- Trifft das KI-Assistenzsystem vollautomatisch rechtlich wirksame oder beeinträchtigende Entscheidungen über Personen auf der Basis verarbeiteter personenbezogener Daten (untersagt nach Art. 22 DSGVO)?
  - D.h. findet beispielsweise eine automatisierte Analyse der Arbeitsleistung, Gesundheit, Aufenthaltsorte oder Interessen statt?

- Wurde eine Datenschutzfolgenabschätzung durchgeführt (Art. 35 DSGVO)?

- Führt das KI-Assistenzsystem zur Ungleichbehandlung von Personen (s. GiBG und BEinstG)?

- Wurde eine Betriebsvereinbarung abgeschlossen (§ 96a ArbVG)?

Zum Nachlesen empfehlen wir:

- Die folgenden Kapitel in diesem Buch:
  - Gesetzliche Rahmenbedingungen für den Einsatz von KI-Assistenzsystemen am Arbeitsplatz
  - KI im Betrieb: Wer soll wie mitreden?
- Das Buch *Beschäftigtendatenschutz. Handbuch für die betriebliche Praxis*, herausgegeben von Susanne Haslinger, Andreas Krisch und Thomas Riesenecker-Caba[12]
- Den Arbeiterkammer-Bericht *Künstliche Intelligenz in der Arbeitswelt* von Eva Angerler, Martina Chlestil und Michael Heiling[13]

LITERATURVERZEICHNIS

Albrecht, Thorben, und Christian Kellermann. 2020. *Künstliche Intelligenz und die Zukunft der digitalen Arbeitsgesellschaft: Konturen einer ganzheitlichen Technikfolgenabschätzung.* Working Paper Forschungsförderung. *https://www.boeckler.de/de/faust-detail.htm?sync_id=HBS-007902.*

Angerler, Eva, Martina Chlestil, und Michael Heiling. 2021. *Künstliche Intelligenz in der Arbeitswelt. https://wien.arbeiterkammer.at/interessenvertretung/arbeitdigital/policypapers/Policy_Paper_KI_in_der_Arbeitswelt.pdf.*

Christl, Wolfie. 2021. *Digitale Überwachung und Kontrolle am Arbeitsplatz. Von der Ausweitung betrieblicher Datenerfassung zum algorithmischen Management? https://crackedlabs.org/dl/CrackedLabs_Christl_UeberwachungKontrolleArbeitsplatz.pdf.*

Haslinger, Susanne, Andreas Krisch, und Thomas Riesenecker-Caba (Hrsg.) 2020. *Beschäftigtendatenschutz. Handbuch für die betriebliche Praxis.* 2. Aufl.

High-Level Expert Group on AI. 2020. *Assessment List for Trustworthy Artificial Intelligence (ALTAI) for Self-Assessment. https://doi.org/10.2759/002360.*

Riesenecker-Caba, Thomas, und Franz Astleithner. 2021. *Verarbeitung personenbezogener Beschäftigtendaten und Grenzen betrieblicher Mitbestimmung in einer digitalisierten Arbeitswelt. https://www.forba.at/wp-content/uploads/2021/06/Verarbeitung-persbez-Daten-und-MitbestimmungFORBA-Bericht2021_DigiFonds.pdf.*

Schröter, Welf. 2019. Der mitbestimmte Algorithmus. Arbeitsweltliche Kriterien zur sozialen Gestaltung von Algorithmen und algorithmischen Entscheidungssystemen. In *Der mitbestimmte Algorithmus. Gestaltungskompetenz für den Wandel der Arbeit,* herausgegeben von Welf Schröter, 101–50.

Schröter, Welf (Hrsg.) 2019. *Der mitbestimmte Algorithmus. Gestaltungskompetenz für den Wandel der Arbeit.* Mössingen-Talheim: Talheimer Verlag.

Winter, Philip Matthias, Sebastian Eder, Johannes Weissenböck, Christoph Schwald, Thomas Doms, Tom Vogt, Sepp Hochreiter, und Bernhard Nessler. 2021. *White Paper – Trusted Artificial Intelligence: Towards Certification of Machine Learning Applications. https://www.tuv.at/wp-content/uploads/2022/03/Whitepaper_Trusted-AI_TUeV-AUSTRIA_JKU.pdf.*

12 Haslinger, Krisch, und Riesenecker-Caba, 2020.

13 Angerler, Chlestil, und Heiling, 2021.

Darf`s ein bisserl mehr sein? Hintergrundwissen über die KI für Detailverliebte

Überblick über Infrastruktur, Entwicklungsumgebungen und Verbrauchsmodelle von KI-Systemen

DIESES KAPITEL IST FÜR SIE VON INTERESSE, WENN …

… Sie mitreden können möchten, wenn Ihre IT-Abteilung über KI-Systeme diskutiert.

… Sie bereits ein grundlegendes Verständnis für KI mitbringen und sich einen Grundstock der wesentlichen technischen Grundbegriffe aneignen möchten.

Technologiestapel für KI

KI-Entwickler*innen stehen zahlreiche Werkzeuge (engl. Tools) und Techniken zur Verfügung. Dieser technische Grundstock wird als *moderner Technologiestapel* bezeichnet.

Obwohl sich die Algorithmen in den letzten 20 Jahren wenig veränderten, haben sich die Tools einer bedeutenden Entwicklung unterzogen. Heutzutage müssen Entwickler*innen nicht mehr alles von Grund auf aufbauen. Sie können Plattformen für Maschinelles Lernen, Bibliotheken, Computing- und Datenplattformen benutzen. Diese Softwaretools können zum Teil in Form von Dienstleistungen erworben werden, manchmal auch in Kombination mit der passenden Hardware. Ein moderner Technologiestapel besteht, aus der Sicht von Entwickler*innen, aus zwei Komponenten: der *Infrastruktur* und der *Entwicklungsumgebung*.[1] Diese werden im Folgenden am Beispiel des Maschinellen Lernens (ML) erklärt.

Infrastruktur

Die Infrastruktur bildet die Grundlage des Ablaufs des Maschinellen Lernens. Dazu gehören Ressourcen, Techniken, Prozesse, Plattformen und Werkzeuge, die dazu benötigt werden, Daten zu speichern, zu verarbeiten, die ML-Modelle zu entwickeln und mit ihnen zu operieren. Die Algorithmen des Maschinellen Lernens selbst werden auch zur Infrastruktur gezählt. Die Wahl des Modells des Maschinellen Lernens entscheidet darüber, welche Daten aufbereitet, Werkzeuge benutzt und welche Komponenten miteinander verknüpft werden.

Die Infrastruktur besteht aus folgenden Komponenten:

- **Compute:** Die Rechenpower, die genutzt wird, um KI/ML-Algorithmen zu bedienen. Dafür werden unterschiedliche Server, virtuelle Maschinen und spezialisierte Hardware (z. B. Grafikkarten oder Neuronale-Netzwerk-Chips) verwendet.

