

The background of the book cover is a photograph of a person's hands typing on a laptop. Overlaid on this image are several digital and AI-themed graphics: a semi-transparent red rectangle in the upper left containing the series title, a white box with subtitle text, a glowing robot head icon with a speech bubble saying 'Can i help you?', and various faint digital code and network patterns.

Band 14

**UNIVERSITY
SOCIETY
INDUSTRY**

Beiträge zum lebensbegleitenden
Lernen und Wissenstransfer

Agnes Raschauer, Nino Tomaschek (Hrsg.)

KI trifft Arbeit

Wie künstliche Intelligenz
die Arbeitswelt verändert

WAXMANN

University – Society – Industry

Beiträge zum lebensbegleitenden
Lernen und Wissenstransfer

herausgegeben vom
Postgraduate Center der Universität Wien

Band 14

Agnes Raschauer, Nino Tomaschek (Hrsg.)

KI trifft Arbeit

Wie künstliche Intelligenz
die Arbeitswelt verändert



Waxmann 2025
Münster · New York

Bibliografische Informationen der Deutschen Nationalbibliothek

Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über <http://dnb.dnb.de> abrufbar.

University – Society – Industry, Band 14

ISSN 2194-9530

Print-ISBN 978-3-8188-0053-6

E-Book-ISBN 978-3-8188-5053-1

<https://doi.org/10.31244/9783818850531>

Waxmann Verlag GmbH, 2025

Steinfurter Straße 555, 48159 Münster

www.waxmann.com

info@waxmann.com

Umschlaggestaltung: Anne Breitenbach, Münster

Umschlagbild: © khunkornStudio | Adobe Stock

Satz: Roger Stoddart, Münster

Dieses E-Book ist unter der Lizenz CC BY-NC-SA 4.0 open access verfügbar.



Diese Lizenz gilt nur für das Originalmaterial. Alle gekennzeichneten Fremdinhalte (z.B. Abbildungen, Fotos, Zitate etc.) sind von der CC-Lizenz ausgenommen und für deren Wiederverwendung ist es ggf. erforderlich, weitere Nutzungsgenehmigungen beim jeweiligen Rechteinhaber einzuholen.

Inhalt

Agnes Raschauer & Nino Tomaschek

Wegbereiter der Arbeitswelt von morgen?

KI-Anwendungen zwischen Hype und Schreckgespenst 7

Teil I: Transformativ oder überschätzt? Auswirkungen auf Arbeitsmarkt, Berufe und Qualifikationen

Wenke Klingbeil-Döring

Künstliche Intelligenz und der Arbeitsmarkt

Einflussgrößen, Entwicklungen und ethische Überlegungen 17

*Alexander Schmölz, Pia-Zoe Hahne, Sebastian Klocker &
Wolfgang Katzian*

KI in der Arbeitswelt

Der Mensch im Spannungsfeld von Entfremdung und
digitalem Humanismus 29

Julia Bock-Schappelwein & Sandra Bilek-Steindl

Berufe im Wandel

Das Zusammenspiel von künstlicher Intelligenz und
kognitiven Tätigkeiten 41

Jan-Ove Wiesner, Philipp Besinger, Benedikt Fuchs & Sebastian Schlund

Künstliche Intelligenz in der österreichischen Arbeitswelt:

Status quo und Perspektiven 53

*Viktoria A. Krenn, Torsten Ullrich, Elena Wiegelmann, Tarim Stüwe &
Eva Eggeling*

Auswirkungen von künstlicher Intelligenz auf den Arbeitsmarkt –

Schwerpunkt Kärnten 73

Teil II: Vertrauen, Transparenz und Partizipation: Herausforderungen für den KI-Einsatz

Julia Eisner

KI in der Unternehmenskommunikation: vom Tool zum Teammitglied? 99

Erich Prem

Look who's talking

Sprecheridentität und Diskurskontrolle in der Arbeit mit generativer KI 111

Lejla Rizvanovik

Algorithmische Diskriminierung am Arbeitsplatz

Fairness, Transparenz, Datenschutz und kontrafaktische

Erklärungen in KI-gestützten ADM-Systemen 121

Teil III: Impulse zur Gestaltung in konkreten Arbeitsfeldern

Titus Udrea & Helene Baumgartner

Strategische Vorausschau und Technikfolgenabschätzung

Cybersicherheit in der KI-gestützten Arbeitswelt proaktiv gestalten 139

Stefanie Spielberger

Veränderte Spielregeln

Wie künstliche Intelligenz die Anforderungen an Kompetenzen,

Vertrauen und (Zusammen-)Arbeit in der Arbeitswelt verändert 153

Barbara Geyer & Rita Stampfl

Künstliche Intelligenz an Hochschulen

Eine systemische Analyse der KI-Integration in Hochschulen

mit dem Viable System Model (VSM) 169

Samuel Eke & Nils Hofmann

Autonomy Against the Machine

Clinical Decision-Making and the Politics of AI 183

Petra Schön

PLAInLaw4U. Rechtlicher Klartext im (beruflichen) Alltag

Mit KI auf dem Weg zu mehr Verständlichkeit, Teilhabe

und Rechtsstaatlichkeit 199

Josef Sawetz

Befreit KI die Welt von Arbeit? 213

Autorinnen und Autoren 227

Herausgeberin und Herausgeber 231

Wegbereiter der Arbeitswelt von morgen?

KI-Anwendungen zwischen Hype und Schreckgespenst

Agnes Raschauer & Nino Tomaschek

Künstliche Intelligenz ist in der Arbeitswelt angekommen. Dieser Befund zieht sich durch die Auseinandersetzung mit dem Einsatz von KI-Technologie im Arbeitsleben. Er verweist einerseits auf die vielfältigen bereits existierenden Einsatzmöglichkeiten in unterschiedlichsten Branchen, Arbeitsbereichen und Tätigkeitsfeldern und ist andererseits mit der schieren Geschwindigkeit der technologischen Entwicklungen verknüpft, mit der immer neue oder weiterentwickelte Anwendungen eingeführt werden. So wird von KI-Revolution gesprochen, ein „Paradigmenwechsel“ (Eisner in diesem Band, S. 99) ausgemacht und die spezifische „disruptive“ Qualität der Veränderungen für die Arbeitswelt (Bock-Schappelwein & Bilek-Steindl in diesem Band, S. 41) hervorgehoben.

Während der griffige Name „künstliche Intelligenz“ sicher zur diskursiven Strahlkraft beiträgt, werden Expert*innen nicht müde, auf seine Unschärfe hinzuweisen. So bezeichnet KI keine Technologie an sich, sondern umfasst ein Spektrum von (informatischen) Verfahren im Kontext von maschinellem Lernen – etwa Large-Language-Modelle oder Deep Learning und generative KI-Anwendungen. Ganz generell gilt als Künstliche Intelligenz „die Fähigkeit einer Maschine, menschliche Fähigkeiten wie logisches Denken, Lernen, Planen und Kreativität zu imitieren“ (Europäisches Parlament, 2020). Gerade aber diese Analogie zu menschlicher Intelligenz wird kritisch gesehen, da mit KI primär statistische Modelle einhergehen. Ergebnisse stehen nicht für „wahre“ oder korrekte Antworten, sondern wahrscheinliche Anschlussoptionen.

Für den Einsatz von KI-Systemen im Arbeitsleben ist die Möglichkeit, extrem große Datenmengen zu verarbeiten und daraus Muster oder Vorhersagen zu generieren, zentral. Durch die Auslagerung spezifischer Tätigkeiten an KI-Anwendungen entstehen Möglichkeiten zur Optimierung und Effizienzsteigerung von Arbeitsprozessen. Damit können eine erhöhte Produktqualität oder die Verbesserung von Arbeitsschutz und -sicherheit verbunden sein. Problematisch wird die fehlende Transparenz und Nachvollziehbarkeit von Entscheidungen beziehungsweise Ergebnissen eingeschätzt. Immer wieder kann gezeigt werden, dass KI-Anwendungen Vorurteile reproduzieren oder ihnen verzerrende Effekte eingeschrieben sind. Des Weiteren produzieren KI-Systeme Fehler. Textbasierten KI-Tools werden z.B. sogenannte Halluzinationen nachgewiesen, fehlerhafte Informationen, die als korrekte Ergebnisse dargestellt werden.

Auf gesellschaftlicher Ebene stellen sich mit dem Einsatz von künstlicher Intelligenz grundlegende Fragen zur Sicherung von Grund- und Menschenrechten etwa hinsichtlich des Schutzes persönlicher Daten, zur Cybersicherheit oder der Möglichkeit demokratischer Gestaltung von Technologie, die mit Profitinteressen entwickelt, vertrieben und eingesetzt wird. Zudem eröffnet die steigende Anwendung Herausforderungen aus ökologischer Perspektive, da der Betrieb von KI sehr energieintensiv ist, sowie hinsichtlich globaler Verteilungsungerechtigkeit, weil die Arbeitszusam-

menhänge in Entwicklung, Vertrieb und Instandhaltung ausbeuterische Verhältnisse im globalen Süden stützen können (Wangari & Vaidyanathan, 2025).

Wie in Teil I dieser Publikation gezeigt wird, gibt es bislang primär Prognosen zu den Auswirkungen der KI-Integration auf den (österreichischen) Arbeitsmarkt. Die konstatierten Veränderungen sind jedoch umfassend. So wird spekuliert, dass ganze Branchen verschwinden, Tätigkeitsfelder von KI übernommen und Berufsprofile transformiert werden könnten. Wie sich der Wandel jedoch konkret und empirisch begründet gestaltet, muss erst im Detail gezeigt werden. Dies trifft auch auf den Umgang mit KI in Unternehmen zu. Während die 2024 in Kraft getretene KI-Verordnung der Europäischen Union Unternehmen in die Pflicht nimmt, ihre Mitarbeiter*innen für den Einsatz von KI – wenn dieser arbeitsrelevant ist – zu schulen und mit Chancen und Risiken im Umgang vertraut zu machen (Unternehmensserviceportal, o. D.), zeigt eine aktuelle Studie der Statistik Austria, dass Wissen über KI und KI-Kompetenz noch deutlich ausbaufähig sind (Statistik Austria, 2025).

Im 14. Band der Publikationsreihe „University – Society – Industry: Beiträge zum lebensbegleitenden Lernen und Wissenstransfer“ fokussieren wir auf die Auswirkungen des Einsatzes von KI-Anwendungen im Arbeitsleben. Anhand empirischer Daten und konkreter Arbeitsbereiche diskutieren wir, welche Chancen und Risiken die Integration von KI in Arbeitsprozesse birgt. Welche Veränderungen für Berufsfelder und Branchen sind tatsächlich auszumachen? Aber auch: Welche sozialen, ethischen und rechtlichen Fragestellungen gehen mit den Transformationen einher? Schließlich stellen sich mit dem Einsatz von KI nicht nur technologisch-operative Fragen. Vielmehr müssen mit der „Abgabe“ von Tätigkeiten an maschinelle Akteur*innen auch Überlegungen zu Autor*innenschaft, Partizipation und Verantwortung angestellt werden.

Wer kann oder soll in einem Arbeitsprozess für ein Produkt verantwortlich gehalten werden, wenn es durch eine KI-Anwendung erstellt wurde? Wie können Mitarbeiter*innen verantwortungsvoll Entscheidungen treffen, wenn sich die Grundlagen für diese Entscheidungen ihrer Nachvollziehbarkeit entziehen?

„KI trifft Arbeit. Wie künstliche Intelligenz die Arbeitswelt verändert“ ist die Begleitpublikation zum offenen Weiterbildungsprogramm „University Meets Industry“ (uniMind), einer am Postgraduate Center der Universität Wien verankerten Initiative, die den Austausch zwischen Universitäten, Unternehmen, gesellschaftlichen Institutionen und der Öffentlichkeit fördern möchte. In interaktiven Workshops und praxisorientierten Lectures diskutieren Wissenschaftler*innen aktuelle Forschungsfragen mit der interessierten Öffentlichkeit, Teilnehmer*innen aus Organisationen, Unternehmen, öffentlichen Institutionen und NGOs. 2025 standen die Aktivitäten unter dem Überthema „KI trifft Arbeit“. Bei den Veranstaltungen und auch durch die begleitende Publikation sollen der Austausch von heterogenen Perspektiven, die Vernetzung von Wissensräumen und das transdisziplinäre Lernen gefördert werden.

Thematisch gliedert sich der Band in drei Abschnitte: Teil I hat eine Bestandsaufnahme und Einordnung der viel zitierten Veränderungsprozesse des Arbeitsmarkts durch die voranschreitende Integration von KI zum Ziel. Dabei wird analysiert, wie sich Branchen, Berufsfelder und Anforderungsprofile transformieren und wie die ausge-machten Auswirkungen aus ethisch-sozialer Perspektive eingeschätzt werden können. In Teil II stehen Vertrauen, Transparenz und Partizipation als zentrale Momente der

aktuellen Transformationsprozesse von KI-vermittelten Arbeitswelten im Zentrum. Teil III widmet sich dem Einsatz von KI-Anwendungen in spezifischen Arbeitskontexten und diskutiert Impulse zur Gestaltung einer Integration von KI – etwa in den Bereichen Sicherheitsmanagement, Hochschulen oder Medizin.

Teil I: Transformativ oder überschätzt? Auswirkungen auf Arbeitsmarkt, Berufe und Qualifikationen

Der vorliegende Band wird von *Wenke Klingbeil-Döring* mit einem Beitrag zu dem aktuellen Transformationsgeschehen in der Arbeitswelt eröffnet, bei dem sie an den postulierten tiefgreifenden Veränderungen durch die Integration von KI ansetzt. Neben einer Darstellung, was unter Künstlicher Intelligenz verstanden werden kann, seziert sie Befunde eines umgreifenden, KI-basierten Wandels – etwa von Arbeit 4.0 zu 5.0 – und hinterfragt, welche Einflussgrößen konkret Veränderungsdynamiken wirken. Dabei plädiert sie für die integrierte Betrachtung von technologischen, ökonomischen, sozialen und ethischen Aspekten. Für Klingbeil-Döring lässt sich KI als Werkzeug fassen, das einer qualitativen Transformation von Arbeit den Weg bereiten kann.

Als ambivalent charakterisieren *Alexander Schmölz*, *Pia-Zoe Hahne*, *Sebastian Klocker* und *Wolfgang Katzian* die mit KI in Verbindung stehenden Veränderungen der Arbeitswelt, „die zwischen Entfremdung [...] und Humanisierung [...] oszillieren“ (Schmölz et al. in diesem Band, S. 29). Sie legen theoretische Positionen zu Entfremdung und Humanisierung dar und zeichnen nach, auf welche Weise der Einsatz von KI im Arbeitskontext den vorgestellten Momenten von Kontrolle, Überwachung und mangelnder Anerkennung auf der einen Seite und Autonomie, Partizipation und verbesserten Arbeitsbedingungen auf der anderen entsprechen kann. In diesem Zusammenhang stellen die Autor*innen den Ansatz des Digitalen Humanismus vor und plädieren für einen kritisch-humanistischen Blick auf Digitalisierung sowie für die gesellschaftliche Gestaltbarkeit von technologischen Entwicklungen.

Wie sich nun Veränderungsprozesse am Arbeitsmarkt konkret darstellen können, legen *Julia Bock-Schappelwein* und *Sandra Bilek-Steindl* dar. Die Autorinnen heben die atemberaubende Geschwindigkeit des mit KI in Verbindung stehenden Wandels hervor, weswegen sie die Effekte, die mit der Integration von KI einhergehen, als folgenschwerer einschätzen, als dies bei vorangegangenen technologiebasierten Transformationsprozessen der Fall war. Neu seien auch die Auswirkungen, die der Einsatz von KI speziell für kognitive Tätigkeiten habe. Bock-Schappelwein und Bilek-Steindl analysieren nach einer begrifflichen Klärung von Künstlicher Intelligenz mögliche Auswirkungen auf Berufe und zeigen auf, welche Tätigkeiten besonders von einer Substitution durch KI betroffen sein können.

Ebenfalls empirisch widmen sich *Jan-Ove Wiesner*, *Philipp Besinger*, *Benedikt Fuchs* und *Sebastian Schlund* dem Einsatz von KI auf dem österreichischen Arbeitsmarkt. Sie präsentieren die Ergebnisse einer Online-Befragung von 108 Unternehmen in allen neun Bundesländern, die zwischen September 2024 und März 2025 durchgeführt wurde. Im Unterschied zu einer vorangegangenen Untersuchung aus dem Jahr 2021 können ein Anstieg des Einsatzes im operativen Geschehen sowie die

Steigerung der Akzeptanz von KI-Technologie im Arbeitsumfeld ausgemacht werden. Diese lässt sich jedoch nach Tätigkeitsbereich und auch Alter des Unternehmens differenzieren. Die Autoren konstatieren, dass der Einsatz von KI-Tools weiterhin nur wenig strategisch betrieben wird und ein „instrumentelles Verständnis“ vorherrscht, in dem KI für Optimierung und Kostenreduktion eingesetzt wird, in die Gestaltung grundlegender organisationaler Prozesse und Strukturen aber keinen Eingang findet.

Im Beitrag von *Viktoria A. Krenn, Torsten Ullrich, Elena Wiegelmann, Tarim Stüwe* und *Eva Eggeling* wird der Fokus auf das Bundesland Kärnten und zu erwartende Effekte für Arbeitsmarkt, Branchenentwicklung und Qualifikationsprofile gelegt. Für den Kärntner Arbeitsmarkt, der durch demografischen Wandel sowie ungleichmäßige Branchenverteilung geprägt ist, existiert bis dato noch keine Forschung dazu. Im Rahmen einer vom Kärntner Arbeitsmarktservice (AMS Kärnten) in Auftrag gegebenen Studie führten die Autor*innen eine Literaturanalyse zu den Effekten von KI-Integration für Arbeitsmarktentwicklung, eine Auswertung von Jobausschreibungen sowie eine Umfrage mit 176 Personen durch, um Risiken und Potenziale sowie Forschungsbedarf zu identifizieren. Bei ihrer Ergebnispräsentation heben sie die Bedeutung der Gestaltung von Aus- und Weiterbildung hervor, um entsprechende Kompetenzen zu forcieren.

Teil II: Vertrauen, Transparenz und Partizipation: Herausforderungen für den KI-Einsatz

Die zunehmende Anwendung von generativen KI-Systemen in der Unternehmenskommunikation markiert *Julia Eisner* zu Folge einen „Paradigmenwechsel“ (in diesem Band, S. 99), der einerseits klassische menschenzentrierte Kommunikationsmodelle herausfordert und über die Konzeption von KI-Anwendungen als aktiven Teilnehmer*innen an Kommunikationsgeschehen Fragen von Agency, Partizipation und Verantwortung auf neue Weise stellt. Andererseits verändere sich Unternehmenskommunikation grundlegend – Stichwort zunehmende Hybridisierung kommunikativer Prozesse. Die Autorin untersucht diesen Wandel von kommunikativer und professioneller Praxis und legt dar, was sich Unternehmen von der Integration von KI erwarten und welche Risiken damit einhergehen können.

Das Thema Verantwortung greift auch *Erich Prem* in seinem Beitrag auf, der auf Autor*innenschaft und Verantwortung bei der Anwendung von generativer KI, im speziellen von großen Sprachmodellen, im Arbeitsleben fokussiert. Prem skizziert die Beobachtung, dass die Urheber*innenschaft von Texten nicht mehr direkt menschlichen Akteur*innen zugeschrieben werden kann, im Unternehmenskontext Verantwortung jedoch vielfach den Mitarbeitenden übertragen wird. Gleichzeitig stellt im Humanismus gerade Autor*innenschaft die Voraussetzung dar, um überhaupt verantwortlich gehalten werden zu können. In seinem Beitrag skizziert Prem nun, welche philosophischen Fragestellungen sich in der KI-geprägten Arbeitswelt mit Bezug auf Autor*innenschaft, Diskursordnung und weiterführend auch in Hinblick auf Transparenz und gesellschaftliche Teilhabe stellen.

Transparenz und Partizipation stehen im Zentrum des Kapitels von *Lejla Rizvanovic*. Am Beispiel von Systemen automatisierter Entscheidungsfindung (ADM-Systemen), die etwa in Bewerbungsverfahren oder bei Leistungsbewertungen immer

häufiger zum Einsatz kommen, legt die Autorin dar, wie sich algorithmische Diskriminierung gestalten kann. Auf Basis einer rechtsdogmatischen Analyse von Artikel 22 DSGVO und relevanter EuGH-Judikatur verfolgt Rizvanovik die Frage, wie Transparenz und Partizipation bei automatisierter Entscheidungsfindung verbessert werden können. Sie stellt kontrafaktische Erklärungen als Transparenzinstrument und eine Möglichkeit, Intransparenz und diskriminierenden Effekten entgegenzuwirken, vor.

Teil III: Impulse zur Gestaltung in konkreten Arbeitsfeldern

Titus Udrea und *Helene Baumgartner* schreiben den mit der Integration von Künstlicher Intelligenz einhergehenden Transformationen ein strukturelles Moment zu, weil nicht nur Qualifikationsanforderungen und Beschäftigungsprofile im Wandel begriffen sind, sondern auch die Art und Weise der Arbeitsorganisation, das Treffen von Entscheidungen in Organisationen sowie das Risikomanagement. Gerade mit Blick auf die Gestaltung der Sicherheitsarchitektur machen die Autor*innen eine doppelte Herausforderung aus: Während KI-Systeme neue, gravierende Sicherheitsrisiken eröffnen, ermöglichen sie gleichzeitig eine verbesserte Verteidigung gegen Angriffe und Bedrohungsszenarien. Udrea und Baumgartner argumentieren für das Neu-Denken von Governance-Modellen, die auf strategischer Vorausschau und Technikfolgenabschätzung gründen und partizipative Prozesse zur KI-Gestaltung, organisationskulturelle Dimensionen und nichttechnische Risiken miteinbeziehen.

Stefanie Spielberger nimmt die Veränderungen durch die gesteigerte Verfügbarkeit von generativer KI speziell im Hinblick auf hybride Arbeit in den Blick. Sie argumentiert, dass Unternehmen durch Wandel von innerbetrieblichen Abläufen, Kommunikation und Zusammenarbeit herausgefordert sind, vielerorts den Wandel jedoch kaum strategisch begleiten. Für eine gelingende Integration von KI in Unternehmen wird AI Literacy als neues Kompetenzfeld relevant, das die Autorin konzeptionell fundiert, sowie die Art und Weise der Kollaboration zwischen Mitarbeiter*innen und KI. Sie identifiziert Vertrauen als zentrales Element in dieser Gleichung und entwirft vier Handlungsempfehlungen für Unternehmen, um die Integration von KI in organisationale Prozesse positiv zu gestalten.

Wie manifestiert sich der Einsatz von Künstlicher Intelligenz an Hochschulen? Mit welchen Effekten und Herausforderungen sehen sich Bildungs- und Forschungsstätten konfrontiert? *Barbara Geyer* und *Rita Stampfl* analysieren die Veränderungsprozesse aus systemtheoretischer Perspektive anhand des Viable System Model (VSM), das sie auf Hochschulen anwenden. Der Einsatz von KI an Hochschulen wird anhand von fünf Systemen, (1) primäre Aktivitäten, (2) Koordination, (3) Steuerung und Ressourcen, (4) Strategische Planung und (5) Identität und Werte, durchdekliniert. Sie argumentieren, es handle sich dabei nicht allein um die Einführung neuer Technologien und Anwendungen, vielmehr gehe damit ein grundlegender Wandel von Aufgaben und Selbstverständnis von Hochschulen einher. Zusätzlich zur technischen Umsetzung komme auch der kritischen Reflexion und bewussten Gestaltung zentrale Bedeutung zu.

In der Medizin sind KI-Anwendungen ebenfalls angekommen und beeinflussen etwa, wie Diagnosen gestellt oder administrative Prozesse in Krankenhäusern

und der Umgang mit Patient*innen gestaltet werden. *Samuel Eke* und *Nils Hofmann* diskutieren in ihrem Beitrag das Konzept von ärztlicher Autonomie und legen dar, inwiefern autonomes (verantwortliches) Handeln durch den Einsatz von KI-Anwendungen herausgefordert wird. Zentrale Momente dabei sind etwa Nicht-Erklärbarkeit und Intransparenz von durch KI gefällten Entscheidungen sowie der unternehmerische Kontext, in dem die Tools entwickelt, vertrieben und weiterentwickelt werden. Ärzt*innen verlieren in diesem Kontext mitunter die Möglichkeit, ihr Handeln Patient*innen gegenüber umfänglich zu begründen. Die Autoren skizzieren damit einhergehende Risiken und ethische Implikationen und führen das Konzept von politischer Autonomie ein, um Partizipation und Verantwortung bei der Anwendung von KI zu erhöhen.

Petra Schön stellt mit dem Pilotprojekt PLAIInLaw4U eine KI-Anwendung vor, die den Zugang zu und die Verständlichkeit von Rechtstexten verbessern soll. Hintergrund ist die hohe Komplexität und Spezifität von rechtlicher Sprache etwa in Gesetzestexten oder bei Verordnungen, die von Personen ohne juristische Ausbildung vielfach kaum oder nur mit großer Anstrengung verstanden werden können. Das beschneidet das Wissen von Bürger*innen über ihre Rechte und Pflichten und stellt somit eine demokratiepolitische Problematik dar. Daran setzt PLAIInLaw4U an, ein interdisziplinäres Open-Source-KI-Sprachmodell für den DACH-Raum, das juristische Texte für Nicht-Jurist*innen übersetzen und damit barrierefrei zugänglich machen will.

Josef Sawetz fragt abschließend: „Befreit KI die Welt von Arbeit?“ Werden durch KI geleistete Tätigkeiten menschliche Arbeit zukünftig eher unterstützen und ergänzen oder ersetzen? Zur Analyse dieser Frage skizziert der Autor Erkenntnisse unterschiedlicher Disziplinen wie der Evolutionsbiologie, der Ökonomie oder der KI-Forschung. Ob der Einsatz von KI in eine Utopie oder Dystopie führen wird, hinterfragt Sawetz und beschreibt neue Organisationsformen, benötigte Kompetenzen und Arbeitsanforderungen. Abschließend entwirft er den Arbeitsmarkt in fünf Jahren als Zukunftsszenario.

An einer lesenswerten Publikation sind viele Personen beteiligt, die wir an dieser Stelle hervorheben möchten. Ein großes Dankeschön gilt allen Autor*innen für die aktuellen Analysen und Diskussionsbeiträge zu den Veränderungsprozessen in einer KI-vermittelten Arbeitswelt. Des Weiteren möchten wir Lena Zauchner für die Mitgestaltung der Recherche von möglichen Beitragsthemen sowie die Arbeiten zur Koordination der Artikel und Buchgestaltung danken. Für das exzellente Lektorat und die gute Kooperation geht ein großes Dankeschön an Marisa Tasser. Erwähnen möchten wir zudem die professionelle, langjährige Zusammenarbeit mit dem Waxmann Verlag.

Abschließend sei das Engagement aller Partnerorganisationen des uniMind-Projekts betont. Wir bedanken uns herzlich bei der Wirtschaftskammer Wien, der Arbeiterkammer Wien und der Wiener Stadtwerke GmbH. Über die Zusammenarbeit und gemeinsame Ausgestaltung des Jahresthemas ermöglichen sie den Austausch zwischen Universität/Forschung und interessierter Öffentlichkeit, Unternehmen und Organisationen. Dadurch bereiten sie den Weg für partizipatives Voneinander-Lernen und dienen als zentrale Multiplikator*innen, um neue Erkenntnisse in die Gesellschaft zu tragen, aber auch Resonanz und Feedback wieder an die Wissenschaftstreibenden zurückzuspielen.

Künstliche Intelligenz, so möchten wir dieses Vorwort beenden, prägt berufliche Tätigkeiten und Arbeitsprozesse heute auf vielfältige Weise. Während der Einsatz in manchen Kontexten nicht mehr wegzudenken ist und die Funktionsweise der betreffenden Bereiche grundlegend zu transformieren vermag, sind anderenorts tatsächliche Auswirkungen im Vergleich zu den prognostizierten Umwälzungen eher marginal. Für die Auseinandersetzung mit KI im Arbeitsleben ist in jedem Fall die aktive und kritische Gestaltung ihres Einsatzes zentral. Dies inkludiert, Potenziale und Risiken zu analysieren, ethische und soziale Gesichtspunkte mitzudenken und Einsatzgebiete wie -bedingungen klar zu definieren.

Literatur

- Unternehmensserviceportal. (o. D.). *AI Act: Was Unternehmen jetzt wissen müssen*. <https://www.usp.gv.at/aktuelles/newsliste/ai-act-was-unternehmen-jetzt-wissen-muessen.html>
- Europäisches Parlament. (2020). *Was ist künstliche Intelligenz und wie wird sie genutzt?* <https://www.europarl.europa.eu/topics/de/article/20200827STO85804/was-ist-kunstliche-intelligenz-und-wie-wird-sie-genutzt>
- Statistik Austria. (2025). *Standard-Dokumentation Metainformationen zur Statistik über den IKT-Einsatz in Haushalten*. https://www.statistik.at/fileadmin/shared/QM/Standarddokumentationen/B_2/std_b_ikt-einsatz_in_haushalten_ab_2024.pdf
- Wangari, S. & Vaidyanathan, G. (2025, 21. April). How Big Tech hides its outsourced African workforce. *Rest Of World*. <https://restofworld.org/2025/big-tech-ai-labor-supply-chain-african-workers/>

„University Meets Industry“ wurde von der UNESCO-Kommission als UN-Dekaden-Projekt „Bildung für nachhaltige Entwicklung“ ausgezeichnet. Die Auszeichnung erfolgt für Projekte, die einen Beitrag zur Profilierung einer nachhaltigen Bildungsarbeit leisten und dabei alle drei Nachhaltigkeitsdimensionen – Ökonomie, Ökologie und Soziales – gemeinsam berücksichtigen.



Ausgezeichnet von der
Österreichischen UNESCO-Kommission

Teil I:
Transformativ oder überschätzt?
Auswirkungen auf Arbeitsmarkt,
Berufe und Qualifikationen

Künstliche Intelligenz und der Arbeitsmarkt

Einflussgrößen, Entwicklungen und ethische Überlegungen

Wenke Klingbeil-Döring

Von Substitution zu Transformation: eine neue Perspektive auf KI und Arbeit

Künstliche Intelligenz (KI) zählt zu den wichtigsten Treibern des aktuellen Wandels von Arbeit (World Economic Forum, 2025). Sie schickt sich an, die Entwicklung an der uns inzwischen recht vertrauten Wegmarke der sogenannten „Arbeit 4.0“ in eine neue Richtung zu lenken. Was genau ihr Einfluss bewirkt, ist aber nach wie vor unklar. Zwar wächst die Zahl der Studien zu den Auswirkungen von KI auf Arbeit und den Arbeitsmarkt mit ihrer bemerkbar zunehmenden Verbreitung, jedoch handelt es sich dabei größtenteils noch um Prognosen zu in der Regel recht kompakt umrissenen Aspekten. Diese Unschärfe ist einerseits Folge der offenen und dynamischen Natur der Entwicklung selbst, andererseits aber auch des häufig zu engen Fokus der Untersuchungen auf die unmittelbaren Effekte von KI auf Arbeitsprozesse, Berufsbilder und Beschäftigung. Im Fokus auf quantifizierbare Wandlungsmarker überblenden die allermeisten Untersuchungen, dass hier ein viel breiteres Gefüge von Einflussfaktoren innerhalb der gesellschaftlichen und individuellen Bedeutung von Arbeit als tragende Institution und Praxis der Arbeitsgesellschaft wirkt; dass Arbeit sich also *auch*, aber nicht nur unter dem Einfluss von KI wandelt, und selbst Anforderungen – herkommend von ihrer genuinen Rolle – in den Wandel einbringt. So fehlt eine integrierte Perspektive, die neben technologischen auch soziale Aspekte, neben der ökonomischen auch die ethische Dimension des laufenden Wandels von Arbeit in den Blick nimmt. Diese Perspektive soll hier – auch im Aufzeigen bestehender diskursiver Leerstellen – entwickelt werden. Wenn verstanden ist, dass Arbeit ihre am Menschen orientierte Gestaltbarkeit und ihre weitreichenden normativen Implikationen auch und gerade unter dem Einfluss eines Technologiesprungs bewahrt, bei dem es um Logik, Effizienz, Machbarkeit und Funktionalität geht, lassen sich weitaus verlässlichere Beschreibungen des KI-induzierten Wandels von Arbeit finden.

Neue Wandlungsdynamiken: von Arbeit 4.0 zu Arbeit 5.0?

Wo Phänomene oder Prozesse beschrieben werden sollen, über die sich noch wenig sagen lässt, ist das Nachverfolgen möglicher Kontinuitäten ein verlässlich vielversprechendes Vorgehen. Hier ist das die sogenannte Arbeit 4.0. Das Schlagwort zeigt an: Wir haben es mit einem technologiegetriebenen Wandel im Ausmaß einer vierten industriellen Revolution zu tun. Digitalisierung, Automatisierung und Vernetzung formen im Zusammenspiel in einem beispiellosen Tempo und in nichtgekannter Intensität diese neue Version von Arbeit: Digitale, vernetzte Arbeitsmittel haben neue Kommunikations- und Kooperationsformen etabliert, Arbeit verdichtet, sie örtlich, zeitlich und organisatorisch flexibilisiert und individualisiert (Bauer & Hofmann,

2018; Hackl et al., 2017). Kompetenzanforderungen sind dazugekommen oder haben sich verändert, ebenso Berufsprofile, Tätigkeiten und Geschäftsfelder, die durch globale Kooperationen und Ideenaustausch heute in vielen Branchen stark internationalisiert sind (BMAS, 2017). Arbeitnehmende und Unternehmen kennen Arbeit 4.0 als dynamisch, chancenreich und herausfordernd. Dabei bleibt Arbeit als gesellschaftliche Institution und Praxis konstitutiv, tragend und gestaltend (Hackl et al., 2017; Jähnichen, 2018). Arbeit prägt – das ist ihre Rolle in der Arbeitsgesellschaft – nicht nur ökonomische Prozesse, sondern auch soziale Teilhabe, individuelle Identität und gesellschaftliche Anerkennungsverhältnisse. Sie ist nicht nur Mittel zum Einkommenserwerb, sondern daneben zentrales Element des Selbstverständnisses von Individuen und der Verfassung gesellschaftlicher Ordnungen (Klingbeil-Döring, 2019). Schon immer berühren Wandlungsprozesse von Arbeit nicht nur technische oder arbeitsmarktpolitische Fragen, sondern ganz entscheidend auch kulturelle, ethische und politische Grundannahmen (Castel, 2000; Eggebrecht et al., 1980).

Das gilt auch, wenn nun KI hinzukommt und neue Dynamiken, potenziell sogar Eigeninitiative einbringt. Dennoch lassen sich die bekannten Folgen der bisherigen, jeweils industrielle Revolutionen auslösenden Technologiesprünge auch für eine bloß formale Beschreibung nur bedingt als Prognoseschablone auf die gegenwärtige Entwicklung legen. So gerät der Versuch, die Auswirkungen des Einflusses von KI auf Arbeit mittels Zuschreibung einer Kontinuität nachzuvollziehen, hier erst einmal ins Stocken. Denn: Diesmal bewirkt eben nicht eine bestimmte Technologie in klar bestimmbar Tätigkeitsbereichen in erster Linie eine sichtbare Effektivierung und dann auch soziale Veränderungen. Das, was den auf der jetzigen Entwicklungsstufe von Arbeit wirksamen Technologien zugrunde liegt und was wir als KI wahrnehmen und beschreiben, entzieht sich unseren etablierten Dimensionierungen gesellschaftlicher und wirtschaftlicher Transformationskräfte und erschwert genau deshalb die vielfach vorhandenen Prognosebemühungen. Dies hängt eng mit dem Wesen der KI selbst zusammen – also mit ihrer konzeptuellen Verfasstheit und den ihr inhärenten Eigenschaften. Dabei sind diese keineswegs abschließend geklärt, auch, weil KI sich ständig weiterentwickelt. Umstritten ist zudem, ob die Zuschreibung von „Intelligenz“ für KI-Systeme überhaupt angemessen oder vielleicht sogar irreführend ist (Lenzen, 2020).

Mit dem Begriff „Künstliche Intelligenz“ verbinden sich in der öffentlichen Wahrnehmung häufig bestimmte Vorstellungen und Konzepte. In alltagsweltlichen Diskursen wird KI oftmals als eine quasi-autonome Entität mit spezifischen Eigenschaften oder Fähigkeiten aufgefasst. Bisweilen kommt es uns so vor, als seien Technologien und Tools wie KI-gestützte Software, Chatbots, Roboter oder selbstfahrende Autos eigenständig handelnde Akteure. Die zunehmende Präsenz von KI in verschiedenen Lebensbereichen nährt diese Vorstellung. Sie schwingt mit, wenn wir darüber nachdenken, wie KI unsere Arbeit und dann auch den Arbeitsmarkt verändert. Tatsächlich hat sie mit dem, was KI tatsächlich ausmacht oder kennzeichnet, nur am Rande bzw. an der Oberfläche ihres unmittelbaren Einflusses auf unsere Lebensrealität zu tun.

Die KI-Forschung begreift ihren Gegenstand als „[...] ein Teilgebiet der Informatik mit dem Ziel, Maschinen zu befähigen, Aufgaben ‚intelligent‘ auszuführen. Dabei ist weder festgelegt, was ‚intelligent‘ bedeutet, noch welche Techniken zum Einsatz kommen“ (Döbel et al., 2018, S. 8). Das „Projekt“ KI, so ergänzt die Wissenschafts-

journalistin Manuela Lenzen diese allgemeine Beschreibung, „umfasst ein ganzes Bündel unterschiedlicher und immer wieder weiter oder neu entwickelter Verfahren“ (Lenzen, 2020, S. 11). Folglich sind nicht bestimmte Technologien KI, sondern umgekehrt KI die Basis von Technologien, die so operieren, dass kognitive Prozesse als Berechnungsprozesse innerhalb künstlicher neuronaler Netze (KNN) umgesetzt werden. In diesem Sinne sind die Systeme „intelligent“, wobei nicht davon gesprochen werden kann, dass sie menschliche Intelligenz nachahmen (Lenzen, 2020, S. 11). Die Systeme arbeiten auf der Basis von Daten; maschinelles Lernen ist notwendig und inhärent datenbasiert. Was eine KI ausgibt, beruht auf Beispielen und Statistiken. Wie leistungs- und generalisierungsfähig sie dabei ist, hängt von der Qualität und Quantität der Daten ab, mit denen sie operiert.

Diese Eigenschaft ist für die Beschreibung und Bewertung des Einflusses von KI auf Arbeit entscheidend: In ihrem Tun ist KI eben nicht autonom, sondern auf vielfältige Weise auf menschliche Eingriffe und Entscheidungen angewiesen – sei es bei der Auswahl und Aufbereitung der Trainingsdaten, der Festlegung von Modellarchitekturen und Zielgrößen oder bei der Interpretation und Bewertung ihrer Ergebnisse im jeweiligen Anwendungskontext (Poretschkin et al., 2021). Die KI-Forscherin Kate Crawford (2024) spricht in diesem Zusammenhang von der materiellen Seite der KI und nimmt damit deren Abhängigkeit von menschlicher Arbeit in den Blick. Das gilt, so muss Crawfords Einschätzung hier ergänzt werden, schon für die Frage, ob in einem bestimmten Arbeitskontext KI überhaupt zum Einsatz kommt. Ob es sinnvoll oder wünschenswert ist, KI-Systeme einzusetzen, hängt davon ab, ob sie tatsächlich zur Lösung konkreter und dann letztlich menschlicher Probleme beitragen können. Unternehmen setzen KI ein, wo sie einen funktionalen, sozialen oder ökonomischen Bedarf sehen, den KI besser bewältigen kann als bestehende Mittel. Darüber hinaus muss der Einsatz gesellschaftlich legitimiert sein – etwa im Hinblick auf Datenschutz, Fairness oder Teilhabe.

Auch in der prominenten Unterscheidung zwischen sogenannter „schwacher“ und „starker“ KI drückt sich aus, dass KI gegenwärtig und auf allen vom jetzigen Standpunkt aus denkbaren Entwicklungsstufen nicht ohne den Menschen auskommt. Als schwache KI sind jene Systeme qualifiziert, die spezifische Problemstellungen auf der Grundlage zuvor definierter Regeln und Muster bearbeiten. Innerhalb dieses vorgegebenen Rahmens sind sie lernfähig, d.h. sie können sich durch selbstständige Datenanpassung und -verarbeitung weiterentwickeln. Demgegenüber bezeichnet starke KI ein bislang hypothetisches Entwicklungsstadium, in dem künstliche Systeme kognitive Fähigkeiten in einem dem Menschen vergleichbaren – oder sogar überlegenen – Maß aufweisen. Solche Systeme wären prinzipiell in der Lage, vollständig autonom zu agieren und sich menschlicher Kontrolle zu entziehen (Giering & Kirchner, 2021). Derzeit lassen sich Überlegungen zum Einfluss von KI auf Arbeit sinnvoll nur auf schwache KI beziehen – also auf jene Systeme, die bereits existieren oder zumindest technisch realisierbar erscheinen. Das sind spezialisierte Systeme, die klar umrissene Aufgaben lösen, etwa Spracherkennung, Musteranalyse oder Entscheidungsunterstützung. Solche KI kann nur sinnvoll innerhalb der von Menschen gesetzten Ziele, Rahmenbedingungen und Anwendungskontexte handeln. Sie ist ein Werkzeug, das menschliches Handeln ergänzt, unterstützt oder erweitert (Graus et al., 2021), und bleibt damit an das Tun sowie an die Erfordernisse und Seinsbedingungen des Menschen gebunden.

Nun lässt sich konkretisieren, welche Transformationsdynamik KI bezogen auf Arbeit absehbar entfaltet und wie sie dann deren Entwicklung bestimmt. Weil Daten und die Erfordernisse des Menschen die Grundlage maschinellen Lernens und damit des Operierens von KI sind, gelten die Prägekräfte und sichtbaren normativen Implikationen von Arbeit 4.0 auch unter KI-Einfluss. Aber die unterstellte Kontinuität besteht nur bedingt: Weil KI-Systeme nicht ausschließlich auf festen Regeln und vorab definierten Abläufen basieren, sondern in der Lage sind, aus Daten zu lernen und sich an neue Muster anzupassen, ist davon auszugehen, dass KI sehr viel tiefgreifender auf Arbeit wirken wird (Acemoglu & Restrepo, 2020) und sich für weit mehr Aufgaben und in weit mehr Bereichen einsetzen lässt (Graus et al., 2021). So können KI-Systeme etwa auch solche Aufgaben übernehmen, die eine komplexere Denkleistung oder eine höhere Qualifikation erfordern (Heinen et al., 2017). Weil sie zudem selbstlernend und vernetzt sind, können sich ihre Fähigkeiten und ihr Einsatzbereich theoretisch ständig erweitern, ohne dass es für jede neue Aufgabe oder Anwendung einer bewussten, von außen gesteuerten Weiterentwicklung oder Erweiterung bedarf. Dabei agieren bereits jetzt viele KI-Systeme im Hintergrund, sind also nicht unbedingt sichtbar Teil des Arbeitsprozesses.

Leerstellen der Debatte

Außerdem, so lässt sich im weiter gefassten Blickwinkel auf den Wandel von Arbeit sagen, wirft das Eindringen Künstlicher Intelligenz in die Arbeitswelt nicht zuletzt existenzielle Fragestellungen auf, die sich in ihrer Qualität grundlegend von den sozialen Fragen früherer technologischer Umbrüche unterscheiden. Diese neuen Problemstellungen überschreiten auf den ersten Blick den engeren Bereich der Arbeit, berühren ihn jedoch zentral. Die beschriebenen Dynamiken, die KI in den laufenden Transformationsprozess von Arbeit einbringt, und die in der Debatte oft zu enge Perspektive auf sie begründen – so lässt sich bis hierhin festhalten – die vielfachen Schwierigkeiten der Schilderung, aber auch der Steuerung des Einflusses von KI auf Arbeit:

- 1) Es fehlt ein historisches Referenzmodell für die Qualität dieser Entwicklung, da es sich nicht um einen begrenzten technologischen Wandel im herkömmlichen Sinne handelt.
- 2) Darüber hinaus besteht bislang kein hinreichend geklärtes Verständnis davon, was KI ist und welche Wirkpotenziale ihr tatsächlich innewohnen; dies ist weiterhin Gegenstand kontroverser Debatten.
- 3) Die zukünftige technische Entwicklung lässt sich kaum verlässlich prognostizieren – weder in Bezug auf die konkrete Ausgestaltung noch auf die Reichweite künftiger KI-Anwendungen. Was werden die Systeme können und wofür lassen sie sich einsetzen? Inwieweit die Entwicklung und der Einsatz von KI in der Arbeitswelt voranschreiten, ist maßgeblich von den allgemeinen wirtschaftlichen und gesellschaftlichen Rahmenbedingungen abhängig. Etwa: Welche Bedarfe haben Unternehmen? Wo lassen sich die Systeme sinnvoll einsetzen? Können Mitarbeitende hinreichend qualifiziert werden?

- 4) Insbesondere die sozialen Einflussgrößen und die normativen Implikationen, die Arbeit selbst mitbringt, finden in der Debatte bislang kaum Berücksichtigung. Dabei ist offenkundig, dass Arbeit und die Arbeitsgesellschaft nicht ausschließlich durch das technisch Machbare oder wirtschaftlich Zweckmäßige bestimmt werden. Vielmehr wirkt eine Vielzahl von Einflussfaktoren auf die Entwicklung und Gestaltung von Arbeit ein (World Economic Forum, 2025), darunter demografische Veränderungen, Migrationsbewegungen sowie sich wandelnde gesellschaftliche Wertvorstellungen in Bezug auf Arbeit. Hinzu kommen rechtliche Rahmenbedingungen auf nationaler wie internationaler Ebene sowie globale und ökologische Herausforderungen. Diese sozialen, rechtlichen und politischen Einflussgrößen stehen in einem wechselseitigen Verhältnis zu technologischen Entwicklungen, da sie maßgeblich die Bedingungen mitbestimmen, unter denen sich KI-Systeme überhaupt entfalten und eingesetzt werden können.

Die Debatte um den Einfluss von KI auf Arbeit kann nur zu tragfähigen Beschreibungen gelangen, wenn sie das Zusammenwirken von Technologie und menschlichem Verhalten in den Blick nimmt (Richthofen et al., 2021). Das schließt auch die Lebensbedingungen in der Arbeitsgesellschaft und die normativen Implikationen von Arbeit selbst ein. Gerade einer gestaltungsleitenden Betrachtung genügt es nicht, technologische Entwicklungen und deren Auswirkungen auf Arbeit und den Arbeitsmarkt isoliert zu betrachten. Um die Veränderungen anforderungsgerecht und verantwortungsvoll zu steuern, bedarf es eines Verständnisses, das die gezeigten Einflussgrößen und deren Reziprozität explizit und vor dem Hintergrund berechtigter normativer Ansprüche und ethischer Überlegungen berücksichtigt. Neben der Technologie selbst müssen Menschen und Gesellschaften als Akteure in den Blick genommen werden, die technologische Systeme anwenden, gestalten und regulieren und für deren Lebensgestaltung Arbeit eine ganz bestimmte Rolle spielt.

Studien und Prognosen: Wie wird KI Arbeit verändern?

Welche Erkenntnisse lassen sich aus der aktuellen Debatte über die Auswirkungen von KI auf Arbeit für diese Perspektive ableiten? Die wissenschaftliche Beschäftigung mit den arbeitsweltbezogenen Folgen von KI steht derzeit noch am Anfang. Es lässt sich nicht verlässlich absehen, welchen Einfluss KI-Technologien künftig auf Arbeit haben werden; erst recht lassen sich auf Basis ihres gegenwärtigen Verbreitungsgrads und der betrieblichen Umsetzung keine belastbaren Aussagen treffen. Zwar hat sich die Nutzung von KI in deutschen Unternehmen von 13,3 Prozent im Jahr 2023 auf 27 Prozent im Jahr 2024 branchenübergreifend verdoppelt und ist damit in über drei Viertel der Firmen bereits Alltag, konkret geplant oder aktiv diskutiert, jedoch variieren der Einsatz und die Implementierung stark (Kerkhof et al., 2024). In verschiedenen Branchen gibt es deutliche Unterschiede in der Nutzung von KI. Besonders Unternehmen aus der Werbung, Marktforschung oder Informationstechnologie bzw. Anbieter von Informationsdienstleistungen setzen KI intensiv ein. Insgesamt zeigt sich, dass v.a. industrielle und bestimmte Dienstleistungsunternehmen KI nutzen, während Handel, Gastronomie und Baugewerbe noch zurückhaltender sind (Kerkhof et al., 2024). Aktuell wird KI v.a. für Datenanalyse und

Prozessoptimierung eingesetzt, generative KI-Systeme wie ChatGPT finden häufig in der Kundenbetreuung, im Marketing oder in Forschung und Entwicklung Anwendung (Bitkom, 2024). Die Unternehmen sind hier bislang noch zögerlich (Bitkom, 2024), aber zunehmend auf der Suche nach Fachkräften, die über Fähigkeiten im Umgang mit generativer KI verfügen (Büchel & Engler, 2025). Insgesamt aber sind die Einsatzbereiche von KI noch begrenzt, etwa, weil geeignetes Personal fehlt (Büchel & Engler, 2025), rechtliche Bedenken bestehen oder die Unternehmen individuelle Lösungen benötigen, deren Entwicklung aufwendig und teuer ist (Giering & Kirchner, 2021). So hält KI in Unternehmen langsam und inkrementell Einzug, ein disruptiver Wandel findet nicht statt und ist unter den gegenwärtigen Bedingungen auch nicht zu erwarten (Giering, 2022).

Abgesehen von den Untersuchungen, die die tatsächliche Verbreitung von KI-Technologien in Unternehmen evaluieren, sind die allermeisten Studien dazu, wie KI Arbeit und den Arbeitsmarkt verändern wird, Prognosen. Sie treffen – zumeist in der engen Perspektive einer Modellierung des technisch und wirtschaftlich Erwartbaren – Tendenzaussagen für eine tatsächlich viel breiter beeinflusste und dabei noch weitestgehend offene und gestaltbare Entwicklung. Mit ihrer Studie *The future of employment: How susceptible are jobs to computerisation?* legten der Ökonom Carl Frey und der Professor für Maschinelles Lernen Michael Osborne im Jahr 2013 die erste umfassende Untersuchung zum Einfluss von KI auf den Arbeitsmarkt und auf die Entwicklung und die Nachfrage bestimmter Tätigkeiten und Berufe vor (Frey & Osborne, 2013). Ihrer mathematischen Modellierung zufolge weisen 47 Prozent aller bestehenden Arbeitsplätze in den USA ein hohes Automatisierungspotenzial innerhalb eines Zeitraums von zehn bis zwanzig Jahren nach dem Erhebungszeitpunkt der Studie auf. In den folgenden Jahren würden, so die Autoren weiter, unter der Voraussetzung einer Weiterentwicklung der Technologie – hier stützen sich Frey und Osborne auf Expertenaussagen – Schritt für Schritt weitere Arbeitsplätze betroffen sein, insbesondere in niedrigen und mittleren Einkommensgruppen in den Bereichen Dienstleistung, Produktion, Logistik und Transport.

Die eindrücklichen Zahlen der Studie, ihre ambivalente Rezeption (Lenzen, 2018) und wohl nicht zuletzt das von Frey und Osborne prognostizierte Szenario einer „technologischen Arbeitslosigkeit“ (Kellermann & Petring, 2019) regten in der Folge weitere Untersuchungen an. So ließ das Bundesministerium für Arbeit und Soziales im Jahr 2015 die Übertragbarkeit der Ergebnisse von Frey und Osborne auf den Arbeitsmarkt der Bundesrepublik Deutschland prüfen (Bonin et al., 2015). Anders als Frey und Osborne legten die Autoren den Fokus nicht auf Berufe, sondern auf Tätigkeitsprofile bzw. Tätigkeitsstrukturen und gingen in ihrer Prognose von einer weniger dynamischen Weiterentwicklung der Technologien aus. So weist die Studie insgesamt eine etwas geringere Automatisierungswahrscheinlichkeit durch KI aus, die mit zunehmender Qualifikation und höherem Einkommen signifikant abnimmt. Damit relativieren die Autoren die von Frey und Osborne ermittelten Werte in gewissem Maße und sehen zudem „[d]ie Gesamtbeschäftigung [...] nicht zwangsläufig gefährdet“ (Bonin et al., 2015, S. ii). Denn, so ihre Schlussfolgerung: „Selbst wenn Automatisierung unmittelbar zu Arbeitsplatzverlusten führt, entstehen durch den Wandel zugleich neue Arbeitsplätze, beispielsweise bei der Herstellung der neuen Technologien“ (Bonin et al., 2015, S. ii).

Insgesamt, so formuliert Oliver Giering (2022, S. 59), werden die in der Forschung zum Einfluss von KI auf Arbeit und den Arbeitsmarkt existierenden negativen Prognosen, die von Arbeitsplatzverlusten ausgehen, von solchen in Zweifel gezogen, die „viele nicht-technische, soziale und arbeitspolitische Faktoren im Zusammenhang mit der Verbreitung, Einführung und Nutzung von KI nicht ausreichend berücksichtigt“ sehen. Das gilt auch nach der Einführung generativer KI-Systeme wie ChatGPT. Zwar schreiben verschiedene Studien dieser Entwicklungsstufe aufgrund ihrer Fähigkeit, eigenständig neue Inhalte zu erzeugen, ein höheres disruptives Potenzial zu als früheren KI-Technologien – insbesondere in Tätigkeitsfeldern, die bislang als weitgehend resistent gegenüber Automatisierung galten, etwa aufgrund ihres geringen Anteils an Routinetätigkeiten –, zugleich scheint aber auch hier die Fähigkeit hervorstechen, menschliche Arbeit zu unterstützen (Arntz et al., 2025). Tatsächlich spricht viel dafür, dass KI potenziell dazu in der Lage ist, neue Arbeitsplätze zu schaffen und bestehende zu verbessern (Lenzen, 2018); angesichts der gezeigten engen Verknüpfung maschineller und menschlicher Arbeit ist diese Entwicklungsrichtung sehr wahrscheinlich. Überdies lässt sie sich im Rückblick für alle bisherigen größeren Technologiesprünge seit der Industrialisierung feststellen: Wo Maschinen Tätigkeiten übernahmen, erschlossen sich neue Geschäftsfelder mit neuen Berufen und Beschäftigungsformen. Technologie hat auf keiner Entwicklungsstufe Arbeit vernichtet, wohl aber ihre Form, ihren Ort und ihren Inhalt verändert. Was das für Arbeit und über sie hinaus bedeutet, hängt davon ab, wie die betroffenen Gesellschaften auf diesen Wandel reagieren – idealerweise durch Bildung, Innovation und sozialpolitische Gestaltung. Denn – auch das zeigt der Rückblick auf vergangene Technologiesprünge – der Wandel kann mit tiefgreifenden sozialen Verwerfungen einhergehen, wenn er v.a. an ökonomischen Rationalitäten und dem technisch Machbaren ausgerichtet erfolgt, normative und soziale Implikationen aber unberücksichtigt bleiben.

Diese Perspektive gewinnt umso mehr an Bedeutung, als auch die ökonomischen Erwartungen im Hinblick auf die Auswirkungen von KI auf die gesamtwirtschaftliche Produktivität bislang verhalten ausfallen. So ist ein bedeutendes gesamtwirtschaftliches Produktivitätswachstum durch das zunehmende Eindringen von KI in Arbeit einer aktuellen Prognose zufolge nicht zu erwarten (Demary et al., 2025). Zwar könne der Einsatz von KI auf der Ebene einzelner Arbeitsplätze oder Tätigkeiten zu Produktivitätssteigerungen führen, doch deren Realisierung hänge stark von betriebspezifischen Voraussetzungen ab und sei kein automatischer Effekt. Insgesamt, so fasst die Studie zusammen, spreche „derzeit wenig dafür, dass sich hinter dem Produktivitätspotenzial von KI ein arbeitssparender technischer Fortschritt verbirgt. Vielmehr deuten die derzeitigen Befunde darauf hin, dass KI in einem komplementären Verhältnis zum Faktor Arbeit steht“ (Demary et al., 2025, S. 4), KI menschliche Arbeit also unterstützt, ergänzt oder erweitert. Diese Einschätzung korrespondiert mit der Perspektive der Beschäftigten, die aber die Ambivalenz der Durchdringung von Arbeit durch KI betont (Arntz et al., 2025; Fleck et al., 2022): Der subjektiven Einschätzung, mithilfe von KI die eigene Arbeitsproduktivität steigern zu können, stehen Befürchtungen gegenüber, die sich auf eine potenzielle Zunahme von Arbeitsbelastung und -anforderungen beziehen. Zentrale Erkenntnis ist dabei, dass KI-Anwendungen insbesondere dann auf eine hohe Akzeptanz stoßen und als entlastend erlebt werden, wenn sie nahtlos in bestehende Arbeitsmittel und

-abläufe integriert sind und somit als unterstützend statt als störend oder disruptiv wahrgenommen werden. Das zeigt: Das Interesse der Beschäftigten liegt weniger in einer Ersetzung als vielmehr in einer qualitativen Verbesserung und Entlastung ihrer Arbeit.

Eine im August 2024 veröffentlichte Untersuchung dazu, welche beruflichen Kompetenzen künftig gefragt sind (Chui et al., 2024), bestätigt die Prognose einer in erster Linie qualitativen Transformation von Arbeit unter dem Einfluss von KI. Sie geht davon aus, dass bestimmte Job-Profile auf dem Arbeitsmarkt künftig stärker, andere weniger gefragt sind. Demnach könnten MINT- und Gesundheitsberufe am europäischen Arbeitsmarkt bis 2030 um bis zu 30 Prozent wachsen, was etwa sieben Millionen zusätzliche Stellen bedeuten würde. Gleichzeitig wird ein Rückgang der Nachfrage in Berufsbereichen mit hohem Automatisierungspotenzial um 300.000 bis 5 Millionen Stellen erwartet, insbesondere in der Gastronomie, Produktion, im Kundenservice, Verkauf und der administrativen Unterstützung. Insgesamt geht die Studie wie auch der *Future Jobs Report 2025* des Weltwirtschaftsforums (World Economic Forum, 2025) von einem moderaten Nettoanstieg der Beschäftigung aus. Weitaus stärker ins Gewicht fällt jedoch die prognostizierte Zunahme an beruflichen Umstiegsprozessen: Die Ergebnisse legen nahe, dass die Automatisierung infolge des KI-Einsatzes in den kommenden Jahren den zentralen Treiber für berufliche Veränderungen darstellen wird (World Economic Forum, 2025). Beschäftigte sind zunehmend auf neue Kompetenzen angewiesen und werden – analog zu allgemeinen Entwicklungen im Zuge der Digitalisierung (BMAS, 2017) – ihre Fähigkeiten künftig in kürzeren Zyklen anpassen und weiterentwickeln müssen, um mit dem technologischen Wandel Schritt zu halten. Dies unterstreicht die wachsende Bedeutung von Bildung und beruflicher Weiterbildung, deren Relevanz sowohl auf Seiten der Unternehmen als auch der Beschäftigten zunehmend erkannt wird (Bitkom Akademie & HRpepper, 2024a). Auch hier kommt der sozialpolitischen Gestaltung eine zentrale Rolle zu, um die Transformationsprozesse arbeitsmarkt- und gesellschaftsverträglich zu begleiten (Kellermann & Petring, 2019).

KI als Impuls einer qualitativen Transformation von Arbeit

Die hier entwickelte Perspektive auf den Einfluss von KI auf Arbeit und Arbeitsmarkt lässt – unter Einbezug des aktuellen Umsetzungsstands von KI in Unternehmen, absehbarer Entwicklungen und Untersuchungen zu ihren Potenzialen – folgenden Schluss zu: KI wird die Arbeitswelt nicht radikal umgestalten, aber in zunehmendem Maße integriert werden, primär als Werkzeug, das menschliche Arbeit ergänzt, aber auf sie angewiesen bleibt. Tätigkeiten und Beschäftigungsverhältnisse werden sich verändern, jedoch vorwiegend in qualitativer Hinsicht, sodass von einer qualitativen Transformation gesprochen werden kann. KI ersetzt Tätigkeiten nicht einfach, sondern modifiziert sie – einhergehend mit neuen Anforderungen an menschliche Fachkompetenz, Aufsicht und ethische Reflexion.

Für die Bewertung und Gestaltung der Implementierung von KI in Arbeitsprozesse ist diese Einsicht zentral, und zwar aus zwei Perspektiven. Mit Blick auf KI selbst wird deutlich: Da KI-Systeme nicht autonom agieren, muss sich der Fokus der Debatte von einer potenziellen Substitution menschlicher Arbeit hin zu einer

tiefgreifenden Transformation von Arbeitsprozessen verschieben. Die Abhängigkeit technologischer Systeme von menschlichen Entscheidungen verdeutlicht, dass KI keineswegs ein neutraler oder objektiver Faktor im Arbeitskontext ist. Die Implementierung von KI in der Arbeitswelt ist auf allen Ebenen gestaltbar. Damit zusammenhängend macht der Fokus auf die Rolle und Bedeutung von Arbeit in der Arbeitsgesellschaft klar: Um sozialen Verwerfungen oder übermäßigen Belastungen der Beschäftigten vorzubeugen – wie sie bereits in früheren Technologiesprüngen und auch im Übergang zu Arbeit 4.0 zu beobachten waren –, muss sich diese Gestaltung an den Erfordernissen von Arbeit selbst orientieren. Entscheidend ist, Arbeit als tragende Institution und zentrale Praxis der Arbeitsgesellschaft zu erkennen und unter KI-Einfluss zu stärken. Dann lassen sich die Chancen nutzen, die KI eröffnet. Dazu zählen nicht allein die Kompensation von Fachkräftemangel, der Erhalt der Wettbewerbsfähigkeit oder die Steigerung individueller Arbeitsproduktivität, sondern insbesondere auch die Verwirklichung der normativen Rolle von Arbeit. Tatsächlich birgt ausgerechnet KI einiges Potenzial, Arbeit „menschlich“, d. h. im Sinne des Menschen zu gestalten: Sie ist in der Lage, von einseitiger und entfremdender Arbeit zu entlasten, Mitgestaltung und Autonomie zu fördern – angefangen bei der Technikgestaltung selbst –, kann Lernen und Weiterbildung erleichtern und neue Formen der Zusammenarbeit realisieren, die Individualität und Kreativität Raum geben. Inwieweit und auf welche Weise sich dieses Potenzial ausschöpfen lässt, bleibt eine Frage für zukünftige Forschung.

Literatur

- Acemoglu, D. & Restrepo, P. (2020). The wrong kind of AI? Artificial intelligence and the future of labour demand. *Cambridge Journal of Regions Economy and Society*, 13(1), 25–35. <https://doi.org/10.1093/cjres/rsz022>
- Arntz, M., Baum, M., Brüll, E., Dorau, R., Hartwig, M., Lehmer, F., Matthes, B., Meyer, S.-C., Schlenker, O., Tisch, A. & Wischniewski, S. (2025). *Digitalisierung und Wandel der Beschäftigung (DiWaBe 2.0): Eine Datengrundlage für die Erforschung von Künstlicher Intelligenz und anderer Technologien in der Arbeitswelt*. Bundesanstalt für Arbeitsschutz und Arbeitsmedizin. <https://www.baua.de/DE/Angebote/Publikationen/Berichte/F2573>
- Bauer, W. & Hofmann, J. (2018). Arbeit, IT und Digitalisierung. In J. Hofmann (Hrsg.), *Arbeit 4.0 – Digitalisierung, IT und Arbeit* (S. 1–16). Springer Vieweg. https://doi.org/10.1007/978-3-658-21359-6_1
- Bitkom. (2024, 16. Oktober). *Erstmals beschäftigt sich mehr als die Hälfte der Unternehmen mit KI* [Pressemeldung]. <https://www.bitkom.org/Presse/Presseinformation/Erstmals-beschaeftigt-Haelfte-Unternehmen-KI>
- Bitkom Akademie & HRpepper (Hrsg.). (2024). *Weiterbildung 2024: Alles nur KI?* <https://bitkom-akademie.de/weiterbildungsstudie-2024>
- Bonin, H., Gregory, T. & Zierahn, U. (2015). *Übertragung der Studie von Frey/Osborne (2013) auf Deutschland: Endbericht* (Kurzexpertise Nr. 57). ZEW – Zentrum für europäische Wirtschaftsforschung. https://www.bmas.de/SharedDocs/Downloads/DE/Publikationen/Forschungsberichte/fb-455.pdf?__blob=publicationFile&v=2
- Büchel, J. & Engler, J. (2025). *Generative KI: Ein Wachstumstrend*. *IW-Kurzbericht 39/2025*. Institut der deutschen Wirtschaft. https://www.iwkoeln.de/fileadmin/user_upload/Studien/Kurzberichte/PDF/2025/IW-Kurzbericht_2025-Generative-KI.pdf

- Bundesministerium für Arbeit und Soziales (BMAS). (Hrsg.). (2017). *Weißbuch Arbeiten 4.0*. https://www.bmas.de/SharedDocs/Downloads/DE/Publikationen/a883-weissbuch.pdf?__blob=publicationFile&v=2
- Castel, R. (2000). *Die Metamorphosen der sozialen Frage: Eine Chronik der Lohnarbeit*. UVK.
- Chui, M., Durth, S., Hazan, E., Huyghues-Despointes, R., Madgavkar, A., Maor, D., Singh Dandona, G. & Smit, S. (2024). *Eine neue Zukunft der Arbeit: Der Wettlauf um die Einführung von KI in Europa – welche Fähigkeiten jetzt gefragt sind*. McKinsey Global Institute. <https://www.mckinsey.de/~media/mckinsey/locations/europe%20and%20middle%20east/deutschland/news/presse/2024/2024%20-%2005%20-%2023%20mgi%20genai%20future%20of%20work/mgi%20report%20genai%20future%20of%20work%20deutsch.pdf>
- Crawford, K. (2024). *Atlas der KI: Die materielle Wahrheit hinter den neuen Datenimperien*. C. H. Beck.
- Demary, V., Grömling, M., Kestermann, C., Scheufen, M., Seele, S., Stettes, O. & Trenz, M. (2025). *Wie wird KI die Produktivität in Deutschland verändern? Gutachten im Auftrag des Gemeinschaftsausschusses der Deutschen Gewerblichen Wirtschaft*. Institut der deutschen Wirtschaft. <https://www.iwkoeln.de/studien/vera-demary-michael-groemling-christian-kestermann-marc-scheufen-stefanie-seele-oliver-stettes-marco-trenz-wie-wird-ki-die-produktivitaet-in-deutschland-veraendern.html>
- Döbel, I., Leis, M., Manuel, M. V., Neustroev, D., Petzka, H., Riemer, A., Rüping, S., Voß, A., Wegele, M. & Welz, J. (2018). *Maschinelles Lernen. Eine Analyse zu Kompetenzen, Forschung und Anwendung*. Fraunhofer-Gesellschaft. https://www.bigdata.fraunhofer.de/content/dam/bigdata/de/documents/Publikationen/Fraunhofer_Studie_ML_201809.pdf
- Eggebrecht, A., Flemming, J., Meyer, G., Müller, A. V., Oppolzer, A., Paulinyi, A. & Schneider, H. (1980). *Geschichte der Arbeit: Vom Alten Ägypten bis zur Gegenwart*. Kiepenheuer & Witsch.
- Fleck, L., Graus, E. & Klinger, M. (2022). *Verändert Künstliche Intelligenz die Zukunft unserer Arbeit? Wahrnehmungen von betroffenen Arbeitnehmer:innen* (ROA External Reports No. ai:economics Kurzdossier December 2022). ROA. https://www.denkfabrik-bmas.de/fileadmin/Downloads/Publikationen/Veraendert_Kuenstliche_Intelligenz_die_Zukunft_unserer_Arbeit.pdf
- Frey, C. B. & Osborne, M. A. (2013). *The Future of Employment: How Susceptible are Jobs to Computerization?* University of Oxford.
- Giering, O. (2022). Künstliche Intelligenz und Arbeit: Betrachtungen zwischen Prognose und betrieblicher Realität. *Zeitschrift für Arbeitswissenschaft*, 76, 50–64. <https://doi.org/10.1007/s41449-021-00289-0>
- Giering, O. & Kirchner, S. (2021). Künstliche Intelligenz am Arbeitsplatz. Forschungsstand, Konzepte und empirische Zusammenhänge zu Autonomie. *SozW Soziale Welt*, 72(4), 551–588. <https://doi.org/10.5771/0038-6073-2021-4>
- Graus, E., Özgül, P. & Steens, S. (2021). *Künstliche Intelligenz: Die Zukunft der Arbeit anhand von Erkenntnissen aus der Unternehmenspraxis gestalten* (ROA External Reports Nr. November 2021). ROA. https://www.denkfabrik-bmas.de/fileadmin/Downloads/Publikationen/Kuenstliche_Intelligenz_Die_Zukunft_der_Arbeit_anhand_von_Erkenntnissen_aus_der_Unternehmenspraxis_gestalten.pdf
- Hackl, B., Wagner, M., Attmer, L. & Baumann, D. (2017). *New Work: Auf dem Weg zur neuen Arbeitswelt. Management-Impulse, Praxisbeispiele, Studien*. Springer Gabler. <https://doi.org/10.1007/978-3-658-16266-5>
- Heinen, N., Heuer, A. & Schautschick, P. (2017). Künstliche Intelligenz und der Faktor Arbeit. *Wirtschaftsdienst*, 97(10), 714–720. <https://doi.org/10.1007/s10273-017-2203-5>
- Jähnichen, T. (2018). Arbeitsethik 4.0 – Sozialethische Herausforderungen der digitalen Vernetzung der Arbeitswelt. In R. Pelikan & J. Rehm (Hrsg.), *Arbeit im Alltag 4.0*. –

- Wie Digitalisierung ethisch zu lernen ist. Beiträge aus betrieblicher, arbeitsmarktpolitischer und theologischer Sicht* (S. 35–55). LIT Verlag.
- Kellermann, C. & Petring, A. (2019). *Künstliche Intelligenz und Arbeit: Gesellschaftliche Dimensionen einer Technikfolgenabschätzung* (WISO direkt 18/2019). Friedrich-Ebert-Stiftung, Abteilung Wirtschafts- und Sozialpolitik. <https://library.fes.de/pdf-files/wiso/15738.pdf>
- Kerkhof, A., Licht, T., Menkhoff, M. & Wohlrabe, K. (2024). Die Nutzung von Künstlicher Intelligenz in der deutschen Wirtschaft. *Ifo Schnelldienst*, 77(8), 39–43. <https://www.ifo.de/sites/default/files/docbase/docs/sd-2024-08-kerkhof-et-al-ki-in-unternehmen.pdf>
- Klingbeil-Döring, W. (2019). *Der moralische Wert der Arbeit: Zu den Bedingungen und zum Gehalt eines neuen Verständnisses von Arbeit* [Dissertation, Humboldt-Universität zu Berlin]. <https://doi.org/10.18452/19837>
- Lenzen, M. (2018). *Künstliche Intelligenz: Was sie kann & was uns erwartet*. C. H. Beck.
- Lenzen, M. (2020). *Künstliche Intelligenz: Fakten, Chancen, Risiken*. C. H. Beck.
- Poretschkin, M., Schmitz, A., Akila, M., Adilova, L., Becker, D., Cremers, A. B., Hecker, D., Houben, S., Mock, M., Rosenzweig, J., Sicking, J., Schulz, E., Voss, A. & Wrobel, S. (2021). *Leitfaden zur Gestaltung Vertrauenswürdiger Künstlicher Intelligenz. KI-Prüfkatalog*. <https://www.iais.fraunhofer.de/de/publikationen/studien/2021/ki-pruef-katalog.html>
- Richthofen, G., Gümüşay, A.A. & Send, H. (2021). Künstliche Intelligenz und die Zukunft von Arbeit. In R. Altenburger & R. Schmidpeter (Hrsg.), *CSR und Künstliche Intelligenz. Management-Reihe Corporate Social Responsibility* (S. 353–366). Springer Gabler. https://doi.org/10.1007/978-3-662-63223-9_19
- World Economic Forum. (2025). *The Future of Jobs Report 2025*. <https://www.weforum.org/publications/the-future-of-jobs-report-2025/>

KI in der Arbeitswelt

Der Mensch im Spannungsfeld von Entfremdung und digitalem Humanismus

Alexander Schmölz, Pia-Zoe Hahne, Sebastian Klocker & Wolfgang Katzian

1. Einführung

Die Arbeitswelt bietet einen direkten Spiegel zur Veränderung der Gesellschaft durch neue Technologien: Die industrielle Revolution hat die Arbeitsbedingungen grundlegend verändert. Die Entwicklung und der Einsatz von neuen Technologien wie Künstliche Intelligenz (KI) scheinen für einen neuen Umbruch zu sorgen. Der Einsatz von KI in der Arbeitswelt birgt jedoch ambivalente Potenziale, die zwischen Entfremdung (Kassem, 2024; Yates, 2022) und Humanisierung (Fromm, 1980; Kleinöder et al., 2019) oszillieren. In diesem Beitrag wird untersucht, wie KI-Technologien sowohl Prozesse der Entfremdung als auch der Humanisierung von Arbeit vorantreiben können. Die Digitalisierung führt zu erheblichen Veränderungen in den Arbeitsprozessen, deren Ausmaß und Auswirkungen nicht nur technischer, sondern auch sozialer und ethischer Natur sind. KI-Werkzeuge haben das Potenzial, Arbeitsprozesse effizienter zu gestalten, die Produktivität zu steigern, die Arbeit zu erleichtern und Arbeitszeit zu verkürzen. Gleichzeitig entstehen jedoch neue Formen der Entfremdung und Kontrolle, die vor allem in den Bereichen unsichtbarer Arbeit und algorithmischer Steuerung deutlich werden.

Zunächst wird auf Basis theoretischer Überlegungen zu Entfremdung und Humanisierung von Arbeit verdeutlicht, dass Entfremdung vor allem durch Kontrollverlust, Überwachung und fehlende Anerkennung der Arbeit hervorgerufen wird. Ein zentrales Beispiel ist dabei die unsichtbare Arbeit (Hahne, 2025), die zur Datenproduktion und Klassifizierung für KI-Systeme notwendig ist. Diese Arbeit ist weitgehend anonymisiert, schlecht bezahlt und strukturell ausgeblendet (Crawford, 2021; Gray & Suri, 2019; Irani, 2015). Die Betroffenen haben keine Kontrolle über das Produkt ihrer Arbeit und sind von jeglicher Anerkennung ausgeschlossen. Diese dystopischen Entwicklungen beinhalten die totale Institutionalisierung (Goffman, 1973) der Arbeit durch automatisierte und algorithmisierte Überwachungs- und Kontrollsysteme. Diese umfassende Reglementierung führt zu einem digitalen Burnout und verstärkt bestehende Ungleichheiten (Schmoelz & Bauer, 2021). Im Gegensatz dazu wird in einem nächsten Schritt die Humanisierung von Arbeit als Prozess beschrieben, bei dem Technologien zur Förderung von Autonomie, Mitbestimmung und besseren Arbeitsbedingungen eingesetzt werden. Partizipative Gestaltung digitaler Systeme und die Stärkung von Mitbestimmungsrechten stehen hierbei im Vordergrund. Gewerkschaften spielen eine wichtige Rolle, um den digitalen Wandel im Sinne des Digitalen Humanismus mitzugestalten (Katzian & Klocker, 2023). Die Perspektive des Digitalen Humanismus (Nida-Rümelin & Staudacher, 2024; Schmoelz, 2023; Werthner et al., 2019) betont die Möglichkeit, kreative Freiräume zu schaffen, digitale Infrastrukturen partizipativ zu gestalten und Inklusion zu fördern. Ein solches Ver-

ständnis von Digitalisierung zielt darauf ab, technologische Entwicklungen aktiv im Sinne menschlicher Werte und sozialer Gerechtigkeit zu gestalten.

Durch die Verknüpfung des digitalen Humanismus mit der Theorie und Analyse der Entfremdung wird ein kritischer digitaler Humanismus entwickelt, der in zwei bestehenden Studien begründet ist. Ein zentraler Ansatzpunkt für den kritischen digitalen Humanismus liegt in der Erkenntnis, dass technologische Entwicklungen niemals neutral sind, sondern stets gesellschaftlich geformt werden. Christian Fuchs (2022) betont, dass Digitalisierung nicht nur als technischer Prozess verstanden werden darf, sondern immer auch als soziales Phänomen, das mit Machtverhältnissen und ökonomischen Interessen verwoben ist. Aus dieser Perspektive ist Digitalisierung ein ideologisches Projekt, das sowohl zur Unterdrückung als auch zur Befreiung beitragen kann. Alexander Schmölz (2023) stellt klar, dass der zentrale Gegensatz zwischen Neoliberalismus und Digitalem Humanismus in der unterschiedlichen Gewichtung von Effizienz, Leistung und Menschlichkeit liegt. Der Neoliberalismus basiert auf Prinzipien wie Wettbewerb, Effizienz und Performativität, die zunehmend auch auf digitale Infrastrukturen übertragen werden. Dieser Ansatz verstärkt Machtasymmetrien, indem wirtschaftliche Leistungsfähigkeit als zentrales Kriterium für soziale Teilhabe definiert wird. Digitale Infrastrukturen, die auf diesem Prinzip beruhen, dienen primär der Kontrolle, Überwachung und Effizienzsteigerung. Im Gegensatz dazu betont der Digitale Humanismus die Bedeutung von Teilhabe, Mitbestimmung und sozialer Gerechtigkeit. Anstatt Technologie nur als Werkzeug zur Steigerung wirtschaftlicher Effizienz zu betrachten, wird sie als Mittel zur Förderung der Urteilskraft, Autonomie, Kreativität und Inklusion gesehen. Der Digitale Humanismus fordert eine bewusste Gestaltung digitaler Systeme, die nicht nur technische, sondern auch soziale und ethische Dimensionen berücksichtigt. Diese konträren Ansätze machen deutlich, dass ein kritischer Digitaler Humanismus notwendig ist, um der Dominanz neoliberaler Diskurse entgegenzuwirken und alternative Visionen einer gerechten digitalen Gesellschaft zu ermöglichen. Der Digitale Humanismus verfolgt hier das Ziel, Technologien so zu gestalten, dass sie zur Förderung sozialer Gerechtigkeit, Demokratie und individueller Freiheit beitragen. Dies bedeutet, dass technologische Systeme aktiv im Sinne eines kritischen digitalen Humanismus gestaltet werden müssen, um die Entfremdung von Arbeit zu überwinden und Möglichkeiten zur Humanisierung von Arbeit zu eröffnen. Dabei ist zentral, dass hier der Humanismus nicht kritisch gewendet wird, sondern die Kritische Theorie in den Humanismus eingewoben und damit humanistisch gewendet wird. Das bedeutet, dass die zentralen kritischen Dimensionen herangezogen und Bestandteil des humanistischen Theoriegebäudes werden, ohne in den Fatalismus der Dialektik der Aufklärung (Horkheimer & Adorno, 1969) zu verfallen. Der Humanismus bleibt als Leittheorie am Menschen orientiert und zukunftsweisend. Das Potenzial des Menschen, Unterdrückung zu überwinden und Freiheit zu schaffen, bleibt tragfähiges Axiom. Damit bleibt die Möglichkeit für gutes Leben im Schlechten bestehen und es wird verdeutlicht, dass die Zukunft der Arbeit im Kontext von KI kein naturgegebener und zwangsläufiger Prozess ist, sondern maßgeblich von persönlichen, gemeinschaftlichen und gesellschaftlichen Entscheidungen abhängt und von Gründen geleitet ist. Ein kritisch-humanistischer Ansatz zur Digitalisierung ist notwendig, um Entfremdung zu verhindern und zugleich die Potenziale der KI zur Humanisierung von Arbeit auszuschöpfen.

2. Entfremdung durch KI-Technologien

Das Konzept der Entfremdung hat in der Kritischen Theorie eine lange Geschichte. Marx (2009 [1844]) verwendet das Konzept der Entfremdung, um das Problem zu beschreiben, welches für die Arbeiter:innen durch die Industrialisierung entstanden ist. Laut Marx entfremden sich Arbeitnehmer:innen von Arbeitsprozessen, von anderen Arbeitnehmer:innen, vom Endprodukt ihrer Arbeit und letztendlich von sich selbst als Mensch und von der Menschlichkeit. Während eine Person vor der Industrialisierung in der Lage war, den gesamten Produktionsprozess von Anfang bis Ende zu überblicken, sind Arbeitnehmer:innen ab der Industrialisierung oft nur noch für kleine Teilprozesse zuständig, ohne die Verbindung zu übergreifenden Prozessen erkennen zu können. Die Entfremdung von anderen Arbeiter:innen findet statt, da aktuelle Arbeitspraktiken nicht auf Kommunikation und Austausch ausgerichtet sind, sondern sich durch Isolation der Arbeitenden auszeichnen. Während ein:e Arbeiter:in früher mit Kolleg:innen zusammensaß, sich über die verschiedenen Phasen des Produktionsprozesses unterhielt und diskutierte, wird Austausch jetzt teilweise aktiv verhindert, da dieser die Produktivität verringern könnte. Durch die Entfremdung vom Arbeitsprozess und von anderen Arbeitnehmer:innen werden Arbeiter:innen auch von dem Endprodukt entfremdet, zu dem sie beigetragen haben: Obwohl das Produkt, das die Arbeiterin herstellt, für das Funktionieren des Endprodukts von großer Bedeutung sein kann, erhält sie keine aktive Wertschätzung und verschwindet so in der Unkenntlichkeit. Da die Arbeit für viele ein wesentlicher Teil der eigenen Identität ist, führt die Entfremdung von der eigenen Arbeit zur Entfremdung von der eigenen Identität (Marx, 2009 [1844]). Was die Arbeit entfremdet, ist also das Fehlen von Wertschätzung, menschlichen Werten und Mitbestimmung. Diese Unterscheidung verdeutlicht auch, warum Entfremdung zutiefst antihumanistisch ist.

Die Entfremdung von Arbeit durch KI-Technologien zeigt sich besonders deutlich in der sogenannten unsichtbaren Arbeit, die zur Datenproduktion und Klassifizierung für KI-Systeme notwendig ist (Crawford, 2021; Gray & Suri, 2019; Hahne, 2025; Irani, 2015; Newlands, 2021). Hahne betont, dass diese unsichtbare Arbeit eine grundlegende Voraussetzung für das Funktionieren von KI ist, aber gleichzeitig in ihrer gesellschaftlichen Wahrnehmung weitgehend ausgeblendet wird. Diese Form der Arbeit ist oftmals schlecht bezahlt, anonymisiert und unter prekären Bedingungen organisiert. Die Arbeiter:innen, die diese Tätigkeiten verrichten, haben keine Kontrolle über das Produkt ihrer Arbeit und werden von jeglicher Anerkennung ausgeschlossen. Hahne (2025) beschreibt, dass unsichtbare Arbeit insbesondere in Niedriglohnsektoren und durch digitale Plattformen organisiert wird, bei denen die Arbeitsbedingungen durch algorithmische Kontrolle und fehlende Mitbestimmung geprägt sind. Diese digitale Schattenarbeit ist gekennzeichnet durch ihre Fragmentierung und die fehlende Transparenz ihrer Strukturen. Die Entfremdung manifestiert sich hier als ein Zustand, in dem die Arbeitenden ihrer eigenen Arbeit entfremdet werden, weil sie weder ihren Beitrag zur Gesamtleistung erkennen noch Einfluss auf die Bedingungen ihrer Tätigkeit ausüben können. Besonders problematisch ist die Tatsache, dass unsichtbare Arbeit oft als selbstverständlich erachtet wird. Tätigkeiten wie das Klassifizieren von Daten, die Überprüfung von Inhalten oder das Antrainieren von Modellen für maschinelles Lernen bleiben weitgehend unsichtbar und werden in offiziellen Statistiken und gesellschaftlichen Debatten kaum thematisiert.

Hahne (2025) verdeutlicht, dass diese Form der Arbeit nicht nur strukturell unsichtbar gemacht wird, sondern auch aktiv durch unternehmerische Strategien verschleiert wird, um Produktionsprozesse als autonom und selbstgesteuert darzustellen, obwohl sie in erheblichem Maße von menschlicher Arbeit abhängig sind. Solche Arbeitsformen, die als digitale Schattenarbeit bezeichnet werden, sind charakteristisch für die Plattformökonomie (Kassem, 2024) und werden oft durch undurchsichtige Mechanismen der Kontrolle und Überwachung verstärkt (Yates, 2022).

Die Plattformökonomie von Amazon veranschaulicht eindrücklich, wie Entfremdung von Arbeit durch digitale Technologien verstärkt wird. Die Arbeit in Amazons Lagerhäusern basiert auf einer hochgradig rationalisierten und automatisierten Organisation des Arbeitsprozesses, die an Tayloristische Prinzipien angelehnt ist. Die Arbeitsabläufe sind in verschiedene Schritte unterteilt, die strikt voneinander getrennt werden: Inbound (Empfang, Registrierung und Lagerung von Produkten) und Outbound (Kommissionierung und Verpackung). Die Arbeitenden sind gezwungen, monotone und sich ständig wiederholende Aufgaben in hohem Tempo auszuführen, während ihre Produktivität durch Überwachungstechnologien kontinuierlich gemessen wird. Diese Struktur führt zu einer Entfremdung von der eigenen Arbeitstätigkeit, da die Arbeitenden kaum Kontrolle über den gesamten Produktionsprozess haben und ihre Tätigkeit auf einzelne, mechanisch ausgeführte Aufgaben reduziert wird. Die technologische Überwachung erfolgt sowohl durch soziale Kontrolle, beispielsweise durch Vorgesetzte, als auch durch algorithmische Systeme, die Arbeitsleistung kontinuierlich überwachen und bewerten. Darüber hinaus werden die Arbeitenden von dem Produkt ihrer Arbeit entfremdet. Ihre Interaktion mit den Waren ist auf kurze Momente beschränkt, in denen sie die Produkte registrieren, lagern, kommissionieren oder verpacken. Der Zweck und Wert der Waren sind für sie irrelevant, da ihre Arbeit auf maximale Effizienz und Geschwindigkeit ausgerichtet ist. Diese Formen der Entfremdung werden durch die strikte Kontrolle des Arbeitsprozesses verstärkt. Die Einführung von Leistungskennzahlen wie Units Per Hour (UPH) legt fest, wie viele Waren pro Stunde bearbeitet werden müssen. Wer diese Leistung nicht erbringt, wird als „Underperformer“ eingestuft und mit disziplinarischen Maßnahmen konfrontiert. Letztlich werden die Arbeitenden nicht nur von ihrer Arbeit und deren Produkten entfremdet, sondern auch von ihrem eigenen kreativen Potenzial. Ihre Tätigkeit wird zu einem mechanischen Prozess, der keinen Raum für Eigeninitiative oder kreative Gestaltung lässt. Diese Art der Arbeitsorganisation dient nicht der Förderung menschlicher Freiheiten, sondern reduziert Menschen auf Funktionen innerhalb eines größeren, algorithmisch gesteuerten Systems (Kassem, 2024).