- **Data***:* Daten sind eine wichtige Komponente eines Systems des Maschinellen Lernens, da das System nur so gut ist wie die Daten, mit denen es trainiert wird. Es gibt eine große Auswahl an Datenplattformen: strukturierte und nicht-strukturierte Datenbasen, Big-Data-Plattformen, verwaltete Datenbasen oder cloudbasierte Datenbasen. Oft werden auch frei verfügbare Datensätze mit eigenen kombiniert.

1 Subramanian, 2018.

- **Algorithmen des Maschinellen Lernens (ML-Algorithmen):** Es gibt drei Kategorien von ML-Algorithmen: *überwachtes Lernen* (auch „beaufsichtigtes Lernen“ genannt), *nicht überwachtes Lernen* (auch „unbeaufsichtigtes Lernen“) und *bestärkendes Lernen*. Bei überwachtem Lernen bekommt das System die Kategorien, die es Eingabedaten zuordnen soll, vorgegeben und erhält während des Lernvorgangs ein Feedback dazu, ob die Eingabedaten der richtigen Kategorie zugeordnet wurden oder nicht. Anhand des Feedbacks lernt das System, korrekte Zuordnungen zu erzeugen. Bei nicht überwachtem Lernen werden Eingabedaten daraufhin untersucht, ob es Regelmäßigkeiten gibt, anhand welcher sich Kategorien bilden lassen. Bei bestärkendem Lernen schließlich wird das System optimiert, indem es nach dem Trial-and-Error-Prinzip lernt, welches Verhalten zu einer maximalen Belohnung führt. Überwachtes und nicht überwachtes Lernen ist epistemisch opak, d. h., es lässt sich schwer nachvollziehen, wie ein Ergebnis zustande kam bzw. welche Muster in einem Datensatz als relevant erachtet wurden und welche nicht.

- **Plattform des Maschinellen Lernens:** Plattformen und Frameworks können die Trainingsdaten normalerweise von unterschiedlichen Quellen heranziehen, bieten eine Reihe von Trainingsalgorithmen an und unterstützen dabei mehrere Programmiersprachen. Häufig verwendete ML-Plattformen sind z. B. scikit-learn, TensorFlow oder PyTorch.

Wo wird gerechnet?

Die Datenübertragung ist ein weiteres wichtiges Element der Infrastruktur. Diese ist zentral für das Training, die Anwendung und die Optimierung des Modells. Üblicherweise werden Verbindungen von einer Datenquelle zu sogenannten Pipelines (dt. Leitungssysteme) hergestellt. Pipelines werden benutzt, um Workflows zu automatisieren und sind besonders relevant, wenn nicht alle Daten auf demselben physischen System, sondern zum Beispiel auf mehreren Computern verteilt sind oder in der Cloud verarbeitet werden.

In der *Trainingsphase* eines KI-Assistenzsystems sind Faktoren wie die *entstehenden Kosten*, die *Zweckmäßigkeit der Daten und des Prozesses*, die *Bestimmungen* hinsichtlich der Daten (wie z. B. ihre Lokalisierung) sowie die *Sicherheit* von großer Bedeutung. Es gilt abzuwägen, ob es günstiger und zweckmäßiger ist, wenn das Training vor Ort oder auf einer Cloud stattfindet. Diese Überlegungen hängen auch davon ab, wie viele Modelle verwendet werden sowie von der Größe und Art der Daten und inwiefern die Infrastruktur automatisiert werden kann.

In der Phase der *Anwendung* muss zwischen *Performanz* und *Latenz* balanciert werden, wobei die Hardware als limitierender Faktor zu betrachten ist. ML-Modelle, die schnelle Antworten oder eine niedrige Latenz voraussetzen, sollten lokale oder Edge-Infrastrukturen verwenden und dafür optimiert werden, auf lokaler Hardware mit geringem Energieverbrauch zu laufen. Modelle, die Latenz tolerieren, können Cloud-Infrastrukturen einsetzen, die bei Bedarf skaliert werden können.

Entwicklungsumgebung

Die Entwicklungsumgebung ist ein Softwarepaket und ein wichtiges Hilfsmittel bei der Entwicklung von Algorithmen. In diesem Zusammenhang sprechen wir von *Bibliotheken*, der *integrierten Entwicklungsumgebung*, der *Visualisierung* und dem *Workflow*.

- **Bibliotheken:** Unterschiedliche Bibliotheken können zusammen mit den ML-Plattformen verwendet werden, um z. B. erweiterte mathematische Operationen anzuwenden (NumPy), oder um eine spezifische kognitive Fähigkeit hinzuzufügen, wie Computer Vision (OpenCV), die Übersetzung von Sprache (OpenNMT), etc.

- **Integrierte Entwicklungsumgebungen (engl. Integrated Development Environment):** Entwicklungsumgebungen sind Werkzeuge zum Schreiben, Ausführen und Analysieren von Programmen (z. B. PyCharm, Matlab, Jupyter, etc.). Sie erleichtern die Arbeit, da sie häufige Abläufe bequem durchführbar machen.

- **Visualisierung:** Die Visualisierung von Daten spielt eine bedeutende Rolle im Technologiestapel. Dafür werden z. B. Matlab, seaborn, Facets, oder komplette Datenanalyseplattformen wie Tableau verwendet.

- **Workflow:** So genannte Workflow-Tools im Technologiestapel ermöglichen den Austausch, die Kollaboration und Automatisierung unter den Entwickler*innen. Mit der steigenden Zahl an Entwickler*innen, die sich mit KI/ML-Algorithmen beschäftigen, wird es immer wichtiger, die Kommunikation untereinander aufrechtzuerhalten. Einige Werkzeuge dafür sind beispielsweise Jupyter, Anaconda, GitHub oder VSTS.

Verbrauchsmodelle

Anbieter*innen von cloudbasierten Dienstleistungen für KI-Fähigkeiten stellen den Entwickler*innen je nach Bedarf unterschiedliche Verbrauchsmodelle zur Verfügung. Diese Modelle können grob in unterschiedliche Abstufungen unterteilt werden. **Je größer der Anteil an Dienstleistungen ist, den man einkauft, umso weniger Knowhow muss man selbst aufbauen. Gleichzeitig sinkt aber die Kontrolle, die über die Abläufe ausgeübt werden kann.**

Im Folgenden besprechen wir vier Varianten. Die Variante „i“ bietet den maximalen Überblick und Kontrolle von der Benutzer*innenseite, während mit jeder weiteren Abstufung mehr Verantwortung in die Hände der Dienstleister gelegt wird. Die Abkürzung „AIaaS“ steht für den engl. Begriff *„Artificial Intelligence as a Service“*.