Darüber hinaus zeigt sich Entfremdung auch auf einer strukturellen Ebene. Die kontinuierliche Automatisierung und Digitalisierung von Arbeitsprozessen führen zu einer Verlagerung von Machtstrukturen, in denen die Interessen der Beschäftigten oft wenig berücksichtigt werden. Wie Katzian und Klocker (2023) betonen, ist der zunehmende Einsatz von KI in der Arbeitswelt nicht nur ein technologischer, sondern auch ein politischer Prozess, der durch Fragen von Mitbestimmung und Kontrolle geprägt wird. Ein weiteres Problemfeld betrifft die Abhängigkeit von algorithmisch gesteuerten Systemen, die Entscheidungen ohne menschliches Zutun treffen können. Hier entstehen neue Machtverhältnisse, bei denen Beschäftigte den Entscheidungen von Systemen ausgeliefert sind, deren Funktionsweise sie weder nachvollziehen noch beeinflussen können. Dies verstärkt die strukturelle Entfremdung und führt zu einer

Entmenslichung von Arbeit. Yates (2022) betont, dass diese Art der Automatisierung vor allem in Niedriglohnsektoren zu einer Verstärkung von Machtungleichheiten führt.

Diese Entwicklungen der Entfremdung führen in Extremfällen zur totalen Institutionalisierung der Arbeit. Hier wird an Goffmans Konzept der „totalen Institution“ (Goffman, 1973) angeknüpft, die im digitalen Zeitalter durch automatisierte und algorithmisierte Überwachungs- und Kontrollsysteme sowie umfassende Reglementierung von Arbeit verstärkt zur Realität wird und in der die Individuen ihrer Autonomie beraubt und durch umfassende Kontrollmechanismen gesteuert werden. Die vollständige Erfassung aller Arbeitsschritte durch digitale Systeme erinnert an Benthams Konzept des Panoptikums (Bentham, 1791), das durch Überwachung Macht und Kontrolle sichert. Dies zeigt sich besonders deutlich in Szenarien, in denen KI-Technologien zur Überwachung und Kontrolle eingesetzt werden. Algorithmische Entscheidungssysteme, die ohne menschliche Mitbestimmung agieren, schaffen neue Hierarchien und Abhängigkeiten. Besonders problematisch ist dabei die Möglichkeit, dass Arbeitnehmer:innen zu bloßen Datenlieferanten degradiert werden, deren Arbeit unsichtbar bleibt und deren Einfluss auf Arbeitsprozesse minimal ist.

Dabei wird die Gefahr eines individuellen und gesellschaftlichen digitalen Burnouts real, da die Überwachung und Bewertung von Arbeitsleistung durch KI-gestützte Systeme zu einer ständigen Erreichbarkeit und einem erhöhten Leistungsdruck führen können. Dieser Zustand führt zu einem digitalen Burnout und verstärkt bestehende Ungleichheiten, da insbesondere geringqualifizierte Beschäftigte unter den Folgen der Digitalisierung leiden. Der digitale Burnout ist dabei nicht nur ein individuelles, sondern ein strukturelles Problem. Die ständige Verfügbarkeit, die durch digitale Technologien gefördert wird, verwischt zunehmend die Grenzen zwischen Arbeit und Freizeit. Dies hat zur Folge, dass Erholungsphasen verkürzt werden und ein Zustand der chronischen Erschöpfung entsteht. Auch psychische Belastungen nehmen zu, da Arbeitnehmer:innen sich gezwungen sehen, immer produktiver zu arbeiten, um den digitalen Anforderungen gerecht zu werden. Dieser Zustand führt zu einer Degradierung von sinnstiftender Arbeit und zu einem Verlust an Autonomie. Digitale Technologien, die ursprünglich zur Erleichterung von Arbeit gedacht waren, können somit genau das Gegenteil bewirken, wenn sie ohne soziale und ethische Kontrolle eingesetzt werden. Der digitale Burnout ist daher ein Symptom einer zunehmenden Entfremdung, die aus der unreflektierten Anwendung von KI-Technologien resultiert (Schmoelz & Bauer, 2021).

3. Digitaler Humanismus und die Humanisierung von Arbeit durch KI-Technologien

Der Humanismus betont im allgemeinen die Entfaltung von spezifisch menschlichen Eigenschaften (Schmoelz, 2020) und stellt damit die menschliche Autorenschaft und Freiheit (Nida-Rümelin, 2016) sowie Empathie und Authentizität (Hutterer, 1998) in den Mittelpunkt. Humanisierung der Arbeit bezeichnet dabei die Gestaltung von Arbeitsbedingungen, bei denen Fremdbestimmung möglichst durch Eigenentscheidungen ersetzt wird, um Zwangsverhältnisse abzubauen und Freiheit zur Entfaltung menschlicher Eigenschaften zu ermöglichen. Ziel ist es, Arbeitsorganisationen

so zu gestalten, dass Zwänge vermieden werden, um demokratische Strukturen und Freiheit zu fördern. Dabei geht es sowohl um die Reduzierung physischer und psychischer Belastungen als auch um die kritische Auseinandersetzung mit hierarchischen Befehlsstrukturen, die Autonomie und Mitbestimmung einschränken (Müller, 2019). Bei der Humanisierung von digitaler Arbeit ist somit der Prozess gemeint, bei dem digitale Technologien zur Förderung von Autonomie, Mitbestimmung und besseren Arbeitsbedingungen eingesetzt werden. Hier wird insbesondere Fromms Ansatz der Humanisierung der Technik relevant, der darauf abzielt, Technologien in den Diensten der menschlichen Entwicklung einzusetzen (Fromm, 1980). Die partizipative Gestaltung digitaler Systeme und die Stärkung von Mitbestimmungsrechten sind zentrale Forderungen des Digitalen Humanismus. Die Entwicklungen im Bereich des Digitalen Humanismus (Nida-Rümelin & Staudacher, 2024; Schmoelz, 2023; Werthner et al., 2019) betonen die Potenziale im digitalen Raum, menschliche Autorenschaft in den Mittelpunkt zu stellen und kreative Freiräume zu schaffen sowie Inklusion zu fördern. Digitale Systeme haben das Potenzial, Arbeitsprozesse nicht nur effizienter, sondern auch humaner zu gestalten. Dies zeigt sich insbesondere in der Möglichkeit, Arbeit flexibler und selbstbestimmter zu organisieren. Der Digitale Humanismus fordert, dass Technologien so gestaltet werden, dass sie die Würde des Menschen achten und zur Förderung seiner Potenziale beitragen. Ein zentraler Aspekt dieser Bestrebungen ist die Ermöglichung von Mitbestimmung und Teilhabe. Durch digitale Werkzeuge können Mitarbeitende stärker in Entscheidungsprozesse eingebunden werden, was zu einer Demokratisierung von Arbeit führen kann. Wie zum Beispiel

- digitale Abstimmungsplattformen in der Projektarbeit: In Unternehmen werden zunehmend digitale Tools wie kollaborative Abstimmungsplattformen oder digitale Konsenssysteme eingesetzt, etwa in agilen Teams. Mitarbeitende können so in Echtzeit über Arbeitsprozesse, Zeitpläne oder die Priorisierung von Aufgaben mitentscheiden. Anstatt dass Entscheidungen ausschließlich durch das Management getroffen werden, ermöglichen diese Tools eine breite Beteiligung aller Teammitglieder – unabhängig von ihrer hierarchischen Position. Das stärkt die Mitbestimmung und führt zu transparenteren und demokratischeren Entscheidungsprozessen.
- partizipative Gestaltung von KI-Systemen im Betrieb: Bei der Einführung eines KI-gestützten Schichtplanungssystems werden Beschäftigte aktiv in die Entwicklung einbezogen – etwa durch Feedbackrunden, Nutzer:innentests oder die Mitgestaltung der Entscheidungskriterien (z.B. soziale Bedürfnisse, familiäre Verpflichtungen). Die Mitarbeitenden bringen ihre Erfahrungen und Bedarfe ein, wodurch die KI nicht nur effizient, sondern auch fair und mitarbeiter:innenorientiert konfiguriert werden kann. Diese Form der Mitgestaltung stärkt nicht nur die Akzeptanz, sondern auch die Selbstwirksamkeit der Beschäftigten und macht betriebliche Entscheidungen transparenter.

Dabei spielen nicht nur technologische, sondern auch soziale und rechtliche Rahmenbedingungen eine wichtige Rolle. Die demokratische Partizipation ist ein wesentlicher Faktor, um Entfremdung zu verhindern und Humanisierung zu fördern.

Vor diesem Hintergrund argumentieren Katzian und Klocker (2023), dass digitale Technologien nicht zwangsläufig zu Entfremdung führen müssen, sondern auch zur Humanisierung der Arbeitsbedingungen beitragen können. Sie heben hervor, dass insbesondere durch Mitbestimmung, partizipative Gestaltung und kollektive Aushandlungsprozesse menschenzentrierte Arbeitsstrukturen geschaffen werden können. Ein zentraler Punkt ist die Einbindung von Gewerkschaften als Akteure des digitalen Wandels. Es wird betont, dass Gewerkschaften eine bedeutende Rolle dabei spielen, die Digitalisierung aktiv mitzugestalten und im Sinne der Beschäftigten zu beeinflussen. Durch Kollektivverträge, Betriebsvereinbarungen und die Ausarbeitung ethischer Richtlinien können Gewerkschaften dazu beitragen, menschenzentrierte Arbeitsmodelle im digitalen Zeitalter zu etablieren. Des Weiteren wird die Bedeutung von Bildung und Weiterbildung im digitalen Zeitalter herausgestrichen. Nur durch gezielte Förderung digitaler Kompetenzen können Beschäftigte in die Lage versetzt werden, eigenverantwortlich mit neuen Technologien umzugehen und diese aktiv zu gestalten. Dies umfasst sowohl technische Fähigkeiten als auch soziokulturelle Kompetenzen, um die Auswirkungen der Digitalisierung kritisch reflektieren zu können. Die Autoren betonen zudem, dass der Digitale Humanismus ein normatives Konzept darstellt, das sich gegen technokratische Steuerungssysteme richtet und stattdessen eine menschenzentrierte Gestaltung anstrebt. Ziel ist es, technologische Entwicklungen so zu gestalten, dass sie nicht nur ökonomische Effizienz, sondern auch soziale Teilhabe und individuelle Entfaltung ermöglichen. Dies bedeutet, dass die Potenziale von KI gezielt zur Entlastung von monotonen oder gefährlichen Tätigkeiten eingesetzt werden können, während gleichzeitig Kreativität, Autonomie und Zusammenarbeit gefördert werden. Ein zentraler Aspekt dieser Bestrebungen ist die Ermöglichung von Mitbestimmung und Teilhabe. Die Partizipation ist ein wesentlicher Faktor, um Entfremdung zu verhindern und Humanisierung zu fördern. Schließlich wird betont, dass digitale Technologien nicht nur als Werkzeuge zur Produktivitätssteigerung betrachtet werden dürfen, sondern auch als Mittel zur Verbesserung der Lebensqualität. So ist die kritische Reflexion der technologischen Entwicklung notwendig, um sicherzustellen, dass diese im Sinne der Humanisierung gestaltet wird. Die Gestaltung der Zukunft der Arbeit mit KI muss daher als ein sozialer Aushandlungsprozess verstanden werden, an dem verschiedene Akteur:innen aktiv beteiligt sind.

Ein konkretes Beispiel ergibt sich aus neuen Entwicklungen von datengetriebenem Projektmanagement. Projektmanager:innen setzen immer mehr KI-Tools und Datenanalysen ein, um ihre Projekte effizienter und effektiver umzusetzen. Bei all diesen Bestrebungen ist es zentral, dass der Mensch nicht aus den Augen verloren wird. Im datengetriebenen Projektmanagement kann der Digitale Humanismus zunehmend als Orientierungspunkt betrachtet werden, um Technologien menschenzentriert und ethisch verantwortlich einzusetzen. Erfolgreiche Projekte der Zukunft basieren auf einer Kombination aus Technologie, emotionaler Intelligenz, Anpassungsfähigkeit und Zusammenarbeit. Dies bedeutet, dass digitale Werkzeuge nicht nur zur Effizienzsteigerung genutzt werden sollten, sondern auch zur Förderung sozialer Teilhabe und Mitbestimmung. Ein zentrales Prinzip dabei ist die Integration menschlicher Eigenschaften in datengetriebene Entscheidungsprozesse. Während Algorithmen wertvolle Analysen liefern können, bleibt es essenziell, dass Menschen die Interpretation und Nutzung dieser Daten kontrollieren. So wird vermieden, dass Entscheidungen vollständig automatisiert getroffen werden, ohne Berücksichtigung

sozialer und ethischer Aspekte. Projektmanager:innen müssen dabei sicherstellen, dass datengetriebene Systeme nicht nur auf Effizienz, sondern auch auf Kreativität und Zusammenarbeit ausgerichtet werden. Digitale Werkzeuge wie KI-basierte Projektmanagement-Tools ermöglichen es, große Datenmengen schnell zu verarbeiten. Doch ihre humanistische Anwendung erfordert, dass Mitarbeitende in die Entwicklung und Anpassung dieser Systeme eingebunden werden. Ein weiteres Beispiel ist die Bedeutung emotionaler Intelligenz in datengetriebenen Projekten. Während Algorithmen Prozesse optimieren können, ist die menschliche Fähigkeit, Beziehungen aufzubauen und Empathie zu zeigen, von zentraler Bedeutung, um ein Arbeitsumfeld zu schaffen, das sowohl produktiv als auch menschlich ist (Longhurst & Woojin, 2023). Der Digitale Humanismus kann im Projektmanagement daher ein Gleichgewicht zwischen Effizienz und Menschlichkeit fördern, indem digitale Systeme so gestaltet werden, dass sie Transparenz fördern, Mitbestimmung ermöglichen und Raum für Kreativität bieten. Zusammenfassend dient der Digitale Humanismus als Kompass im datengetriebenen Projektmanagement, indem er sicherstellt, dass technologische Entwicklungen nicht nur auf Grundlage von Effizienzoptimierung erfolgen, sondern auch auf Basis ethischer und sozialer Kriterien.

Ein weiteres Beispiel ist das Potenzial menschlicher Deliberation im Prozess der agilen Softwareentwicklung. Agile Softwareentwicklung bietet einen vielversprechenden Rahmen zur Integration ethischer Deliberationen im Sinne des Digitalen Humanismus. Wie Zuber et al. (2024) beschreiben, lassen sich ethische Überlegungen am besten in iterative Entwicklungsprozesse wie Scrum einbetten, da diese auf kontinuierlichem Lernen, Anpassungsfähigkeit und Team-Empowerment beruhen. Die Autor:innen heben hervor, dass ethische Deliberationen in agilen Prozessen in drei Modi stattfinden sollten:

1. Disclosive Contemplation – Identifikation von Werten und ethischen Anforderungen in einer transparenten und nachvollziehbaren Weise
2. Weighing Contemplation – Abwägung konkurrierender ethischer Anforderungen im Entscheidungsprozess, z. B. zwischen kreativer Freiheit im Entwicklungs- und Designprozess und Benutzerfreundlichkeit beim Produkt
3. Applicative Contemplation – praktische Umsetzung ethischer Prinzipien in technische Funktionen.

Ein konkretes Beispiel hierfür ist die agile Entwicklung einer Software oder eines LLM, bei der ethische Anforderungen von Anfang an in den Produkt-Backlog integriert werden. Dies bedeutet, dass bereits in der Planungsphase Werte wie Datenschutz, Transparenz und Fairness als zentrale Anforderungen definiert werden. Während jedes Sprints wird in regelmäßigen Abständen überprüft, ob diese Anforderungen angemessen berücksichtigt wurden. Falls notwendig, werden sie im Dialog mit den Beteiligten angepasst und weiterentwickelt. Diese kontinuierliche Reflexion geschieht idealerweise nicht nur auf technischer Ebene, sondern auch durch regelmäßige ethische Deliberationen innerhalb des Entwicklerteams. Ein sogenannter „Ethics Sprint“ könnte gezielt eingesetzt werden, um ethische Fragen zu diskutieren, bevor neue Features entwickelt oder bestehende geändert werden. Zusätzlich können retrospektive Meetings genutzt werden, um ethische Herausforderungen im Entwicklungsprozess zu identifizieren und für zukünftige Sprints zu berücksichtigen. Durch

diese iterative und reflektierte Vorgehensweise wird nicht nur die technische Qualität des Produkts verbessert, sondern auch seine ethische Verträglichkeit. Dies steht im Einklang mit den Prinzipien des Digitalen Humanismus, der darauf abzielt, technologische Entwicklungen so zu gestalten, dass sie dem Menschen dienen und soziale Gerechtigkeit fördern. Diese kontinuierliche Einbindung ethischer Überlegungen ermöglicht es, agile Prozesse mit den Prinzipien des Digitalen Humanismus zu verbinden, um menschenzentrierte digitale Werkzeuge zu entwickeln.

Beim Ansatz des Digitalen Humanismus werden KI-Technologien und KI-Entwicklungsprozesse demnach so gestaltet, dass sie zur Humanisierung von Arbeit beitragen. Schmoelz und Bauer (2021) betonen, dass dies durch die Förderung von Partizipation, Transparenz und Inklusion erreicht werden kann. Digitale Systeme, die auf demokratischen Prinzipien beruhen und den Mitarbeitenden Mitbestimmungsrechte einräumen, haben das Potenzial, Arbeitsprozesse gerechter und menschlicher zu gestalten. Ein zentraler Aspekt des digitalen Humanismus ist die Möglichkeit, repetitive, gefährliche oder belastende Tätigkeiten zu automatisieren, um den Menschen Raum für kreative und sinnstiftende Arbeit zu schaffen. Dabei kommt es darauf an, dass Technologien nicht nur zur Effizienzsteigerung eingesetzt werden, sondern auch zur Förderung von sozialer Teilhabe und individueller Entfaltung. Dies erfordert eine bewusste Gestaltung von Arbeitsprozessen, bei der digitale Werkzeuge als Mittel zur Unterstützung menschlicher Potenziale genutzt werden. Weiterhin wird hervorgehoben, dass die aktive Gestaltung digitaler Systeme durch Partizipationsmöglichkeiten der Beschäftigten eine entscheidende Rolle spielt. Der Digitale Humanismus wird hierbei als normatives Leitbild verstanden, das darauf abzielt, technologische Entwicklungen im Sinne von sozialer Gerechtigkeit und individueller Freiheit zu gestalten. Ein solcher Ansatz kann dazu beitragen, die Potenziale von KI zur Humanisierung von Arbeit auszuschöpfen und gleichzeitig Entfremdungsprozesse zu verhindern.

4. Fazit

Der Beitrag hat gezeigt, dass der Einsatz von Künstlicher Intelligenz (KI) in der Arbeitswelt ein Spannungsfeld darstellt, das zwischen Entfremdung und Humanisierung oszilliert. Die Potenziale von KI zur Steigerung von Effizienz und Produktivität stehen im Gegensatz zu den Risiken von Entfremdung, die durch unsichtbare Arbeit, Überwachungsmechanismen und algorithmische Kontrolle verstärkt werden. Besonders deutlich wird diese Problematik in der Diskussion um unsichtbare Arbeit. Die systematische Anonymisierung und Prekarisierung dieser Tätigkeiten führen zu einer Entfremdung, bei der die Arbeiter:innen nicht nur von den Produkten ihrer Arbeit, sondern auch von jeglicher gesellschaftlicher Anerkennung ausgeschlossen werden. In dieser Perspektive zeigt sich, dass Entfremdung vor allem dort entsteht, wo Arbeitsprozesse nicht transparent und Machtstrukturen ungleich verteilt sind. Die zunehmende Automatisierung und digitale Kontrolle führt zu einer Verschärfung dieser Entfremdungsprozesse. Die Gefahr der Degradierung von Arbeitenden zu bloßen Datenlieferanten ist eine reale Bedrohung, die in einer technologisch bestimmten Arbeitswelt zu sozialer Ausgrenzung und Ungleichheit führen kann. Dem gegenüber steht die Vorstellung des Digitalen Humanismus. Diese Perspektive sieht in KI-Tech-

nologien nicht nur Werkzeuge der Effizienzsteigerung, sondern auch Mittel zur Humanisierung von Arbeit. Durch Mitbestimmung, Transparenz und Teilhabe können digitale Werkzeuge so gestaltet werden, dass sie individuelle und gemeinschaftliche Entfaltung ermöglichen. Der digitale Humanismus verdeutlicht, dass die Zukunft der Arbeit mit KI kein determiniertes Schicksal ist, sondern aktiv gestaltet werden kann. Die Entwicklung hängt also davon ab, wie gesellschaftliche Akteure die Technologien gestalten und regulieren. Der Mensch steht im Spannungsfeld von Entfremdung und Humanisierung und kann diesen Prozess durch bewusste gesellschaftliche Gestaltung beeinflussen. Es wird deutlich, dass politische, soziale und ethische Rahmenbedingungen notwendig sind, um die Potenziale von KI zur Humanisierung von Arbeit auszuschöpfen und gleichzeitig Entfremdung zu verhindern. Abschließend lässt sich feststellen, dass eine humanistische Gestaltung von KI-Arbeit nicht nur eine Frage technischer Lösungen ist, sondern vor allem eine Frage von Demokratie, Teilhabe und Gerechtigkeit. Die Gestaltung der Zukunft der Arbeit erfordert daher eine kritische Auseinandersetzung mit den gesellschaftlichen Bedingungen, unter denen KI-Technologien eingesetzt werden. Ein humanistischer Ansatz, der die Bedürfnisse und Rechte der Menschen in den Mittelpunkt stellt, ist unerlässlich, um die Potenziale der Digitalisierung positiv zu gestalten. Der Digitale Humanismus bietet hierbei einen Rahmen, um technologische Entwicklungen kritisch zu reflektieren und im Sinne von Demokratie, Teilhabe und Menschlichkeit zu gestalten. Während Entfremdungsszenarien auf zunehmende Kontrolle und Überwachung hinweisen (Goffman, 1973; Schmoelz & Bauer, 2021), betont der digitale Humanismus die Möglichkeit, durch KI kreative Freiräume zu schaffen und Partizipation zu fördern. Der Digitale Humanismus als Konzept strebt danach, technologische Entwicklungen aktiv im Sinne menschlicher Werte und sozialer Gerechtigkeit zu gestalten (Nidarümelin & Staudacher, 2024; Schmoelz, 2023; Werthner et al., 2019).

Mit dem Einweben der Kritischen Theorie in den Digitalen Humanismus entsteht ein kritischer Digitaler Humanismus. Der kritische Digitale Humanismus setzt sich mit der gesellschaftlichen Gestaltung technologischer Entwicklungen auseinander und geht davon aus, dass Digitalisierung nie neutral ist, sondern stets von sozialen, politischen und wirtschaftlichen Interessen geprägt und somit auch Ausdruck politischer Kräfteverhältnisse ist. Dieser Ansatz verweist darauf, dass technologische Fortschritte nicht nur als technische Prozesse verstanden werden dürfen, sondern auch als soziale Phänomene, die mit Machtstrukturen und Ideologien verknüpft sind. Digitalisierung kann daher sowohl als Instrument der Befreiung wie auch der Unterdrückung dienen. Im Gegensatz zum neoliberalen Paradigma, das auf Wettbewerb, Effizienz und Leistungssteigerung fokussiert ist, orientiert sich der Digitale Humanismus an Werten wie sozialer Gerechtigkeit, Teilhabe und demokratischer Mitbestimmung. Neoliberale Konzepte setzen wirtschaftliche Effizienz als Maßstab für gesellschaftliche Teilhabe und entwickeln technologische Infrastrukturen, die primär Kontrolle und Leistungsoptimierung fördern. Der kritische Digitale Humanismus hingegen zielt darauf ab, digitale Systeme so zu gestalten, dass sie menschliche Autonomie, Kreativität, Teilhabe und Emanzipation unterstützen. Um dies erfolgreich durchsetzen zu können, müssen jedoch parallel die Kräfteverhältnisse so gestaltet werden, dass sich die Digitalisierung zu Gunsten aller Menschen entwickeln kann. Das erfordert unter anderem die Stärkung der Gewerkschaften. Denn ein kritischer digitaler

Humanismus wird sich nicht durch den „Good Will“ einiger weniger oder die richtigen Entscheidungen einzelner Personen praktisch vollziehen lassen.

Der kritische Digitale Humanismus ist ein Ansatz, der technologische Entwicklungen aus einer humanistischen und zugleich kritisch-reflexiven Perspektive betrachtet. Er integriert die zentralen Prinzipien der Kritischen Theorie – insbesondere die Auseinandersetzung mit Machtstrukturen, Entfremdung und sozialer Ungleichheit – in ein humanistisches Theoriegebäude. Ziel ist es, technologische Systeme so zu gestalten, dass sie soziale Gerechtigkeit, Ko-Kreativität und individuelle Freiheit fördern. Der Beitrag verdeutlicht, dass die Zukunft der Arbeit im Kontext von KI kein naturgegebener Prozess ist, sondern maßgeblich von politischen, sozialen und technologischen Entscheidungen abhängt. Der Ansatz des Digitalen Humanismus ist notwendig, um Entfremdung zu verhindern und zugleich die Potenziale zur Humanisierung von Arbeit auszuschöpfen. Die Frage, wie KI-Arbeit gestaltet wird, ist letztlich eine Frage nach sozialer Gerechtigkeit und Teilhabe. Es liegt an den Akteur:innen der Gesellschaft, Technologien nicht nur effizient, sondern auch menschlich zu gestalten.

Literatur

- Bentham, J. (1791). *Panopticon or The Inspection House*. Payne.
- Crawford, K. (2021). *Atlas of AI: Power, politics, and the planetary costs of artificial intelligence*. Yale University Press.
- Fromm, E. (1980). *Die Revolution der Hoffnung. Für eine Humanisierung der Technik*. Rowohlt.
- Fuchs, C. (2022). *Digital humanism: A philosophy for 21st century digital society*. Emerald Publishing Limited.
- Goffman, E. (1973). *Asyle: Über die soziale Situation psychiatrischer Patienten und anderer Insassen*. Suhrkamp.
- Gray, M. L. & Suri, S. (2019). *Ghost Work: How to Stop Silicon Valley from Building a New Global Underclass*. HarperCollins Publishers.
- Hahne, P.-Z. (2025). Invisible Labour: Who Keeps the Algorithm Running? *Philosophy and Theory of Artificial Intelligence*, in Print.
- Horkheimer, M. & Adorno, T. W. (1969). *Dialektik der Aufklärung. Philosophische Fragmente*. S. Fischer Verlag.
- Hutterer, R. (1998). *Das Paradigma der humanistischen Psychologie*. Springer.
- Irani, L. (2015). Difference and Dependence among Digital Workers: The Case of Amazon Mechanical Turk. *South Atlantic Quarterly*, 114(1), 225–234. <https://doi.org/10.1215/00382876-2831665>
- Kassem, S. (2024). *Work and alienation in the platform economy: Amazon and the power of organization*. Bristol University Press.
- Katzian, W. & Klocker, S. (2023). Gute Arbeit im digitalen Zeitalter. In G. Krause (Hrsg.), *Die Praxis des Digitalen Humanismus* (S. 19–27). Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-42946-1_2
- Kleinöder, N., Müller, S. & Uhl, K. (Hrsg.). (2019). *»Humanisierung der Arbeit«: Aufbrüche und Konflikte in der rationalisierten Arbeitswelt des 20. Jahrhunderts*. transcript.
- Longhurst, R. & Woojin, C. (2023, 6. November). What the Next Generation of Project Management Will Look Like. *Harvard Business Review*. <https://hbr.org/2023/11/what-the-next-generation-of-project-management-will-look-like>
- Marx, K. (2009). *Ökonomisch-philosophische Manuskripte*. Suhrkamp.

- Müller, M. (2019). Die IG Metall im Diskurs um die Humanisierung des Arbeitslebens. In N. Kleinöder, S. Müller & K. Uhl (Hrsg.), »*Humanisierung der Arbeit*« (S. 255–276). transcript.»*Humanisierung der Arbeit*«
- Newlands, G. (2021). Lifting the curtain: Strategic visibility of human labour in AI-as-a-Service. *Big Data & Society*, 8(1), 20539517211016026. <https://doi.org/10.1177/20539517211016026>
- Nida-Rümelin, J. (2016). *Humanistische Reflexionen*. Suhrkamp.
- Nida-Rümelin, J. & Staudacher, K. (2024). Philosophical Foundations of Digital Humanism. In H. Werthner, C. Ghezzi, J. Kramer, J. Nida-Rümelin, B. Nuseibeh, E. Prem & A. Stanger (Hrsg.), *Introduction to Digital Humanism* (S. 17–30). Springer Nature. https://doi.org/10.1007/978-3-031-45304-5_2
- Schmoelz, A. (2020). Die Conditio Humana im digitalen Zeitalter. Zur Grundlegung des Digitalen Humanismus und des Wiener Manifests. *MedienPädagogik. Zeitschrift für Theorie und Praxis der Medienbildung*, 20, 208–234. <https://doi.org/10.21240/mpaed/00/2020.11.13.X>
- Schmoelz, A. (2023). Digital Humanism, Progressive Neoliberalism and the European Digital Governance System for Vocational and Adult Education. *Journal of Adult and Continuing Education*, 9(2), 735–759. <https://doi.org/10.1177/14779714231161449>
- Schmoelz, A. & Bauer, V. (2021). Digitalisierung, Humanismus und die Zukunft von Arbeit und Berufsbildung: Zentrale Prognosen aus der Berufsbildungsforschung. In R. Löffler, P. Schlögl & A. Schmoelz (Hrsg.), *50 Jahre Berufsbildungsforschung in Österreich im Spannungsfeld von Wissenschaft, Politik und Praxis* (S. 221–236). wbv. <http://dx.doi.org/10.3278/6004915w221>
- Werthner, H., Lee, E., Akkermans, H., Vardi, M., Ghezzi, C., Magnenat-Thalmann, N., Nowotny, H., Hardman, L., Stock, O., Larus, J., Aiello, M., Nardelli, E., Stampfer, M., Frauenberger, C., Ortiz, M., Reichl, P., Schiaffonati, V., Tsigkanos, C., Aspray, W. ... Nalis-Neuner, I. (2019). *Wiener Manifest für Digitalen Humanismus*. https://www.informatik.tuwien.ac.at/dighum/wp-content/uploads/2019/07/Vienna_Manifesto_on_Digital_Humanism_DE.pdf
- Yates, M. D. (2022). *Work work work: Labor, alienation, and class struggle*. Monthly Review Press.
- Zuber, N., Gogoll, J., Kacianka, S., Nida-Rümelin, J. & Pretschner, A. (2024). Value-Sensitive Software Design: Ethical Deliberation in Agile Development Processes. In H. Werthner, C. Ghezzi, J. Kramer, J. Nida-Rümelin, B. Nuseibeh, E. Prem & A. Stanger (Hrsg.), *Introduction to Digital Humanism: A Textbook* (S. 339–358). Springer Nature. https://doi.org/10.1007/978-3-031-45304-5_22

Berufe im Wandel

Das Zusammenspiel von künstlicher Intelligenz und kognitiven Tätigkeiten

Julia Bock-Schappelwein & Sandra Bilek-Steindl

1. Einführung

Large Language Models (LLM) wie ChatGPT haben die Nutzung von künstlicher Intelligenz (KI) innerhalb kürzester Zeit mit enormer Geschwindigkeit vorangetrieben (Bünnagel, 2024). Damit verbunden sind weitreichende Implikationen für die Wirtschaft und insbesondere für den Arbeitsmarkt. Die Fähigkeiten der KI haben aber nicht nur zu beeindruckenden technologischen Durchbrüchen geführt, sondern stellen auch die Grundlagen traditioneller Arbeitsweisen und Arbeitsprozesse in Frage.

Obwohl auch schon früher Wellen technologischer Veränderungen Auswirkungen auf den Arbeitsmarkt hatten, dürfte sich diesmal die Nutzung von KI disruptiver auf den Arbeitsmarkt auswirken, allein schon aufgrund der Geschwindigkeit und weil diesmal Beschäftigungsbereiche unter Druck geraten, die bislang nicht als automatisierbar und damit als den Menschen vorbehalten galten wie etwa kognitive Tätigkeiten (Cazzaniga et al., 2024). Bick et al. (2024) zeigen für die USA, dass sich die generative künstliche Intelligenz schneller verbreitet als ältere digitale Technologien. Die Adoptionsrate von künstlicher Intelligenz beläuft sich zwei Jahre nach der Markteinführung bereits auf 39,5 Prozent und ist damit fast doppelt so hoch wie im Fall des Internets, das nach zwei Jahren eine Rate von 20 Prozent aufwies. In Österreich verdoppelte sich der Anteil der Unternehmen, die KI einsetzen, in nur einem Jahr von 10,8 Prozent auf 20,3 Prozent (2023/24), wie die Erhebung über den Einsatz von Informations- und Kommunikationstechnologien (IKT) in Unternehmen zeigt (Statistik Austria, 2023, 2024).

In der Vergangenheit beschränkte sich die Automatisierung von Prozessen auf Tätigkeiten, die klar vorgegebenen Regeln folgten, wohingegen nun aus der Beobachtung von Eingabedaten und entsprechenden Outputs automatisch Regeln abgeleitet werden, was zur Folge hat, dass mehr Typen von Tätigkeiten automatisiert werden können als in der Vergangenheit (Tolan et al., 2021). Während bei früheren Automatisierungs- und Digitalisierungswellen insbesondere manuelle Routinetätigkeiten unter Druck gerieten (siehe dazu beispielsweise Bock-Schappelwein, 2016), werden durch den Einsatz von KI besonders kognitive Tätigkeiten adressiert.

Vor diesem Hintergrund widmet sich der Beitrag der Frage, auf welche Berufe KI in welcher Art und Weise potenziell einwirken wird. Ziel ist, mithilfe unterschiedlicher Merkmalsausprägungen auf Berufsebene herauszuarbeiten, welche Berufe besonders exponiert erscheinen. Dies umfasst zum einen die Nachfrage nach neuen Tätigkeitsfeldern im Zuge des verstärkten Technologieeinsatzes und zum anderen die mögliche Substitution von Arbeitskräften durch KI. Damit können Perspektiven für den österreichischen Arbeitsmarkt aufgezeigt werden.

Der nachfolgende Abschnitt 2 gibt einen kurzen Überblick zu KI-Technologien und ordnet ihre Anwendungen gemäß bestehender Definitionen bzw. Kategorisierung-

gen ein. Abschnitt 3 umfasst aktuelle Literatur hinsichtlich der Auswirkungen von KI auf Beschäftigung, Produktivität, Branchen, Tätigkeiten und Berufe. Abschnitt 4 bezieht sich auf Berufsfelder, die von KI betroffen sind, und ordnet diese gemäß ihren Tätigkeitsschwerpunkten ein. Im Hinblick auf Berufe mit Schwerpunkt auf kognitiven Routinetätigkeiten, die sich im Jahr 2024 auf rund ein Fünftel der unselbständig Beschäftigten in Österreich beliefen, zeigt sich, dass diese unterschiedlich von KI betroffen sind. Beispielsweise sind Tätigkeiten, die von Bürokräften durchgeführt werden, eher negativ von KI betroffen, wohingegen Gesundheitsberufe oder auch spezifische technische Berufe tendenziell weniger negativ betroffen sein dürften. Abschließend werden Schlussfolgerungen formuliert.

2. Künstliche Intelligenz – eine Begriffsbestimmung

Künstliche Intelligenz (KI) umfasst ein breites Spektrum an Technologien, die es Maschinen ermöglichen sollen, kognitive Fähigkeiten des Menschen nachzuahmen, zu interpretieren, zu handeln und zu lernen (Cazzaniga et al., 2024). Sie basiert auf programmierten Abläufen oder kann durch maschinelles Lernen erzeugt werden, wofür große Datenmengen und hohe Rechenleistung benötigt werden. Maschinelles Lernen zeichnet sich dadurch aus, dass ein Algorithmus durch Wiederholung erlernt, selbständig eine Aufgabe zu erfüllen. Diese Technologie kann in den Bereichen Bild-, Muster- und Spracherkennung, semantische Spracherkennung sowie zur Prozessoptimierung eingesetzt werden. Ein Untergebiet von maschinellem Lernen sind neuronale Netze, die dadurch trainiert werden, dass immer wieder Daten vorgelegt werden und mit der Wiederholung die Daten exakter zugeordnet werden können (Fraunhofer-Institut für Kognitive Systeme IKS, o.D.).

Generative KI basiert auf dem Prinzip des maschinellen Lernens, das es Maschinen ermöglicht, aus vorhandenen Daten zu lernen. Die Nutzung von KI kann eine Erleichterung bei Datenauswertungen oder bei der Herstellung von Beziehungen zwischen Daten, bei der Gefahrenfrüherkennung oder bei der Risikoabwägung darstellen, oder ebenso beim Personalisieren von Kompetenzentwicklung (Bünnagel, 2024). Die generative KI (GenAI) enthält Systeme wie Sprachmodelle, die neue Inhalte (z.B. Texte, Bilder) durch Lernen aus umfangreichen Trainingsdaten erstellen können. Mithilfe von KI können riesige Datenmengen verarbeitet, Muster erkannt und Entscheidungen getroffen werden (Cazzaniga et al., 2024). Large Language Models (LLM) wie etwa ChatGPT (Generative Pre-Trained Transformer) basieren auf sehr großen Datenmengen und auf maschinellem Lernen.

In der Literatur finden sich unterschiedliche Definitionen bzw. Kategorisierungen von KI, wobei De Witt (2024) zwischen starker und schwacher KI unterscheidet. Starke KI zeichnet sich dadurch aus, dass sie Probleme untersucht und eigenständig neue bzw. kreative Lösungen finden kann. Diese Version beschreibt eine menschenähnliche oder menschliche Fähigkeiten übertreffende KI. Die schwache KI dagegen besitzt keine Kreativität und kann nicht eigenständig lernen. Vielmehr durchforstet sie große Datenmengen oder untersucht sie auf Muster. Ein Anwendungsfall ist das maschinelle Lernen (ML). Der Deutsche Ethikrat (2023) unterscheidet neben schwacher und starker KI noch spezielle versus allgemeine KI sowie enge versus breite KI. Ihm zufolge gelten

als enge KI [...] dabei Anwendungen, welche menschliche Fähigkeiten in einer Domäne simulieren, um spezifische Aufgaben zu erfüllen. Breite KI erweitert das Spektrum ihrer Anwendbarkeit über einzelne Domänen hinaus. Der Begriff der starken KI wird für die Vision einer Künstlichen Intelligenz verwendet, die jenseits der möglicherweise perfekten Simulation menschlicher Kognition auch über mentale Zustände, Einsichtsfähigkeit und Emotionen verfügen würde (Deutscher Ethikrat, 2023, S. 19).

Die Abgrenzung zwischen starker und schwacher KI, spezieller und allgemeiner KI sowie enger und breiter KI kann dem Ethikrat zufolge erstens entlang der Breite der Fähigkeiten, über die die KI verfügt, erfolgen und zweitens, ob die „Simulation von Intelligenz mit Intelligenz gleichzusetzen ist oder ob es einen kategorischen Unterschied zwischen der Simulation von Verständnis und genuinem (z.B. sprachlichen) Verständnis gibt, der für ‚echte‘ Intelligenz essenziell ist“ (Deutscher Ethikrat, 2023, S. 121). Enge KI-Anwendungen zeichnen sich dem Ethikrat zufolge dadurch aus, dass die Simulation menschlicher Fähigkeiten in einer Domäne oder die Verwendung von Verfahren wie maschinelles Lernen für die Erfüllung von Aufgaben oder zur Problemlösung verwendet werden. Demgegenüber beschränkt sich die breite KI nicht auf eine Domäne und die starke KI wiederum beschreibt die Vision einer KI, „die jenseits der möglicherweise perfekten Simulation menschlicher Kognition auch über mentale Zustände, Einsichtsfähigkeit und Emotionen verfügen würde“ (Deutscher Ethikrat, 2023, S. 123).

3. Auswirkungen von KI

Die Auswirkungen von KI auf den Arbeitsmarkt sind vielfältig und können sich in unterschiedlichen Formen manifestieren. Aufgrund der rasanten technologischen Entwicklungen sind diese Effekte jedoch schwer vorhersehbar und bislang noch nicht umfassend erforscht. Aktuelle Literatur beschäftigt sich unter anderem mit Auswirkungen auf Beschäftigung, Produktivität, Firmen, Branchen, Tätigkeiten und Berufsfelder (für einen Überblick siehe beispielsweise Felten et al., 2021). Felten et al. (2021) verweisen etwa darauf, dass Fortschritte bei KI dazu führen, dass Arbeitskräfte verdrängt werden, sich Zuständigkeiten im Beruf ändern und ebenso Strategien, organisationales Lernen, Management und Performance von Firmen beeinflusst werden.

3.1 Auswirkungen auf Beschäftigung und Produktivität

In der Literatur wird sowohl auf eine Substitution von Arbeit durch Kapital bzw. KI als auch auf ein komplementäres Verhältnis zwischen diesen beiden Produktionsfaktoren verwiesen, wobei einerseits das Automatisierungspotenzial auf dem Arbeitsmarkt mit Sorge betrachtet wird und andererseits die positiven Auswirkungen von KI auf die Arbeitsproduktivität hervorgehoben werden, wenn sie beispielsweise das

menschliche Beurteilungsvermögen ergänzt und zu Effizienzsteigerungen führt (Ab-rardi et al., 2022; Agrawal et al., 2018).

Aus ökonomischer Sicht erhöhen Zeit- und Kosteneinsparungen die Produktivität. Die Auswirkungen von KI werden in der empirischen Literatur sowohl auf Ebene von Berufsfeldern, Tätigkeiten (siehe dazu Abschnitt 3.2), von Unternehmen als auch auf der makroökonomischen Ebene untersucht.

Acemoglu (2025) überträgt die produktivitätssteigernden Effekte, welche durch Kosteneinsparungen im Bereich der Tätigkeiten gemessen werden, auf die makroökonomische Ebene, um die gesamtwirtschaftlichen Auswirkungen zu ermitteln. Dabei wird gezeigt, dass der Produktivitätsfortschritt auch in der Gesamtwirtschaft und im BIP zu sehen ist, wobei der Umfang der makroökonomischen Folgen noch weiter zu untersuchen ist.

Auch auf Unternehmensebene weisen empirische Studien einen positiven Zusammenhang zwischen Produktivität und KI-Nutzung auf (Calvino & Fontanelli, 2023; Czarnitzki et al., 2023; Marioni et al., 2024) bzw. erwartet die überwiegende Mehrheit von Unternehmen in Deutschland gemäß den Ergebnissen von Unternehmensbefragungen positive Produktivitätseffekte von KI (Kerkhof et al., 2024). Die Datengrundlage für die Nutzung von KI in obigen Studien sind Unternehmensumfragen und KI-Patentdaten. Produktivitätssteigernde Effekte treten nicht immer sofort auf; häufig müssen sich Unternehmen zunächst an die neue Technologie anpassen, neue Geschäftsprozesse entwickeln, Erfahrungen sammeln und ihre Belegschaft schulen (Brynjolfsson et al., 2021). Die positiven Effekte von KI auf die Arbeitsproduktivität sind auch im Zusammenspiel mit der Nutzung anderer digitaler Technologien zu sehen, welche diese verstärken bzw. erst ermöglichen (Calvino & Fontanelli, 2023).

3.2 Auswirkungen auf Berufsfelder und Tätigkeiten

Anders als frühere technologische Transformationsphasen wie Automatisierung oder Digitalisierung, als insbesondere manuelle Routineberufe im unteren und mittleren Qualifikationssegment unter Druck gerieten, dürfte die KI diesmal dazu führen, dass Berufe im oberen Qualifikationssegment von diesem Transformationsprozess betroffen sind bzw. höher bezahlte Tätigkeiten sowie Angestelltenberufe und weniger Arbeiter:innenberufe (Cazzaniga et al., 2024; Hampole et al., 2025; Tolan et al., 2021), und innerhalb der Angestelltenberufe vor allem solche mit wenig sozialer Interaktion (Engberg et al., 2024). Umgekehrt können, wie von Milanez (2023) angemerkt, durch den Einsatz von KI auch direkt Arbeitsplätze entstehen, weil KI-Technologien entwickelt, aktualisiert und gewartet sowie der Umgang mit KI-Technologien geschult werden müssen. Es ist daher nicht auszuschließen, dass durch den Einsatz von KI die Tätigkeiten am Arbeitsplatz umorganisiert werden (Milanez, 2023). Lane und Williams (2023) erwähnen darüber hinaus, dass die Auswirkungen des Einsatzes von KI am Arbeitsplatz von der konkreten Anwendung beeinflusst werden. Wie schon in der Vergangenheit können durch den Einsatz von KI gänzlich neue Aufgaben bzw. Tätigkeiten entstehen bzw. die Ausübung bestehender Tätigkeiten erleichtert werden oder auch Tätigkeiten gänzlich wegbrechen (Acemoglu & Restrepo, 2018, 2020; Lane et al., 2023). Entsprechend differenziert werden auch die Auswirkungen auf die

Beschäftigung ausfallen, in Abhängigkeit davon, wie stark sie der KI ausgesetzt sind (Felten et al., 2018, 2021, 2023).

Deming et al. (2025) zufolge dürften Aufgaben wie das Verfassen von Geschäftsplänen, Artikelüberschriften oder das Schreiben von Softwarecodes von der KI übernommen werden, wohingegen Aufgaben wie Analysieren, Entscheidungsfindung, Abwägen zwischen widersprüchlichen Perspektiven oder der Umgang mit der Belegschaft an Bedeutung gewinnen. Dadurch sollten sich auch die Erwartungen der Unternehmen an die Arbeitskräfte erhöhen. Schnell und Salvi (2024) zufolge kann der Einsatz von KI auch dazu beitragen, dass Arbeitskräfte mit bislang unterdurchschnittlicher Performance die höchsten Produktivitätsfortschritte erlangen sowie die Leistungen von Arbeitskräften merklich gesteigert oder auch die Genauigkeit und Arbeitsqualität verbessert werden. Ähnlich zeigen Brynjolfsson et al. (2023), dass bei Tätigkeiten im Bereich des Kundendienstes der Einsatz von KI vor allem bei geringer qualifizierten und noch weniger erfahrenen Arbeitskräften zu Produktivitätssteigerungen führt.

Autor (2024) wiederum verweist darauf, dass die KI auch dazu beitragen kann, den Wert des menschlichen Fachwissens im mittleren Qualifikationssegment zu erhöhen. Da die KI bei der Entscheidungsfindung unterstützen kann, kann sie für Arbeitskräfte mit komplementären Kompetenzen Entscheidungsfindungsaufgaben übernehmen, die bislang Expert:innen wie etwa Ärzt:innen bei der Befundung oder Rechtsanwält:innen bei der Vertragserstellung vorbehalten waren. Kuhn und Seibert (2025) finden, dass das Substituierbarkeitspotenzial durch den Einsatz neuer Technologien zwar noch immer bei Hilfstätigkeiten oder mittleren Tätigkeiten am höchsten ist, jedoch das Potenzial bei Expert:innenberufen u. a. durch KI am stärksten gestiegen ist.

Besonders tangiert sind kognitive Tätigkeiten bzw. Berufe, die Urteilsvermögen, Problemlösungsfähigkeit oder Dateninterpretation erfordern (Cazzaniga et al., 2024). David Autor betont in einem Interview, dass insbesondere Berufe, die sich auf Routinetätigkeiten wie etwa Produktions-, Büro- und viele administrative Hilfstätigkeiten konzentrieren, betroffen sein werden. Entscheidend ist seiner Einschätzung nach allerdings die Rolle des Fachwissens, das erfahrene Arbeitskräfte haben. Technologische Veränderungen können die Relevanz des Fachwissens in einem Beruf verringern oder umgekehrt auch die Anforderungen an das Fachwissen im Beruf erhöhen. Es zählt aber nicht das Ausmaß der automatisierten Tätigkeiten, sondern vielmehr die Anforderungen an die Fachkenntnisse in diesen wegfallenden Tätigkeiten. Sofern Tätigkeiten, die Fachwissen erfordern, wegfallen, und dadurch der Anteil der Tätigkeiten, die kein solches Fachwissen erfordern, steigt, erhöht sich die Ersatzbarkeit erfahrener Arbeitskräfte und umgekehrt, wenn Routinetätigkeiten automatisiert werden, erhöht sich die Relevanz des fachlichen Erfahrungswissens einer Arbeitskraft (Winters & Latner, 2025).

4. Betroffene Berufsfelder

Hinweise dazu, welche Berufe potenziell von der KI tangiert werden können, liefert die von Felten et al. (2021) entwickelte Messung, ob ein Beruf potenziell der KI ausgesetzt ist, und die von Pizzinelli et al. (2023) entwickelte Messung zu den Kom-

plementaritätspotenzialen, die die erwartbaren substitutiven und komplementären Effekte von KI abbildet, d.h. ob die Arbeitskräfte in einem Beruf bei ihren Arbeitsinhalten durch KI unterstützt oder konkurrenziert werden (Schnell & Salvi, 2024). Felten et al. (2021) verwenden dafür KI-Anwendungen, die von Electronic Frontier Foundation (EFF) in einem Projekt zur Messung des KI-Fortschritts definiert und beschrieben wurden (d.s. abstract strategy games, real-time video games, image recognition, visual question answering, image generation, reading comprehension, language modelling, translation, speech recognition, instrumental track recognition; die Definition dazu findet sich in Felten et al., 2021, S. 2199), und verbinden diese mit Daten von O*NET (Occupational Information Network) in den USA zu den beruflichen Fähigkeiten. Diesen Daten zufolge sind Tänzer:innen, Fitnesstrainer:innen, Bauhilfskräfte oder auch Bauarbeiter:innen am wenigsten stark dem Einsatz von KI ausgesetzt, wohingegen Berater:innen, Finanzprüfer:innen oder auch Versicherungsmathematiker:innen am stärksten ausgesetzt sind. Pizzinelli et al. (2023) verwenden ebenfalls die Daten von O*NET zum Arbeitskontext¹ und Fertigkeiten. Aus der Gegenüberstellung dieser beiden Ansätze, wie stark ein Beruf potenziell der KI ausgesetzt ist sowie den Komplementaritätspotenzialen, kann ein spezifischer Beruf in einer der folgenden vier Gruppen verortet werden (Pizzinelli et al., 2023, S. 8; Schnell & Salvi, 2024, S. 5):

- Geringe Betroffenheit und geringe Komplementarität („wenig tangiert“): kaum Auswirkungen auf den Beruf. Diese Berufe stehen nicht in einem direkten Wettbewerb mit der KI, können diese aber auch nicht zum eigenen Vorteil nutzen (Schnell & Salvi, 2024).
- Geringe Betroffenheit und hohe Komplementarität („begünstigt“): kaum Auswirkungen des Einsatzes von KI aufgrund geringen Spielraums für die Interaktion mit menschlichen Fähigkeiten auf die Nachfrage, wiewohl sich KI unterstützend auf die Arbeit auswirken kann.
- Hohe Betroffenheit und geringe Komplementarität („gefährdet“): höheres Substituierbarkeitsrisiko; weniger Nachfrage in diesem Beruf nicht auszuschließen. Arbeitskräfte werden vom KI-Einsatz stärker tangiert, können aber daraus nur einen geringen zusätzlichen Nutzen ziehen (Schnell & Salvi, 2024).
- Hohe Betroffenheit und hohe Komplementarität („profitierend“): KI stellt eine unterstützende Technologie dar, weshalb unter der Bedingung, dass Zugang zu Technologie und entsprechende Kompetenzen und Fertigkeiten im Umgang mit dieser Technologie verfügbar sind, Produktivitätsgewinne erwartbar sind.

Schnell und Salvi (2024) sowie Jestl (o. D.) zufolge befinden sich in der Gruppe der profitierenden Berufe etwa Führungskräfte, Lehrkräfte, Ingenieur:innen oder auch Jurist:innen; in der Gruppe der begünstigten Berufe beispielsweise Ärzt:innen, Betreuungskräfte oder auch Baufachkräfte. Zu den wenig tangierten Berufen zählen etwa Assistenzberufe im Gesundheitswesen, Bürokräfte mit Kund:innenkontakt, Maschinenbedienung, Montageberufe oder auch Reinigungsberufe. Als gefährdete Berufe

¹ Der Arbeitskontext wird mit 57 Variablen beschrieben, wovon 11 ausgewählt und zu fünf Gruppen zusammengefasst wurden, die als besonders relevant erachtet wurden, dass KI menschliche Tätigkeiten übernimmt.

werden beispielsweise IKT-Techniker:innen oder auch allgemeine Bürokräfte eingestuft.

Die hier identifizierten Berufsfelder können gemäß ihrer Tätigkeitsschwerpunkte zugeordnet werden. Hierfür wird auf die Unterscheidung gemäß Bock-Schappelwein (2016) zurückgegriffen, welche Berufe nach analytischen und interaktiven Nicht-Routinetätigkeiten, kognitiven Routinetätigkeiten, manuellen Routinetätigkeiten und manuellen Nicht-Routinetätigkeiten unterscheidet. Damit lassen sich Berufe bzw. Berufsfelder identifizieren, in denen einerseits kognitive Routinetätigkeiten einen hohen Stellenwert einnehmen und andererseits, aufbauend auf den Daten von Felten et al. (2021) und Pizzinelli et al. (2023)², Hinweise zu finden sind, wie stark diese der Nutzung von KI ausgesetzt sind und ob Komplementaritätspotenzial besteht.

Kognitive Routinetätigkeiten sind in der Kategorisierung von Bock-Schappelwein (2016) sehr breit gefasst und beinhalten sowohl klassische Bürotätigkeiten, spezifische Buchhaltungsaufgaben als auch technische Aufgaben. Sie zeichnen sich dadurch aus, dass Technologie nicht nur ersetzend, sondern auch ergänzend bzw. bei den Aufgaben unterstützend wirken kann. Berufe mit Schwerpunkt auf kognitive Routinetätigkeiten finden sich unter Techniker:innen und gleichrangigen nichttechnischen Berufen in den Bereichen ingenieurtechnische Fachkräfte, Assistenzberufe im Gesundheitswesen, nichtakademische betriebswirtschaftliche und kaufmännische Fachkräfte, Verwaltungsfachkräfte und Informations- und Kommunikationstechniker:innen sowie in den Büroberufen (Bock-Schappelwein, 2016).

Die technischen und gleichrangigen nichttechnischen Berufe zeichnen sich in der Kategorisierung von Bock-Schappelwein und Egger (2023), welche mit Blick auf die Auswirkungen von Digitalisierung auf die Beschäftigung auf den Arbeiten von Fossen und Sorgner (2019) beruht, dadurch aus, dass sie durchwegs vom Einsatz digitaler Technologien profitieren. Dennoch gibt es auch Tendenzen, dass die menschliche Arbeitskraft dadurch ersetzt werden kann. Besonders bei den Büroberufen ist es nicht auszuschließen, dass diese durch den Einsatz digitaler Technologien durchwegs unter Druck geraten dürften.

Auch in den Daten von Felten et al. (2021) und Pizzinelli et al. (2023) zeigt sich ein ähnliches Bild, wonach sich in den technischen und gleichrangigen nichttechnischen Berufen mit Schwerpunkt auf kognitive Routinetätigkeiten das gesamte Spektrum von Betroffenheit von KI und Komplementaritätspotenzial widerspiegelt, d. h. darin finden sich Berufe, die vom KI-Einsatz profitieren oder wo kaum Auswirkungen erwartbar erscheinen wie etwa in Gesundheitsberufen, oder auch solche Berufsfelder, die sowohl profitieren als auch gefährdet sein können wie beispielsweise Fachkräfte in der öffentlichen Verwaltung. Bei den Büroberufen dürften hingegen Arbeitskräfte vom KI-Einsatz stärker tangiert werden, können daraus aber nur einen geringen zusätzlichen Nutzen ziehen.

Hinsichtlich Relevanz für den österreichischen Arbeitsmarkt lässt sich aus der Mikrozensus-Arbeitskräfteerhebung von Statistik Austria für die unselbständige Beschäftigung ableiten, dass im Jahresdurchschnitt 2024 rund ein Fünftel der un-

2 Die Daten zu den Berufen beruhen auf der SOC-Klassifikation (Standard Occupational Classification), die Daten zum Tätigkeitsschwerpunkt auf der ISCO-Klassifikation (International Standard Classification of Occupations). Mithilfe einer Korrespondenztabelle zwischen SOC und ISCO auf Ebene der SOC-6-Steller (SOC 2010) und ISCO-4-Steller (ISCO 08) können diese beiden Datensätze miteinander verknüpft werden. Details dazu finden sich beispielsweise in Bock-Schappelwein und Egger (2023).

selbständig Beschäftigten in Österreich in Berufen mit Schwerpunkt auf kognitive Routinetätigkeiten beschäftigt war. Wiewohl dieser Anteil seit Mitte der 1990er-Jahre relativ stabil ist, so zeigt die rezente Beschäftigungsentwicklung dennoch eine merklich eingetrübte Dynamik im Vergleich zu den analytischen und interaktiven Nicht-Routinetätigkeiten, die etwa Forschen, Analysieren, Planen und Führung sowie kommunikationsspezifische Faktoren wie das Verhandeln, Lehren, Managen oder Präsentieren umfassen. Besonders häufig sind Frauen in Berufen mit Schwerpunkt auf kognitive Tätigkeiten beschäftigt, insbesondere in den Gesundheitsberufen, aber auch in den vom Einsatz von KI besonders betroffenen Büroberufen.

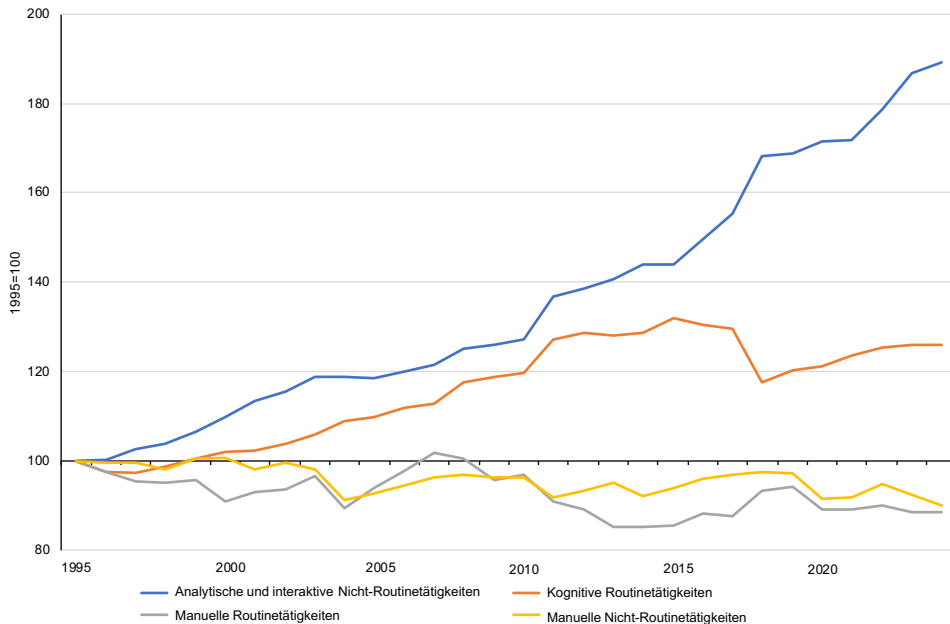


Abb. 1: Unselbständig Beschäftigte nach Tätigkeitsschwerpunkt in Österreich (1995–2024)

Quelle: Statistik Austria: Mikrozensus-Arbeitskräfteerhebung, WIFO-Berechnungen. 2010/2011 Bruch in der Datenreihe. ISCO-88 1995-2010, ISCO-08 2011-2017. Bruch in der Datenreihe 2003/04: Umstellung in der Mikrozensus-Arbeitskräfteerhebung. Darstellung in Anlehnung an Bock-Schappelwein (2016). ISCO-Umstellung 2017/18: Umkodierung ISCO-3-steller 322 nach 222

5. Fazit

Obschon Berufe mit Schwerpunkt auf kognitive Tätigkeiten wie etwa Büroberufe besonders stark vom Einsatz von KI betroffen sein können, so sind die Auswirkungen auf die Beschäftigten dennoch unter dem Blickwinkel zu betrachten, wodurch sich die wegfallenden Tätigkeiten auszeichnen, insbesondere, wie viel Fachwissen darin enthalten ist. Sofern durch den Einsatz von KI Fachwissen ersetzt wird und die spezifischen Fähigkeiten der Arbeitskräfte weniger benötigt werden, können Arbeitskräfte eher unter Druck geraten als in dem Fall, dass durch den Einsatz von KI Routinearbeit ersetzt wird und das Fachwissen der Arbeitskraft an Bedeutung gewinnt.

In einem solchen Umfeld werden lebensbegleitende Berufsbildungs- und Beratungsangebote immer essenzieller, weshalb auch im Erwachsenenalter unterschiedliche Möglichkeiten geboten werden sollten, um einen breiten Zugang zu Beratungs- und Orientierungsangeboten sicherzustellen. Ebenso sollte der Zugang zu und die Absolvierung von entsprechenden Aus-, Weiter- und Höherqualifizierungen wie etwa Praxiserprobung vor Beginn einer Ausbildung auch für Erwachsene (Schnuppertage oder Trainingswochen), zielgruppenspezifische Kommunikation und individuelle Beratungssettings, ausbildungsbegleitende Unterstützungsformate wie Peer-Gruppen, Beratungen oder Coachings, Teilzeitmodelle für die berufliche Aus- und Weiterbildung von Erwachsenen (auch in der Berufsschule) und entsprechende Finanzierungsmodelle unterstützt werden (Bock-Schappelwein et al., 2025).

Literatur

- Abrardi, L., Cambini, C. & Rondi, L. (2022). Artificial intelligence, firms and consumer behavior: A survey. *Journal of Economic Surveys*, 36(4), 969–991. <https://doi.org/10.1111/joes.12455>
- Acemoglu, D. (2025). The simple macroeconomics of AI. *Economic Policy*, 40(121), 13–58. <https://doi.org/10.1093/epolic/eiae042>
- Acemoglu, D. & Restrepo, P. (2018). Artificial Intelligence, Automation, and Work. In A. Agrawal, J. Gans & A. Goldfarb (Hrsg.), *The Economics of Artificial Intelligence: An Agenda* (S. 197–236). University of Chicago Press. <http://www.nber.org/chapters/c14027>
- Acemoglu, D. & Restrepo, P. (2020). The wrong kind of AI? Artificial intelligence and the future of labour demand. *Cambridge Journal of Regions Economy and Society*, 13(1), 25–35. <https://doi.org/10.1093/cjres/rsz022>
- Agrawal, A., Gans, J. S. & Goldfarb, A. (2018). Human Judgment and AI Pricing. *AEA Papers And Proceedings*, 108, 58–63. <https://doi.org/10.1257/pandp.20181022>
- Autor, D. (2024). *Applying AI to Rebuild Middle Class Jobs* (NBER Working Paper 32140). National Bureau of Economic Research. <https://doi.org/10.3386/w32140>
- Bick, A., Blandin, A. & Deming, D. (2024). *The Rapid Adoption of Generative AI* (NBER Working Paper 32966). National Bureau of Economic Research. <https://doi.org/10.3386/w32966>
- Bock-Schappelwein, J. (2016). Digitalisierung und Arbeit. In M. Peneder, J. Bock-Schappelwein, M. Firgo, O. Fritz & G. Streicher (Hrsg.), *Österreich im Wandel der Digitalisierung* (S. 110–126). WIFO.
- Bock-Schappelwein, J., Bergmann, N., Dornmayr, H., Löffler, R., Schmid, K. & Steiner, M. (2025). *Too late?! – Berufs(aus)bildung im Erwachsenenalter* (WIFO Research Briefs 6/2025). <https://www.wifo.ac.at/publication/pid/58161354>
- Bock-Schappelwein, J. & Egger, A. (2023). *Arbeitsmarkt und Beruf 2030. Rückschlüsse für Österreich* (AMS report 173). Arbeitsmarktservice Österreich.
- Brynjolfsson, E., Li, D. & Raymond, L. (2023). *Generative AI at Work* (NBER Working Paper 31161). National Bureau of Economic Research. <https://doi.org/10.3386/w31161>
- Brynjolfsson, E., Rock, D. & Syverson, C. (2020). The Productivity J-Curve: How Intangibles Complement General Purpose Technologies. *American Economic Journal: Macroeconomics*, 13(1), 333–372. <https://doi.org/10.1257/mac.20180386>
- Bünningel, W. (2024). *Wissenskommunikation, maschinelles Lernen und Sprachmodelle: Wie KI das Wissensmanagement verändert*. Springer Gabler. <https://doi.org/10.1007/978-3-662-69913-3>

- Calvino, F. & Fontanelli, L. (2023). *A portrait of AI adopters across countries: Firm characteristics, assets' complementarities and productivity* (OECD Science, Technology and Industry Working Papers No. 2023/02). OECD Publishing. <https://doi.org/10.1787/0fb79bb9-en>
- Cazzaniga, M., Jaumotte, F., Li, L., Melina, G., Panton, A. J., Pizzinelli, C., Rockall, E. J. & Tavares, M. M. (2024). Gen-AI. Artificial Intelligence and the Future of Work. *IMF Staff Discussion Note*, 2024(001). <https://doi.org/10.5089/9798400262548.006>
- De Witt, C. (2024). Künstliche Intelligenz in der Berufsbildung. Technologische Entwicklungen, didaktische Potenziale und notwendige ethische Standards. *Berufsbildung in Wissenschaft und Praxis*, 53(1), 8–12.
- Czarnitzki, D., Fernández, G. P. & Rammer, C. (2023). Artificial intelligence and firm-level productivity. *Journal of Economic Behavior & Organization*, 211, 188–205. <https://doi.org/10.1016/j.jebo.2023.05.008>
- Da Silva Marioni, L., Rincon-Aznar, A. & Venturini, F. (2024). Productivity performance, distance to frontier and AI innovation: Firm-level evidence from Europe. *Journal of Economic Behavior & Organization*, 228, 106762. <https://doi.org/10.1016/j.jebo.2024.106762>
- Deming, D. J., Ong, C. & Summers, L. H. (2025). *Technological Disruption in the Labor Market* (NBER Working Paper 33323). National Bureau of Economic Research. <https://doi.org/10.3386/w33323>
- Deutscher Ethikrat. (2023). *Mensch und Maschine – Herausforderungen durch Künstliche Intelligenz* (Stellungnahme). <https://www.ethikrat.org/publikationen/stellungnahmen/mensch-und-maschine/>
- Engberg, E., Görg, H., Lodefalk, M., Javed, F., Längkvist, M., Monteiro, N. P., Nordås, H. K., Pulito, G., Schroeder, S. & Tang, A. (o.D.). *AI Unboxed and Jobs: A Novel Measure and Firm-Level Evidence from Three Countries* (IZA DP No. 16717). IZA Institute of Labor Economics. <https://www.iza.org/publications/dp/16717/ai-unboxed-and-jobs-a-novel-measure-and-firm-level-evidence-from-three-countries>
- Felten, E. W., Raj, M. & Seamans, R. (2018). A Method to Link Advances in Artificial Intelligence to Occupational Abilities. *AEA Papers and Proceedings*, 108, 54–57. <https://doi.org/10.1257/pandp.20181021>
- Felten, E., Raj, M. & Seamans, R. (2021). Occupational, industry, and geographic exposure to artificial intelligence: A novel dataset and its potential uses. *Strategic Management Journal*, 42(12), 2195–2217. <https://doi.org/10.1002/smj.3286>
- Felten, E. W., Raj, M. & Seamans, R. (2023). *How will Language Modelers like ChatGPT Affect Occupations and Industries?* <https://dx.doi.org/10.2139/ssrn.4375268>
- Fossen, F. & Sorgner, A. (2019). Mapping the Future of Occupations: Transformative and Destructive Effects of New Digital Technologies on Jobs. *Foresight and STI Governance*, 13(2), 10–18. <https://doi.org/10.17323/2500-2597.2019.2.10.18>
- Fraunhofer-Institut für Kognitive Systeme IKS. (o.D.). *Künstliche Intelligenz (KI) und maschinelles Lernen*. <https://www.iks.fraunhofer.de/de/themen/kuenstliche-intelligenz.html>
- Hampole, M., Papanikolaou, D., Schmidt, L. D. W. & Seegmiller, B. (2025). *Artificial Intelligence and the Labor Market* (NBER Working Paper 33509). National Bureau of Economic Research. <https://doi.org/10.3386/w33509>
- Jestl, S. (o.D.). *Potentielle Effekte von Künstlicher Intelligenz am Wiener Arbeitsmarkt: Kurzanalyse der Stadt Wien – Wirtschaft, Arbeit und Statistik*. Stadt Wien – Wirtschaft, Arbeit und Statistik. <https://www.wien.gv.at/wirtschaft/standort/pdf/kurzanalyse-effekte-ki-arbeitsmarkt.pdf>
- Kerkhof, A., Licht, T., Menkhoff, M. & Wohlrabe, K. (2024). Die Nutzung von Künstlicher Intelligenz in der deutschen Wirtschaft. *Ifo Schnelldienst*, 77(8), 39–43. <https://www.ifo.de/sites/default/files/docbase/docs/sd-2024-08-kerkhof-et-al-ki-in-unternehmen.pdf>

- Kuhn, S. & Seibert, H. (2025). *Digitalisierung der Arbeitswelt: Durch künstliche Intelligenz sind inzwischen auch viele Expertentätigkeiten ersetzbar* (IAB-Regional 1/2025). IAB – Institut für Arbeitsmarkt- und Berufsforschung. https://doku.iab.de/regional/BB/2025/regional_bb_0125.pdf
- Lane, M. & Williams, M. (2023). *Defining and classifying AI in the workplace* (OECD Social, Employment and Migration Working Papers No. 290). OECD Publishing. <https://doi.org/10.1787/59e89d7f-en>
- Lane, M., Williams, M. & Broecke, S. (2023). *The impact of AI on the workplace: Main findings from the OECD AI surveys of employers and workers* (OECD Social, Employment and Migration Working Papers No. 288). OECD Publishing. <https://doi.org/10.1787/ea0a0fe1-en>
- Milanez, A. (2023). *The impact of AI on the workplace: Evidence from OECD case studies of AI implementation* (OECD Social, Employment and Migration Working Papers No. 289). OECD Publishing. <https://doi.org/10.1787/2247ce58-en>
- Pizzinelli, C., Panton, A. J., Tavares, M. M., Cazzaniga, M. & Li, L. (2023). Labor Market Exposure to AI: Cross-country Differences and Distributional Implications. *IMF Working Papers*, 2023(216), 1. <https://doi.org/10.5089/9798400254802.001>
- Schnell, P. & Salvi, M. (2024). *Zukunftssichere Berufe?: Wie künstliche Intelligenz den Schweizer Arbeitsmarkt verändert* (avenir suisse analyse). avenir Suisse. <https://www.avenir-suisse.ch/publication/zukunftssichere-berufe/>
- Statistik Austria. (2023). *IKT-Einsatz in Unternehmen 2023. Einsatz von Informations- und Kommunikationstechnologien in Unternehmen*. <https://www.statistik.at/services/tools/services/publikationen/detail/1785>
- Statistik Austria. (2024). *IKT-Einsatz in Unternehmen 2024. Einsatz von Informations- und Kommunikationstechnologien in Unternehmen*. <https://www.statistik.at/services/tools/services/publikationen/detail/2088>
- Tolan, S., Pesole, A., Martínez-Plumed, F., Fernández-Macías, E., Hernández-Orallo, J. & Gómez, E. (2021). Measuring the Occupational Impact of AI: Tasks, Cognitive Abilities and AI Benchmarks. *Journal of Artificial Intelligence Research*, 71, 191–236. <https://doi.org/10.1613/jair.1.12647>
- Winters, J. & Latner, J. (2025, 9. Januar). Does automation replace experts or augment expertise? The answer is yes. *IAB-Forum*. <https://www.iab-forum.de/en/does-automation-replace-experts-or-augment-expertise-the-answer-is-yes/>

Künstliche Intelligenz in der österreichischen Arbeitswelt: Status quo und Perspektiven

Jan-Ove Wiesner, Philipp Besinger, Benedikt Fuchs & Sebastian Schlund

1. Einleitung

Künstliche Intelligenz (KI) hat sich in den letzten Jahren zu einem bedeutenden technologischen Treiber entwickelt und durchdringt zunehmend alle Bereiche von Wirtschaft und Gesellschaft. Seit Durchführung der ersten umfassenden Studie zum Einsatz von KI in Österreich im Jahr 2021 (Fuchs et al., 2022) haben sich sowohl die technologischen Möglichkeiten als auch die Anwendungsfelder von KI weiterentwickelt. Insbesondere generative KI sowie neue Herausforderungen in den Bereichen Datensicherheit und Ethik erfordern eine Aktualisierung und Erweiterung der bisherigen Erkenntnislage.

Im Rahmen der vorliegenden Neuauflage der Studie zum Einsatz von KI in österreichischen Unternehmen wird ein aktuelles und differenziertes Bild hinsichtlich der Implementierung, der Potenziale sowie der Herausforderungen von KI-Anwendungen vermittelt. Aufbauend auf den Ergebnissen der Erhebung aus dem Jahr 2021 werden Entwicklungen und Veränderungen im Implementierungsgrad analysiert. Dabei liegt ein besonderer Fokus auf der Analyse, inwieweit sich bestehende Unterschiede in den Implementierungsgraden und Herausforderungen zwischen Großunternehmen und kleinen sowie mittleren Unternehmen (KMU) fortgesetzt oder verändert haben und welche Fortschritte bei der strategischen Verankerung sowie im Kompetenzaufbau erzielt wurden. Dazu zählen unter anderem technologische Fortschritte im Bereich generativer KI-Systeme und neue regulatorische Rahmenbedingungen wie der EU AI Act (Europäisches Parlament und Rat der Europäischen Union, 2024). Zudem wird der Wandel in der Akzeptanz von KI in Unternehmen und Gesellschaft untersucht, der auch durch öffentliche Debatten über potenzielle Risiken von KI beeinflusst wird.

Die Relevanz dieser Studie ergibt sich aus der zentralen Bedeutung von KI für die Wettbewerbsfähigkeit des Wirtschaftsstandorts Österreich (BMK & BMDW, 2021). Für Unternehmen, Politik und Forschung ist es essenziell zu verstehen, wo Österreich im internationalen Vergleich steht, welche Potenziale noch nicht ausgeschöpft werden und wo Handlungsbedarf besteht. Die Studie liefert dafür eine fundierte empirische Grundlage und identifiziert konkrete Ansatzpunkte zur Förderung und Gestaltung des KI-Einsatzes.

Die Ergebnisse der Studie verdeutlichen die Etablierung von KI in Bereichen wie Informationstechnologie und Kundenservice, während ursprüngliche Erwartungen umfassender Arbeitsplatzautomatisierung bisher nicht eintrafen. Zudem treten neue, bedeutende Fragen zur ökologischen Nachhaltigkeit und Nachvollziehbarkeit algorithmischer Entscheidungen in den Vordergrund. Diese Befunde ermöglichen eine differenzierte Einschätzung des transformativen Potenzials von KI und fundieren strategische Entscheidungen sowie zukünftige Investitionen. Darüber hinaus erweitern sie die Forschung, liefern praxisrelevante Impulse für die KI-Nutzung, schließen

eine Lücke in der österreichischen Datenlage und ermöglichen evidenzbasierte Diskussionen über Chancen, Risiken und Rahmenbedingungen.

2. Überblick über den Datensatz

Der zwischen September 2024 und März 2025 erhobene Datensatz bietet eine empirische Grundlage für die Analyse des Status quo der Nutzung, Implementierung und Wahrnehmung von KI in österreichischen Unternehmen. Im Rahmen einer freiwilligen Onlinebefragung nahmen Mitarbeitende aus 108 Unternehmen unterschiedlicher Branchenzugehörigkeit und Größenklassen aus allen neun Bundesländern Österreichs teil. Die regionale Verteilung der teilnehmenden Unternehmen gewährleistet eine Abdeckung aller neun Bundesländer Österreichs (siehe Abbildung 1), wobei Wien (29% der Teilnehmenden), Steiermark (22%) und Oberösterreich (12%) am stärksten repräsentiert sind. Im Vergleich dazu sind flächen- und bevölkerungsmäßig kleinere Bundesländer wie Burgenland und Vorarlberg mit 3% und 2% der Teilnehmenden weniger stark vertreten.

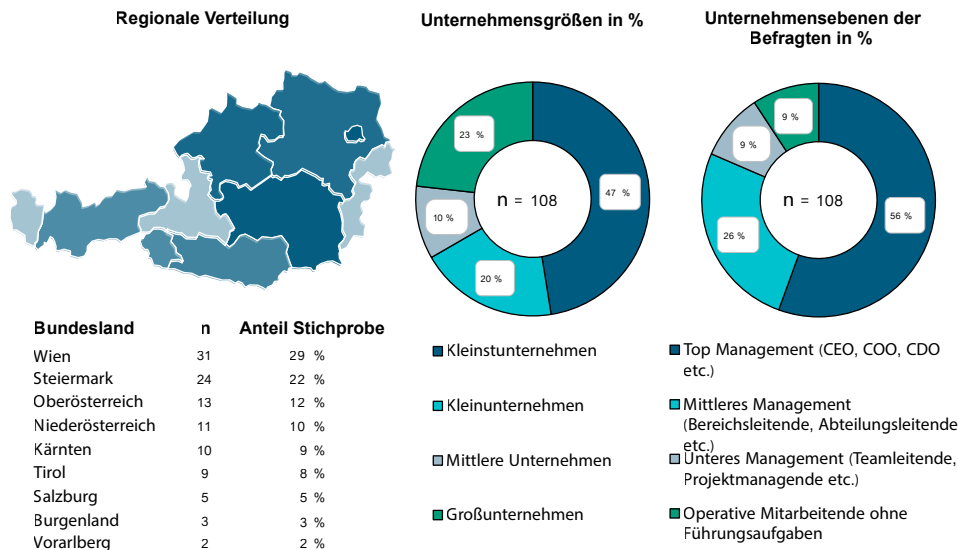


Abb. 1: Darstellung der regionalen Verteilung der repräsentierten Unternehmen in Österreich sowie der prozentualen Verteilung nach Unternehmensgröße und Hierarchieebene der Befragten

Die Stichprobe umfasst Unternehmen verschiedener Größenklassen, deren Einteilung nach den Kriterien des Bundesministeriums für Wirtschaft, Energie und Tourismus (o.D.) erfolgte. KMU stellen mit etwa 77% den größten Teil der Stichprobe dar. Von der Gesamtstichprobe entfallen 47% auf Kleinstunternehmen (weniger als 10 Mitarbeitende), 20% auf Kleinunternehmen (10–49 Mitarbeitende) und 10% auf mittlere Unternehmen (50–249 Mitarbeitende). Großunternehmen (mehr als 250 Mitarbeitende und mehr als 50 Mio. Euro Umsatz) sind mit rund 23% vertreten.

Die Branchenzugehörigkeit der Unternehmen wurde gemäß der österreichischen Klassifikation der wirtschaftlichen Tätigkeiten (Statistik Austria, 2025) bestimmt, die von Statistik Austria zur einheitlichen Zuordnung von Unternehmen zu Wirtschaftszweigen verwendet wird. Die Verteilung der Branchenanteile der Befragten ist in Abbildung 2 dargestellt. Die größten Anteile entfallen auf Information und Kommunikation (30%) und Dienstleistungen (20%). Weitere Branchen sind das Baugewerbe (9%), der Maschinenbau (6%), die Möbelherstellung (6%), das Gesundheits- und Sozialwesen (6%) sowie die öffentliche Verwaltung (5%).

Die Befragten sind mit 56% mehrheitlich im Top-Management (CEO, COO, CDO etc.) vertreten (siehe Abbildung 1). Weitere Anteile entfallen auf das mittlere Management (Bereichsleitende, Abteilungsleitende etc.) mit 26%, das untere Management (Teamleitende, Projektmanagende etc.) mit 9% sowie operative Mitarbeitende ohne Führungsaufgaben mit 9%.

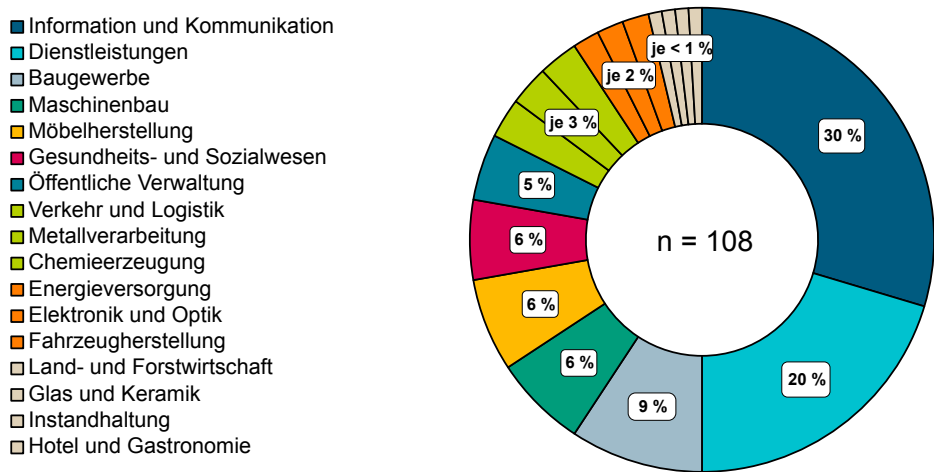


Abb. 2: Darstellung der Branchenverteilung der befragten österreichischen Unternehmen

3. Ergebnisse der Befragung

Dieses Kapitel enthält eine Darstellung der zentralen Ergebnisse der Erhebung zur Anwendung Künstlicher Intelligenz in österreichischen Unternehmen. Es werden der aktuelle Stand der KI-Implementierung, die strategischen Herausforderungen, die Anwendungsbereiche, die Veränderungen in Akzeptanz und Wahrnehmung sowie branchenspezifische Unterschiede dargestellt.

3.1 KI-Implementierung

Die Verteilung des KI-Einsatzes in unterschiedlichen Entwicklungsstadien wird in Abbildung 3 veranschaulicht. Es lässt sich erkennen, dass sich der größte Anteil der befragten Unternehmen (29%) im Stadium des „operativen Einsatzes“ befindet. Dies

impliziert die Integration von KI-Applikationen in bestehende Geschäftsprozesse über Pilotprojekte hinaus. Dies korrespondiert mit der weltweiten Entwicklung. Gemäß dem AI-Index-Report 2025 gaben im Jahr 2024 global 78% der Unternehmen aus der Stichprobe an, KI operativ zu nutzen, verglichen mit 55% im Vorjahr (Maslej et al., 2025). Gleichzeitig lässt die Analyse erkennen, dass 19% der befragten österreichischen Unternehmen das Thema KI gegenwärtig als „nicht relevant“ einstufen. Ferner erachten 14% KI als relevant, planen jedoch gegenwärtig keine konkreten Implementierungsschritte.

Zwischen diesen Stadien lassen sich weitere Phasen differenzieren. So befinden sich 15% der Unternehmen in der Planungsphase, in der KI-Strategien und -Projekte konzipiert, aber noch nicht praktisch umgesetzt werden. Weitere 18% befinden sich in der Pilot- bzw. Testphase, in der konkrete KI-Anwendungen im begrenzten Umfang erprobt und hinsichtlich ihrer Machbarkeit und ihres Nutzens bewertet werden. 6% der Unternehmen befinden sich in der Implementierungsphase.

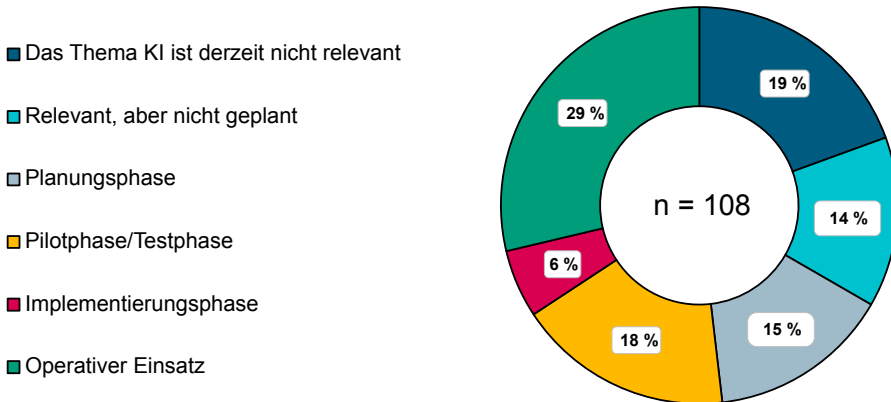


Abb. 3: Verteilung des KI- Implementierungsgrads bei den befragten österreichischen Unternehmen

Die Analyse der KI-Implementierungsgrade differenziert nach Unternehmensgröße verdeutlicht strukturelle Unterschiede im Umgang mit KI (siehe Abbildung 4). Während Großunternehmen tendenziell in den fortgeschritteneren Phasen der Pilotierung (29%) oder des operativen Einsatzes (32%) von KI anzutreffen sind, zeigt sich bei Kleinst- und Kleinunternehmen eine große Diskrepanz. Einerseits stuft ein erheblicher Anteil dieser Unternehmen (ca. 35%) das Thema KI als noch nicht relevant für die eigene Geschäftstätigkeit ein, andererseits berichten rund 37% von einem bereits aktiven und operativen Einsatz von KI-Technologien.



Abb. 4: KI-Implementierungsgrad in Relation zur Unternehmensgröße

Die Analyse des KI-Implementierungsgrads differenziert nach Unternehmensalter (siehe Abbildung 5) zeigt einen Zusammenhang zwischen dem Alter eines Unternehmens und dem Umgang mit KI. Jüngere Unternehmen (1–10 Jahre) führen mit einem höheren Anteil von 58% im operativen KI-Einsatz, während dieser Anteil mit zunehmendem Unternehmensalter kontinuierlich abnimmt. Unternehmen älter als 51 Jahre nutzen KI operativ nicht und Unternehmen älter als 31 Jahre stufen KI häufiger als „nicht relevant“ oder als „relevant, aber nicht geplant“ ein.



Abb. 5: KI-Implementierungsgrad in Relation zum Unternehmensalter

Im branchenspezifischen Vergleich ergeben sich Unterschiede hinsichtlich Implementierung von KI (siehe Abbildung 6). Der IT-Sektor weist mit 44 % operativ tätigen Unternehmen und 13 % in aktiver Implementierungsphase die höchsten Anteile auf. In anderen Branchen zeigen sich abweichende Muster. Im Baugewerbe überwiegt die Planungsphase mit 30 %, gefolgt von 20 % der Unternehmen, die das Thema KI derzeit als nicht relevant einstufen.