- **Variante „i“ – KI-Stapel:** Der KI-Stapel ist das Referenzverbrauchsmodell. Jede Komponente einer Infrastruktur (ML-Plattform, Algorithmen, Compute, Daten) wird hier von der benutzenden Person aufgesetzt und verwaltet. Die benutzende Person baut, trainiert und verwendet die ML-Modelle selbst. Sie ist verantwortlich dafür, alle Komponenten der Entwicklungsumgebung zu installieren und entsprechend einzusetzen. Dieses Verbrauchsmodell kann vor Ort und in einer privaten Cloud verwendet werden. Das Maß an Kontrolle ist hoch und potentieller Missbrauch kann höchstens im eigenen Unternehmen auftreten.

- **Variante „ii“ – AIaaS:** Die User*innen verwenden hierbei die Modelle, Algorithmen, Datenspeicher, Compute-Ressourcen und bei Bedarf auch die Hardware, die sie vom Anbieter kaufen. Die Infrastruktur und die Komponenten werden dabei nicht selbstständig installiert und verwaltet. Die als Service angebotenen ML-Fähigkeiten, wie z. B. Google Cloud ML Engine oder Amazon AWS zählen hierzu.

- **Variante „iii“ – Verwaltete AIaaS:** Diese Variante bildet die nächste Stufe, bei der die User*innen sich noch weniger mit der Infrastruktur auseinandersetzen müssen. Der Anbieter übernimmt auch die Verwaltung der Algorithmen. Services wie Google Cloud AutoML, Amazon SageMaker oder Azure ML Studio gehören in diese Kategorie.

- **Variante „iv“ – Cognition-aaS:** Bei diesen Verbrauchsmodellen erwerben die Benutzer*innen komplexe kognitive Fähigkeiten als fertige Dienste. Sollte z. B. eine Anwendung für Videoüberwachung entwickelt werden, können Fähigkeiten zur visuellen Erkennung gekauft werden (z. B. Amazon Rekognition Video, Google Vision AI, Azure Computer Vision, etc.).

LITERATURVERZEICHNIS

Subramanian, Sriram. 2018. *Modern AI Stack and AI as a Service Consumption Models. https://medium.com/clouddon/modern-ai-stack-ai-service-consumption-models-f9957dce7b25.*

Wie KI-Assistenzsysteme ihre Umwelt wahrnehmen: Sensoren

DIESES KAPITEL IST FÜR SIE VON INTERESSE, WENN …

… Sie wissen möchten, wie KI-Assistenzsysteme über verschiedene Sensoren ihre Umgebung wahrnehmen.

… Sie mehr über die Vorteile und Risiken verschiedener Sensoren erfahren möchten.

Abbildung 1 bietet einen Überblick über verschiedene Sensoren von KI-Assistenzsystemen am Arbeitsplatz, unterteilt nach ihrer Platzierung, den Typen/Verfahren und dem Ziel der Erfassung. Es wird unterschieden zwischen Sensoren, die in der Umgebung verbaut sind (z. B. Überwachungskameras in öffentlichen Räumen) und mobilen Sensoren, die temporär an Personen oder Objekten angebracht werden (z. B. Dashcams der Polizei). Häufige sensorische Ziele sind die (Teil-)Erkennung bzw. Verfolgung des Menschen (z. B. Gesichtserkennung am Flughafen), von Geräten (z. B. Ortung des Mobiltelefons bei Verlust), Bauteilen (z. B. Verunreinigungsprüfung in der Aluminiumproduktion), Maschinen (z. B. Sicherstellung, dass Roboterzonen leer sind) und digitalen Zuständen (z. B. Gewährleistung von Antwortzeit bei Supportanfragen). Die Typen reichen von elektromagnetisch zu haptisch/mechanisch.

Je nach Typ bergen die Sensoren unterschiedliche Nutzen und Risiken, die wir nun näher betrachten wollen. Sie finden im Folgenden eine Beschreibung einzelner Sensorsysteme, deren technischer Vor- und Nachteile sowie eine Darstellung von Risiken für die Privatsphäre. Grundsätzlich gilt: Je mehr Informationsgehalt mit einem Sensor erfasst wird, umso größer die Potentiale seiner Anwendungsmöglichkeiten, aber umso höher auch das Risiko für die Privatsphäre (vgl. Abbildung 2).

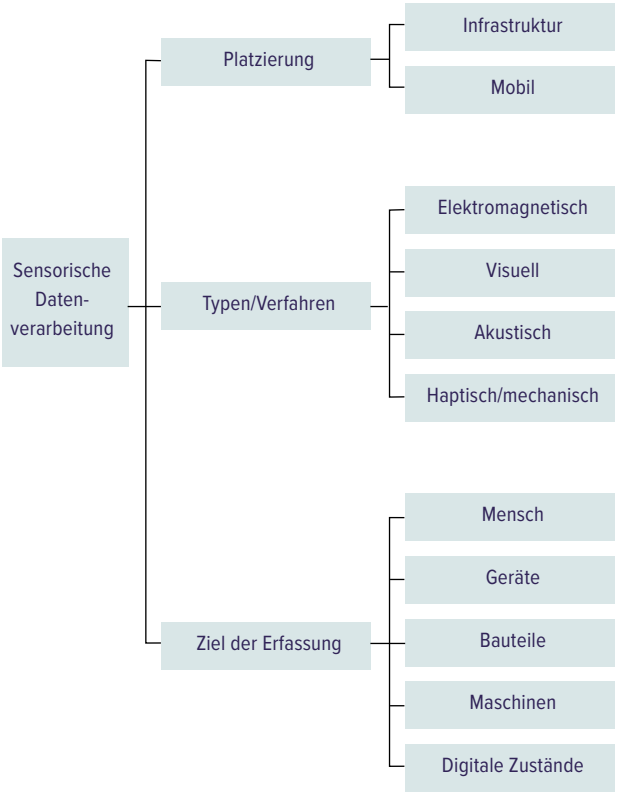


Abbildung 1: Baumdiagramm der sensorischen Datenverarbeitung (nach Huber u. a., 2021), wobei Sensoren jeweils nach Platzierung, dem Typ/Verfahren und dem Ziel der Erfassung unterteilt werden

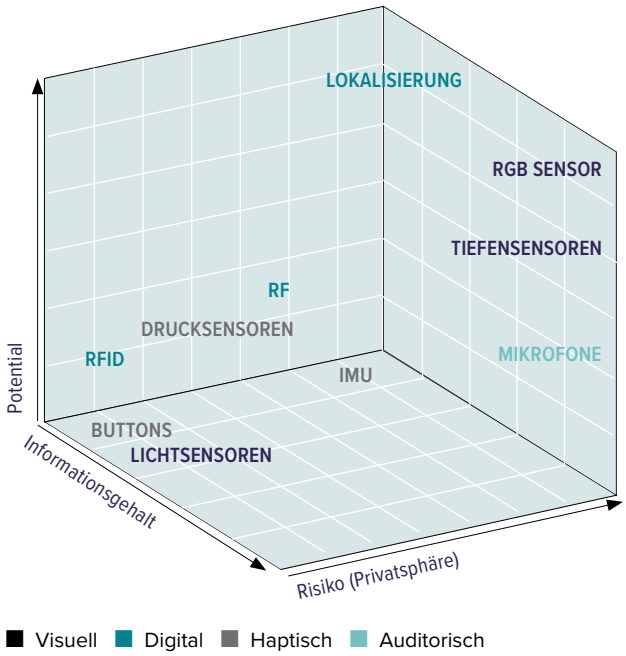


Abbildung 2: Gegenüberstellung von Potential, Risiken und Informationsgehalt von unterschiedlichen Sensortypen, farblich nach Gruppen unterteilt