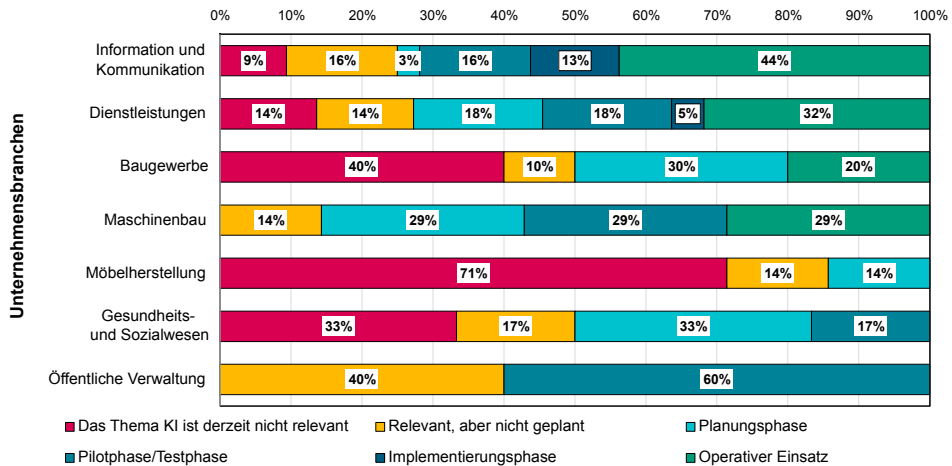


Abb. 6: KI-Implementierungsgrad in den unterschiedlichen Branchen mit mehr als fünf Befragten aus der jeweiligen Branche

Eine durchgeführte Korrelationsanalyse ergab einen schwachen positiven Zusammenhang zwischen dem KI-Implementierungsgrad und angegebener Umsatzentwicklung in den letzten drei Jahren ($r = .11$, $p = .459$). Dieser Zusammenhang war statistisch nicht signifikant ($p > .05$), was darauf hindeutet, dass in der vorliegenden Stichprobe keine Evidenzen dafür gefunden wurden, dass ein höherer KI-Implementierungsgrad mit einer signifikant stärkeren Umsatzentwicklung einhergeht.

Des Weiteren zeigte sich ein schwacher negativer Zusammenhang zwischen dem KI-Implementierungsgrad und der angegebenen Entwicklung der Mitarbeitendenanzahl in den letzten drei Jahren ($r = -.14$, $p = .364$). Auch dieser Zusammenhang war statistisch nicht signifikant ($p > .05$), was bedeutet, dass es keine signifikanten Hinweise darauf gibt, dass ein höherer KI-Implementierungsgrad mit einer deutlichen Reduktion der Mitarbeitendenanzahl verbunden ist. Die Ergebnisse werden in Abbildung 7 dargestellt.

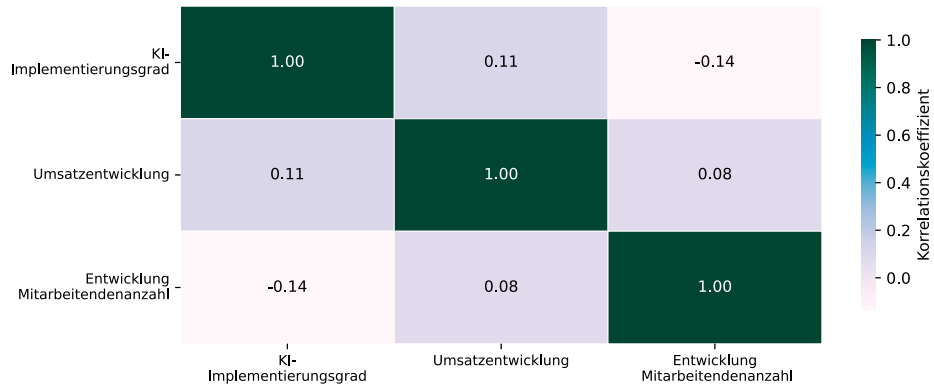


Abb. 7: Korrelationsmatrix zur Analyse der Zusammenhänge von Umsatzentwicklung und Entwicklung der Mitarbeitendenanzahl der letzten drei Jahre mit dem KI-Implementierungsgrad

3.2 Strategische Verankerung von KI

Die Analyse des aktuellen Status der Strategieentwicklung ist in Abbildung 8 dargestellt. Eine Betrachtung der Entwicklung zeigt, dass im Jahr 2021 der Anteil der Unternehmen, die keine explizite Strategie zur Implementierung von KI verfolgten, bei etwa 55 % lag (Fuchs et al., 2022). Zum aktuellen Zeitpunkt sind es 40 % der befragten Unternehmen, die keine KI-Strategie verfolgen. Gleichzeitig geben 14 % der Unternehmen an, sich in der Planungsphase der Strategieentwicklung zu befinden, und 14 % führen eine aktive Strategieentwicklung durch. Des Weiteren verfolgen 9 % der Unternehmen festgelegte Strategien in einzelnen Bereichen, während 14 % angeben, ihre Strategien weitgehend umzusetzen. Bei 9 % der Befragten ist der Stand der strategischen Planung nicht bekannt.

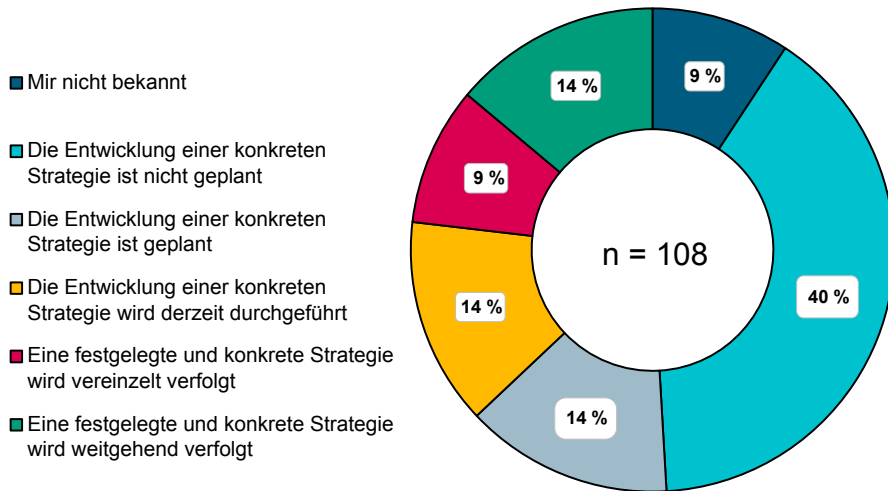


Abb. 8: Verteilung des Status der Strategieentwicklung bei den befragten österreichischen Unternehmen

Neben der aktuellen Strategieentwicklung ist die Einschätzung der zukünftigen Bedeutung von KI ein wichtiger Aspekt der strategischen Überlegungen. Um diese Einschätzung zu erfassen, wurden die Befragten gebeten, die Bedeutung von KI für ihr Unternehmen in fünf Jahren auf einer Skala von 0 (sehr unbedeutend) bis 10 (sehr bedeutend) zu bewerten. Die Verteilung der Antworten, aufgeschlüsselt nach Unternehmensebenen, ist in Abbildung 9 dargestellt. Der arithmetische Mittelwert der Gesamteinschätzungen liegt bei 6,75, wobei das mittlere und untere Management mit einem Mittelwert von 7,7 die höchsten durchschnittlichen Potenzialbewertungen abgaben, gefolgt von Befragten aus dem operativen Bereich ohne Führungsaufgaben (6,8) und dem Top-Management (6,2).

Zur Analyse der Beziehungen zwischen der Strategieentwicklung für KI, der Einschätzung der zukünftigen Bedeutung von KI und weiteren relevanten Faktoren wurde eine Korrelationsmatrix erstellt. Die Korrelationsanalyse in Abbildung 10 zeigt einen mittleren positiven Zusammenhang zwischen der Strategieentwicklung und dem KI-Implementierungsgrad ($r = .39$, $p < .001$). Dieser Zusammenhang ist statistisch signifikant. Ebenso wurde ein mittlerer positiver Zusammenhang zwischen dem KI-Implementierungsgrad und der Bedeutung von KI in fünf Jahren festgestellt ($r = .48$, $p < .001$) sowie ein mittlerer positiver Zusammenhang zwischen der Strategieentwicklung und Bedeutung von KI in fünf Jahren ($r = .33$, $p = .002$). Des Weiteren wurde ein sehr hoher positiver Zusammenhang zwischen dem generellen KI-Einfluss auf die Marktposition und dem Einfluss generativer KI auf die Marktposition festgestellt ($r = .76$, $p < .001$), was darauf hindeutet, dass Unternehmen, die einen starken Einfluss von KI auf ihre Marktposition sehen, diesen Effekt maßgeblich auch generativer KI zuschreiben.

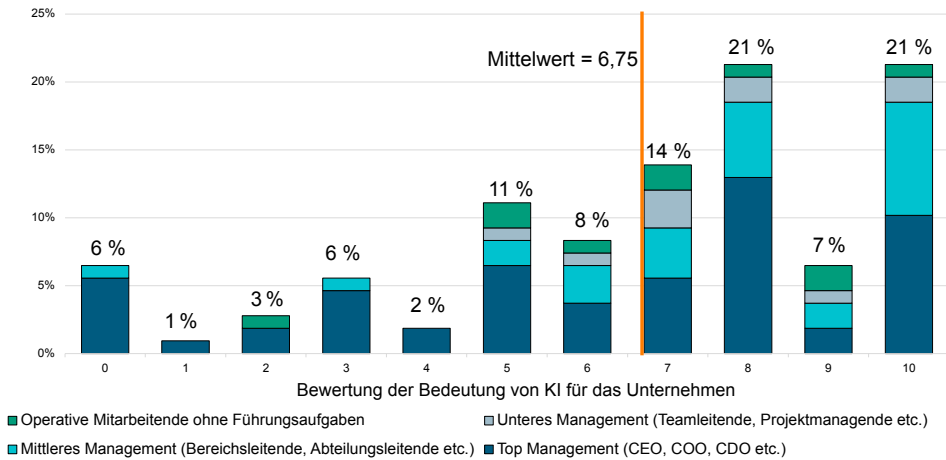


Abb. 9: Einschätzung der zukünftigen Bedeutung von KI für das Unternehmen in fünf Jahren, differenziert nach Unternehmensebenen (0 = sehr unbedeutend, 10 = sehr bedeutend)

Im Hinblick auf das Unternehmensalter zeigten sich keine statistisch signifikanten Zusammenhänge mit der Strategieentwicklung ($r = -.15$, $p = .170$), dem KI-Implementierungsgrad ($r = -.15$, $p = .175$), der Bedeutung von KI in fünf Jahren ($r = -.02$, $p = .863$), dem generellen KI-Einfluss auf die Marktposition ($r = -.06$, $p = .590$) oder dem Einfluss generativer KI auf die Marktposition ($r = -.13$, $p = .238$), was darauf schließen lässt, dass das Alter der Unternehmen in dieser Stichprobe keinen direkten Einfluss auf diese KI-bezogenen Aspekte zu haben scheint.

Ähnlich verhielt es sich mit der Unternehmensklasse, die keine statistisch signifikanten Zusammenhänge mit der Strategieentwicklung ($r = .05$, $p = .643$), dem KI-Implementierungsgrad ($r = -.18$, $p = .092$), der Bedeutung von KI in fünf Jahren ($r = .09$, $p = .433$), dem generellen KI-Einfluss auf die Marktposition ($r = -.06$, $p = .603$) oder dem Einfluss generativer KI auf die Marktposition ($r = -.07$, $p = .503$) aufwies, was darauf schließen lässt, dass die Unternehmensgröße in dieser Stichprobe keinen direkten signifikanten Einfluss auf die genannten KI-bezogenen Variablen zu haben scheint.

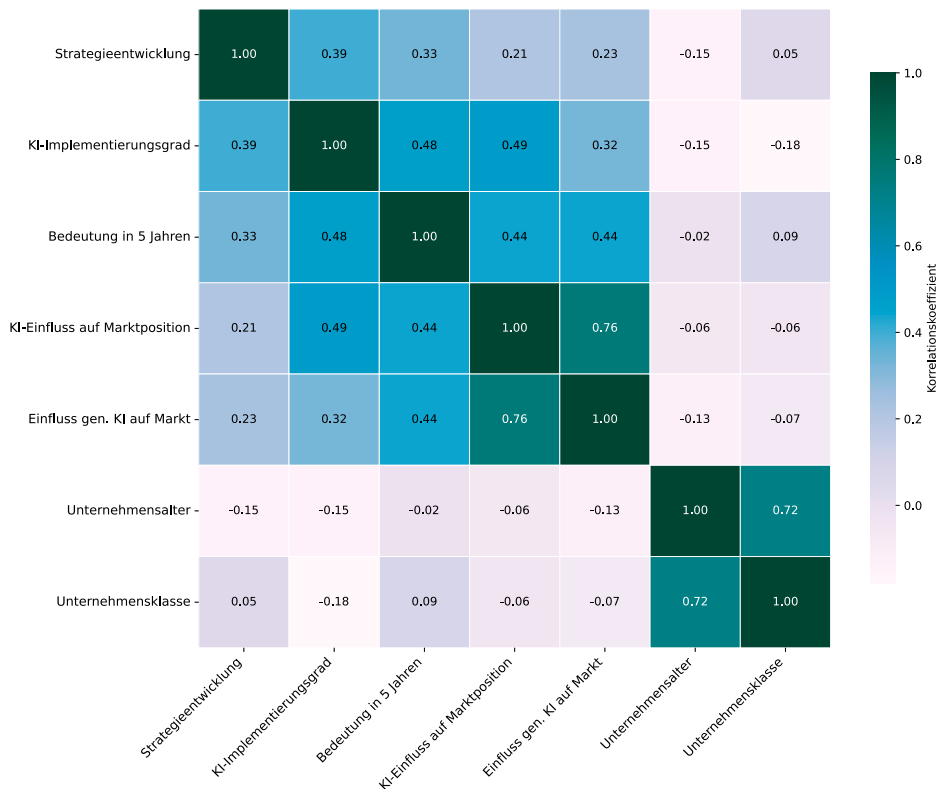


Abb. 10: Darstellung der Korrelationskoeffizienten zwischen verschiedenen Variablen zur strategischen Verankerung und Bedeutung von KI

3.3 Einschätzung des KI-Potenzials

Um die aktuelle Einschätzung des Potenzials von KI in unterschiedlichen Unternehmensbereichen zu erfassen, bewerteten die Teilnehmenden der Befragung das Potenzial von KI für den jeweiligen Unternehmensbereich. Die in Abbildung 11 dargestellten Ergebnisse verdeutlichen die Variabilität in der Potenzialbewertung für unterschiedliche Unternehmensbereiche. Ein hohes Potenzial („eher hoch“ oder „sehr hoch“) wird vor allem im Marketing und in der IT gesehen. Im Marketing geben 74% der Befragten ein hohes Potenzial an (31% „eher hoch“, 43% „sehr hoch“), während es in der IT 67% sind (27% „eher hoch“, 40% „sehr hoch“). In mehreren weiteren Bereichen liegt der Anteil der Befragten, die ein hohes Potenzial sehen, ebenfalls über 50%. Dies betrifft die Administration (59% „eher hoch“ oder „sehr hoch“), die Dienstleistungen (56% „eher hoch“ oder „sehr hoch“), die Forschung & Entwicklung (55% „eher hoch“ oder „sehr hoch“), den Kundendienst (54% „eher hoch“ oder „sehr hoch“) und die Planung (53% „eher hoch“ oder „sehr hoch“).

Ein geringes Potenzial („sehr gering“ oder „eher gering“) wird am häufigsten für die Montage (43%) genannt. Es folgen die Fertigung (22%) und die Beschaffung (20%).

Die Unsicherheit bei der Potenzialbewertung („kann ich nicht beurteilen“) ist in der Montage (34%), Fertigung (31%) und Logistik (25%) am größten.

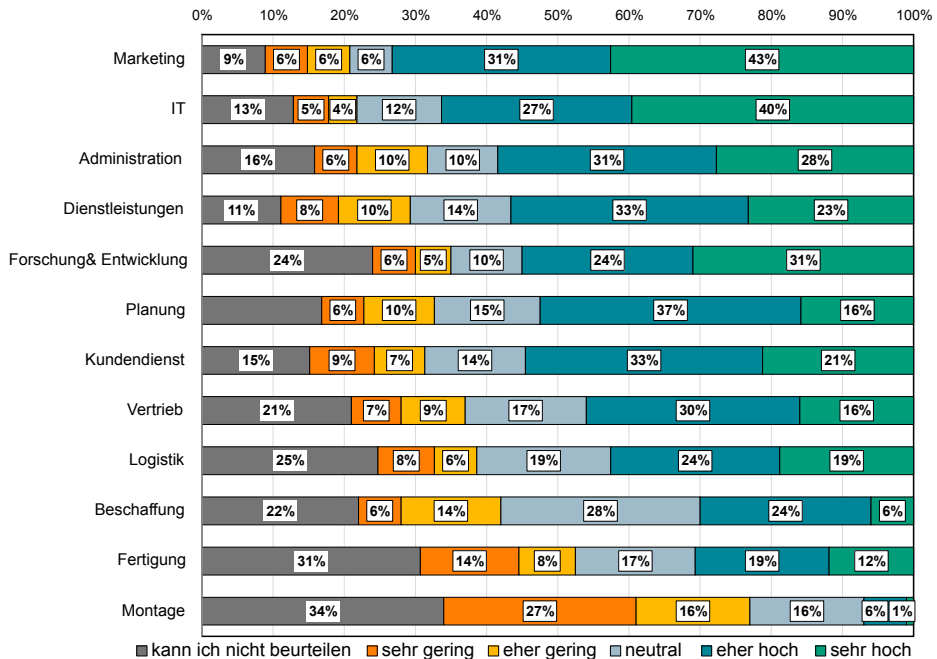


Abb. 11: Wahrgenommene Potenziale in unterschiedlichen Unternehmensbereichen

3.4 Veränderung in Akzeptanz und Wahrnehmung

Die Akzeptanz und Wahrnehmung von KI in österreichischen Unternehmen haben sich seit der Befragung von Fuchs et al. aus dem Jahr 2021 verändert. Der Anteil jener Unternehmen, die KI als derzeit nicht relevant einstufen, ist von 36 % im Jahr 2021 auf 19 % gesunken. Diese Entwicklung spiegelt sich auch in der grundsätzlichen Einstellung gegenüber der Einführung von KI-Anwendungen zur Unterstützung von Mitarbeitenden wider. Die aktuelle Befragung zeigt, dass diese überwiegend positiv bewertet wird (siehe Abbildung 12). Auf einer Skala von 0 bis 10 beträgt der Mittelwert etwa 7,3, wobei dieser 2021 noch bei etwa 6,5 lag (Fuchs et al., 2022). 57 % der Befragten vergeben einen Wert von 8 oder größer, während 10 % einen Wert kleiner als 2 angeben. Der Wert 10 wurde mit 31 % am häufigsten vergeben.

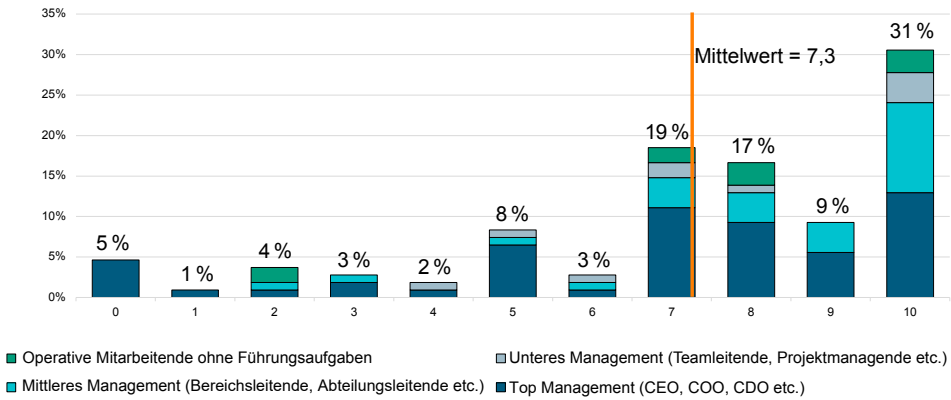


Abb. 12: Bewertung der KI-Unterstützung für Mitarbeitende durch verschiedene Unternehmensebenen (0 = „sehr negativ“ bis 10 = „sehr positiv“)

3.5 Herausforderungen und Risiken

Zur besseren Einschätzung der Hürden bei Auswahl, Implementierung und Anwendung von KI bewerteten die teilnehmenden Unternehmen eine vorgegebene Auswahl möglicher Herausforderungen. Die Ergebnisse sind in Abbildung 13 dargestellt. Besonders häufig wurde der Datenschutz (z. B. DSGVO) genannt, den 52 % der Befragten (n=56) als Hemmnis identifizierten. Betrachtet man die einzelnen Implementierungsphasen, so zeigt sich, dass der Datenschutz in allen Phasen als relevant wahrgenommen wird, jedoch besonders im „operativen Einsatz“ (25 %) eine große Rolle spielt. Ebenfalls von Bedeutung ist das Management von KI-bezogenen Risiken, das von 51 % der Unternehmen (n=55) als Herausforderung genannt wurde. Auch hier zeigt sich ein Trend, dass das Risikomanagement im „operativen Einsatz“ (25 %) im Vergleich zu früheren Phasen an Bedeutung gewinnt. Ein weiterer wichtiger Aspekt betrifft die fehlenden Kompetenzen der Mitarbeitenden. Dies wurde von 45 % der Befragten (n=49) als Herausforderung genannt. Die Identifizierung passender Use Cases wird vor allem in den frühen Phasen („Planungsphase“: 28 %, „Pilot-/Testphase“: 19 %) als Herausforderung wahrgenommen. Die Datenverfügbarkeit, -qualität und -aufbereitung gewinnt mit fortschreitender Implementierung an Bedeutung und wird im „operativen Einsatz“ (33 %) am häufigsten genannt. Die Integration von KI in bestehende Prozesse und IT-Landschaften wird besonders häufig (30 %) als Herausforderung genannt, wo sich das Unternehmen im operativen Einsatz befindet und die Implementierungsphase bereits durchlaufen wurde.

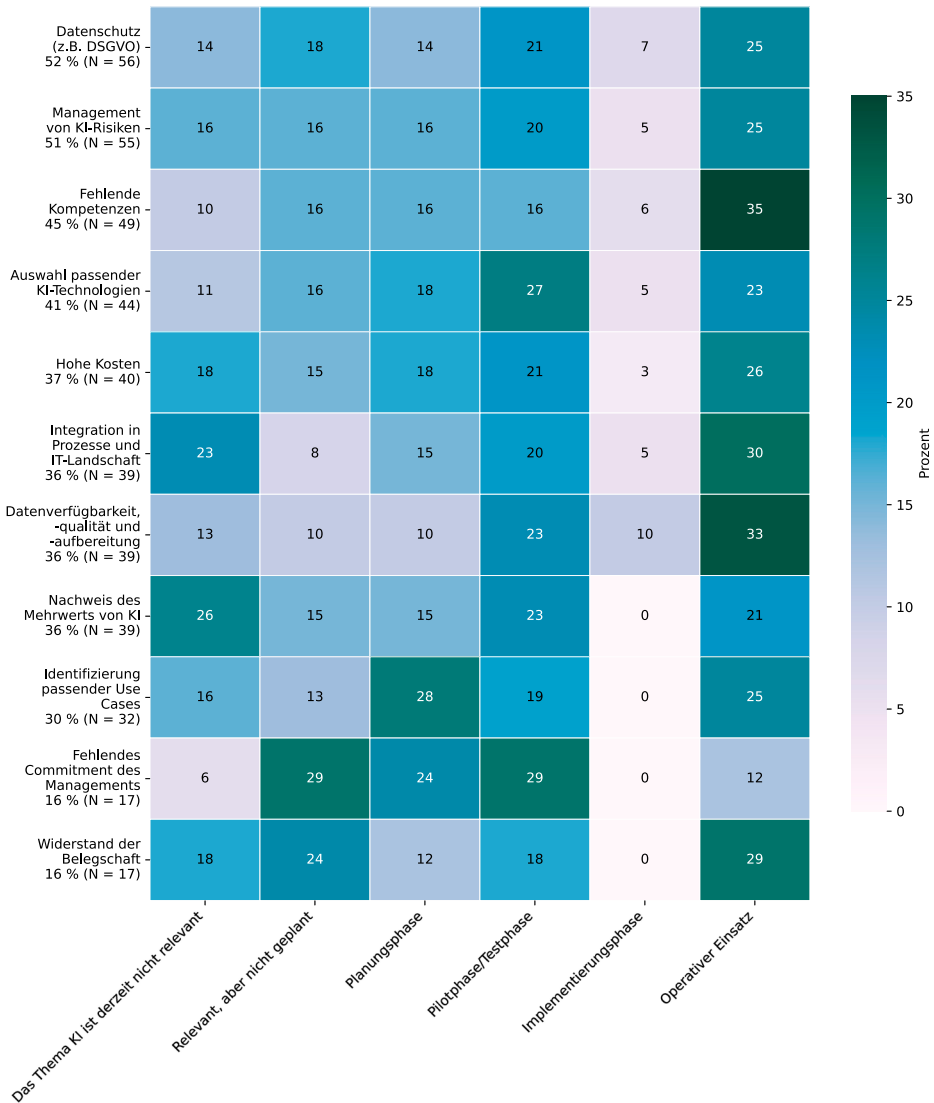


Abb. 13: Wahrgenommene Herausforderungen bei der KI-Implementierung in Abhängigkeit vom Implementierungsgrad

Um die Komplexität der Herausforderungen und die vielfältigen Bewältigungsstrategien der Unternehmen besser zu verstehen, wurde eine Netzwerkanalyse durchgeführt. Die in Abbildung 14 ersichtliche Analyse visualisiert Beziehungen zwischen verschiedenen Begriffen, die im Zusammenhang mit den von Unternehmen genannten Strategien zur Bewältigung von Herausforderungen bei der Implementierung von KI auftreten. Die Positionierung und die Verbindungen der Knoten zeigen die beobachteten Assoziationen zwischen den verschiedenen Maßnahmen und Konzepten. Ihre Größe spiegelt die Häufigkeit ihrer Nennungen in den freien Textantworten wi-

der. So bietet diese Analyse Einblicke in die Struktur der Bewältigungsansätze und die Zusammenhänge zwischen den verschiedenen Handlungsfeldern.

Im Zentrum des Netzwerks stehen die Begriffe „KI“, „Schulung“, „Workshop“ und „Mitarbeiter“. Diese Fokussierung auf bildungsorientierte Maßnahmen korrespondiert unmittelbar mit den Ergebnissen der Studie von Fuchs et al. (2022), die „fehlende Kompetenzen bei Mitarbeitenden“ als eine der Hauptbarrieren (40% der Nennungen) für die erfolgreiche KI-Implementierung identifizierte. Begriffe wie „Strategie“, „Management“, „Unternehmen“ und „Abteilung“ sind im Netzwerk präsent, aber weniger zentral als „Schulung“ und „Mitarbeiter“. Die Verbindung zwischen „Abteilung“, „Thema“ und „Projekt“ zeigt, dass die Auseinandersetzung mit KI-bezogenen Themen und Projekten häufig auf Ebene einzelner Abteilungen innerhalb der Unternehmen stattfindet. Die Verbindungen zwischen „Prozess“, „extern“, „Unternehmen“, „Suche“ und „Beratung“ zeigen, dass österreichische Unternehmen bei der Bewältigung von Herausforderungen im KI-Implementierungsprozess auch externe Ressourcen und Expertise in Betracht ziehen, indem sie nach externer Beratung suchen.

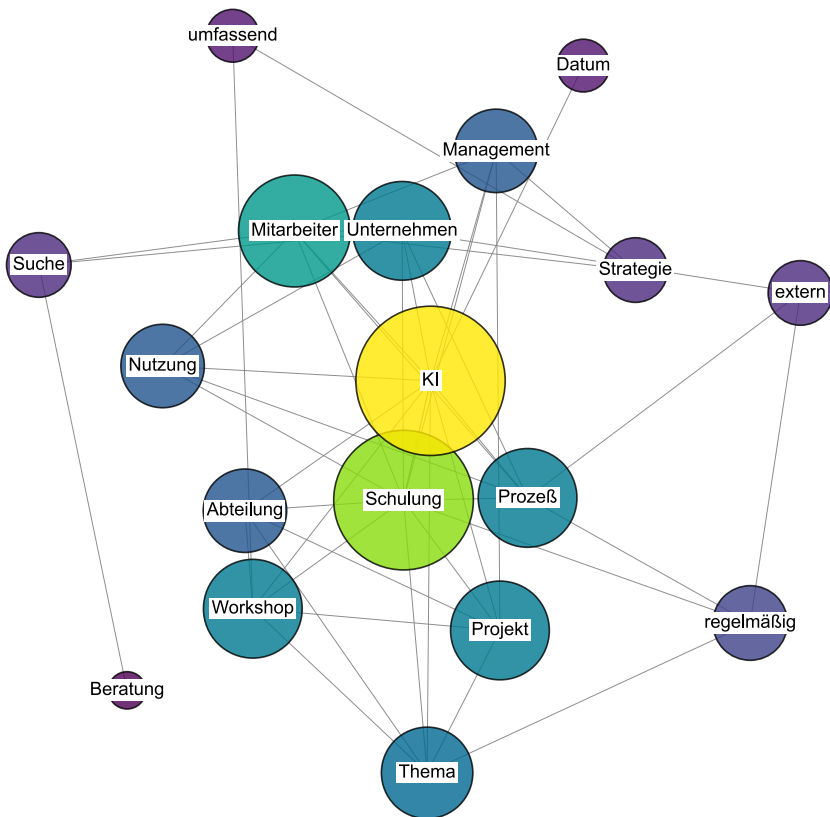


Abb. 14: Visualisierung der Beziehungen zwischen Strategien und Maßnahmen zur Bewältigung von KI-Implementierungsherausforderungen in österreichischen Unternehmen

4. Perspektiven und Handlungsempfehlungen

Die Analyse der Ergebnisse zeigt einen Wandel im unternehmerischen Denken bezüglich KI. Der Rückgang der Unternehmen, die KI als nicht relevant einstufen, deutet auf einen wachsenden Innovations- und Digitalisierungsdruck hin. Österreichische Unternehmen erkennen zunehmend die Notwendigkeit, sich mit KI auseinanderzusetzen, um ihre langfristige Wettbewerbsfähigkeit zu sichern.

Gleichzeitig wird eine differenzierte Potenzialbewertung von KI in verschiedenen Unternehmensbereichen deutlich. Während KI-Anwendungen in unterstützenden Bereichen wie IT und Marketing aufgrund ihres sichtbaren und quantifizierbaren Nutzens auf höhere Akzeptanz stoßen, deutet die Zurückhaltung in traditionelleren operativen Bereichen darauf hin, dass die Herausforderung in der Entwicklung konkreter und überzeugender Use Cases liegt. In diesen Bereichen sind entsprechende Anwendungsfälle seltener oder weniger offensichtlich.

Die identifizierten Herausforderungen, insbesondere der Mangel an KI-Kompetenzen, die hohen Kosten und der unklare Mehrwert, verdeutlichen die Komplexität der KI-Implementierung. Hierbei spielt für die Befragten der Datenschutz eine wesentliche Rolle und wird branchenübergreifend als größte Hürde wahrgenommen. Die zentrale Rolle von „Schulung“ bei der Bewältigung dieser Herausforderungen unterstreicht die Notwendigkeit eines umfassenden Kompetenzaufbaus. Dieser Kompetenzaufbau sollte neben technischem Know-how auch organisatorische und prozessbezogene Aspekte berücksichtigen.

Zudem deutet die unterschiedliche Gewichtung der Herausforderungen in den einzelnen Implementierungsphasen auf einen dynamischen Prozess hin. Die Prioritäten verschieben sich im Zeitverlauf, was eine flexible und anpassungsfähige Herangehensweise bei der Planung und Umsetzung von KI-Projekten erfordert. Ein weiterer wichtiger Aspekt ist die unterschiedliche Perspektive der Befragten aus verschiedenen Unternehmensebenen. Das Ergebnis, dass das mittlere und untere Management die Bedeutung von KI für das Unternehmen in fünf Jahren tendenziell höher einschätzt als das Top-Management (Mittelwert von 7,7 vs. 6,2), deutet darauf hin, dass die strategische Priorisierung und Investitionen in KI-Technologien unterschiedlich wahrgenommen werden.

Um die skizzierten Herausforderungen zu bewältigen und die Potenziale von KI bestmöglich zu nutzen, sind gezielte Maßnahmen verschiedener Akteure erforderlich. Diese Maßnahmen können auch dazu beitragen, die Prinzipien der „Positive Impact Production“ (Sobottka et al., 2024) zu unterstützen, die eine nachhaltige und gesellschaftlich positive Produktion anstrebt. Tabelle 1 fasst die Handlungsempfehlungen für die Implementierung von KI-Anwendungen zusammen.

Tab. 1: Handlungsempfehlungen für die Implementierung von KI-Anwendungen an Wirtschaft, Wissenschaft und Gesellschaft

Akteure	Handlungsempfehlungen
Unternehmen (insbesondere KMU)	– Entwicklung einer klaren und umfassenden KI-Strategie, die sowohl kurz- als auch langfristige Ziele und Prioritäten setzt – Erhöhung der Investitionen in die Aus- und Weiterbildung der Mitarbeitenden, um die notwendigen KI-Kompetenzen aufzubauen – Auf- und Ausbau eines effektiven Managements von KI-bezogenen Risiken, insbesondere im Hinblick auf Datenschutz und ethische Aspekte
Politik	– Schaffung von Rahmenbedingungen, die die KI-Adaption in der Wirtschaft fördern, beispielsweise durch Anreize für Investitionen in KI oder den Abbau bürokratischer Hürden – Bereitstellung von Fördermitteln für Forschung und Entwicklung im Bereich KI, um Innovationen voranzutreiben – Entwicklung von Richtlinien und Leitlinien, die den ethischen Einsatz von KI-Technologien gewährleisten
Forschungseinrichtungen	– Durchführung von anwendungsbezogener Forschung zu den Auswirkungen von KI auf österreichische Unternehmen, insbesondere in Bezug auf branchenspezifische Anforderungen und KMU-Bedürfnisse – Entwicklung von KI-Lösungen und Bewertungsmetriken, die speziell auf die Bedürfnisse von KMU zugeschnitten sind – Etablierung von Partnerschaften mit Unternehmen, um gemeinsame Forschungsprojekte durchzuführen und den Wissenstransfer zu fördern
Bildungseinrichtungen	– Integration von KI-relevanten Inhalten in Lehrpläne, um Studierende und Auszubildende auf die Anforderungen des KI-Arbeitsmarktes vorzubereiten – Angebot von Weiterbildungsprogrammen für Fachkräfte, um bestehende Kompetenzen im Bereich KI auszubauen – Förderung des interdisziplinären Dialogs zwischen verschiedenen Fachbereichen, um ein umfassendes Verständnis der KI-Technologie zu ermöglichen
Technologische Anbieter	– Entwicklung benutzerfreundlicher und intuitiv bedienbarer KI-Lösungen, die auch von KMU leicht implementiert werden können – Fokussierung auf transparente und erklärbare KI-Modelle, um das Vertrauen in die Technologie zu stärken – Bereitstellung von umfassendem Support und Schulung für Unternehmen, um die KI-Implementierung zu begleiten

Die branchenspezifischen Unterschiede im KI-Implementierungsgrad werfen ebenfalls wichtige Fragen nach den zugrundeliegenden Faktoren auf. Die fortgeschrittene Implementierung im IT-Sektor könnte auf dessen Nähe zu den KI-Technologien und den dortigen Innovationsdruck zurückzuführen sein. Die beobachtete Zurückhaltung in traditionelleren Branchen deutet möglicherweise auf spezifische Hemmnisse wie mangelnde Ressourcen, fehlendes Know-how oder eine geringere wahrgenommene Dringlichkeit hin. Diese Heterogenität birgt das Risiko einer wachsenden digitalen Kluft zwischen den Branchen. Daher ist die Entwicklung branchenspezifischer Strategien entscheidend, um sicherzustellen, dass alle Wirtschaftszweige von den Potenzialen der KI profitieren können.

Schließlich spiegelt die starke Fokussierung auf das Potenzial von KI für Prozessverbesserungen ein eher instrumentelles Verständnis von KI wider. KI wird primär als Werkzeug zur Effizienzsteigerung und Kostensenkung wahrgenommen, weniger als Möglichkeit zur grundlegenden Transformation von Geschäftsmodellen oder zur Stärkung der Mitarbeitenden. Diese Tendenz zeigt sich beispielsweise in der hohen Einschätzung des KI-Potenzials in Bereichen wie IT und Marketing, wo der Fokus oft auf Automatisierung von Aufgaben, Datenanalyse für gezieltere Kampagnen und ähnlichen Effizienzsteigerungen liegt. Auch die Tatsache, dass die Herausforderung „Nachweis des Mehrwerts von KI“ von vielen Unternehmen genannt wird, deutet darauf hin, dass der ökonomische Nutzen von KI im Vordergrund steht. Diese Sichtweise könnte Chancen für eine ganzheitlichere Nutzung von KI ungenutzt lassen. Es ist daher wichtig, auch die Potenziale von KI für die Unterstützung und Entlastung der Mitarbeitenden sowie für die Entwicklung neuer, innovativer Dienstleistungen zu erschließen.

Aus den Erkenntnissen ergeben sich für die Forschung mehrere relevante Untersuchungsfelder. Ein zentrales Thema ist die Entwicklung von Methoden zur Messung und Bewertung der Effektivität von KI-Systemen. Unternehmen benötigen verlässliche Metriken zur Quantifizierung des Mehrwerts von KI-Investitionen in unterschiedlichen Dimensionen von Produktivitätssteigerung und Kosteneinsparung über Qualitätsverbesserung und Mitarbeitendenzufriedenheit bis hin zum Innovationspotenzial. Ein weiteres wichtiges Forschungsfeld betrifft die Entwicklung von KI-Systemen, die spezifisch auf die Bedürfnisse kleiner und mittlerer Unternehmen (KMU) zugeschnitten sind. Es deutet sich an, dass KMU bei der Implementierung von KI-Lösungen tendenziell zurückbleiben. Hier sind Forschungsansätze gefragt, die kostengünstige, skalierbare und leicht integrierbare KI-Lösungen für KMU entwickeln. Schließlich sollte die Forschung verstärkt die ethischen und gesellschaftlichen Implikationen fortschrittlicher KI-Systeme thematisieren. Mit zunehmender Autonomie und Entscheidungsfähigkeit von KI-Systemen gewinnen Fragen der Verantwortlichkeit, Transparenz und ethischen Grenzen an Bedeutung. Hier erscheinen interdisziplinäre Forschungsansätze notwendig, die technologische Entwicklungen mit ethischen und rechtlichen Überlegungen in Einklang bringen.

5. Zusammenfassung

Die vorliegende Neuauflage der KI-Studie 2025 bietet eine differenzierte Analyse der Entwicklung und Implementierung von KI in österreichischen Unternehmen. Die in Tabelle 2 dargestellten Ergebnisse verdeutlichen sowohl Fortschritte als auch weiterhin bestehende Herausforderungen im Vergleich zur Erhebung von 2021.

Ein wesentlicher Fortschritt zeigt sich im Implementierungsgrad von KI-Anwendungen. Während 2021 lediglich 9% der Unternehmen KI im operativen Betrieb nutzten, hat sich dieser Anteil bis 2025 auf 29% erhöht. Dieser Zuwachs spiegelt die wachsende Bedeutung und Akzeptanz von KI-Technologien innerhalb der österreichischen Wirtschaft wider. Ebenso hat sich die strategische Ausrichtung der Unternehmen verändert. Im Jahr 2021 gaben noch 55% der Befragten an, keine Entwicklung einer KI-Strategie begonnen zu haben, während 2025 nun die Mehrheit der Unternehmen zumindest in Ansätzen eine solche Strategie verfolgt. Dies deutet auf eine reifere und proaktivere Herangehensweise an die KI-Integration hin.

Die Studie dokumentiert auch eine Ausweitung der Anwendungsbereiche von KI. Neben den etablierten Einsatzgebieten wie IT, Dienstleistungen und Forschung & Entwicklung finden KI-Lösungen vermehrt Anwendung in Bereichen wie Marketing, Personalwesen und Logistik. Dies unterstreicht das zunehmende Verständnis für die vielfältigen Einsatzmöglichkeiten von KI in unterschiedlichen Unternehmensfunktionen.

Allerdings verdeutlichen die weiterhin als eines der Haupthindernisse genannten fehlenden Mitarbeitendenkompetenzen (von 45% der Unternehmen identifiziert) die Notwendigkeit gezielter Maßnahmen zur Unternehmens- und Mitarbeitendenentwicklung im Bereich KI. Trotz der wachsenden Akzeptanz von KI auf strategischer und operativer Ebene bleibt der Aufbau des notwendigen Know-hows im Umgang mit KI-Technologien eine zentrale Aufgabe. Die Studie identifiziert somit einen klaren Bedarf an Forschungsarbeiten und Unterstützungsangeboten, die auf die Entwicklung von KI-Kompetenzen in der Belegschaft und die strategische Verankerung von KI in den Unternehmensstrukturen abzielen.

Tab. 2: Vergleich der Studienergebnisse zum KI-Einsatz in österreichischen Unternehmen von 2022 und 2025

Vergleichsbereich	Studie Fuchs et al., 2022	Aktuelle Studie 2025	Entwicklung
KI-Strategieentwicklung	55 % der befragten Unternehmen gaben an, keine Strategieentwicklung zu verfolgen.	40 % der befragten Unternehmen gaben an, keine Strategieentwicklung zu verfolgen.	Rückgang der Unternehmen ohne Strategieentwicklung.
KI-Implementierungsgrad (operativer Betrieb)	9 % der befragten Unternehmen haben KI-Anwendungen im operativen Betrieb.	29 % der befragten Unternehmen haben KI-Anwendungen im operativen Betrieb.	Deutlicher Anstieg des Anteils der Unternehmen, die KI operativ nutzen.
Akzeptanz von KI-Anwendungen zur Mitarbeitendenunterstützung (Mittelwert)	Mittelwert von ca. 6,5 (auf einer Skala von 0–10).	Mittelwert von ca. 7,3 (auf einer Skala von 0–10).	Zunahme der positiven Bewertung von KI-Anwendungen zur Mitarbeitendenunterstützung.
Einstufung KI als nicht relevant für das eigene Unternehmen	36 % der befragten Unternehmen geben an, dass das Thema aktuell keine Relevanz hat.	19 % der befragten Unternehmen geben an, dass das Thema aktuell keine Relevanz hat.	Deutlicher Rückgang der Unternehmen, für die KI nicht relevant ist.
Fehlende Kompetenzen als Herausforderung	Fehlende Kompetenzen der Mitarbeitenden werden von 40 % der Unternehmen als eine der Hauptbarrieren identifiziert.	Fehlende Kompetenzen der Mitarbeitenden werden von 45 % der Unternehmen als eine der Hauptbarrieren identifiziert.	Fehlende Kompetenzen der Mitarbeitenden bleiben eine der Hauptbarrieren für die operative KI-Nutzung.

Literatur

- Bundesministerium für Klimaschutz, Umwelt, Energie, Mobilität, Innovation und Technologie (BMK) & Bundesministerium für Digitalisierung und Wirtschaftsstandort (BMDW). (2021). *Strategie der Bundesregierung für Künstliche Intelligenz. Artificial Intelligence Mission Austria 2030 (AIM AT 2030)*. <https://www.digitalaustria.gv.at/Strategien/Strategie-der-Bundesregierung-f-r-K-nstliche-Intelligenz--AIM-AT-2030-.html>
- Bundesministerium für Wirtschaft, Energie und Tourismus (BMWET). (o.D.). *KMU in Österreich*. Bundesministerium für Wirtschaft, Energie und Tourismus. <https://www.bmwet.gv.at/Services/Zahlen-Daten-Fakten/KMU-in-%c3%96sterreich.html>
- Europäisches Parlament & Rat der Europäischen Union. (2024). *Verordnung (EU) 2024/1689 des Europäischen Parlaments und des Rates vom 13. Juni 2024 zur Festlegung harmonisierter Vorschriften für künstliche Intelligenz und zur Änderung der Verordnungen (EG) Nr. 300/2008, (EU) Nr. 167/2013, (EU) Nr. 168/2013, (EU) 2018/858, (EU) 2018/1139 und (EU) 2019/2144 sowie der Richtlinien 2014/90/EU, (EU) 2016/797 und (EU) 2020/1828 (Verordnung über künstliche Intelligenz)*. Amtsblatt der Europäischen Union. <https://eur-lex.europa.eu/eli/reg/2024/1689/oj/deu>

- Fuchs, B., Schumacher, A., Eggeling, E. & Schlund, S. (2022). *Künstliche Intelligenz in Österreichs Unternehmen. Fraunhofer Austria KI-Studie*. Fraunhofer Austria Research GmbH. <https://doi.org/10.24406/austria-n-647745>
- Maslej, N., Fattorini, L., Perrault, R., Gil, Y., Parli, V., Kariuki, N., Capstick, E., Reuel, A., Brynjolfsson, E., Etchemendy, J., Ligett, K., Lyons, T., Manyika, J., Niebles, J. C., Shoham, Y., Wald, R., Walsh, T., Hamrah, A., Santarlasci, L. ... Oak, S. (2025). *The AI Index 2025 Annual Report*. AI Index Steering Committee, Institute for Human-Centered AI, Stanford University.
- Sobottka, T., Ansari, F., Lingitz, L., Hold, P., Nowak, M. & Rudolf, P. (2024). *Positionspapier Positive Impact Production: Nachhaltiger Wohlstand durch positive ökologische und gesellschaftliche Wirkung des Produktionssystems*. Fraunhofer Austria Research GmbH.
- Statistik Austria. (2025). *ÖNACE 2025 – Klassifikation der Wirtschaftstätigkeiten*. https://www.statistik.at/KDBWeb/kdb_VersionAuswahl.do?NAV=DE&VersID=10531&EXT=J&KDBtoken=

Auswirkungen von künstlicher Intelligenz auf den Arbeitsmarkt – Schwerpunkt Kärnten

Viktoria A. Krenn, Torsten Ullrich, Elena Wiegelmann, Tarim Stüwe & Eva Eggeling

1. Ausgangslage

Die Geschichte der Künstlichen Intelligenz (KI) lässt sich historisch weit zurückverfolgen und nahm in den 1940er-Jahren mit der Entwicklung des programmierbaren Digitalcomputers eine konkrete technische Gestalt an. Trotz intensiver Forschungsaktivitäten seit den 1950er-Jahren erfolgte der Fortschritt in Etappen bis zur industriellen Etablierung von maschinellem Lernen und Deep Learning in den 2000ern. Die Veröffentlichung von ChatGPT Ende 2022 trug zu einem wesentlichen Durchbruch bei und markierte den Beginn der breiten öffentlichen Wahrnehmung von KI. Inzwischen gilt KI als eine Schlüsseltechnologie mit enormem Einfluss auf Wirtschaft und Gesellschaft. Anders als klassische Automatisierungssysteme ist sie in der Lage, aus Daten zu lernen, sich an neue Bedingungen anzupassen und autonome Entscheidungen zu treffen. Dadurch transformiert KI die Arbeitswelt auf fundamentale Weise; während standardisierte, repetitive Tätigkeiten zunehmend automatisiert werden, entstehen gleichzeitig neue Berufsprofile und Tätigkeitsfelder. Parallel dazu gewinnen kreative, soziale und analytische Kompetenzen an Bedeutung. Die damit einhergehenden Veränderungen des Arbeitsmarkts eröffnen sowohl Potenziale als auch Herausforderungen, besonders im Hinblick auf Qualifizierung, Weiterbildung und Umschulung.

Im Rahmen einer systematischen Literaturrecherche konnten knapp 100 Studien aus unterschiedlichen Regionen weltweit identifiziert werden, die sich mit den Auswirkungen von KI auf den Arbeitsmarkt befassen. Um Aktualität zu gewährleisten, beschränkte sich die Literaturanalyse auf die Jahre 2016 bis 2024 (siehe Übersicht im Anhang A). Von den identifizierten Studien beziehen sich 14 explizit auf Europa und sechs davon auf den deutschsprachigen Raum (Bock-Schappelwein & Egger, 2023; Dengler & Matthes, 2021; Dornmayr & Riepl, 2023; Fregin et al., 2023; Fuchs et al., 2022; Specht, 2024).

Die aktuelle globale Situation wird seitens des World Economic Forum stetig beobachtet und im „Future of Jobs Report“ regelmäßig veröffentlicht (World Economic Forum, 2023, 2025). Insbesondere in den beiden letzten Reports wurde die Rolle von KI hervorgehoben. Demnach identifizierten 86 Prozent der Arbeitgebenden KI als einen Haupttreiber für Unternehmenstransformation. Technische Fähigkeiten im Kontext von KI und Big Data sind von Platz 15 der Fähigkeitenrangliste („core skills“) 2023 auf Platz 11 im Jahr 2025 vorgerückt und gleichzeitig von Platz 7 der „skills on the rise“ auf Platz 1 im selben Zeitraum.

Auch die Organisation für wirtschaftliche Zusammenarbeit und Entwicklung (OECD) unterstreicht in ihrer Analyse (Manca, 2023), dass die Entwicklung von Fähigkeiten in Bereichen wie maschinelles Lernen, Skriptsprachen und Softwareentwicklungsprinzipien und parallel dazu kognitive Kompetenzen wie kreatives Problemlösen wichtiger werden. Die Nachfrage nach KI-Kompetenzen steigt dabei deutlich rascher als die nach anderen Qualifikationen. Beispielsweise zeigt sich im

Vereinigten Königreich eine siebenfach erhöhte Nachfrage bei maschinellem Lernen und in den Vereinigten Staaten sogar eine 12- bis 21-fach erhöhte Nachfrage. Arbeitgebende und Unternehmen reagieren auf den sich verändernden Kompetenzbedarf hauptsächlich, indem sie bestehende Beschäftigte durch Umschulung oder Höherqualifizierung fördern (Lane et al., 2023). Gleichzeitig berichteten die umgeschulten Beschäftigten häufiger von positiven Effekten von KI auf ihre Arbeitsbedingungen, was den Wert von Investitionen in Mitarbeiterschulungen unterstreicht. Mangelnde Qualifikation und fehlende Kompetenzen werden neben den hohen Technologiekosten und staatlicher Regulierung als die größten Hindernisse in der Einführung von KI angegeben (Lane et al., 2023).

Eine Studie des Instituts der deutschen Wirtschaft konnte allerdings keinen belastbaren Zusammenhang zwischen der Einführung von KI und der Mitarbeiterentwicklung auf Unternehmensebene im Zeitraum von 2019 bis 2022 nachweisen (Hammermann et al., 2023), der jedoch maßgeblich von der Corona-Pandemie beeinflusst wird. Gleichzeitig gibt es erste Forschungsergebnisse KI-Chatbots betreffend, die niedrige Produktivitätsgewinne in den Jahren 2023/2024 aufzeigen und damit die Erwartungen eines schnellen Arbeitsmarktwandels und starker kurzfristiger Effekte relativieren (Humlum & Vestergaard, 2025). Diese Ergebnisse deuten auf komplexe Effekte hin und hinterfragen, ob potenzielle Produktivitätsgewinne Risiken durch Arbeitsplatzverluste ausgleichen können.

Daten aus der Arbeitskräfteerhebung der Europäischen Union (European Union Labour Force Survey), analysiert von der OECD (Lassébie & Quintini, 2022), zeigen, dass 9 Prozent der Beschäftigten in Berufen tätig sind, in denen mindestens 25 Prozent der Tätigkeiten hochgradig automatisierbar sind. Wobei Länder wie Ungarn, Lettland, Slowakei und Tschechien einem erhöhten Risiko ausgesetzt sind, dass Beschäftigung durch Automatisierung beeinträchtigt wird. Länder wie die Niederlande, die Schweiz und Norwegen weisen dagegen ein deutlich verringertes Risiko auf, was unterstreicht, dass die Risiken und Chancen der Automatisierung und KI nicht nur auf EU- oder europäischer Ebene betrachtet werden sollten, sondern auch innerhalb einzelner Länder und Regionen.

Derzeit existiert lediglich eine limitierte Anzahl an Forschungsarbeiten, welche sich mit den potenziellen Auswirkungen von KI auf den österreichischen Arbeitsmarkt befassen (Bock-Schappelwein & Egger, 2023; Fuchs et al., 2022). Dies lässt den Schluss zu, dass weiterer Forschungsbedarf in diesem Bereich – sowohl auf Bundes- als auch auf Landesebene – besteht. Die österreichische Arbeitswelt sieht sich mit Herausforderungen und Chancen konfrontiert, die maßgeblich durch technologische Innovationen wie KI, aber auch durch demografische Veränderungen geprägt sind. Um den Anforderungen in Bereichen wie Daten und KI oder anderen Themenkomplexen wie grüne Wirtschaft und Pflegewirtschaft gerecht zu werden, sind gezielte Weiterbildungsinitiativen unumgänglich (Bock-Schappelwein & Egger, 2023). Österreichische Unternehmen erkennen zunehmend die Möglichkeiten, die KI als Produktivitätstreiber bietet (Fuchs et al., 2022). In der Praxis gibt es jedoch erhebliches Verbesserungspotenzial bei der strategischen Priorisierung von KI, insbesondere aufgrund mangelnder Mitarbeiterkompetenzen. Die Einführung von KI führt zu erheblichen Veränderungen in den Qualifikationsanforderungen am Arbeitsmarkt. Gleichzeitig eröffnet der frühzeitige und umfassende Einsatz von KI-Technologien

die Möglichkeit, neue Berufsbilder zu schaffen, Arbeitsprozesse zu optimieren und die Wettbewerbsfähigkeit österreichischer Unternehmen langfristig zu sichern.

Hinsichtlich möglicher Veränderungen des Kärntner Arbeitsmarktes durch KI liegen bislang keine Studien vor. Diese Arbeit ist ein Auszug einer Studie, die das Kärntner Arbeitsmarktservice (AMS Kärnten) in Auftrag gegeben hat, um die aktuelle Situation in Kärnten multidimensional aufzugreifen und damit die Integration von KI am Arbeitsmarkt zu erleichtern.

1.1 Fokus Kärntner Arbeitsmarkt

Speziell in Kärnten gibt es aufgrund der demografischen Verschiebung, Distanz zu Wirtschaftszentren, saisonalen Beschäftigungsschwankungen sowie einer ungleichmäßigen Branchenverteilung eine besondere Ausgangssituation. Die regionale Wirtschaft ist in hohem Maße von Branchen wie Elektrotechnik und Maschinenbau (ÖNACE Klasse C, siehe Tabelle 4) geprägt. Es besteht jedoch ein prognostizierter Bedarf von etwa 1.600 Stellen in technischen, naturwissenschaftlichen und medizinischen Berufen sowie 2.100 Stellen in sonstigen wissenschaftlichen Berufen (Aigner-Walder et al., 2024). Dem gegenüber steht bis 2040 eine berechnete Lücke von 50.000 erwerbsfähigen Personen im Alter zwischen 15 und 64 Jahren (Schmid & Nowak, 2023). Diese ist auf starke Bevölkerungsrückgänge zurückzuführen: Demografisch liegt Kärntens Bevölkerungswachstum in den letzten 20 Jahren mit 2 Prozent deutlich unter dem österreichischen Durchschnitt von 11,7 Prozent (Aigner-Walder et al., 2024, S. 2). Auch innerhalb Kärntens manifestiert sich ein ungleich verteiltes Wachstum. Während für Städte wie Klagenfurt (4,4%) und Villach (2,8%) ein moderates Wachstum vorliegt, muss für ländliche Gebiete wie Hermagor ein Rückgang von bis zu -7,8 Prozent antizipiert werden (Aigner-Walder et al., 2024, S. 3). Bei den meisten Prognosemodellen (Schmid & Nowak, 2023) wird bis 2040 insgesamt von einem Bevölkerungsrückgang von 3.400 bis 20.000 Personen ausgegangen. Gleichzeitig werden je nach Prognosevariante (Benchmark- vs. Trendvariante) auch für die meisten Qualifikationsabschlüsse Rückläufe vorhergesagt (Schmid & Nowak, 2023). Am problematischsten ist die Lage bei den Lehrabschlüssen, bei denen in jedem Fall bis 2040 ein Minus von 18.000 erwartet wird. Bei Fortschreibung der aktuellen Qualifikationsstruktur (Benchmarkvariante) ist zusätzlich mit einem Rückgang von jeweils 6.000–8.000 bei BMS-, BHS-, Tertiär- und Pflichtschulabschlüssen zu rechnen. Sollten sich die Qualifizierungstrends der letzten 50 Jahre fortsetzen (Trendvariante) ist zumindest bei den BHS- und Tertiärabschlüssen eher mit einer Zunahme zu rechnen. Denn speziell die Tertiärabschlussrate steigt in ganz Österreich kontinuierlich an (Bock-Schappelwein & Egger, 2023).

Hinsichtlich dieser Veränderungen der Bevölkerungs- und Qualifikationsstrukturen bringt KI zusätzliche Herausforderungen für den Arbeitsmarkt und die Erwerbstätigen mit sich. Denn es gilt, die Kompetenzen und Fähigkeiten, die im Zusammenhang mit KI erforderlich werden, zu identifizieren und zeitnahe zu fördern.

2. Veränderungen am Arbeitsmarkt

2.1 Veränderung der Qualifikationsprofile

Bei der Transformation der Qualitätsprofile wird insbesondere die Bedeutung neuer technischer Fähigkeiten hervorgehoben (Alekseeva et al., 2021; Georgieff & Hyee, 2021; Lassebie & Quintini, 2022; Manca, 2023; Milanez, 2023; OECD, 2023; Squicciarini & Nachtigall, 2021; Verma et al., 2021). Dazu zählen Kenntnisse über KI-Technologien (z.B. maschinelles Lernen, Datenanalyse, KI-Tools), Programmierfähigkeiten, Kompetenzen in den Bereichen Big Data, Statistik und Datenmodellierung (z.B. Regression, Klassifikation, Clusteranalyse) sowie grundlegende Datenkompetenz. Parallel dazu werden aber auch kognitive und soziale Fähigkeiten gefragter. Im Vordergrund stehen Problemlösungskompetenz, kritisches Denken, Entscheidungsfähigkeit wie auch Kommunikation, Teamarbeit, Empathie, Ethik und Anpassungsfähigkeit. Eine Übersicht der wichtigsten Kompetenzen ist in Tabelle 1 zu finden.

Tab. 1: Aus der Literatur (Anhang A) exzerpierte Kompetenzen und Fähigkeiten in den Bereichen Technik, Kognition und Social Skills

Fähigkeitsbereich	Wichtige Kompetenzen
Technische Kompetenzen	Kenntnisse über KI-Technologien, Erfahrung mit Tools wie TensorFlow, PyTorch, Programmierung (Python, Java, R), Datenanalyse, Statistik, Modellierung, maschinelles Lernen, Technische Grundlagen (Mathematik/Statistik)
Kognitive Fähigkeiten	Problemlösungs- und Entscheidungsfähigkeit, strategisches und analytisches/logisches Denken, Kreativität und Innovationsfähigkeit, Prozessoptimierung und Automatisierungsverständnis
Soziale Fähigkeiten	Teamarbeit und klare Kommunikation, Zusammenarbeit in interdisziplinären Teams, Selbstorganisation und Motivation, Anpassungsfähigkeit, Führungsverantwortung, Empathie, proaktive Haltung und ethisches Denken

Gleichzeitig spielt die Automatisierbarkeit von Tätigkeiten eine große Rolle (Lassebie & Quintini, 2022). Demnach sind Verhandlungsgeschick, gesellschaftliche Wahrnehmung, Fürsorge, Technologiedesign und Überzeugungsarbeit zentrale Hürden für die Automatisierung. Auch die Themen „komplexes Problemlösen“ und „aktives Zuhören“ stellen besondere Herausforderungen für KI-Systeme dar. Zwar können KI-Systeme einzelne Teilaspekte unterstützen, aber besonders das Verstehen emotionaler Reaktionen, gezielte Einflussnahme oder die Anpassung von Technologien an individuelle Bedürfnisse erfordert weiterhin menschliches Handeln.

Um diese Qualifikationsprofile zu fördern und die Potenziale von KI verantwortungsvoll zu nutzen, sind entsprechende gesellschaftliche Rahmenbedingungen unabdingbar. Aus- und Weiterbildung spielen dabei eine entscheidende Rolle. KI-Kompetenzen müssen bereits in der Schule vermittelt werden, um die nächste Generation optimal auf eine zunehmend technologiegestützte Arbeitswelt vorzubereiten. Dies ist umso wichtiger, da ein Drittel der Bevölkerung im Alter von 16 bis 74 Jah-

ren angibt, über kein grundlegendes digitales Wissen zu verfügen (Bock-Schappellwein & Egger, 2023). Die Einführung KI-gestützter Prozesse am Arbeitsmarkt führt zusätzlich zu neuen Anforderungen, die gezielte Maßnahmen zur Umschulung und Qualifizierung durch Weiterbildung notwendig machen.

Aufschluss über die aktuellen Veränderungen und das Entstehen neuer Qualifikations- und Arbeitsplatzprofile in Kärnten und in Österreich liefert die Datenbank AMS eJob-Room¹: Am 10.02.2025 waren in Kärnten 9882 Arbeitsplätze ausgeschrieben. Die Stellenausschreibungen wurden automatisiert analysiert (mittels Web-Scraping-Skript und der Library Selenium) und bereinigt. 6681 Datensätze waren vollständig und 117 Arbeitsplätze, respektive 1,75 Prozent, enthielten Anforderungsprofile mit mindestens einem KI-Begriff (siehe Tabelle 2). Deren Häufung ist in Abbildung 1 dargestellt und zeigt vorrangig Stellen im Kontext von Automatisierung und Digitalisierung, jedoch noch kaum Stellen mit einem direkten KI-Bezug. Im Österreichsschnitt liegt Kärnten damit an vorletzter Stelle (siehe Tabelle 3).

Tab. 2: Alphabetische Übersicht der Suchbegriffe mit Bezug zu Künstlicher Intelligenz, die bei der Analyse von Stellenbeschreibungen berücksichtigt wurden

Algorithmic	Digitalisierung	Neural Networks
Artificial Intelligence	Entscheidungsbäume	Neuronale Netze
Automation	GPT	Pattern Recognition
Automatisierung	IBM Watson	Produktautomatisierung
Bildverarbeitung	Image Processing	Prompt Engineering
Chat	Keras	Random Forests
Computer Vision	Künstliche Intelligenz	Robotik
Data Analysis	Machine Learning	Speech Recognition
Data Science	Machine Vision	Spracherkennung
Datenanalyse	Maschinelles Lernen	Text Mining
Datenwissenschaft	Mustererkennung	Textanalyse
Decision Trees	Natural Language Processing	Tokenization
Deep Learning	Natürliche Sprachverarbeitung	Visual Analysis

1 <https://jobs.ams.at>

Tab. 3: Vergleich der Häufigkeiten von KI-Begriffen in Stellenausschreibungen der neun Bundesländer

Bundesland	Anzahl der Stellenausschreibungen	Anzahl der KI-Begriffe in Stellenbeschreibungen	Prozent-satz (%)
Burgenland	1776	24	1,35
Kärnten	6681	117	1,75
NOE	9550	204	2,14
OOE	9261	509	5,50
Salzburg	6571	150	2,28
Steiermark	8263	234	2,83
Tirol	8566	166	1,94
Vorarlberg	3776	91	2,41
Wien	9469	428	4,52

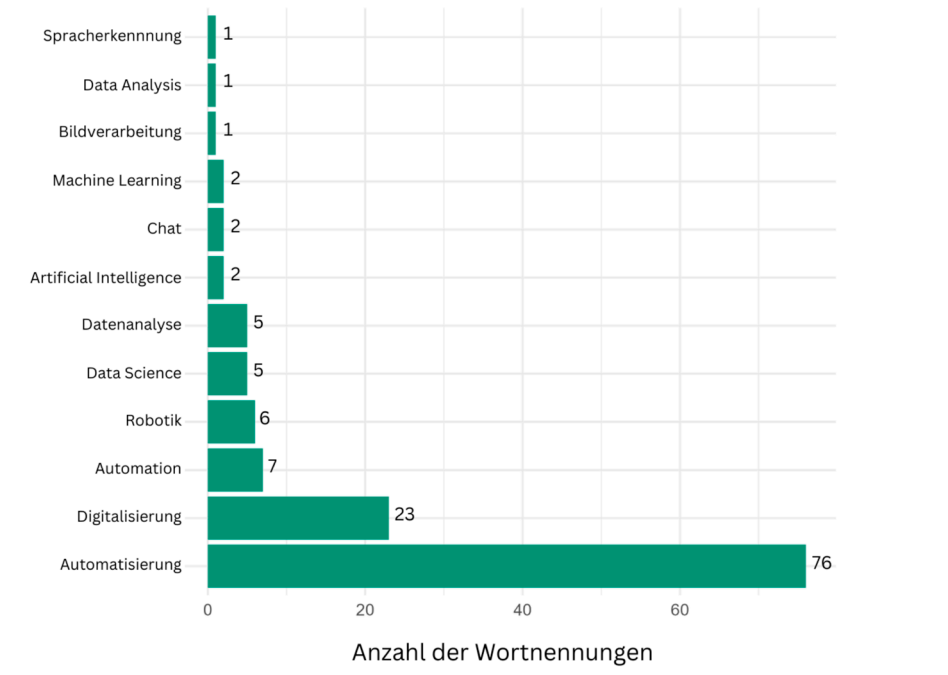


Abb. 1: Häufigkeit KI-bezogener Begriffe in den Stellenbeschreibungen im Raum Kärnten

2.2 Veränderung Branchenzusammensetzung

Obwohl KI in den Stellenausschreibungen Kärntens derzeit noch spärlich aufscheint, hat der Einsatz von KI-Systemen bereits zu einer teilweisen Automatisierung von Routinetätigkeiten im Arbeitsalltag geführt. Es ist zu erwarten, dass der Einfluss von KI auf den Arbeitsmarkt weiter zunehmen und dadurch Berufsprofile grundlegend verändern bzw. neue Berufsprofile schaffen wird.

Eine Prüfung der gelisteten Studien (Bock-Schappelwein & Egger, 2023; Bordot, 2022; Brynjolfsson et al., 2023; Chui et al., 2023; Dell’Acqua et al., 2023; Dornmayr & Riepl, 2023; Fregin et al., 2023; Fuchs et al., 2022; Georgieff & Hyee, 2021; Green & Lamby, 2023; Grienberger et al., 2024; Hammermann et al., 2023; Hazan et al., 2024; Johnson et al., 2021; Käufer & Frein, 2024; Lane & Saint-Martin, 2021; Lane et al., 2023; Lassébie & Quintini, 2022; Manca, 2023; Milanez, 2023; OECD, 2023; Paolillo et al., 2022; Singla et al., 2024; Squicciarini & Nachtigall, 2021; Tolan et al., 2021; World Economic Forum, 2023) identifizierte insgesamt 214 sich wandelnde Berufe. Hierbei ist jedoch zu beachten, dass es sich nicht um einen isolierten KI-Effekt handelt, sondern um Veränderungen, die auch durch Automatisierung, Robotik sowie demografische und gesellschaftliche Parameter beeinflusst werden. Referenziert werden die Berufe im Folgenden auf die neue ÖNACE 2025 Klassifizierung A bis V (Tabelle 4), zur vereinfachten Lesbarkeit werden im weiteren Verlauf wertfreie Abkürzungen verwendet. Bei 122 Berufen wird ein zunehmender Bedarf an Arbeitsplätzen prognostiziert, während bei 52 Berufsgruppen die Nachfrage abnimmt. Für 40 Berufsgruppen ist eine signifikante Veränderung oder Weiterentwicklung der primären Tätigkeiten wahrscheinlich (siehe Abbildung 2). Dadurch ergibt sich gesamt eine positive Job-Nettobilanz.

Auf Branchenebene wird wenig überraschend der stärkste Zuwachs an Arbeitsplätzen in der neu eingeführten Klasse K *Telekommunikation* erwartet. KI treibt diese Zunahme stark voran, da Berufe in der Informatik und Softwareentwicklung direkt an den Themenkreis KI anknüpfen. KI lässt auch neue Berufsprofile in der Klasse K entstehen (z.B. KI-Beratungsfachkräfte, KI-Entwickler). Für die Klassen Q *Bildung*, R *Gesundheit* und S *Kultur* weisen alle Nennungen auf eine steigende Nachfrage hin. Die Klasse C *Waren*, die in Kärnten den größten Beschäftigungssektor bildet, erfährt sowohl einen starken Anstieg als auch Abfall der Nachfrage. Daher ist eine deutliche Umstrukturierung dieser Branche zu erwarten. Diese Veränderungen werden durch das hohe Automatisierungspotenzial durch KI und innovative Software begünstigt (Fregin et al., 2023). In geringerem Ausmaß ist dies auch für die Klassen F *Bau*, G *Handel*, H *Verkehr*, I *Tourismus*, J *Verlagswesen*, M *Grundstücke* und O *Wirtschaftliche Dienste* der Fall. Lediglich in den Klassen A *Landwirtschaft*, L *Finanz* und N *Wissenschaftliche Dienste* werden Berufe überwiegend im Kontext sinkender Nachfrage genannt, wodurch sich für diese Klassen eine negative Job-Nettobilanz ergibt. Die verbleibenden Klassen weisen keine Nennungen auf.

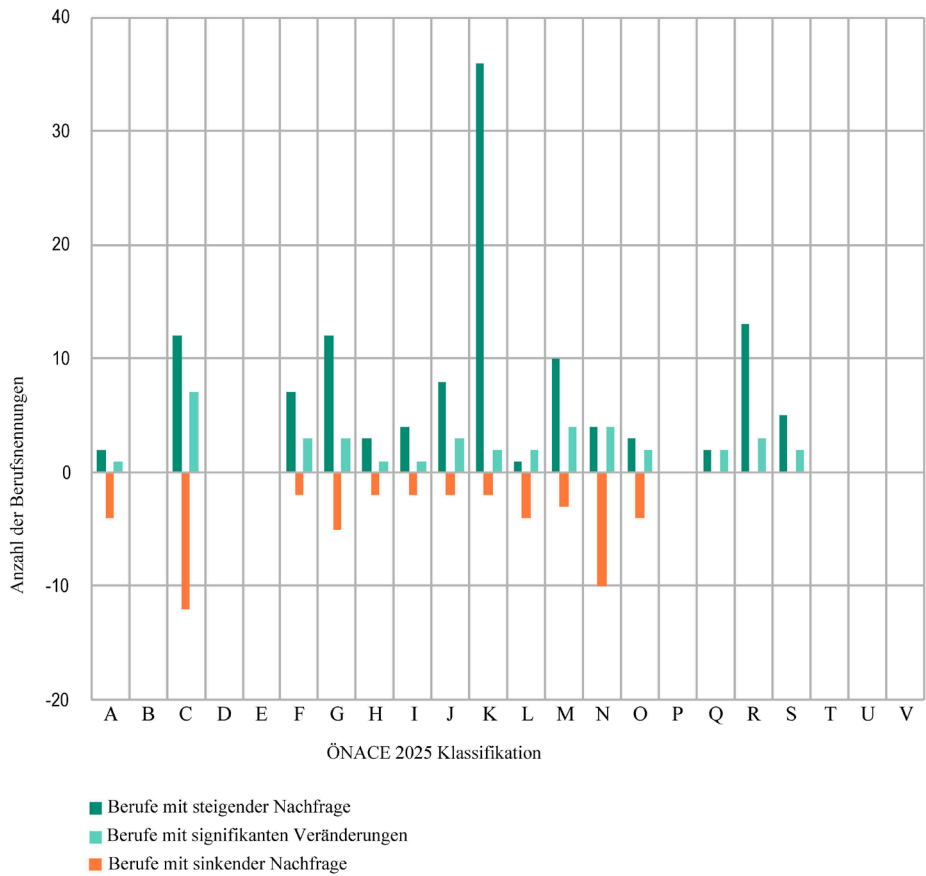


Abb. 2: Veränderung der Anzahl an Arbeitsplätzen unter Berücksichtigung von Berufen mit steigender Nachfrage (dunkelgrün), sich verändernden Berufsprofilen (hellgrün) und sinkender Nachfrage (orange). Dargestellt ist die Anzahl der Berufsnennungen in der Literatur für die jeweiligen Branchen. Durch die Zu- und Abnahme der Berufsprofile ergibt sich eine insgesamt positive Job-Nettobilanz

Tab. 4: ÖNACE 2025 Wirtschaftsklassen und die Beschäftigungszahlen in Kärnten für das Jahr 2024. Wertfrei sind zur Vereinfachung der Lesbarkeit im Manuskript nur Klassenbuchstaben und Kurzbeschreibungen angeführt

Klasse	Kurzbeschreibung	Beschreibung nach ÖNACE 2025	Beschäftigungszahlen 2024
A	Landwirtschaft	Land- und Forstwirtschaft, Fischerei und Fischzucht	2.024
B	Bergbau	Bergbau und Gewinnung von Steinen und Erden	457
C	Waren	Herstellung von Waren	37.439
D	Energie	Energieversorgung	2.662
E	Wasser	Wasserversorgung, Abwasser- und Abfallentsorgung und Beseitigung von Umweltverschmutzungen	1.373
F	Bau	Bau	16.765
G	Handel	Handel	32.107
H	Verkehr	Verkehr und Lagerei	10.157
I	Tourismus	Beherbergung und Gastronomie	14.232
J	Verlagswesen	Verlagswesen, Rundfunk sowie Erstellung und Verbreitung von Medieninhalten	2.037
K	Telekommunikation	Telekommunikation, Softwareentwicklung, IT-Beratung und sonstige Dienstleistungen der Informationstechnologie und der Computerinfrastruktur	2.037
L	Finanz	Erbringung von Finanz- und Versicherungsdienstleistungen	5.718
M	Grundstücke	Grundstücks- und Wohnungswesen	1.830
N	Wissenschaftliche Dienste	Erbringung von wissenschaftlichen und technischen Dienstleistungen	8.580
O	Wirtschaftliche Dienste	Erbringung von sonstigen wirtschaftlichen Dienstleistungen	11.979
P	Verwaltung	Öffentliche Verwaltung, Verteidigung; Sozialversicherung	37.614
Q	Bildung	Erziehung und Unterricht	4.695
R	Gesundheit	Gesundheits- und Sozialwesen	19.874
S	Kultur	Kunst, Sport und Erholung	1.861
T	Sonstige Dienste	Erbringung von sonstigen Dienstleistungen	4.863
U	Privathaushalte	Private Haushalte mit Hauspersonal sowie Herstellung von Waren und Erbringung von Dienstleistungen durch private Haushalte für den Eigenbedarf ohne ausgeprägten Schwerpunkt	101
V	Exterritoriale Organisation	Exterritoriale Organisationen und Körperschaften	0

3. Umfrage zum Arbeitsmarkt in Kärnten

Umfrageergebnisse, die im Zeitraum Dezember 2024 bis März 2025 erhoben wurden, untermauern die bisherigen Ergebnisse: Die Teilnehmenden setzen sich sowohl aus geladenen Experten und Expertinnen aus den Themenfeldern Künstliche Intelligenz, Demographie und Arbeitsmarkt zusammen (n=18) als auch KI-Interessierten (n=158) aus unterschiedlichen Branchen. Die Stichprobengröße und Verteilung lassen Interpretationen zu, jedoch ist eine Übertragung auf alle Branchen oder die gesamte Kärntner Wirtschaft nicht zulässig. Der Fragebogen besteht aus einer Kombination statistisch auswertbarer wie auch offener Fragen zur qualitativen Interpretation (insgesamt 28 Fragen). Drei Themengebiete werden näher betrachtet:

1. Unternehmensparameter,
2. Branchen und Tätigkeitsfelder und
3. Arbeitsmarkt und Arbeitsweise.

3.1 Unternehmensparameter

Der erste Teil des Fragebogens erfasst die Unternehmensparameter. Unter den Teilnehmenden ist ein großer Anteil an Kleinstunternehmen (44%), was zugleich den hohen Anteil an Geschäftsführenden (49%) erklärt. Kleine, mittlere und Großunternehmen sind zu 16–20 Prozent gleichmäßig vertreten, mit einer Durchmischung der Organisationsebene zwischen Operativen (11%), mittlerem Management (22%) und oberem Management (18%). Bei der Verteilung der Branchen zeigt sich ebenfalls eine gute Durchmischung, die in großen Teilen die Anzahl der unselbstständigen Beschäftigten in Kärnten widerspiegelt. Die Branche K *Telekommunikation* ist aufgrund der Thematik überproportional repräsentiert (siehe Abbildung 3).

Um insbesondere die Erfahrungen von klein- und mittelständischen Betrieben (KMU) zu evaluieren, werden an diese Unternehmensgruppen weiterführende Fragen zu den Themen Herausforderungen und Vorteile für KMU gerichtet (n=141, der 176 Gesamtteilnehmenden). Obwohl 79 Prozent der Teilnehmenden von Vorteilen durch KI für KMU in Kärnten ausgehen, werden als die drei größten Herausforderungen bei der Einführung von KI fehlendes Know-how und Fachwissen (59%), finanzielle und personelle Ressourcen für Infrastruktur/KI-Tools und Weiterbildung (21%) sowie unklare Rechts- und Datenschutzbelange (14%) genannt.

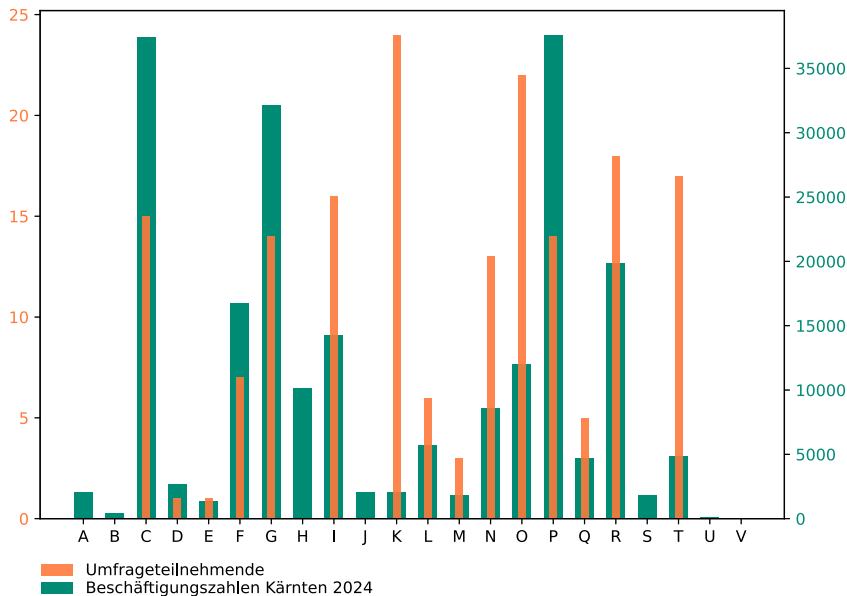


Abb. 3: Die Branchenzugehörigkeit der Umfrageteilnehmenden spiegelt in großem Maße die Beschäftigungszahlen 2024 wider. Dabei ist zu berücksichtigen, dass sich die Klassifikation der Klassen J und K in ÖNACE ab 2025 ändert. Daher sind diese beiden Klassen als kumulierte Summe zu interpretieren

3.2 Branchen und Tätigkeitsfelder

Im zweiten Abschnitt werden die Veränderungen der Branchen und Tätigkeitsfelder ermittelt. Mehrheitlich gehen die Teilnehmenden von tiefgreifenden Veränderungen ihrer Branche durch KI in den kommenden fünf bis zehn Jahren aus (80%), ganz besonders in Großunternehmen (94%) und im oberen Management (89%). Diese betreffen, so die formulierte Annahme, verschiedene Unternehmensbereiche und führen zu vermehrter Automatisierung. Davon am stärksten betroffen sind, über alle Branchen hinweg, Administration, Verwaltung, IT, Kundenservice und Marketing. Auch in der Produktion und in der Datenverarbeitung werden große Veränderungen erwartet. Eine Aufschlüsselung der quantitativen Ergebnisse auf Branchenebene ist für Branchen mit mehr als zehn Datensätzen in Tabelle 5 angeführt.

Zu den allgemeinen Trends offenbaren sich hierbei branchenspezifische Unterschiede. Speziell die Branchen G *Handel* und I *Tourismus* erwarten in den nächsten fünf bis zehn Jahren weniger Veränderungen als die Branchen K *Telekommunikation* und N *Wissenschaftliche Dienste*.

Von der Entstehung neuer Berufsprofile gehen 57 Prozent der Teilnehmenden aus. Eine erhöhte Erwartungshaltung ist bei mittleren Unternehmen (82%) und der operative Ebene (84%) festzustellen. Besonders häufig werden neue Berufsprofile im Zusammenhang mit KI-Entwicklung, KI-Beratung und KI-Management genannt. Auf Branchenebene liegt in den Branchen K *Telekommunikation* und N *Wissenschaftliche Dienste* eine deutlich höhere Erwartungshaltung vor als z.B. in den Branchen G *Handel* und R *Gesundheit*.

Tab. 5: Auswertung der Ergebnisse auf Branchenebene für Branche mit mehr als 10 Teilnehmenden. Klasseneinteilung entsprechend ÖNACE 2025 Wirtschaftsklassen (siehe Tabelle 4).

	C	G	I	K	N	O	P	R	T
Teilnehmende* ¹	15	14	16	24	13	22	14	18	17
Veränderung Arbeitsweise* ² (%)	80	64	62	100	100	73	86	78	76
Neue Berufskategorien* ³ (%)	53	29	62	79	77	64	50	33	65
Umschulung / Qualifikationen* ⁴ (%)	93	57	69	79	100	73	71	72	76
Investitionsbereitschaft* ⁵ (1–10)	5,4	4,4	3,9	6,5	6,6	6,4	5,9	4,5	4,3
KI bereits implementiert* ⁶ (%)	47	71	44	96	77	77	64	50	71
KI gegen Fachkräftemangel* ⁷ (%)	47	79	25	71	62	64	57	25	53

*1 Anzahl der Umfrageteilnehmenden

*2 Ja-Antworten in Prozent auf die Frage: Glauben Sie, dass KI die Arbeitsweise in Ihrer Branche in den nächsten 5-10 Jahren erheblich verändern wird?

*3 Ja-Antworten in Prozent auf die Frage: Können Sie sich vorstellen, dass in Ihrer Branche durch KI völlig neue Berufskategorien entstehen?

*4 Ja-Antworten in Prozent auf die Frage: Wird die Einführung von KI eine signifikante Umschulung oder neue Qualifikationen für bestehende Stellen in Ihrer Branche erfordern?

*5 Antworten auf einer Skala 1–10 auf die Frage: Wie würden Sie die allgemeine Bereitschaft Ihrer Branche beschreiben, in KI-Weiterbildungen zu investieren?

*6 Ja-Antworten in Prozent auf die Frage: Haben Sie in Ihrem Unternehmen bereits KI-Technologien implementiert?

*7 Ja-Antworten in Prozent auf die Frage: Könnte KI dazu beitragen den Fachkräftemangel auf dem Kärntner Arbeitsmarkt zu verringern?

Obwohl sich 72 Prozent der Teilnehmenden einig sind, dass signifikante Umschulungen und neue Qualifikationen notwendig sind, wird die allgemeine Bereitschaft, in Weiterbildung zu investieren, nur als moderat (Wert 5,5 von 10) eingestuft. Die Branche N *Wissenschaftliche Dienste* zeigt, kurz gefolgt von Branche C *Waren*, die höchste Bereitschaft, in KI-Weiterbildung zu investieren. Branchen, die von einer größeren Veränderung der Arbeitsweise in den nächsten fünf bis zehn Jahren ausgehen, weisen zudem eine höhere Bereitschaft auf, in KI-Weiterbildung zu investieren. Weiterbildungen werden besonders in den Bereichen KI- und IT-Grundlagen, Ethik- und Rechtsgrundlagen sowie Datenschutz und Datenanalyse als notwendig erachtet.

KI-Technologien sind in 66 Prozent der teilnehmenden Unternehmen implementiert. Insbesondere in den Bereichen Marketing und Kundenservice werden bereits vermehrt generative KI-Modelle (GenAI) und Sprachmodelle zur Texterstellung, Content-Produktion und Automatisierung eingesetzt. Zudem wird KI auch schon in den Bereichen Personalwesen/Verwaltung, Vertrieb, Produktion und Logistik genutzt. Während einige Unternehmen bereits großflächig KI-gestützte Lösungen verwenden (Branche K *Telekommunikation* zu 96%), befinden sich andere noch in der Test- oder Pilotphase. Unternehmen, die KI noch weniger nutzen, sehen Einsatzbereiche in der Automatisierung, Effizienzsteigerung und Kundenkommunikation. Viele sehen Potenzial in der Prozessoptimierung, insbesondere in der Buchhaltung, im Control-

ling, in der Produktionsplanung und in der Personalverwaltung. Doch auch im Marketing und im Design wird KI als nützlich erachtet.

Kundenkommunikation ist ein weiterer wesentlicher Bereich, in dem KI, beispielsweise mit Chatbots oder bei der automatisierten Beantwortung von E-Mails und Schriftverkehr, unterstützen kann. Einige Unternehmen möchten KI in Forschung und Entwicklung (R&D) sowie für Datenanalyse und statistische Auswertungen einsetzen. Während viele Teilnehmende den Nutzen der Technologie erkennen, gibt es auch Stimmen, die einen menschlichen Kontakt in der Kundeninteraktion bevorzugen oder keinen aktuellen Bedarf für KI in ihrem Unternehmen sehen.

Bei der Frage, ob KI die Gesamtzahl der Arbeitsplätze in den Branchen verändern wird, erwarten 61 Prozent der Teilnehmenden eine gleichbleibende Anzahl an Arbeitsplätzen. Besonders die Branchen I *Tourismus*, O *Wirtschaftliche Dienste* und N *Wissenschaftliche Dienste* gehen von gleichbleibenden Zahlen aus.

Eine Reduktion an Arbeitsplätzen erwarten 41 Prozent der Teilnehmenden, insbesondere in den Branchen C *Waren* (53 %) und G *Handel* (43 %). Umgelegt auf Tätigkeitsprofile sind die Einsparungen insbesondere bei repetitiven und standardisierten Aufgaben wie Sachbearbeitung und administrative Tätigkeiten zu verorten. Dies betrifft Backoffice-Tätigkeiten, Übersetzungsdienste, Kundenservice, aber auch den Kreativ- und Marketingbereich (Text-, Grafikerstellung und Content Creation). In der Softwareentwicklung sind insbesondere einfache Programmieraufgaben und Qualitätskontrollen durch KI substituierbar. Allerdings wird betont, dass Berufe mit persönlichem Kundenkontakt, handwerkliche Tätigkeiten und spezialisierte Beratung nicht ersetzt werden können, sondern durch KI nur ergänzt werden.

Nur wenige gehen von einem Anstieg der Anzahl an Arbeitsplätzen aus (8 %). Den größten Zuwachs erwarten die Branche T *Sonstige Dienste* (24 %) und K *Telekommunikation* (17 %) durch die Entstehung neuer spezialisierter Berufsgruppen.