Elektromagnetische Sensoren

Elektromagnetische Sensoren ermöglichen die *Lokalisation* (bspw. Ortung via GPS) und *Identifizierung* (bspw. Türöffnung via RFID) von Objekten und Personen. Elektromagnetische Sensoren messen die Ausbreitungsgeschwindigkeit von elektromagnetischen Wellen zwischen zwei Punkten. Dadurch können Informationen wie die Nähe von Objekten zueinander oder die Position von Objekten an sich abgeleitet werden.

- **GPS:** GPS (engl. *Global Positioning System*) ermöglicht die Positionsbestimmung mittels Satellit. Technische Risiken sind die potentiell geringe Genauigkeit der Technologie (die Auflösung kann sich im 10-Meter-Bereich befinden) und deren potentiell eingeschränkte Funktionsfähigkeit (GPS funktioniert nicht innerhalb von Gebäuden). Für die Privatsphäre von Personen besteht das Risiko der Identifizierung des Aufenthaltsorts und der -dauer.

- **RFID:** Bei der Identifizierung mittels RFID[1] (engl. *Radio-Frequency Identification*) müssen sich ein RFID-Chip und ein Lesegerät räumlich eng zueinander befinden. Der Informationsgehalt ist binär (vorhanden oder nicht). Wenn mobile Objekte mit RFID versehen werden und mehrere Lesegeräte passieren, kann dadurch Tracking realisiert werden. Dies wird beispielsweise auf Produktionsstraßen genutzt, um den Zusammenbau von Maschinen zu überwachen. Auch ist es möglich, Berechtigungen mit der ID des RFID-Chips oder direkt am Chip eines Objektes zu hinterlegen (z. B. für ein Türschloss). Die Technologie ist sehr etabliert und günstig, lässt aber je nach Anwendungsfall auch Rückschlüsse auf Personen zu.

- **Radartechnologie:** Üblicherweise verwenden auf elektromagnetischen Wellen basierende Verfahren zwei aktive Geräte. Aber auch passive Verfahren, bei denen die Veränderung eines definierten elektromagnetischen Feldes von Sender zu Empfänger gemessen wird, sind möglich (engl. *RF-Radar Sensing*). Befinden sich Objekte in dem definierten elektromagnetischen Feld, kommt es zu Signaländerungen. Mithilfe von Algorithmen könnten damit sogar Personen[2] oder Gesten von Personen[3] erkannt und verortet werden. Aufgrund der Störanfälligkeit von elektromagnetischen Signalen sind sicherheitskritische Anwendungen jedoch nicht zu empfehlen.

Visuelle Sensoren

- **1D:** Unter 1D-Systemen kann man einfache Lichtsensoren verstehen, die nur das Auftreffen von Licht einer bestimmten Wellenlänge messen. Ein klassisches Beispiel sind Infrarot-Lichtschranken, die rückmelden können, ob Licht auf den Infrarot-Sensor trifft oder nicht. Auch lassen sich mit 1D-Systemen Lokalisationslösungen realisieren, sofern der Sender mobil ist[4] und es mehrere Empfänger gibt. Dabei können über Muster (Ein/Aus-Wechsel) komplexe Informationen übermittelt werden, wie z. B. Identifikationsdaten ähnlich zu

1 Maurtua u. a., 2007.
2 Guo u. a., 2019.
3 Avrahami u. a., 2018.
4 Want u. a., 1992.

RFID. Bei entsprechend komplexen Mustern steigt der Zeitbedarf für die Übertragung oder es muss die sogenannte Lese-/Schreibfrequenz bei den Geräten erhöht werden, was fehleranfällig ist.

- **2D**: Unter 2D-Systemen verstehen wir üblicherweise monochrome oder mehrfarbige Bildsensoren. Letztere werden oft als Computerkamera bezeichnet und befinden sich beispielsweise als Webcam in den meisten Laptops. Mit 2D-Systemen aufgenommene Bilder werden häufig mittels Bildverarbeitungsalgorithmen automatisiert ausgewertet. Meist kommen klassische Algorithmen zum Einsatz. Aber auch mit Deep Learning trainierte neuronale Netze finden immer häufiger Anwendung (s. Abschnitt A, Kapitel KI-Assistenzsysteme – was ist das überhaupt?). Deep-Learning-Algorithmen können beliebige Objekte (inkl. Personen) in den aufgenommenen Bildern verorten[5]. Der Nachteil von 2D-Systemen ist, dass mit diesen Kameras vieles aufgezeichnet werden kann, was für den jeweiligen Anwendungsfall nicht unbedingt relevant ist, aber die Privatsphäre von aufgenommenen Personen bedroht. Darüber hinaus sind Kameras und die weiterverarbeitenden Algorithmen störanfällig für Lichteffekte und Verdeckung. Deshalb können sie nicht ohne Weiteres für sicherheitskritische Anwendungen eingesetzt werden.

- **3D:** Bei 3D-Sensoren wird zusätzlich zum Bild die Distanz jedes Bildpunktes relativ zum Sensor festgehalten. Da 3D-Systeme üblicherweise im Infrarotbereich arbeiten, handelt es sich jedoch um ein monochromes Bild, das Tiefeninformation beinhaltet. Das bedeutet, dass 3D-Konturen sichtbar sind, aber der optische, für den Menschen verarbeitbare Eindruck fehlt. Das Technologieökosystem rund um 3D-Sensoren ist nicht ganz so ausgereift wie das von 2D-Verfahren. Darüber hinaus sind die zur Auswertung genutzten Algorithmen komplexer und rechen- sowie transferintensiver. Gleichzeitig sind 3D-Systeme weniger störanfällig für Lichteffekte und schützen die Privatsphäre etwas besser, da sich die Auflösung der Systeme nicht für Detailerkennung eignet, aber sehr wohl, um zum Beispiel über abstrakte Skeletterkennung Bewegungsabläufe zu erfassen[6].

Akustische Sensoren

- **Mikrofone:** Mikrofone eignen sich zur akustischen Wahrnehmung, sowohl im hörbaren wie auch nicht hörbaren Bereich. Beispielsweise kann über die Verarbeitung von Maschinengeräuschen die Art der Tätigkeit bestimmt werden[7]. Auch kann gesprochene Sprache mithilfe von komplexen Computermodellen (z. B. Hidden-Markov-Modelle) und neuronalen Netzwerken analysiert werden. Analog zu 2D-Bildsensoren sind Mikrofone kostengünstig, breit verfügbar und anwendbar. Aufgrund der Möglichkeit, gerichtete Mikrofone zu verwenden, kann man auch den Aufnahmebereich einschränken, was vor allem für sehr laute Umgebungen sinnvoll ist. Ungerichtete Mikrofone ohne Rauschunterdrückung hingegen sind sehr störanfällig. Analog zu den bildgebenden Verfahren besteht auch bei Mikrofonen die Möglichkeit, Informationen mitzuhören, die nicht für ein Assistenzsystem relevant sind, was aber die Privatsphäre von Personen verletzen kann.

Haptische/mechanische Sensoren

- **Mechanisch:** Mechanische Bewegungssensoren messen die Bewegung eines Objekts, das mit dem Sensor fest verbunden ist. Ein einfaches Beispiel ist der Beschleunigungssensor, der in modernen Smartphones verbaut ist und über die Messung der Erdbeschleunigung die Anzeige des Displays entsprechend anpasst. Üblicherweise wird dafür eine inertiale Messeinheit (Inertial Measurement Unit, IMU) verwendet, die auf drei orthogonalen Achsen jeweils die *Beschleunigung*, das *magnetische Feld* und die *Drehrate* misst. Dadurch können beispielsweise mit Sensorobjekten ausgeführte Aktivitäten erkannt werden, wenn die Daten z. B. mit einem neuronalen Netz[8] verarbeitet werden. Bei IMUs kann nur der magnetische Feldsensor direkt gestört werden, zum Beispiel durch von Elektro-

5 Huang u. a., 2017.
6 Rude, Adams, und Beling, 2015.
7 Cheng u. a., 2014.
8 Tao u. a., 2018.

nik erzeugte Magnetfelder. Ansonsten sind diese Sensoren stabil und lassen nur durch sehr gezielte Datenverarbeitung Rückschluss auf die Identität von Nutzenden zu (zum Beispiel mittels Analyse der Gehbewegungen).

- **Haptisch:** Mit haptischen Sensoren lässt sich der physische Kontakt zwischen Objekten klassifizieren, was üblicherweise über Druck geschieht. Die Informationsdichte reicht von einfachem „Kontakt/kein Kontakt", unterschiedlichen Druckstärken bis hin zu Druckmustern. Damit kann man die Veränderung von beweglichen Objekten über die Zeit beobachten und beispielsweise über Bodenplatten registrieren, wo und wie lange sich Menschen in einem Bereich aufhalten[9]. Auch kann beispielsweise erfasst werden, ob Objekte aufgenommen oder losgelassen werden[10]. Einfache „Kontakt/Nicht-Kontakt"-Anwendungen sind gewöhnlich äußerst zuverlässig. Die größte Einschränkung dieser Sensorik ist die geringe Auflösung von drucksensitiven Matten, was jedoch gleichzeitig schonend für die Privatsphäre ist.

SIE HABEN NOCH NICHT GENUG?

In diesem Artikel der Zeitschrift „entwickler.de" wird erklärt, welche Schritte gesetzt werden können, um mithilfe von körpergetragenen Bewegungssensoren und geeigneten maschinellen Lernverfahren erkennen zu können, welche Aktivitäten der Nutzende gerade ausführt.

- *Mit Sensordaten und Maschinellem Lernen Bewegungen erkennen*[11]

ANSPRECHPERSON

DI Michael Haslgrübler, Pro²Future
Experte für sensorische Datenverarbeitung bei Assistenzsystemen
michael.haslgruebler@pro2future.at

LITERATURVERZEICHNIS

Avrahami, Daniel, Mitesh Patel, Yusuke Yamaura, und Sven Kratz. 2018. Below the Surface: Unobtrusive Activity Recognition for Work Surfaces Using RF-radar Sensing. In *23rd International Conference on Intelligent User Interfaces,* 439–51. IUI '18. *https://doi.org/10.1145/3172944.3172962.*

Cheng, Chen-Yang, Tsung-Yin Ou, Tzu-Li Chen, und Yin-Yann Chen. 2014. Transferring Cognitive Apprenticeship to Manufacturing Process Knowledge Management System: A Case Study of Small- and Medium-Sized Coating Industry. VINE 44, Nr. 3: 420–44. *https://doi.org/10.1108/VINE-03-2013-0011.*

Guo, Hangqing, Nan Zhang, Wenjun Shi, ALI-AlQarni Saeed, Shaoen Wu, und Honggang Wang. 2019. Real-Time Indoor 3D Human Imaging Based on MIMO Radar Sensing. In *2019 IEEE International Conference on Multimedia and Expo (ICME),* 1408–13.

Hein, Albert, Frank Krüger, und Thomas Kriste. 2017. Mit Sensordaten und Maschinellem Lernen Bewegungen erkennen. entwickler.de. *https://entwickler.de/iot/mit-sensordaten-und-maschinellem-lernen-bewegungen-erkennen.*

Huang, Jonathan, Vivek Rathod, Chen Sun, Menglong Zhu, Anoop Korattikara, Alireza Fathi, Ian Fischer, u. a. 2017. Speed/Accuracy Trade-Offs for Modern Convolutional Object Detectors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7310–11.

9 Riener, 2010.

10 Malaisé u. a., 2018.

11 Hein, Krüger und Kriste, 2017.

Huber, Jaroslava, Michael Haslgrübler, Martin Schobesberger, Alois Ferscha, Viktorijo Malisa, und Georg Effenberger. 2021. Addressing Worker Safety and Accident Prevention with AI. In *11th International Conference on the Internet of Things*, 150–57. *https://doi.org/10.1145/3494322.3494342*.

Malaisé, Adrien, Pauline Maurice, Francis Colas, François Charpillet, und Serena Ivaldi. 2018. Activity Recognition with Multiple Wearable Sensors for Industrial Applications. In *ACHI 2018 – Eleventh International Conference on Advances in Computer-Human Interactions*. *https://hal.archives-ouvertes.fr/hal-01701996*.

Maurtua, I., P. T. Kirisci, T. Stiefmeier, M. L. Sbodio, und H. Witt. 2007. A Wearable Computing Prototype for Supporting Training Activities in Automotive Production. In 4*th International Forum on Applied Wearable Computing* 2007, 1–12.

Riener, Andreas. 2010. S*ensor-Actuator Supported Implicit Interaction in Driver Assistance Systems.* Wiesbaden: Vieweg+-Teubner.

Rude, Don J., Stephen Adams, und Peter A. Beling. 2015. A Benchmark Dataset for Depth Sensor Based Activity Recognition in a Manufacturing Process. *IFAC-PapersOnLine* 48, Nr. 3: 668–74. *https://doi.org/10.1016/j.ifacol.2015.06.159.*

Tao, Wenjin, Ze-Hao Lai, Ming C. Leu, und Zhaozheng Yin. 2018. Worker Activity Recognition in Smart Manufacturing Using IMU and sEMG Signals with Convolutional Neural Networks. Procedia Manufacturing 26: 1159–66. *https://doi.org/10.1016/j.promfg.2018.07.152*.