Den Einfluss von KI auf die Arbeitszeit schätzen 46 Prozent der Teilnehmenden entweder mit einer Stundenreduktion oder ohne Auswirkungen ein. Besonders die Branchen K *Telekommunikation* und N *Wissenschaftliche Dienste* erwarten zu 63 Prozent keine Veränderungen der Arbeitszeit. In den Branchen R *Gesundheit* und T *Sonstige Dienste* gehen dagegen 60 Prozent von einer Stundenreduktion aus und nur in Branche O *Wirtschaftliche Dienste* erwarten 23 Prozent einen Mehraufwand durch KI.

3.3 Arbeitsmarkt und Arbeitsweise

Der dritte Abschnitt der Umfrage befasst sich mit Fragen zum Einfluss von KI auf den Arbeitsmarkt und die Arbeitsweise. Die Wahrscheinlichkeit, dass KI den Fachkräftemangel in Kärnten verringern kann, wird mehrheitlich negativ eingestuft (51 %). Insbesondere die operative Ebene sieht nur eine 35-prozentige Wahrscheinlichkeit. Auf Branchenebene zeigt sich, dass speziell die Branchen I *Tourismus* (25 %) und R *Gesundheit* (25 %) keine Verbesserung der Fachkräftesituation erwarten. Dies lässt sich durch die menschenzentrierten Berufsprofile erklären. Die Branchen G *Handel* (79 %) und K *Telekommunikation* (71 %) nehmen die Entwicklung auf den Fachkräftemangel durch KI positiv wahr, da durch die Automatisierung und Auslagerung repetitiver Prozesse Fachkräfte gezielter eingesetzt werden können. Unter-

nehmen, die bereits KI im Unternehmen implementiert haben, sind geneigter, eine Entspannung der Fachkräftesituation durch KI zu erwarten (siehe Abbildung 4).

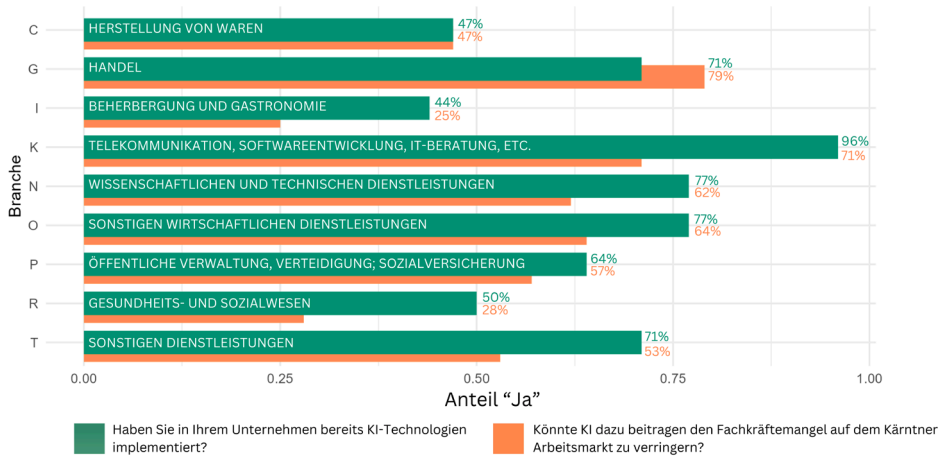


Abb. 4: Antworten auf Branchenebene für die Fragen nach KI-Implementierung (grün) und Verringerung des Fachkräftemangels (orange). Dabei zeigt sich der Zusammenhang, dass Branchen, die bereits vermehrt KI implementiert haben, eher von einer Verringerung des Fachkräftemangels ausgehen

Der zunehmende Einsatz von KI am Arbeitsplatz erfordert eine Neuausrichtung der benötigten Kompetenzen. Die Rückmeldungen der Teilnehmenden sind deckungsgleich mit den Ergebnissen aus Abschnitt 2.1 und lassen sich in dieselben drei Kategorien (Technik, Kognition, Soft Skills, siehe Tabelle 1) unterteilen. Bei den technischen Fähigkeiten werden primär Grundlagenwissen zu KI-Technologien, Datenverarbeitung, Automatisierung sowie Fähigkeiten im Umgang mit KI-Tools und Prompting genannt. Ebenso hervorgehoben wird die Bedeutung analytischen Denkens, Problemlösungskompetenz und die Fähigkeit, KI-generierte Inhalte kritisch zu prüfen und zu hinterfragen. Der menschliche Faktor bleibt bei den Antworten jedoch zentral: Empathie, Kommunikation, Teamfähigkeit und ethisches Urteilsvermögen.

KI beeinflusst auch die wahrgenommene Arbeitsqualität in verschiedenen Dimensionen. In der Auswertung einer Likert-Skala mit den vier zentralen Faktoren Zufriedenheit, Autonomie, Kreativität und Arbeitsbelastung zeigt sich, dass die Mehrheit der Teilnehmenden die Auswirkungen von KI auf die Arbeitsqualität positiv bewertet (50–60 %). Gleichzeitig gibt es Vorbehalte: zwischen 12–25 Prozent der Befragten erwarten negative oder leicht negative Auswirkungen, insbesondere in Bezug auf Kreativität und Autonomie (siehe Abbildung 5).

Im Detail werden die Chancen und Risiken durch KI am Arbeitsmarkt durch die Teilnehmenden wie folgt prognostiziert: Zu den genannten Risiken zählen der mögliche Wegfall von Arbeitsplätzen in standardisierbaren Tätigkeitsbereichen (z. B. Buchhaltung, Datenverwaltung), eine wachsende Kluft zwischen den Qualifikationsniveaus der Beschäftigten sowie Benachteiligung von Personen ohne KI-Kenntnissen. Genannt werden auch psychischer Druck, mit der Leistungsfähigkeit von KI

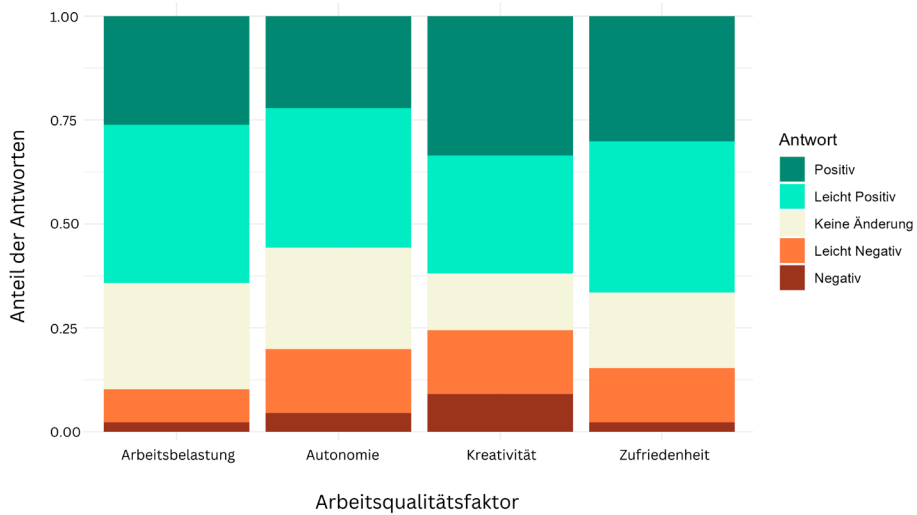


Abb. 5: Auswertung der Likert-Skala für die Arbeitsqualitätsfaktoren Arbeitsbelastung, Autonomie, Kreativität und Zufriedenheit

mitzuhalten, ethische und datenschutzrechtliche Bedenken (u. a. Diskriminierung und Umgang mit sensiblen Daten) sowie der Verlust des menschlichen Faktors und Entfremdung im Arbeitsumfeld.

Gleichzeitig kann KI Routineaufgaben automatisieren, Beschäftigte entlasten und dadurch mehr Raum für kreative und strategische Aufgaben schaffen. Weitere Vorteile sind eine stärkere Flexibilisierung des Arbeitsmarktes durch erleichterte Homeoffice- und Remote-Arbeit sowie personalisierte Weiterbildungsmöglichkeiten durch KI-gestützte Lernplattformen. Zudem kann die Reduktion monotoner Routinetätigkeiten die Arbeitsqualität verbessern. Ein zentraler Vorteil könnte auch die Entspannung des Fachkräftemangels darstellen, etwa durch den Einsatz von KI zur Überbrückung von personellen Engpässen in den Bereichen IT, Gesundheitswesen, Handwerk und insbesondere dem für Kärnten bedeutenden Tourismus.

4. Zusammenfassung

Die rasante Entwicklung der Künstlichen Intelligenz stellt das traditionelle Verständnis von Arbeit, Berufen und den damit verbundenen Anforderungen in Frage. KI übernimmt bereits heute weit mehr komplexe Aufgaben als frühere Technologien, obwohl ihr volles Potenzial noch nicht ausgeschöpft ist. Die Arbeitsmarktauswirkungen könnten daher deutlich umfassender ausfallen als in früheren Innovationsphasen. Dennoch legt die Literaturrecherche nahe, dass die Studienlage im deutschsprachigen Raum und explizit in Österreich derzeit noch begrenzt ist und erhebliches Forschungspotenzial bietet. Eine mögliche Erklärung dafür ist, dass die Folgen eines KI-Effekts auf Branchen und Beschäftigungszahlen von zeitgleichen Geschehnissen wie der Pandemie überlagert und Prognosen dadurch erschwert werden.

Nichtsdestotrotz geben die Ergebnisse Aufschluss über die aktuellen KI-bedingten Entwicklungen am Kärntner Arbeitsmarkt und setzen Impulse für eine zukunftsfähige Anpassung der Arbeitswelt. Kärnten steht aufgrund seiner prognostizierten Bevölkerungsentwicklung und Qualifikationsstruktur besonderen Herausforderungen gegenüber, dem Beschäftigungsrückgang gegenzusteuern. KI bietet dafür erhebliche Potenziale, insbesondere durch Effizienzsteigerung, Fachkräftesicherung und neue Berufsmöglichkeiten. Um die Chancen von KI auf dem Arbeitsmarkt zu maximieren und gleichzeitig das Risiko von Arbeitsplatzverlusten und Diskriminierung zu minimieren, sind Investitionen in die Aus- und Weiterbildung von Erwerbstätigen unumgänglich, auch wenn die Bereitschaft dazu in dieser Umfrage aktuell noch moderat ausfällt. Dies betrifft insbesondere die Branchen, in denen mit starken Veränderungen der Berufsprofile (*C Waren*) und hoher steigender Nachfrage (*K Telekommunikation*) zu rechnen ist. Eine Herausforderung besteht auch darin, die Qualifikationslücke zwischen Beschäftigten mit und ohne KI-Kenntnissen zu schließen. Dies erfordert gezielte Maßnahmen, um soziale Ungleichheiten zu vermeiden und Arbeitnehmende gezielt auf den Wandel vorzubereiten, z.B. durch eine stärkere Verankerung von KI-Grundlagen in der Schulbildung und einen erleichterten Zugang zu Weiterbildungen, um lebenslanges Lernen zu ermöglichen. Dabei müssen neben technischen Fähigkeiten (KI-Kenntnisse, Programmierung, Erfahrung mit KI-Tools) auch kognitive und soziale Kompetenzen wie Problemlösung, Entscheidungsfähigkeit, Teamarbeit und Kommunikationsstärke gefördert werden. Aktuell reflektieren die Stellenausschreibungen diese Bedürfnisse noch nicht, da Automatisierung und Digitalisierung im Vordergrund stehen.

Zu den größten Hindernissen zählen zurzeit noch das fehlende Know-how, die mangelnden finanziellen Ressourcen und die ungewisse Rechtslage. Erwähnenswert ist dahingehend auch, dass die effektive Nutzung von KI-Technologien den kostenintensiven Ausbau einer leistungsfähigen digitalen Infrastruktur erfordert.

Die erfolgreiche Integration von KI erfordert eine interdisziplinäre Zusammenarbeit und eine ganzheitliche Strategie, die wirtschaftliche, ethische und soziale Aspekte gleichermaßen berücksichtigt. Nur so kann KI nachhaltig zur Gestaltung einer zukunftsfähigen Arbeitswelt beitragen.

Anhang A:**Chronologische Auflistung der analysierten Studien zu den Auswirkungen von KI auf den Arbeitsmarkt**

2016	Arntz et al. (2016)
2017	Manyika et al. (2017), Frey & Osborene (2017)
2018	
2019	Kovacs-Ondrejkojovic et al. (2019), Wittpahl (2019)
2020	Acemoglu & Restrepo (2020), Anton et al. (2020), Antón et al. (2020), Clifton et al. (2020), Grennan & Michaely (2020), Ratcheva et al. (2020), Schwabe & Castellacci (2020), de Vries et al. (2020)
2021	Abrardi et al. (2021), Acemoglu (2021), Alekseeva et al. (2021), Braganza et al. (2021), Damoli et al. (2021), Dengler & Matthes (2021), Du & Xie (2021), Georgieff & Hyee (2021), Georgieff & Milanzez (2021) Johnson et al. (2021), Klinova & Korinek (2021), Kong et al. (2021), Lane & Saint-Martin (2021), Lyu & Liu (2021), Mingotto et al. (2021), Squicciarini & Nachtigall (2021), Szabo-Szentgroti et al. (2021), Tolan et al. (2021), Tong et al. (2021), Tschang & Almirall (2021), Verma et al. (2021), Xie et al. (2021)
2022	Acemoglu et al. (2022), Alonso et al. (2022), Bordot (2022), Boustani (2022), Fossen & Sorgner (2022), Fuchs et al. (2022), Klenert et al. (2022), Lassebie & Quintini (2022), Liu et al. (2022), Mirbabaie et al. (2022), Mishra et al. (2022), Nguyen & Vo (2022), Nguyen & Malik (2022), Paolillo et al. (2022), Squicciarini & Staccioli (2022), Su et al. (2022), Vu & Lim (2022), Yang (2022)
2023	Bock-Schappelwein & Egger (2023), Broecke (2023), Brynjolfsson et al. (2023), Chui et al. (2023), Dell'Acqua et al. (2023), Dornmayr & Riepl (2023), Eloundou et al. (2023), Fregin et al. (2024), Gmyrek et al. (2023), Green & Lamby (2023), Hammermann et al. (2023), Hatzius et al. (2023), International Labour Organization (2023), Khogali & Mekid (2023), Klingbeil-Döring (2023), Lane et al. (2023), Manca (2023), Milanecz (2023), OECD (2023), Presbitero & Teng-Calleja (2023), Science Media Center Germany (2023), Touzet (2023), World Economic Forum (2023)
2024	Austria Presse Agentur (2024), Babina et al. (2024), Bachmann et al. (2024), Cazzaniga et al. (2024), CIO (2024), Georgieff (2024), Göpfert (2024), Green (2024), Grienberger et al. (2024), Hazan et al. (2024), Hoffmann & Mehta (2024), Käufer & Frein (2024), Law & Shen (2024), Liu et al. (2024), Nguyen & Casey (2024), OECD (2024), Pratiwi & Sa'ad (2024), Shen & Zhang (2024), Singla et al. (2024), Specht (2024), Zarifhonarvar (2024)

Literatur

- Abrardi, L., Cambini, C. & Rondi, L. (2022). Artificial intelligence, firms and consumer behavior: A survey. *Journal of Economic Surveys*, 36(4), 969–991. <https://doi.org/10.1111/joes.12455>
- Acemoglu, D. (2021). *Harms of AI* (NBER Working Paper 29247). National Bureau of Economic Research. <https://doi.org/10.3386/w29247>
- Acemoglu, D., Autor, D., Hazell, J. & Restrepo, P. (2022). Artificial Intelligence and Jobs: Evidence from Online Vacancies. *Journal of Labor Economics*, 40(S1), 293–340. <https://doi.org/10.1086/718327>
- Acemoglu, D. & Restrepo, P. (2020). The wrong kind of AI? Artificial intelligence and the future of labour demand. *Cambridge Journal of Regions Economy and Society*, 13(1), 25–35. <https://doi.org/10.1093/cjres/rsz022>

- Aigner-Walder, B., Luger, A. & Putz, S. (2024). Arbeits- und Fachkräftemangel in Kärnten. In K. Anderwald, K. Hren & K. Stainer-Hämmerle (Hrsg.), *Kärntner Jahrbuch für Politik 2024* (S. 185–199). Mohorjeva Hermagoras.
- Alekseeva, L., Azar, J., Giné, M., Samila, S. & Taska, B. (2021). The demand for AI skills in the labor market. *Labour Economics*, 71, 102002. <https://doi.org/10.1016/j.labeco.2021.102002>
- Alonso, C., Berg, A., Kothari, S., Papageorgiou, C. & Rehman, S. (2022). Will the AI revolution cause a great divergence? *Journal of Monetary Economics*, 127, 18–37. <https://doi.org/10.1016/j.jmoneco.2022.01.004>
- Anton, E., Behne, A. & Teuteberg, F. (2020). THE HUMANS BEHIND ARTIFICIAL INTELLIGENCE – AN OPERATIONALISATION OF AI COMPETENCIES. In *Proceedings of the 28th European Conference on Information Systems (ECIS)*. https://aisel.aisnet.org/ecis2020_rp/141
- Antón, J., Klenert, D., Fernandez Macias, E., Urzi Brancati, M. C. & Alaveras, G. (2020). *The labour market impact of robotisation in Europe* (Nr. JRC121388). European Commission. <https://publications.jrc.ec.europa.eu/repository/handle/JRC121388>
- Arntz, M., Gregory, T. & Zierahn, U. (2016). *The Risk of Automation for Jobs in OECD Countries. A Comparative Analysis* (OECD Social, Employment and Migration Working Papers No. 189). OECD Publishing. <https://doi.org/10.1787/5jlz9h56dvq7-en>
- Austria Presse Agentur. (2024, 1. April). KI wird sich erst in fünf bis zehn Jahren am Arbeitsmarkt auswirken. *Der Standard*. <https://www.derstandard.at/story/3000000213986/ki-wird-sich-erst-in-fuenf-bis-zehn-jahren-am-arbeitsmarkt-auswirken>
- Babina, T., Fedyk, A., He, A. & Hodson, J. (2024). Artificial intelligence, firm growth, and product innovation. *Journal of Financial Economics*, 151, 103745. <https://doi.org/10.1016/j.jfineco.2023.103745>
- Bachmann, R., Gonschor, M., Lewandowski, P. & Madoñ, K. (2024). The impact of Robots on Labour market transitions in Europe. *Structural Change and Economic Dynamics*, 70, 422–441. <https://doi.org/10.1016/j.strueco.2024.05.005>
- Bock-Schappelwein, J. & Egger, A. (2023). *Arbeitsmarkt und Beruf 2030. Rückschlüsse für Österreich* (AMS report 173). Arbeitsmarktservice Österreich.
- Bordot, F. (2022). Artificial Intelligence, Robots and Unemployment: Evidence from OECD Countries. *Journal of Innovation Economics & Management*, 37(1), 117–138. <https://doi.org/10.3917/jie.037.0117>
- Boustani, N. M. (2022). Artificial intelligence impact on banks clients and employees in an Asian developing country. *Journal of Asia Business Studies*, 16(2), 267–278. <https://doi.org/10.1108/jabs-09-2020-0376>
- Braganza, A., Chen, W., Canhoto, A. & Sap, S. (2021). Productive employment and decent work: The impact of AI adoption on psychological contracts, job engagement and employee trust. *Journal of Business Research*, 131, 485–494. <https://doi.org/10.1016/j.jbusres.2020.08.018>
- Broecke, S. (2023). *Artificial intelligence and labour market matching* (OECD Social, Employment and Migration Working Papers No. 284). OECD Publishing. <https://doi.org/10.1787/2b440821-en>
- Brynjolfsson, E., Li, D. & Raymond, L. (2023). *Generative AI at Work* (NBER Working Paper 31161). National Bureau of Economic Research. <https://doi.org/10.3386/w31161>
- Cazzaniga, M., Jaumotte, F., Li, L., Melina, G., Panton, A. J., Pizzinelli, C., Rockall, E. J. & Tavares, M. M. (2024). Gen-AI: Artificial Intelligence and the Future of Work. *IMF Staff Discussion Note*, 2024(001). <https://doi.org/10.5089/9798400262548.006>
- Chui, M., Hazan, E., Roberts, R., Singla, A., Smaje, K., Sukharevsky, A., Yee, L. & Zemmel, R. (2023). *The economic potential of generative AI. The Next Productivity Frontier*. McKinsey & Company. <https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/the-economic-potential-of-generative-ai-the-next-productivity-frontier>

- CIO. (2024, 23. Mai). Deutscher Arbeitsmarkt: KI verändert bis zu 3 Millionen Jobs in Deutschland. *CIO*. <https://www.cio.de/article/3697721/ki-veraendert-bis-zu-3-millionen-jobs-in-deutschland.html>
- Clifton, J., Glasmeier, A. & Gray, M. (2020). When machines think for us: the consequences for work and place. *Cambridge Journal of Regions, Economy and Society*, 13(1), 3–23. <https://doi.org/10.1093/cjres/rsaa004>
- Damioli, G., Van Roy, V. & Vertesy, D. (2021). The impact of artificial intelligence on labor productivity. *Eurasian Economic Review*, 11, 1–25. <https://doi.org/10.1007/s40821-020-00172-8>
- De Vries, G. J., Gentile, E., Miroudot, S. & Wacker, K. M. (2020). The rise of robots and the fall of routine jobs. *Labour Economics*, 66, 101885. <https://doi.org/10.1016/j.labeco.2020.101885>
- Dell’Acqua, F., McFowland, E., Mollick, E. R., Lifshitz-Assaf, H., Kellogg, K., Rajendran, S., Kraymer, L., Candelon, F. & Lakhani, K. R. (2023). *Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality* (Harvard Business School Technology & Operations Mgt. Unit Working Paper No. 24-013, The Wharton School Research Paper). <https://doi.org/10.2139/ssrn.4573321>
- Dengler, K. & Matthes, B. (2021). *Folgen des technologischen Wandels für den Arbeitsmarkt: Auch komplexere Tätigkeiten könnten zunehmend automatisiert werden* (IAB-KURZBERICHT 13/2021). <https://doku.iab.de/kurzber/2021/kb2021-13.pdf>
- Dornmayr, H. & Riepl, M. (2023). *Unternehmensbefragung zum Arbeits- und Fachkräftebedarf/-mangel. Arbeitskräftenadar 2023* (Nr. ibw-Forschungsbericht Nr. 215). <https://ibw.at/publikationen/id/562/>
- Du, S. & Xie, C. (2021). Paradoxes of artificial intelligence in consumer markets: Ethical challenges and opportunities. *Journal of Business Research*, 129, 961–974. <https://doi.org/10.1016/j.jbusres.2020.08.024>
- Eloundou, T., Manning, S., Mishkin, P. & Rock, D. (2023). *GPTs are GPTs: An Early Look at the Labor Market Impact Potential of Large Language Models*. <https://doi.org/10.48550/arxiv.2303.10130>
- Fossen, F. M. & Sorgner, A. (2022). New digital technologies and heterogeneous wage and employment dynamics in the United States: Evidence from individual-level data. *Technological Forecasting & Social Change*, 175, 121381. <https://doi.org/10.1016/j.techfore.2021.121381>
- Fregin, M.-C., Koch, T., Malfertheimer, V., Özgül, P. & Stops, M. (2023). *Automatisierungspotenziale von beruflichen Tätigkeiten: Künstliche Intelligenz und Software – Beschäftigte sind unterschiedlich betroffen* (IAB-Kurzbericht 21/2023). <https://doi.org/10.48720/IAB.KB.2321>
- Frey, C. B. & Osborne, M. A. (2017). The future of employment: How susceptible are jobs to computerisation? *Technological Forecasting and Social Change*, 114, 254–280. <https://doi.org/10.1016/j.techfore.2016.08.019>
- Fuchs, B., Schumacher, A., Eggeling, E. & Schlund, S. (2022). *Künstliche Intelligenz in Österreichs Unternehmen. Fraunhofer Austria KI-Studie*. <https://doi.org/10.24406/austria-n-647745>
- Georgieff, A. (2024). *Artificial intelligence and wage inequality* (OECD Artificial Intelligence Papers No. 13). OECD Publishing. <https://doi.org/10.1787/bf98a45c-en>
- Georgieff, A. & Hye, R. (2021). *Artificial intelligence and employment* (OECD Social, Employment and Migration Working Papers No. 265). OECD Publishing. <https://doi.org/10.1787/c2c1d276-en>
- Georgieff, A. & Milanez, A. (2021). “What happened to jobs at high risk of automation?” (OECD Social, Employment and Migration Working Papers No. 255). OECD Publishing. <https://doi.org/10.1787/10bc97f4-en>

- Gmyrek, P., Berg, J. & Bescond, D. (2023). *Generative AI and jobs: a global analysis of potential effects on job quantity and quality* (ILO Working Paper 96). ILO. <https://doi.org/10.54394/fhem8239>
- Göpfert, B. (2024, 19. Juni). Der Wandel steht unmittelbar bevor: Generative KI und die Auswirkungen auf den Arbeitsmarkt. *Handelsblatt*. <https://live.handelsblatt.com/der-wandel-steht-unmittelbar-bevor-generative-ki-und-die-auswirkungen-auf-den-arbeitsmarkt/>
- Green, A. (2024). *Artificial intelligence and the changing demand for skills in Canada* (OECD Artificial Intelligence Papers No. 17). OECD Publishing. <https://doi.org/10.1787/1b20cdb6-en>
- Green, A. & Lamby, L. (2023). *The supply, demand and characteristics of the AI workforce across OECD countries* (OECD Social, Employment and Migration Working Papers No. 287). OECD Publishing. <https://doi.org/10.1787/bb17314a-en>
- Grennan, J. & Michaely, R. (2020). *Artificial Intelligence and High-Skilled Work: Evidence from Analysts* (Swiss Finance Institute Research Paper No. 20-84). <https://doi.org/10.2139/ssrn.3681574>
- Grienberger, K., Matthes, B. & Paulus, W. (2024). *Folgen des technologischen Wandels für den Arbeitsmarkt: Vor allem Hochqualifizierte bekommen die Digitalisierung verstärkt zu spüren* (IAB-Kurzbericht 5/2024). <https://doi.org/10.48720/iab.kb.2405>
- Hammermann, A., Monsef, R. P., Stettes, O. & Institut der deutschen Wirtschaft Köln e. V. (2023). *KI und der Arbeitsmarkt: Eine Analyse der Beschäftigungseffekte* (Report IW-Report 55/2023). Institut der deutschen Wirtschaft Köln e.V. <https://www.iwkoeln.de/studien/andrea-hammermann-roschan-pourkhataei-monsef-oliver-stettes-eine-analyse-der-beschaeftigungseffekte.html>
- Hatzius, J., Briggs, J., Kodnani, D. & Pierdomenico, G. (2023). *The Potentially Large Effects of Artificial Intelligence on Economic Growth*. Goldman Sachs Economics Research. <https://www.gspublishing.com/content/research/en/reports/2023/03/27/d64e052b-0f6e-45d7-967b-d7be35fabd16.pdf>
- Hazan, E., Madgavkar, A., Chui, M., Smit, S., Maor, D., Dandona, G. S. & Huyghues-Despointes, R. (2024). *A new future of work: The race to deploy AI and raise skills in Europe and beyond*. McKinsey Global Institute. <https://www.mckinsey.com/mgi/our-research/a-new-future-of-work-the-race-to-deploy-ai-and-raise-skills-in-europe-and-beyond>
- Hoffmann, J. & Mehta, D. (2024). *Industry 4.0: in-depth market analysis* (Market Insights Report). Statista. <https://www.statista.com/study/66974/in-depth-report-industry-40/>
- Humlum, A. & Vestergaard, E. (2025). *Large Language Models, Small Labor Market Effects* (University of Chicago, Becker Friedman Institute for Economics Working Paper No. 2025-56). <https://doi.org/10.2139/ssrn.5219933>
- International Labour Organization. (2023). *Digital Employment Diagnostic Guidelines*. ILO. https://www.ilo.org/sites/default/files/wcmsp5/groups/public/%40ed_emp/documents/publication/wcms_901252.pdf
- Johnson, M., Jain, R., Brennan-Tonetta, P., Swartz, E., Silver, D., Paolini, J., Mamonov, S. & Hill, C. (2021). Impact of Big Data and Artificial Intelligence on Industry: Developing a Workforce Roadmap for a Data Driven Economy. *Global Journal of Flexible Systems Management*, 22(3), 197–217. <https://doi.org/10.1007/s40171-021-00272-y>
- Käuer, T. & Frein, J. (2024). *Trends in der Arbeitswelt* (Statista Trend Report). Statista. <https://de.statista.com/statistik/studie/id/62636/dokument/neue-arbeitswelt/>
- Khogali, H. O. & Mekid, S. (2023). The blended future of automation and AI: Examining some long-term societal and ethical impact features. *Technology in Society*, 73, 102232. <https://doi.org/10.1016/j.techsoc.2023.102232>
- Klenert, D., Fernández-Macías, E. & Antón, J. (2022). Do robots really destroy jobs? Evidence from Europe. *Economic and Industrial Democracy*, 44(1), 280–316. <https://doi.org/10.1177/0143831x211068891>

- Klingbeil-Döring, W. (2023, 29. Juni). Die Auswirkungen von Künstlicher Intelligenz auf den Arbeitsmarkt. *Bundeszentrale für politische Bildung*. <https://www.bpb.de/themen/arbeit/arbeitsmarktpolitik/522513/die-auswirkungen-von-kuenstlicher-intelligenz-auf-den-arbeitsmarkt/>
- Klinova, K. & Korinek, A. (2021). AI and Shared Prosperity. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (AIES '21)* (S. 645–651). Association for Computing Machinery. <https://doi.org/10.1145/3461702.3462619>
- Kong, H., Yuan, Y., Baruch, Y., Bu, N., Jiang, X. & Wang, K. (2021). Influences of artificial intelligence (AI) awareness on career competency and job burnout. *International Journal of Contemporary Hospitality Management*, 33(2), 717–734. <https://doi.org/10.1108/ijchm-07-2020-0789>
- Kovács-Ondrejčovic, O., Strack, R., Antebi, P., Gobernado, A. L. & Lyle, E. (2019). *Decoding Global Trends in Upskilling and Reskilling*. Boston Consultancy Group. <https://www.bcg.com/publications/2019/decoding-global-trends-upskilling-reskilling>
- Lane, M. & Saint-Martin, A. (2021). *The impact of Artificial Intelligence on the labour market: What do we know so far?* (OECD Social, Employment and Migration Working Papers Nr. 256). OECD Publishing. <https://doi.org/10.1787/7c895724-en>
- Lane, M., Williams, M. & Broecke, S. (2023). *The impact of AI on the workplace: Main findings from the OECD AI surveys of employers and workers* (OECD Social, Employment and Migration Working Papers No. 288). OECD Publishing. <https://doi.org/10.1787/ea0a0fe1-en>
- Lassébie, J. & Quintini, G. (2022). *What skills and abilities can automation technologies replicate and what does it mean for workers?* (OECD Social, Employment and Migration Working Papers No. 282). OECD Publishing. <https://doi.org/10.1787/646aad77-en>
- Law, K. K. F. & Shen, M. (2024). How does artificial intelligence shape audit firms? *Management Science*, 71(5), 3641–3666. <https://doi.org/10.1287/mnsc.2022.04040>
- Liu, J., Liu, L., Qian, Y. & Song, S. (2022). The effect of artificial intelligence on carbon intensity: Evidence from China's industrial sector. *Socio-Economic Planning Sciences*, 83, 101002. <https://doi.org/10.1016/j.seps.2020.101002>
- Liu, J., Xu, X., Nan, X., Li, Y., & Tan, Y. (2024). “Generate” the Future of Work through AI: Empirical Evidence from Online Labor Markets [v2]. <https://doi.org/10.48550/arXiv.2308.05201>
- Lyu, W. & Liu, J. (2021). Artificial Intelligence and emerging digital technologies in the energy sector. *Applied Energy*, 303, 117615. <https://doi.org/10.1016/j.apenergy.2021.117615>
- Manca, F. (2023). *Six questions about the demand for artificial intelligence skills in labour markets* (OECD Social, Employment and Migration Working Papers No. 286). OECD Publishing. <https://doi.org/10.1787/ac1bebf0-en>
- Manyika, J., Lund, S., Chui, M., Bughin, J., Woetzel, J., Batra, P., Ko, R. & Sanghvi, S. (2017). *Jobs Lost, Jobs Gained: Workforce Transitions in a Time of Automation*. McKinsey Global Institute.
- Milanez, A. (2023). *The impact of AI on the workplace: Evidence from OECD case studies of AI implementation* (OECD Social, Employment and Migration Working Papers No. 289). OECD Publishing. <https://doi.org/10.1787/2247ce58-en>
- Mingotto, E., Montaguti, F. & Tamma, M. (2021). Challenges in re-designing operations and jobs to embody AI and robotics in services. Findings from a case in the hospitality industry. *Electronic Markets*, 31(3), 493–510. <https://doi.org/10.1007/s12525-020-00439-y>
- Mirbabaie, M., Brünker, F., Frick, N. R. J. M. & Stieglitz, S. (2022). The rise of artificial intelligence – understanding the AI identity threat at the workplace. *Electronic Markets*, 32, 73–99. <https://doi.org/10.1007/s12525-021-00496-x>

- Mishra, S., Ewing, M. T. & Cooper, H. B. (2022). Artificial intelligence focus and firm performance. *Journal of the Academy of Marketing Science*, 50(6), 1176–1197. <https://doi.org/10.1007/s11747-022-00876-5>
- Nguyen, Q. P. & Vo, D. H. (2022). Artificial intelligence and unemployment: An international evidence. *Structural Change and Economic Dynamics*, 63, 40–55. <https://doi.org/10.1016/j.strueco.2022.09.003>
- Nguyen, T. & Casey, D. (2024). *Emerging Tech Impact Radar: 2024*. Gartner Research. <https://www.gartner.com/en/documents/5097331>
- Nguyen, T. & Malik, A. (2021). A Two-Wave Cross-Lagged Study on AI Service Quality: The Moderating Effects of the Job Level and Job Role. *British Journal of Management*, 33(3), 1221–1237. <https://doi.org/10.1111/1467-8551.12540>
- OECD. (2023). *OECD Employment Outlook 2023: Artificial Intelligence and the Labour Market*. OECD Publishing. <https://doi.org/10.1787/08785bba-en>
- OECD. (2024). *Using AI in the workplace* (OECD Artificial Intelligence Papers No. 11). OECD Publishing. <https://doi.org/10.1787/73d417f9-en>
- Paolillo, A., Colella, F., Nosengo, N., Schiano, F., Stewart, W., Zambrano, D., Chappuis, I., Lalive, R. & Floreano, D. (2022). How to compete with robots by assessing job automation risks and resilient alternatives. *Science Robotics*, 7(65). <https://doi.org/10.1126/scirobotics.abg5561>
- Pratiwi, H., & Sa’ad, M. I. (2024). The Impact of Artificial Intelligence (AI) on the Future of Jobs in Computer Science and Information Systems: A Systematic Literature Review. *Information Technology Studies Journal*, 1(1), 41–52. <https://doi.org/10.62207/vdebmf42>
- Presbitero, A. & Teng-Calleja, M. (2023). Job attitudes and career behaviors relating to employees’ perceived incorporation of artificial intelligence in the workplace: a career self-management perspective. *Personnel Review*, 52(4), 1169–1187. <https://doi.org/10.1108/pr-02-2021-0103>
- Ratcheva, V., Leopold, T. A. & Zahidi, S. (2020). *Jobs of Tomorrow: Mapping Opportunity in the New Economy*. World Economic Forum. <https://www.weforum.org/publications/jobs-of-tomorrow-mapping-opportunity-in-the-new-economy/>
- Schmid, K. & Nowak, S. (2023). *Prognose des Arbeitskräfteangebotes in Kärnten: Zentrale Ergebnisse einer aktuellen Studie im Auftrag des AMS Kärnten* (AMS info Nr. 647). Arbeitsmarktservice Österreich.
- Schwabe, H. & Castellacci, F. (2020). Automation, workers’ skills and job satisfaction. *PLOS ONE*, 15(11), e0242929. <https://doi.org/10.1371/journal.pone.0242929>
- Science Media Center Germany. (2023, 12. Mai). *KI und Arbeitsmarkt – welche Effekte sind zu erwarten?* <https://www.sciencemediacenter.de/angebote/23066>
- Shen, Y. & Zhang, X. (2024). The impact of artificial intelligence on employment: the role of virtual agglomeration. *Humanities and Social Sciences Communications*, 11(1), 122. <https://doi.org/10.1057/s41599-024-02647-9>
- Singla, A., Sukharevsky, A., Yee, L., Chui, M. & Hall, B. (2024). *The state of AI in early 2024: Gen AI adoption spikes and starts to generate value*. QuantumBlack, AI by McKinsey. <https://www.mckinsey.com/~media/mckinsey/business%20functions/quantumblack/our%20insights/the%20state%20of%20ai/2024/the-state-of-ai-in-early-2024-final.pdf>
- Specht, F. (2024, 13. März). KI macht zunehmend Hochqualifizierten Konkurrenz. *Handelsblatt*. <https://www.handelsblatt.com/politik/deutschland/arbeitsmarkt-ki-macht-zunehmend-hochqualifizierten-konkurrenz/100023184.html>
- Squicciarini, M. & Nachtigall, H. (2021). *Demand for AI skills in jobs* (OECD Science, Technology and Industry Working Papers No. 2021/03). OECD Publishing. <https://doi.org/10.1787/3ed32d94-en>

- Squicciarini, M. & Staccioli, J. (2022). *Labour-saving technologies and employment levels: Are robots really making workers redundant?* (OECD Science, Technology and Industry Policy Papers No. 124). OECD Publishing. <https://doi.org/10.1787/9ce86ca5-en>
- Su, C., Yuan, X., Umar, M. & Lobonț, O. (2022). Does technological innovation bring destruction or creation to the labor market? *Technology in Society*, 68, 101905. <https://doi.org/10.1016/j.techsoc.2022.101905>
- Szabó-Szentgróti, G., Végvári, B. & Varga, J. (2021). Impact of Industry 4.0 and Digitization on Labor Market for 2030-Verification of Keynes' Prediction. *Sustainability*, 13(14), 7703. <https://doi.org/10.3390/su13147703>
- Tolan, S., Pesole, A., Martínez-Plumed, F., Fernández-Macías, E., Hernández-Orallo, J. & Gómez, E. (2021). Measuring the Occupational Impact of AI: Tasks, Cognitive Abilities and AI Benchmarks. *Journal of Artificial Intelligence Research*, 71, 191–236. <https://doi.org/10.1613/jair.1.12647>
- Tong, S., Jia, N., Luo, X. & Fang, Z. (2021). The Janus face of artificial intelligence feedback: Deployment versus disclosure effects on employee performance. *Strategic Management Journal*, 42(9), 1600–1631. <https://doi.org/10.1002/smj.3322>
- Touzet, C. (2023). *Using AI to support people with disability in the labour market: Opportunities and challenges* (OECD Artificial Intelligence Papers No. 7). OECD Publishing. <https://doi.org/10.1787/008b32b7-en>
- Tschang, F. T. & Almirall, E. (2021). Artificial Intelligence as Augmenting Automation: Implications for Employment. *Academy of Management Perspectives*, 35(4), 642–659. <https://doi.org/10.5465/amp.2019.0062>
- Verma, A., Lamsal, K. & Verma, P. (2021). An investigation of skill requirements in artificial intelligence and machine learning job advertisements. *Industry and Higher Education*, 36(1), 63–73. <https://doi.org/10.1177/0950422221990990>
- Vu, H. T. & Lim, J. (2022). Effects of country and individual factors on public acceptance of artificial intelligence and robotics technologies: a multilevel SEM analysis of 28-country survey data. *Behaviour & Information Technology*, 41(7), 1515–1528. <https://doi.org/10.1080/0144929x.2021.1884288>
- Wittpahl, V. (2018). *Künstliche Intelligenz: Technologien | Anwendung | Gesellschaft*. Springer Vieweg. <https://doi.org/10.1007/978-3-662-58042-4>
- World Economic Forum. (2023). *The Future of Jobs Report 2023*. <https://www.weforum.org/publications/the-future-of-jobs-report-2023/>
- World Economic Forum. (2025). *The Future of Jobs Report 2025*. <https://www.weforum.org/publications/the-future-of-jobs-report-2025/>
- Xie, M., Ding, L., Xia, Y., Guo, J., Pan, J. & Wang, H. (2021). Does artificial intelligence affect the pattern of skill demand? Evidence from Chinese manufacturing firms. *Economic Modelling*, 96, 295–309. <https://doi.org/10.1016/j.econmod.2021.01.009>
- Yang, C. (2022). How Artificial Intelligence Technology Affects Productivity and Employment: Firm-level Evidence from Taiwan. *Research Policy*, 51(6), 104536. <https://doi.org/10.1016/j.respol.2022.104536>
- Zarifhonarvar, A. (2023). Economics of ChatGPT: a labor market view on the occupational impact of artificial intelligence. *Journal Of Electronic Business & Digital Economics*, 3(2), 100–116. <https://doi.org/10.1108/jebde-10-2023-0021>

Teil II:
Vertrauen, Transparenz und Partizipation:
Herausforderungen für den KI-Einsatz

KI in der Unternehmenskommunikation: vom Tool zum Teammitglied?

Julia Eisner

1. Einleitung

Die Integration von Künstlicher Intelligenz (KI) in die Unternehmenskommunikation markiert einen Paradigmenwechsel, der weit über die technologische Ebene hinausreicht. Insbesondere die Entwicklung kommunikativer KI-Systeme führt dazu, dass Maschinen zunehmend nicht nur als Werkzeuge zur Informationsverarbeitung, sondern als eigenständige Akteure innerhalb kommunikativer Prozesse verstanden werden (Guzman & Lewis, 2020; Hepp et al., 2023). Diese Verschiebung bringt neue Chancen, aber auch Herausforderungen mit sich, die technologische, organisationale und ethische Dimensionen gleichermaßen berühren.

Der Aufstieg generativer KI-Modelle wie ChatGPT hat nicht nur die Alltagskommunikation verändert, sondern insbesondere die Arbeit von Kommunikationsprofessionals grundlegend beeinflusst (Banholzer et al., 2023). Systeme wie Chatbots initiieren einen tiefgreifenden Wandel in der Interaktion zwischen Organisationen und Stakeholdern (Hepp et al., 2023; Stieglitz & Wiencierz, 2022) und machen es notwendig, die kommunikativen Fähigkeiten von KI neu zu verstehen (Weller & Lock, 2024). Während die öffentliche Diskussion oft Effizienz- und Profitabilitätsaspekte hervorhebt, richtet die wissenschaftliche Forschung zunehmend den Blick auf die relationalen und sozialen Dimensionen von KI (z.B. Buhmann & Gregory, 2023; Zerfass et al., 2024).

Vor dem Hintergrund dieser Entwicklungen untersucht dieser Beitrag, wie KI heute praktisch in der Unternehmenskommunikation eingesetzt wird, welche Erwartungen Unternehmen mit dem Einsatz dieser Technologien verbinden und welche Spannungsfelder sich daraus ergeben. Im Fokus steht dabei nicht nur die technische Implementierung von KI-Systemen, sondern vor allem die Frage, wie sich Kommunikationspraktiken, Berufsbilder und professionelle Standards im Kontext hybrider Mensch-Maschine-Interaktionen verändern.