Want, Roy, Andy Hopper, Veronica Falcao, und Jonathan Gibbons. 1992. The Active Badge Location System. *ACM Transactions on Information Systems* (TOIS) 10, Nr. 1: 91–102.

KI aus systemischer Sicht und ihre Risiken

Um die systemischen Risiken von KI besser veranschaulichen zu können, wenden wir uns dem Grundaufbau eines KI-Systems zu.

> Heruntergebrochen ist eine KI eine Sammlung technologischer Komponenten, die Daten sammelt, verarbeitet und anhand der Ergebnisse etwas bewirkt. KI-Lösungen können Regeln anwenden, über einen Zeitraum hinweg lernen und sich der Umwelt anpassen.

Schlüsselkomponenten

KI-Systeme bestehen aus drei wesentlichen Schlüsselkomponenten:

- den eingehenden Daten, d. h. dem sogenannten Input
- der Software, d. h. dem Algorithmus
- der Ausgangsentscheidung, dem sogenannten Output[1]

Komponente 1: Input

Die erste Komponente eines KI-Systems besteht aus den zugeführten Daten. Die KI wird mit diesen Daten „gefüttert“, es ist das Material, mit dem sie arbeitet. KI-Systeme werden immer besser darin, neben strukturierten Daten (z. B. Transaktionen) unstrukturierte Inputs zu verarbeiten (z. B. Bilder und Sprache). Dabei können sowohl historische Daten als auch Daten aus der Echtzeit (d. h., aufgenommen mit Sensoren) verarbeitet werden. Ein großes Risiko bei der Erfassung von Daten ist die Aufzeichnung von Daten, die für den Zweck der KI-Anwendung gar nicht zwingend benötigt werden. Das ermöglicht den Missbrauch der Daten für andere Zwecke (bspw. Überwachung). Außerdem kann es zu Fällen kommen, in denen die KI eine Entscheidung trifft, die zwar mit den vorhandenen Trainingsdaten fehlerfrei funktioniert, in der Wirklichkeit jedoch keinen Sinn macht. Ein Beispiel für eine solche widersprüchliche Situation wäre, wenn eine KI mit einem Kamera-Stream (Audio und Video) trainiert wird, die Erkennung jedoch nur auf den Videodaten basieren soll. Beim Aufzeichnen wurde jedoch gesprochen, um die Daten später annotieren zu können. In diesem Fall ist es wahrscheinlich, dass die KI im Training gute Ergebnisse liefert, weil sie die Annotation der Audiospur gelernt hat, diese fehlt jedoch im Echtbetrieb, weshalb die KI nicht funktioniert.

Komponente 2: Algorithmus

Der Algorithmus ist die zweite Komponente eines KI-Systems: dabei handelt es sich um das Berechnungsverfahren, das die Daten verarbeitet. Ein Algorithmus ähnelt vom Prinzip her einem Kochrezept: eine Reihe von Instruktionen, die Schritt für Schritt durchgeführt werden, um ein Ziel zu erreichen. Mehrere Typen von Algorithmen können unterschieden werden:

Wenn Expert*innen dem Computer Trainingsdaten mit den korrekten Eingabe- (Inputs) und Ausgabedaten (Outputs) geben, damit der Algorithmus ein Muster erkennt und Regeln entwickelt, handelt es sich um *beaufsichtigtes Lernen*. Ein Beispiel dafür wäre die Identifizierung von kleinen Variationen in Zellen in MRT[2]-Bildern und die damit verbundene frühzeitige Erkennung von Krebs.

Wenn der Computer ein Datenset zum Trainieren bekommt, das zwar Eingabedaten, aber keine Ausgabeinformation enthält, spricht man von *unbeaufsichtigtem Lernen*. Der Algorithmus muss den besten Weg finden, die Datenpunkte zu gruppieren und die Relationen

1 Canhoto und Clear, 2020.

2 Die Magnetresonanztomographie (MRT) ist ein bildgebendes Verfahren und dient der medizinischen Diagnostik von Körpergewebe und Organen.

zu erkennen (z. B. welche Gegenstände miteinander gekauft werden).

Von *bestärkendem Lernen* spricht man, wenn der Algorithmus ein Datenset zum Trainieren und ein Ziel bekommt und die beste Kombination aus Aktionen finden muss, um an das Ziel zu kommen. Der Algorithmus wird dabei für bestimmte Verfahrensweisen „belohnt" und muss diese Belohnungen maximieren.

Alle Formen der Lernalgorithmen sind mit gewissen Risiken verbunden, wobei ein wesentlicher Aspekt die Berechnung und Gewichtung des statistischen Fehlers ist. Das wird üblicherweise über eine eigene spezielle Funktion modelliert (engl. Loss Function). Damit das System auch korrekt das Richtige tun kann, muss der Abstand zwischen dem akzeptierten und dem tatsächlichen Fehler richtig modelliert werden.

Komponente 3: Output

Das Ergebnis eines Maschinenlern-Prozesses kann im einfachsten Fall z. B. ein Zählpunkt (engl. Score) sein, der keine Funktion hat, bis Analyst*innen ihn für weitere Funktionen verwenden. Andere Systeme können basierend auf den Ergebnissen ihrer Analyse Entscheidungen treffen. Ein Beispiel dafür wäre ein selbstfahrendes Auto, welches fahren, lenken oder bremsen kann, ohne dass ein Mensch intervenieren müsste.

Risiken bestehen unter anderem in der Auswirkung des richtigen oder falschen Ergebnisses auf den Menschen und die Umwelt. Extrembeispiel sind Tode von Nutzer*innen autonomer Fahrzeuge.[3]

Charakteristische KI-Eigenschaften

Genauso wie eine KI aus den Eingabedaten, dem Algorithmus und den Ausgabedaten besteht – ihren Schlüsselkomponenten –, zeichnet sie sich neben zahlreichen anderen Eigenschaften vor allem durch drei besondere, typische Charakteristiken aus. Diese sind die Konnektivität, die kognitive Fähigkeit und die Unwahrnehmbarkeit[4] von KI. Paradoxerweise gehen ausgerechnet von diesen drei Charakteristiken Gefahren für die Eingabedaten, die Software und die Ausgabedaten aus. In Tabelle 1 werden einige Beispiele dafür angeführt.

Eigenschaft 1: Konnektivität

Die Konnektivität ist eine nicht wegzudenkende Eigenschaft von KI, bei der es sich um die Vernetzung der Komponenten der KI handelt. Z. B., wenn ein autonomes Fahrzeug, das sich im Netzwerk mit anderen autonomen Fahrzeugen befindet, einen Fehler macht und diesen korrigiert, lernt es und kommuniziert das neu erworbene Wissen zeitnah mit anderen Fahrzeugen, die sich im Netzwerk befinden. Die Konnektivität setzt voraus, dass die einzelnen Einheiten miteinander kompatibel sind (z. B. Daten im System sind gleich formatiert). Da die Vernetzung in beide Richtungen funktioniert, besteht die Gefahr, dass auch falsche Ausgabedaten schnell und großflächig verbreitet werden. Dabei können verfälschte Daten leicht von außen in das System geraten und ihr Ursprung ist schwer nachvollziehbar.

Eigenschaft 2: Kognitive Fähigkeit

Während ein gewöhnlicher Algorithmus zu einem Ergebnis kommt, indem er einen Prozess an den Eingangsdaten anwendet, kann die KI im Vergleich dazu mehr: sie besitzt eine kognitive Fähigkeit. Sie bildet aus den vorhandenen Eingangsdaten ein Modell, wendet dieses Modell auch an neuen, davor noch unbekannten Daten an und kommt so zu neuen Ergebnissen und Lösungen. Da Ergebnisse von Algorithmen zuerst implementiert werden müssen, um beurteilt werden zu können, liegt die Herausforderung darin, rechtzeitig zu reagieren, wenn ein Algorithmus nicht richtig funktioniert (z. B., wenn ein Algorithmus nicht nachvollziehbare Ausgangswerte produziert).

Eigenschaft 3: Unwahrnehmbarkeit

Die dritte charakteristische Eigenschaft von KI ist ihre Unwahrnehmbarkeit. Meistens bemerken Nutzer*innen nicht viel von einer KI-Anwendung, da sie im Hintergrund läuft. Dabei kann es passieren, dass die Prozesse wenig bis gar nicht kontrolliert oder hinterfragt werden. Wenn die Grenzen, wann welche Daten gesammelt werden, zu verschwimmen beginnen, sind sowohl ethische Probleme als auch die Beeinträchtigung der Privatsphäre die Folge. Sollten die Benutzenden Informationen,

3 Z. B. Yadron und Tynan, 30. Juni 2016.

4 Hoppe, Oktober 2015.

die sie auf einem Bildschirm sehen, automatisch mehr Glaubwürdigkeit schenken (z. B. die personifizierten Informationen, die durch die Algorithmen in den sozialen Netzwerken generiert werden), führt das zu einer falschen Art des Vertrauens in KI.

Tabelle 1 zeigt einige Beispiele, wie die drei Komponenten der KI (eingehende Daten, Software und ausgehende Daten) durch die KI-typischen Eigenschaften (Konnektivität, kognitive Fähigkeit, Unwahrnehmbarkeit) beeinflusst werden könnten.

Komponente	Konnektivität	Kognitive Fähigkeit	Unwahrnehmbarkeit
Eingehende Daten	Verwendung von Daten aus Quellen, die nicht kontrolliert werden können	Datenset kann unpassend für prognostizierendes Profiling sein	Nutzer*in ist nicht informiert, welche Daten gesammelt und verwendet werden
Software	Entscheidung zwischen Standardisierung und Kompatibilität/Fit und Flexibilität	Formeln vereinfachen komplexe Phänomene	Modell ist nicht zugänglich; keine Möglichkeit für Optimierung u. Evaluierung
Output	Fehler und falsche Outputs werden rasch verbreitet	Schwierigkeiten, die Qualität von Vorhersagen zu verifizieren; nicht nachvollziehbare Outputs	Nicht möglich, Outputs zu hinterfragen, sie zu kontrollieren oder zu korrigieren

Tabelle 1: Drei charakteristische Eigenschaften von KI (Konnektivität, kognitive Fähigkeit, Unwahrnehmbarkeit) und drei wichtige Schlüsselkomponenten (eingehende Daten, Software, Output) sowie die damit in direkter Verbindung stehenden Risiken. Tabelle nach Canhoto und Clear, 2020.[5]

5 Canhoto und Clear, 2020.

ANSPRECHPERSON

Jaroslava Huber, MSc, Pro2Future
Expertin für Interaktion mit KI-Systemen
jaroslava.huber@pro2future.at

LITERATURVERZEICHNIS

Canhoto, Ana Isabel, und Fintan Clear. 2020. Artificial Intelligence and Machine Learning as Business Tools: A Framework for Diagnosing Value Destruction Potential. *Business Horizons* 63, Nr. 2. *https://doi.org/10.1016/j.bushor.2019.11.003.*

Hoppe, Thomas. Oktober 2015. Prinzip der Unwahrnehmbarkeit. *Informatik-Spektrum* 38, Nr. 5. *https://doi.org/10.1007/s00287-014-0863-8*

Yadron, Danny, und Dan Tynan. 30. Juni 2016. Tesla Driver Dies in First Fatal Crash While Using Autopilot Mode. *The Guardian*.

Autor*innenverzeichnis

JULIAN ANSLINGER

© Otmar Winterleitner

Mag. Dr. Julian Anslinger arbeitet am *Interdisziplinären Forschungszentrum für Technik, Arbeit und Kultur,* wo er für das Themenfeld Digitalisierung und Künstliche Intelligenz zuständig ist. Disziplinär verortet ist Anslinger in der Sozialpsychologie und Geschlechterforschung. Seine Promotion im Promotionsstudiengang „Intelligente Systeme“ verfasste er am Exzellenzcluster Kognitive Interaktionstechnologie (CITEC) der Universität Bielefeld. Ziel seiner Forschungsarbeiten ist es, soziale und Geschlechtergerechtigkeit in Organisationen, Wissenschaft und Technik zu etablieren. Die Werkzeuge dazu vermittelt er in Lehrveranstaltungen an Studierende der Universität Klagenfurt.

JAROSLAVA HUBER

© Foto Kücher

Jaroslava Huber, MSc absolvierte ihr Studium für Kognitive Neurowissenschaften an der Eötvös Loránd Universität in Budapest. Sie ist wissenschaftliche Mitarbeiterin bei der Pro2Future GmbH, AREA 1, Perception and Aware Systems. Die Area befasst sich mit der Realisierung technischer Systeme mit kognitiven Fähigkeiten in Form von industriellen, sensorbasierten Assistenzsystemen zur optimalen Unterstützung von Arbeiter*innen. Im Rahmen dessen setzt sich Huber speziell mit der kollaborativen Zusammenarbeit zwischen Mensch und Maschine auseinander und ist für den Bereich Interaktion und Kognition zuständig.

MICHAEL HASLGRÜBLER

© Pro2Future

DI Michael Haslgrübler ist Area-Manager am COMET-K1-Zentrum Pro2Future und bearbeitet das Forschungsfeld „Perception and Aware Systems“ gemeinsam mit seinem Team. Nach einem Studium der Informatik an der Johannes Kepler Universität und Stationen in der Wirtschaft ist er im Begriff, sein Doktoratsstudium abzuschließen. Der Schwerpunkt seiner Arbeit ist Digitalisierung von Arbeitsprozessen in der Industrie mit Hilfe von Sensorik, KI-Algorithmen und Aktuatoren, mit der Zielsetzung, computergestützte Systeme als kognitiv zu erleben, die dem Menschen bei seinen Prozessen bestmöglich zur Seite stehen.

ANITA THALER

© IFZ

Mag.a Mag.a Dr.in Anita Thaler ist wissenschaftliche Mitarbeiterin am *Interdisziplinären Forschungszentrum für Technik, Arbeit und Kultur* und leitet den dortigen Forschungsbereich Gender, Science and Technology. Sie studierte ursprünglich Psychologie (zertifizierte Arbeitspsychologin), dann berufsbegleitend Erziehungs- und Bildungswissenschaften und Frauen- und Geschlechterforschung an der Karl-Franzens-Universität Graz und der Alpen-Adria-Universität Klagenfurt. In ihrer Forschung untersucht sie Wechselwirkungen von Wissenschaft, Technologie und Gesellschaft und fokussiert dabei auf Transitions- und Bildungsprozesse in Richtung Nachhaltigkeit und sozialer Gerechtigkeit.

SUSANNE HASLINGER

© Robert Wittek / PRO-GE

Mag.a Susanne Haslinger ist Juristin und in der Produktionsgewerkschaft PRO-GE für den Bereich Sozialpolitik/Grundlagenarbeit zuständig. Sie hat langjährige Erfahrung in der Beratung und Vertretung von Arbeitnehmer*innen und Betriebsrät*innen und ist Mitautorin mehrerer einschlägiger Werke zum Betriebsverfassungsrecht und betrieblichen Datenschutz.

Datenschutzerklärung zum Übernehmen

Für die Einbindung von Arbeitnehmenden und weiteren Interessenvertretungen mittels Interviews, Fragebögen, Workshops oder Ähnlichem ist eine Datenschutz- und Einwilligungserklärung vonnöten, sobald Personen Informationen abgeben, die sie leicht identifizierbar machen. Im Folgenden finden Sie eine Vorlage, die Sie für Ihre Zwecke anpassen können.

Datenschutzerklärung[1]

Projektname:

Beschreibung Projekt:

Konkretes Vorhaben:
(z. B. Durchführung eines Workshops)

Durchführende Institution:

Weitere projektbeteiligte Institutionen: (relevant, wenn personenbezogene Daten an Projektpartner*innen weitergegeben werden sollen)

Interviewer*in/Workshopleitung/etc.:

Interviewdatum/Workshopdatum/etc.:

Wir verarbeiten Ihre personenbezogenen Daten, die unter folgende Datenkategorien fallen:

- Name/Firma,
- Beruf/Berufsbezeichnung
- Kontaktdaten (Telefonnummer, Telefaxnummer, E-Mail-Adresse, etc.)
- ...

sowie besondere Kategorien personenbezogener Daten (gem. Art. 9 DSGVO), nämlich Angaben über:

- Politische Meinung
- Religiöse oder philosophische Überzeugung
- Gewerkschaftszugehörigkeit
- Gesundheit
- Sexualleben
- Biometrische Informationen zur eindeutigen Identifikation

als auch Aufnahmen, nämlich:

- Videoaufzeichnungen
- Fotos
- Screenshots von Webcambildern
- Tonaufnahmen
- ...

Sie stellen uns Daten über sich freiwillig zur Verfügung und wir verarbeiten diese Daten auf Grundlage Ihrer **Einwilligung** zu folgenden Zwecken:

- Kontaktaufnahme
- Workshoporganisation
- Veröffentlichungen auf Social Media
- Wissenschaftliche Zwecke
- Forschungsberichte an den Fördergeber
- Publikationen
- ...

1 Datenschutzerklärung verbleibt bei Empfänger*in.

Ggf. Spezifikation:
Es erfolgt keine Veröffentlichung von personenbezogenen Daten.

oder

Die Ergebnisse werden ausschließlich in anonymisierter Form dargestellt. Das bedeutet: Niemand kann aus den Ergebnissen erkennen, von welcher Person die Angaben gemacht worden sind.

oder

Teile Ihrer Aussagen werden eventuell zitiert in Publikationen, Berichten, im Web oder anderen Ergebnisdarstellungen. Dies geschieht in anonymisierter Form, d. h., ohne Angabe Ihres Namens oder Ihrer Adresse.

Sie können Ihre Einwilligung jederzeit widerrufen. Ein **Widerruf** hat zur Folge, dass wir Ihre Daten ab diesem Zeitpunkt zu oben genannten Zwecken nicht mehr verarbeiten. **Für einen Widerruf wenden Sie sich bitte an:** [...] (*E-Mail-Adresse einsetzen*).

Wir **speichern** Ihre Daten bis zum [...] (*Frist bzw. Kriterien für die Löschung ergänzen*).

Für diese Datenverarbeitung ziehen wir **Auftragsverarbeiter** heran.
Wir geben Ihre Daten an folgende **Empfänger*innen** bzw. Empfängerkategorien weiter [...]
(Empfänger samt Begründung für die Weitergabe ergänzen).
Ihre Daten werden zumindest zum Teil auch **außerhalb der EU bzw. des EWR** verarbeitet, und zwar in [...] (*Staat(en) ergänzen*). Das angemessene Schutzniveau ergibt sich aus einem Angemessenheitsbeschluss der Europäischen Kommission nach Art. 45 DSGVO.

Unsere*n Datenschutzbeauftragte*n erreichen Sie unter: (*E-Mail-Adresse einsetzen*).

Rechtsbehelfsbelehrung

Ihnen stehen grundsätzlich die Rechte auf Auskunft, Berichtigung, Löschung, Einschränkung, Datenübertragbarkeit und Widerspruch zu. Dafür wenden Sie sich an uns. Wenn Sie glauben, dass die Verarbeitung Ihrer Daten gegen das Datenschutzrecht verstößt oder Ihre datenschutzrechtlichen Ansprüche sonst in einer Weise verletzt worden sind, können Sie sich bei der Aufsichtsbehörde beschweren. In Österreich ist die *Datenschutzbehörde* zuständig.

Einwilligung[2]

Ich erkläre hiermit, dass ich mit der Verarbeitung meiner persönlichen Daten, nämlich ... (*exakte Aufzählung der Datenarten, z. B. Name, Adresse etc.*) zum Zweck der ... (*genaue Zweckangabe*) bei ... (*Organisation*) und der Weitergabe an das folgende Unternehmen ... (*genaue Angabe des Dritten, z. B. XY-GmbH in ... PLZ/Adresse, ...*) zum Zweck der ... (*genaue Zweckangabe*) einverstanden bin. Ich hatte Gelegenheit, Fragen zu stellen. Diese wurden vollständig und umfassend beantwortet.

_____

Vorname, Nachname in Druckschrift

_____

Ort, Datum, Unterschrift

2 In doppelter Ausführung.

GEFÖRDERT DURCH
Digifonds
AK WIEN