Theoretisch verortet sich die Analyse im Rahmen des kommunikativen Konstruktivismus (Keller et al., 2013; Knoblauch & Pfadenhauer, 2023), der Kommunikation als zentralen Mechanismus der sozialen Wirklichkeitskonstruktion versteht. In dieser Perspektive wird kommunikative KI nicht bloß als technologische Innovation betrachtet, sondern als sozial konstruiertes Phänomen, das Bedeutungen, Identitäten und soziale Rollen neu formt. Indem Maschinen menschliche Kommunikationsmuster adaptieren und intentionale Zuschreibungen erfahren, werden sie Teil der Prozesse der Sinnkonstruktion in Organisationen.

Ziel dieses Beitrags ist es, die Transformation der Unternehmenskommunikation durch KI differenziert darzustellen und kritisch zu reflektieren. Dabei wird gezeigt, dass die Herausforderungen der KI-Integration nicht allein technischer Natur sind, sondern tief in die sozialen, kulturellen und normativen Strukturen von Organisationen eingreifen (Buhmann & White, 2022; Crawford, 2021). Der Beitrag schließt mit einem Ausblick auf zukünftige Anforderungen an Kommunikationspro-

professionals sowie auf Forschungsfragen, die sich aus der zunehmenden Hybridisierung kommunikativer Prozesse ergeben. Dabei wird insbesondere der Bedarf an neuen Kompetenzen wie AI Literacy und Data Literacy hervorgehoben, um eine verantwortungsvolle Gestaltung dieser Transformationsprozesse zu ermöglichen (Gregory & Virmani, 2020; Soriano & Valdés, 2021).

2. Theoretischer Rahmen

2.1 Historische Entwicklung: von Kybernetik zu automatisierter Kommunikation

Die Automatisierung von Kommunikation ist kein neues Phänomen. Ihre Ursprünge reichen bis in die frühen Entwicklungen der Kybernetik und der Medientheorien der 1970er-Jahre zurück. Bereits in dieser Phase legten Forscher*innen wie Turner (2006) den Schwerpunkt auf die technischen Dimensionen von Kommunikation. Diese technikorientierte Perspektive stand im Gegensatz zur damaligen sozialwissenschaftlichen Forschungstradition, die Kommunikation primär als sinnstiftenden, sozialen Prozess betrachtete (Baecker, 1997).

Diese historische Differenz markiert den Beginn einer bis heute andauernden Entwicklung, die Hepp et al. (2023) in eine dreistufige Evolution der Forschung zu digitaler Kommunikation einteilen. In einer ersten Phase stand die Veränderung von Kommunikations- und Sozialbeziehungen durch digitale Medien im Vordergrund. Die Entstehung des Konzepts der *computervermittelten Kommunikation* (Computer-Mediated Communication, CMC) (Chesebro & Bonsall, 1989) reflektierte die frühen Bemühungen, die Besonderheiten der digitalen Interaktion zwischen Menschen zu verstehen. Parallel dazu etablierte sich die Diskussion über die Informationsgesellschaft als neuer sozialer Formation (Castells, 2000). Die zweite Phase war geprägt von der zunehmenden Auseinandersetzung mit der Datafizierung sozialer Prozesse und Big Data in der Gesellschaft (Dencik & Kaun, 2020; Flensburg & Lomborg, 2021). Große Datenmengen und ihre algorithmische Verarbeitung wurden zum Gegenstand kritischer Auseinandersetzung, was in die Entwicklung der *Critical Data Studies* mündete (Burns et al., 2019; Dalton & Thatcher, 2014).

Seit etwa Beginn der 2020er-Jahre lässt sich eine dritte Phase beobachten – die Automatisierung von Kommunikation. Zentrale Begriffe für das Verständnis dieser Entwicklungen sind Digitalisierung, Datafizierung und Algorithmisierung. Diese Phase ist durch die zunehmende Automatisierung kommunikativer Prozesse und durch den Einsatz von KI in kommunikativen Kontexten gekennzeichnet (Hepp et al., 2023). Dazu zählen Phänomene wie *Robot Journalism* (Carlson, 2016), der Einsatz von *Social Bots* (Gehl & Bakardjieva, 2016) sowie die *algorithmische Inhaltsmoderation* auf digitalen Plattformen (Gorwa et al., 2020). Diese Entwicklung beschreibt eine neue Realität, in der kommunikative Prozesse zunehmend automatisiert und maschinell gestaltet werden.

Die technischen Grundlagen dieser Entwicklungen umfassen sowohl einfache Automatisierungsskripte, wie sie etwa für Social Bots genutzt werden (Veale & Cook, 2018), als auch hochkomplexe Machine-Learning-Systeme (Heuer et al., 2021). Die Kommunikation über diese Systeme basiert auf digitalen Spuren, die

durch Nutzungsdaten und Interaktionsmuster erzeugt werden. Diese digital erzeugten „Spuren“ besitzen eine eigene Materialität, die jedoch weniger sichtbar und greifbar ist als klassische materielle Produktionsprozesse (Burrell, 2016).

Gerade die „unsichtbare“ Materialität (Esposito, 2017) verändert die kommunikativen Prozesse in einer Vielzahl von Feldern, darunter Gesundheit, Justiz, Politik, Journalismus, Kommunikation und Wissenschaft. Dabei agieren diese Prozesse stets eingebettet in zunehmend globalisierte digitale Infrastruktur (Crawford, 2021), die Kommunikationssysteme über nationale Grenzen hinweg vernetzt und transformiert. Kommunikationsautomation beruht somit auf digitalisierten Datenspuren, deren Prozessualität und Undurchsichtigkeit soziale Realitäten auf neue Weise prägen und von diesen geprägt werden. Hepp et al. (2023) betonen deshalb, dass der Blick auf kommunikative Automatisierung eine ganzheitliche Perspektive einfordern muss. Es reicht nicht, technische Phänomene isoliert zu betrachten. Vielmehr müssen die gesellschaftlichen, kulturellen und materiellen Dimensionen der Digitalisierung integriert analysiert werden, um die neuen Formen sozialer Wirklichkeit vollständig zu verstehen.

2.2 Kommunikation als sozialer Konstruktionsprozess

Parallel zur technischen Evolution der Kommunikation entwickelte sich in den Sozialwissenschaften eine theoretische Perspektive, die Kommunikation nicht als reine Informationsübertragung, sondern als zentrale Instanz der sozialen Wirklichkeitskonstruktion versteht. Bereits seit den 1960er-Jahren hat sich der Konstruktivismus als wissenschaftliches Paradigma etabliert und spiegelt damit den gesellschaftlichen Wandel von der Industrie- zur Wissensgesellschaft wider (Keller et al., 2013). Ein zentraler Ausgangspunkt dieser Entwicklung ist das Werk von Peter L. Berger und Thomas Luckmann, *The Social Construction of Reality* (1970). Mit diesem Buch begründeten sie die moderne Wissenssoziologie neu und prägten nachhaltig die Vorstellung, dass gesellschaftliche Wirklichkeit ein Produkt sozialer Interaktionen und Deutungsprozesse ist. Im Gegensatz zu späteren radikal- oder postmodern-konstruktivistischen Strömungen betonten Berger und Luckmann jedoch, dass soziale Konstruktionen nicht beliebig oder völlig relativ seien, sondern innerhalb sozialer Ordnungen strukturiert und stabilisiert werden (Keller et al., 2013, S. 11).

Der kommunikative Konstruktivismus, wie er insbesondere von Knoblauch und Pfadenhauer (2023) diskutiert wird, greift diese Grundidee auf und führt sie konsequent weiter. Im Zentrum steht hier nicht mehr nur das Bewusstsein oder das individuelle Handeln, sondern die kommunikativen Prozesse selbst. Realität ist demnach nicht bloß das Produkt individueller Sinnzuschreibungen, sondern wird in kommunikativen Prozessen sozial erzeugt und stabilisiert. Kommunikation wird damit zur zentralen Instanz, über die Identitäten, soziale Beziehungen, gesellschaftliche Ordnungen und schließlich die gesamte soziale Wirklichkeit hergestellt und reproduziert werden. Medien und Kommunikationsprozesse sind heute nicht mehr nur Mittel zur Übertragung von Informationen, sondern konstitutive Elemente gesellschaftlicher Realität (Keller et al., 2013, Knoblauch & Pfadenhauer, 2023). Kommunikation wird so zur empirisch beobachtbaren Seite des Sozialen – sie bildet die Grundlage für jede Form von Sozialität und sozialen Wandel (Keller et al., 2013).

Die Medien- und Kommunikationswissenschaften haben diese Entwicklungen aufgegriffen. Insbesondere die Theorien zur Mediatisierung und Digitalisierung (Hepp & Hasebrink, 2017) reflektieren die wachsende Allgegenwart von Medien in allen Lebensbereichen. *Mediatisierung* bezeichnet hier das komplexe Wechselspiel zwischen medialen Innovationen und gesellschaftlichem Wandel. Wesentliche Trends dieser Entwicklung sind die fortschreitende Differenzierung von Medienangeboten, die zunehmende Vernetzung sowie die Omnipräsenz medialer Strukturen im Alltag.

In diesem Zusammenhang entstehen neue theoretische Ansätze, die versuchen, die Verschränkung von Individuen, Medien und Kommunikationen konzeptionell zu fassen. Ein Ansatz sind kommunikative Figurationen (Hepp et al., 2023). Figurationen sind soziale Konstellationen, in denen Akteure durch geteilte Relevanzrahmen und Machtverhältnisse miteinander verbunden sind und sich in und durch kommunikative Praktiken konstituieren. Diese Praktiken wiederum sind tief durch medien-spezifische Trends wie Differenzierung, Vernetzung und Datafizierung geprägt. Die Analyse kommunikativer Figurationen erfordert dabei einen Fokus auf die Mesoebene sozialer Organisationen, da sich hier die unterschiedlichen Dynamiken von Medienwandel, Digitalisierung und gesellschaftlichem Strukturwandel konkretisieren (Hepp & Görland, 2024).

Insgesamt eröffnet der kommunikative Konstruktivismus in Kombination mit Ansätzen wie Mediatisierung und kommunikativen Figurationen ein theoretisches Fundament, um die zunehmende Verflechtung von Kommunikation, Medien und Gesellschaft in der digitalen Ära systematisch zu erfassen. Diese Perspektive bildet zugleich die Grundlage für das Verständnis neuer Phänomene wie kommunikative KI, die im nächsten Abschnitt spezifiziert werden. In dieser Tradition werden KI-Systeme nicht als neutrale Werkzeuge, sondern als aktive Mitproduzenten von Kommunikation und somit der sozialen Wirklichkeit betrachtet.

2.3 Kommunikative KI

Die zunehmende Automatisierung kommunikativer Prozesse im Zuge der Digitalisierung verlangt eine Neubestimmung zentraler theoretischer Begriffe. Im internationalen Kontext hat sich das Konzept der kommunikativen Künstlichen Intelligenz (communicative AI) als sensibilisierendes Konzept (sensitizing concept im Sinne von Blumer, 1954) (Andrejevic, 2020; Gambino et al., 2020; Hepp, 2020) innerhalb des neu entstehenden Forschungsfeldes „Automation of Communication“ etabliert (Hepp et al., 2023).

In diesem Rahmen dient das Konzept der kommunikativen KI dazu, die spezifischen Dynamiken und Veränderungen zu erfassen, die auftreten, wenn Systeme wie ChatGPT oder andere KI-basierte Technologien zu eigenständigen Akteuren gesellschaftlicher Kommunikation werden.

Hepp et al. (2023, S. 48) definieren *kommunikative KI* anhand von drei zentralen Kriterien:

- 1) Sie basiert auf verschiedenen Formen der Automatisierung, die gezielt auf Kommunikation ausgerichtet sind.
- 2) Sie ist eingebettet in digitale Infrastrukturen.
- 3) Sie ist untrennbar mit menschlichen Praktiken verflochten.

Diese Definition unterscheidet sich klar von funktionalen Konzepten klassischer KI, die oft auf technische Effizienz oder Automatisierung von Routinetätigkeiten fokussieren. Bei kommunikativer KI steht hingegen kommunikatives Handeln als sozialer und gesellschaftlich relevanter Prozess im Mittelpunkt.

Wie Natale (2021, S. 69–86) herausarbeitet, ist „Intelligenz“ im Kontext von KI keine objektive Eigenschaft, sondern ein sozial konstruierter Zuschreibungsprozess. In Bezug auf kommunikative KI zeigt sich, dass Intelligenz und Agency (Handlungsfähigkeit) technologischer Systeme stark durch kulturelle Bedeutungszuschreibungen geprägt sind (Pfadenhauer & Lehmann, 2022). Systeme wie Chatbots, Sprachassistenten oder algorithmische Plattformen sind tief in komplexe digitale Infrastrukturen eingebettet (Crawford, 2021; Hepp et al., 2023). Diese technische Materialität ist nicht neutral, sondern prägt maßgeblich, welche Kommunikationspraktiken ermöglicht oder begrenzt werden (Pfadenhauer & Grenz, 2017). Aspekte wie Energieverbrauch, Plattformregulierung oder Datenpolitik gehören ebenso zur Analyse dieser neuen Kommunikationsarchitekturen wie die konkreten Interfaces der Systeme selbst (Hepp et al., 2023). Die Materialität kommunikativer KI umfasst damit sowohl die unmittelbar sichtbaren Oberflächen der Kommunikation als auch die unsichtbaren Schichten technischer und sozialer Infrastrukturen (Pfadenhauer & Lehmann, 2022).

3. Kommunikative KI in der Unternehmenskommunikation – aktuelle Entwicklungen und Herausforderungen

3.1 Anwendungsfelder und Erkenntnisse aus der Praxis

Die Integration von KI-Technologien in die Unternehmenskommunikation hat sich in den vergangenen Jahren rasant entwickelt. Besonders die Verfügbarkeit generativer Systeme wie ChatGPT hat die Diskussion um den Einsatz von KI im Kommunikationsbereich erheblich intensiviert (Hepp et al., 2023). Erhebungen und Reports aus der Unternehmenskommunikation wie der *European Communication Monitor 2024/25* (Zerfass et al., 2024) und der *CommTech Index Report 2024/25* (AG CommTech, 2024) zeigen aktuelle Entwicklungen in der Kommunikationsprofession auf. Diese werden nachstehend diskutiert und in Tabelle 1 dargestellt.

Der häufigste Einsatzbereich von KI liegt demnach derzeit noch in der Inspiration für Inhalte und der Content-Erstellung. Generative KI-Systeme werden genutzt, um Entwürfe für Pressemitteilungen, Social-Media-Posts und andere Inhalte zu erstellen. Dabei reicht die Unterstützung von einfachen Textbausteinen bis hin zu komplexen, mehrsprachigen Kampagnentexten, die speziell für verschiedene Märkte lokalisiert werden. Auch die Echtzeit-Übersetzung von Reden oder Unternehmensnews spielt eine wachsende Rolle (z. B. Fekade et al., 2024; Yue et al., 2024).

Ein weiteres Einsatzfeld, das sich beginnt zu entwickeln, betrifft das Monitoring und Issue Management. Kommunikationsabteilungen verwenden KI-gestützte Systeme, um die Berichterstattung über Unternehmen, Marken und Branchen in Echtzeit zu analysieren. Diese Systeme bieten nicht nur eine Tonalitätsanalyse der Medienresonanz, sondern helfen auch dabei, Frühwarnsysteme für potenzielle Reputationskrisen zu etablieren. Predictive Analytics ermöglicht es dabei, Trends und Diskursverschiebungen frühzeitig zu erkennen und strategisch zu adressieren.

Dialogorientierte KI-Systeme wie Chatbots und Voicebots werden zunehmend sowohl in der externen als auch in der internen Kommunikation eingesetzt, wobei ihr Einsatz in der externen Unternehmenskommunikation bislang jedoch noch vergleichsweise begrenzt bleibt. Chatbots übernehmen dort häufig die erste Kontaktaufnahme im Kundenservice, beantworten Anfragen auf Karriereportalen oder unterstützen klassische Serviceprozesse wie häufig gestellte Fragen (FAQs). Im internen Kontext helfen Chatbots dabei, Mitarbeitende bei HR-Prozessen oder IT-Supportanfragen zu begleiten. Deutlich erkennbar ist ein Trend zu spezialisierten und datenschutzkonformen Systemen, die in die bestehende IT-Infrastruktur integriert werden, um den steigenden Anforderungen an Datenschutz und Compliance gerecht zu werden.

Ein weiterer Bereich, in dem KI zunehmend Anwendung findet, ist die Personalisierung von Kommunikation. Durch die Analyse großer Datenmengen und die Zusammenführung von Stakeholder-Profilen ermöglichen KI-Systeme eine stärker individualisierte Ansprache. Unternehmen nutzen KI, um Zielgruppen passgenauer zu segmentieren und Inhalte dynamisch an Interessen und Kommunikationsverhalten anzupassen. Die Personalisierung erstreckt sich dabei nicht nur auf Marketingkommunikation, sondern zunehmend auch auf interne Kommunikationskampagnen oder die direkte Interaktion mit Medienvertreter*innen.

Tab. 1: Anwendungsfelder und Funktionen kommunikativer KI in der Unternehmenskommunikation. Quelle: eigene Darstellung angelehnt an Erkenntnisse aus dem European Communication Monitor 2024/25 (Zerfass et al., 2024) und dem CommTech Index Report 2024/25 (AG CommTech, 2024)

Anwendungsfeld	Funktion der KI
Chatbots & virtuelle Assistenten	Unterstützung von Kund*innenservice, Informationsretrieval, Stakeholderkommunikation
Sprachübersetzung	Ermöglichung interkultureller Kommunikation durch automatische Übersetzung
Spracherkennung	Transkription gesprochener Sprache für interne und externe Kommunikationszwecke
Inhaltserstellung	Generierung von Texten wie Berichten, Social-Media-Posts oder Pressemitteilungen
Social-Media-Analyse	Erfassung von Stimmungen, Trends und öffentlicher Meinung
Personalisierung	Anpassung von Kommunikationsinhalten an Nutzer*innenpräferenzen und -verhalten

Insgesamt zeigt sich, dass KI in der Unternehmenskommunikation derzeit überwiegend als unterstützende Technologie fungiert und primär begleitende Aufgaben übernimmt. Durch die Ergänzung menschlicher Tätigkeiten im Sinne einer Augmentation beschränkt sich ihr Einsatz bislang auf klar definierte, einfache Kommunikationsaufgaben. Allerdings sind erste Tendenzen erkennbar, dass KI zunehmend aktiv an der Gestaltung von Kommunikationsprozessen beteiligt wird und teilautonom agiert. Diese Entwicklungen deuten darauf hin, dass Unternehmen beginnen, über die rei-

ne Unterstützungsfunktion hinauszugehen und KI perspektivisch auch in komplexen Kommunikationssituationen gezielt einzusetzen.

3.2 Was versprechen sich Unternehmen vom Einsatz von KI?

Die Integration von KI-Systemen in die Unternehmenskommunikation ist nicht nur durch technologische Innovationen getrieben, sondern auch durch eine Vielzahl von Erwartungen und Versprechen, die in Praxisdiskursen, Studien und strategischen Visionen artikuliert werden. Unternehmen knüpfen an die Nutzung von KI weitreichende Hoffnungen, die über reine Effizienzsteigerung hinausgehen. Der Blick auf aktuelle empirische Erhebungen und Diskurse zeigt, dass diese Erwartungen in spezifischen Narrativen münden, die selbst Ausdruck sozialer Konstruktionsprozesse sind.

Eine zentrale Erwartung der Unternehmen ist die Hoffnung auf eine höhere Reaktionsgeschwindigkeit und eine kontinuierliche Kommunikationsfähigkeit. KI-gestützte Systeme ermöglichen es Unternehmen, schneller auf Medienanfragen, Kund*innenfeedback oder Krisensituationen zu reagieren. Insbesondere in Zeiten zunehmender Kommunikationsdichte auf sozialen Plattformen wird die Fähigkeit, nahezu in Echtzeit kommunizieren zu können, als strategischer Vorteil verstanden (Zerfass et al., 2024). Die Vorstellung einer „permanenten Kommunikationsbereitschaft“ geht dabei einher mit dem Wunsch nach optimierter Stakeholder-Bindung und einer stärkeren Präsenz im öffentlichen Diskurs.

Darüber hinaus wird KI zunehmend als Treiber für die Personalisierung von Kommunikation betrachtet. Unternehmen versprechen sich von KI-Systemen die Möglichkeit, Inhalte stärker auf die individuellen Bedürfnisse und Präferenzen von Stakeholdern zuzuschneiden. Diese Narrative knüpfen an tief verwurzelte Leitbilder aus Marketing und Public Relations an, wonach personalisierte Kommunikation als besonders effektiv und vertrauensfördernd gilt (AG CommTech, 2024). Neben dieser Hoffnung auf eine stärkere Zielgruppenorientierung spielen auch Innovationsversprechen eine zentrale Rolle, etwa im Bereich datengetriebenen Storytellings oder interaktiver Dialogsysteme. Die Erwartungen richten sich dabei auf eine Erweiterung des kommunikativen Repertoires und die Erschließung neuer, kreativer Möglichkeiten in der Ansprache von Zielgruppen.

Gleichzeitig zeigen kritische Analysen, dass viele dieser Erwartungen normativ überhöht sind und häufig von einem Technikoptimismus getragen werden, der die tatsächlichen Grenzen aktueller KI-Systeme ausblendet. Hepp et al. (2023) weisen darauf hin, dass die Einführung von KI-Systemen oft weniger strategisch erfolgt, sondern eher von opportunistischen Erwartungen an kurzfristige Erfolge getragen wird (Fekade et al., 2024). In vielen Fällen fehlen klare Strategien zur nachhaltigen Integration von KI in Kommunikationsprozesse, was zu einer Überlagerung von Hype und Realität führt. Vor diesem Hintergrund muss konstatiert werden, dass der Einsatz von KI in der Unternehmenskommunikation nicht nur technische und organisatorische Fragen aufwirft, sondern auch eine neue Aushandlung dessen erfordert, was als „gute“ Kommunikation gilt.

3.3 Herausforderungen und Risiken

Die Integration kommunikativer KI in die Unternehmenskommunikation eröffnet nicht nur neue Potenziale, sondern bringt auch Herausforderungen mit sich. Diese lassen sich jedoch nicht einseitig als ethische Probleme klassifizieren, wie es in aktuellen Diskussionen häufig geschieht. Vielmehr zeigen sich Spannungsfelder, die epistemische, organisationale, soziale und ethische Dimensionen gleichermaßen betreffen. Um der Komplexität kommunikativer KI gerecht zu werden, ist eine differenzierte Betrachtung notwendig. In Anlehnung an Buhmann und White (2022) werden ethische Herausforderungen im Umgang mit KI in drei zentrale Bereiche unterteilt: Evidence Concerns, Outcome Concerns und Epistemic Concerns.

Ein zentrales epistemisches Spannungsfeld ergibt sich aus der Intransparenz von KI-Systemen vor allem jenen auf Basis von Large-Language-Modellen (LLM). Viele generative KI-Anwendungen agieren als sogenannte Black-Box-Systeme, deren Entscheidungslogiken selbst für Entwickler*innen nur schwer nachvollziehbar sind (Buhmann & White, 2022; Hepp et al., 2023). Für die Unternehmenskommunikation bedeutet dies eine erhebliche Herausforderung. Wenn Inhalte, Analysen oder Empfehlungen von Systemen produziert werden, deren Funktionsweise nicht transparent ist, stellt sich die Frage, wie Verantwortung und Steuerbarkeit sichergestellt werden können. Dieses Problem betrifft nicht nur juristische Aspekte wie Haftungsfragen, sondern auch die soziale Zuschreibung von Kompetenz sowie die Fehlerkultur innerhalb von Kommunikationsabteilungen. Die epistemische Unschärfe der KI-Entscheidungsprozesse führt so zu einer Destabilisierung tradierter Rollenbilder in der professionellen Kommunikation (Hepp et al., 2023; Yue et al., 2024).

Eine intensiv diskutierte ethische Herausforderung ist die Gefahr algorithmischer Verzerrungen und Diskriminierungen. KI-Systeme können bestehende gesellschaftliche Biases reproduzieren oder neue Formen von Diskriminierung erzeugen, insbesondere wenn Trainingsdaten soziale Ungleichheiten unkritisch fortschreiben (Buhmann & White, 2022; Yue et al., 2024). In der Unternehmenskommunikation kann dies nicht nur Imageschäden verursachen, sondern auch normative Erwartungen an Fairness und Inklusion untergraben.

Auch Fragen des Datenschutzes und der Datensicherheit besitzen hohe ethische Relevanz. Kommunikationsabteilungen, die KI-gestützte Analyse- und Targeting-Tools einsetzen, operieren häufig mit sensiblen Stakeholder-Daten (Buhmann & White, 2022; Yue et al., 2024). Insbesondere in einer zunehmend datensensiblen Öffentlichkeit kann ein Verstoß gegen Datenschutzprinzipien gravierende Reputationsrisiken nach sich ziehen.

Diese strukturellen und ethischen Herausforderungen werden von sozialen Risiken begleitet. Die Integration kommunikativer KI in die Unternehmenskommunikation verändert nicht nur Prozesse und Technologien, sondern hat auch tiefgreifende Auswirkungen auf die Berufsrollen, Kompetenzprofile und Selbstverständnisse von Kommunikationsprofessionals. Traditionelle Tätigkeiten werden zunehmend von KI-gestützten Systemen übernommen oder unterstützt (Zerfass et al., 2024). Dabei verschieben sich Kompetenzanforderungen (Buhmann & White, 2022). Neben grundlegenden digitalen Skills wird insbesondere *AI Literacy* zur Notwendigkeit für Unternehmen (Art. 4, EU AI Act). Diese umfasst sowohl die sichere Bedienung von KI-Systemen als auch die kritische Reflexion ihrer sozialen und ethischen Implika-

tionen. Eng verknüpft damit ist die Notwendigkeit einer ausgeprägten *Data Literacy*. Kommunikationsprofessionals müssen in der Lage sein, datengetriebene Entscheidungen kompetent zu interpretieren, Zusammenhänge kritisch zu reflektieren und die Grenzen algorithmischer Analysen erkennen zu können (Buhmann & White, 2022; Zerfass et al., 2024). *Data Literacy* umfasst somit analytische Fähigkeiten sowie reflexives Bewusstsein.

Ein weiteres soziales Spannungsfeld, das sich aus der Integration von KI ergibt, ist die Anthropomorphisierung von KI-Systemen. Forschungen im Rahmen des CA-SA-Paradigmas (Computers Are Social Actors) zeigen, dass Menschen dazu neigen, KI-Systemen, die menschliche Merkmale wie Profilbilder oder Emojis verwenden, soziale Eigenschaften zuzuschreiben (Guzman & Lewis, 2020; Hepp & Görland, 2024; Muhle, 2024). In der Unternehmenskommunikation manifestiert sich diese Tendenz in der Vorstellung, KI könne als eine Art „Kollege“ bzw. „Kollegin“ fungieren. Diese Zuschreibung schafft einerseits neue Formen von Akzeptanz und Nähe, birgt jedoch andererseits die Gefahr einer Überschätzung technischer Systeme. Dadurch entstehen unrealistische Erwartungshaltungen, die in der Praxis häufig zu Enttäuschungen führen können.

Insgesamt wird deutlich, dass Kommunikationsprofessionals zunehmend vor der Herausforderung stehen, ihre eigene Rolle im Verhältnis zu KI-Systemen neu zu definieren. Dieser Wandel ist keineswegs allein technischer Natur, sondern Ausdruck tiefgreifender sozialer Aushandlungsprozesse (Hepp & Loosen, 2023). In vielen Organisationen zeigt sich dabei eine Diskrepanz zwischen der raschen technologischen Implementierung von KI und der bislang unzureichenden Anpassung professioneller Standards und Ausbildungskonzepte. Um die damit verbundenen Chancen und Risiken angemessen adressieren zu können, ist ein Bewusstsein für die komplexe Verschränkung epistemischer Unsicherheiten, sozialer Dynamiken und ethischer Normen erforderlich. Nur eine differenzierte Betrachtung dieser Spannungsfelder ermöglicht es, die Potenziale kommunikativer KI verantwortungsvoll zu nutzen, ohne die damit einhergehenden Herausforderungen aus dem Blick zu verlieren.

4. Fazit und Ausblick

Die Integration kommunikativer KI in die Unternehmenskommunikation markiert einen Wandel, der weit über technische Innovationen hinausreicht. KI-Systeme werden zunehmend nicht nur als unterstützende Werkzeuge, sondern auch als potenzielle Teilnehmer an kommunikativen Prozessen konstruiert. Diese Entwicklung deutet auf eine Veränderung sowohl der Formen organisationaler Kommunikation als auch der zugrundeliegenden Deutungsmuster hin, auf denen berufliche Rollen und Verantwortlichkeiten basieren.

Die Analyse aktueller Anwendungen zeigt, dass KI heute bereits vielfältig in der Unternehmenskommunikation eingesetzt wird – von der automatisierten Content-Erstellung über dialogorientierte Systeme bis hin zu datenbasiertem Monitoring und personalisierter Stakeholder-Ansprache. Gleichzeitig offenbaren sich hohe Erwartungen an Effizienzsteigerung, Innovationsfähigkeit und neue Formen der Stakeholder-Bindung. Eine kritische Reflexion dieser Entwicklungen macht jedoch deutlich, dass diese Versprechen nicht ungebrochen eingelöst werden können. Vielmehr entstehen

neue epistemische Unsicherheiten, soziale Spannungsfelder und ethische Herausforderungen, die eine differenzierte Betrachtung erforderlich machen. Kommunikationsprofessionals stehen daher vor der Aufgabe, nicht nur technologische Kompetenzen zu entwickeln, sondern auch AI Literacy und Data Literacy als integrale Bestandteile ihres professionellen Selbstverständnisses zu etablieren.

Vor diesem Hintergrund erscheint es sinnvoll, kommunikative KI nicht ausschließlich als technisches Tool zu betrachten, sondern auch als potenziellen aktiven Mitgestalter sozialer Interaktionsprozesse zu reflektieren. KI-Systeme tragen zunehmend zur Bedeutungsproduktion bei, strukturieren Handlungsoptionen und beeinflussen Interaktionsverhältnisse. Erste Tendenzen deuten darauf hin, dass neue Formen hybrider Agency entstehen, in denen menschliche und maschinelle Akteure gemeinsam kommunizieren und agieren (Hepp & Loosen, 2023; Pfadenhauer & Grenz, 2017). Für die zukünftige Forschung wird es daher entscheidend sein, diese hybriden Dynamiken genauer zu untersuchen. Dabei geht es nicht zuletzt um die Frage, wie soziale, organisationale und ethische Standards in einer zunehmend KI-gestützten Kommunikationslandschaft gestaltet und weiterentwickelt werden können.

Literatur

- AG CommTech. (2024). *CommTech Index Report 2024/2025 – AG CommTech*. <https://ag-commtech.de/commtech-index-report-2024-2025/#1724748697409-53720318-fc49>
- Andrejevic, M. (2020). *Automated media*. Routledge. <https://doi.org/10.4324/9780429242595>
- Baecker, D. (1997). Reintroducing communication into cybernetics. *Systemica*, 11, 11–29. <https://ssrn.com/abstract=2200830>
- Banholzer, V., Quest, A. & Rossbach, A. (2023). Künstliche Intelligenz in der Unternehmenskommunikation: Einsatzbereiche, Chancen, Herausforderungen und ethische Leitlinien. In T. Mickeleit & J. Forthmann (Hrsg.), *Erfolgssfaktor CommTech* (S. 231–275). Springer Gabler. https://doi.org/10.1007/978-3-658-40169-6_9
- Berger, P. L. & Luckmann, T. (1970). *Die gesellschaftliche Konstruktion der Wirklichkeit. Eine Theorie der Wissenssoziologie*. Fischer Taschenbuch Verlag.
- Blumer, H. (1954). What is Wrong with Social Theory? *American Sociological Review*, 19(1), 3–10.
- Buhmann, A. & Gregory, A. (2023). Digital corporate communication and artificial intelligence and future roles. In V. Luoma-aho & M. Badham (Hrsg.), *Handbook on Digital Corporate Communication* (S. 281–296). Edward Elgar Publishing. <https://doi.org/10.4337/9781802201963.00031>
- Buhmann, A., & White, C. L. (2022). Artificial Intelligence in Public Relations: Role and Implications. In J. H. Lipschultz, K. Freberg & R. Luttrell (Hrsg.), *The Emerald Handbook of Computer-Mediated Communication and Social Media* (S. 625–638). Emerald Publishing Limited. <https://doi.org/10.1108/978-1-80071-597-420221036>
- Burns, R., Hawkins, B., Hoffmann, A. L., Iliadis, A., & Thatcher, J. (2019). Transdisciplinary approaches to critical data studies. *Proceedings of the Association for Information Science and Technology*, 55(1), 657–660. <https://doi.org/10.1002/pra2.2018.14505501074>
- Burrell, J. (2016). How the machine ‘thinks’: Understanding opacity in machine learning algorithms. *Big Data & Society*, 3(1). <https://doi.org/10.1177/2053951715622512>.

- Carlson, M. (2016). Automated journalism: A posthuman future for digital news? In S. A. Eldridge II & B. Franklin (Hrsg.), *The Routledge companion to digital journalism studies* (S. 226–234). Routledge.
- Castells, M. (2000). Toward a Sociology of the Network Society. *Contemporary Sociology*, 29(5), 693–699. <https://doi.org/10.2307/2655234>
- Chesebro, J. W. & Bonsall, D. G. (1989). *Computer-mediated communication: Human relationships in a computerized world*. University of Alabama Press.
- Crawford, K. (2021). *Atlas of AI*. Yale University Press. <https://doi.org/10.12987/9780300252392>
- Dalton, C. & Thatcher, J. (2014). What Does A Critical Data Studies Look Like, And Why Do We Care? *Society + Space*. <https://www.societyandspace.org/articles/what-does-a-critical-data-studies-look-like-and-why-do-we-care>
- Dencik, L. & Kaun, A. (2020). Datafication and the Welfare State. *Global Perspectives*, 1(1). <https://doi.org/10.1525/gp.2020.12912>
- Esposito, E. (2017). Artificial Communication? The Production of Contingency by Algorithms. *Zeitschrift für Soziologie*, 46(4), 249–265. <https://doi.org/10.1515/zfsoz-2017-1014>
- Fekade, A., Keppeler, N., Mattheus, E., Merz, I. & Wegner, L. (2024). Intelligent communication? Use of generative AI applications in communication agencies. In A. Godulla, C. Buller, V. Freudl, I. Merz, J. Twittenhoff, J. Winkler & L. Zapke (Hrsg.), *The Dynamics of Digital Influence: Communication Trends in Business, Politics and Activism* (S. 31–51). Leipzig. <https://nbn-resolving.org/urn:nbn:de:0168-ss0ar-94685-2>
- Flensburg, S. & Lomborg, S. (2021). Datafication research: Mapping the field for a future agenda. *New Media & Society*, 25(6), 1451–1469. <https://doi.org/10.1177/14614448211046616>
- Gambino, A., Fox, J. & Ratan, R. (2020). Building a Stronger CASA: Extending the Computers Are Social Actors Paradigm. *Human-Machine Communication*, 1, 71–86. <https://doi.org/10.30658/hmc.1.5>
- Gehl, R. W. & Bakardjieva, M. (2016). *Socialbots and Their Friends: Digital Media and the Automation of Sociality*. Routledge. <https://doi.org/10.4324/9781315637228>
- Gregory, A. & Virmani, S. (2020). *The Effects of AI on the Professions: A Literature Repository*. Chartered Institute of Public Relations.
- Gorwa, R., Binns, R. & Katzenbach, C. (2020). Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society*, 7(1), 205395171989794. <https://doi.org/10.1177/2053951719897945>
- Guzman, A. L. & Lewis, S. C. (2020). Artificial intelligence and communication: A Human–Machine Communication research agenda. *New Media & Society*, 22(1), 70–86. <https://doi.org/10.1177/1461444819858691>
- Hepp, A. (2020). Artificial companions, social bots and work bots: communicative robots as research objects of media and communication studies. *Media Culture & Society*, 42(7–8), 1410–1426. <https://doi.org/10.1177/0163443720916412>
- Hepp, A. & Görland, S. O. (2024). Agency in a datafied society: an introduction. *Convergence*, 30(3), 945–955. <https://doi.org/10.1177/13548565241254692>
- Hepp, A. & Hasebrink, U. (2017). Kommunikative Figurationen. Ein konzeptioneller Rahmen zur Erforschung kommunikativer Konstruktionsprozesse in Zeiten tiefgreifender Mediatisierung. *Medien & Kommunikationswissenschaft*, 65(2), 330–347. <https://doi.org/10.5771/1615-634x-2017-2-330>
- Hepp, A. & Loosen, W. (2023). The Interdisciplinarity of HMC: Rethinking Communication, Media, and Agency. In A. L. Guzman, R. McEwen & S. Jones (Hrsg.), *The Sage Handbook of Human–Machine Communication* (S. 12–21). SAGE Publications. <https://doi.org/10.4135/9781529782783.n4>
- Hepp, A., Loosen, W., Dreyer, S., Jarke, J., Kannengießer, S., Katzenbach, C., Malaka, R., Pfadenhauer, M., Puschmann, C. & Schulz, W. (2023). ChatGPT, LaMDA, and the

- Hype Around Communicative AI: The Automation of Communication as a Field of Research in Media and Communication Studies. *Human-Machine Communication*, 6, 41–63. <https://doi.org/10.30658/hmc.6.4>
- Heuer, H., Jarke, J. & Breiter, A. (2021). Machine learning in tutorials – Universal applicability, underinformed application, and other misconceptions. *Big Data & Society*, 8(1). <https://doi.org/10.1177/20539517211017593>
- Keller, R., Knoblauch, H. & Reichertz, J. (2013). Der Kommunikative Konstruktivismus als Weiterführung des Sozialkonstruktivismus – eine Einführung in den Band. In R. Keller, J. Reichertz & H. Knoblauch (Hrsg.), *Kommunikativer Konstruktivismus. Wissen, Kommunikation und Gesellschaft* (S. 9–21). Springer VS. https://doi.org/10.1007/978-3-531-19797-5_1
- Knoblauch, H. & Pfadenhauer, M. (2023). Construction versus Realism? The Unrealized Potential of Communicative Constructivism. *Sociologica*, 17(1), 119–136. <https://doi.org/10.6092/issn.1971-8853/16725>
- Muhle, F. (2024). Robots as addressable non-persons: an analysis of categorial work at the boundaries of the social world. *Frontiers in Sociology*, 9. <https://doi.org/10.3389/fsoc.2024.1260823>
- Natale, S. (2021). *Deceitful media: Artificial intelligence and social life after the Turing test*. Oxford University Press.
- Pfadenhauer, M. & Grenz, T. (2017). Von Objekten zu Objektivierung. *Soziale Welt*, 68(2–3), 225–242. <https://doi.org/10.5771/0038-6073-2017-2-3-225>
- Pfadenhauer, M. & Lehmann, T. (2022). Affects after AI: Sociological perspectives on artificial companionhip. In A. Elliott (Hrsg.), *The Routledge Social Science Handbook of AI* (S. 91–106). <https://doi.org/10.4324/9780429198533-7>
- Soriano, A. S. & Valdés, R. M. T. (2021). Engaging universe 4.0: The case for forming a public relations-strategic intelligence hybrid. *Public Relations Review*, 47(2), 102035. <https://doi.org/10.1016/j.pubrev.2021.102035>
- Stieglitz, S. & Wiencierz, C. (2022). Digitalisierung, Big Data und soziale Medien als Rahmenbedingungen der Unternehmenskommunikation. In A. Zerfaß, M. Piwinger & U. Röttger (Hrsg.), *Handbuch Unternehmenskommunikation* (3. Aufl., S. 289–309). Springer Gabler. https://doi.org/10.1007/978-3-658-22933-7_10
- Turner, F. (2006). *From Counterculture to Cyberculture*. The University of Chicago Press.
- Veale, T. & Cook, M. (2018). *Twitterbots: Making Machines that Make Meaning*. The MIT Press. <https://doi.org/10.7551/mitpress/10859.001.0001>
- Weller, T. & Lock, I. (2024). Automated Communication's Impact on Strategic Communication: Implications from a Systematic Review. *International Journal of Strategic Communication*, 19(1), 13–34. <https://doi.org/10.1080/1553118x.2024.2379501>
- Yue, C. A., Men, L. R., Davis, D. Z., Mitson, R., Zhou, A. & Rawi, A. A. (2024). Public Relations Meets Artificial Intelligence: Assessing Utilization and Outcomes. *Journal of Public Relations Research*, 36(6), 513–534. <https://doi.org/10.1080/1062726x.2024.2400622>
- Zerfass, A., Buhmann, A., Laborde, A., Moreno, A., Romenti, S. & Tench, R. (2024). *European Communication Monitor 2024. Managing tensions in corporate communications in the context of geopolitical crises, artificial intelligence, and managerial learning*. European Public Relations Education and Research Association. <https://www.communicationmonitor.eu/2024/11/23/ecm-european-communication-monitor-2024/>

Look who's talking

Sprecheridentität und Diskurskontrolle in der Arbeit mit generativer KI

Erich Prem

Einsatz und Risiken von generativer KI in Unternehmen

Der rasante Erfolg generativer KI-Sprachsysteme – allen voran ChatGPT und vergleichbare Modelle – hat in kürzester Zeit zu einer weitreichenden Integration dieser Technologie in unternehmerische Prozesse geführt. In Büros auf der ganzen Welt bedienen sich Mitarbeitende solcher Systeme, um alltägliche Aufgaben effizienter zu erledigen: Texte werden zusammengefasst, stilistisch überarbeitet, orthografisch korrigiert, übersetzt oder in andere Kontexte transformiert. Die Kommunikation wird teilautomatisiert erledigt, z.B. durch die automatische Beantwortung von E-Mails. Darüber hinaus werden Präsentationen, Marktanalysen, Projektberichte oder Marketingmaterialien unter Zuhilfenahme von KI erstellt – oft in mehreren Sprachen zugleich (Dam et al., 2024). Dieselben Systeme sind aber längst nicht mehr nur Expertinnen vorbehalten. Auch Laien verwenden sie etwa zur Generierung von Webseiten-Code, zur Erstellung einfacher Softwarelösungen oder zur Automatisierung repetitiver Abläufe. In der Bildungswelt nutzen Schülerinnen und Studierende Chatbots sowohl zur Unterstützung bei Hausaufgaben als auch für ein personalisiertes Sprachtraining. Die Technologie hat sich tief in die Alltagslogik verschiedener gesellschaftlicher Bereiche eingeschrieben.

Allerdings erfolgte die Einführung von Sprachmodellen in die Arbeitsrealität überwiegend ungeplant und unreguliert. In vielen Fällen wurden sie ohne vorherige Schulungen oder strategische Einbettung in bestehende Prozesse implementiert. Dies stellt eine zunehmende Herausforderung für Unternehmen dar, insbesondere in sensiblen Bereichen wie Compliance, Datenschutz oder Sicherheit. Der Einsatz von KI in Unternehmen ist praktisch immer mit Risiken verbunden. Zu diesen Risiken gehören neben den Marktrisiken insbesondere die folgenden (Prem, 2025):

- Innovationsrisiken: Risiken, die aus der mangelnden Akzeptanz von neuen Systemen entstehen oder verpasste Gelegenheiten für neue Produkte
- Technologierisiken: Mangel an robuster und korrekter Funktion, Bedrohungen durch neue und bessere Technologien, Qualitätsprobleme, Restrisiken durch das Fehlen von garantierter oder vollständig überprüfter Funktionalität
- Sicherheitsrisiken: mögliche Gefahren durch Cyberattacken, Datenlecks, unkontrolliertes Systemverhalten
- Öffentlichkeitsrisiken: Verlust an Ansehen und Glaubwürdigkeit in der Öffentlichkeit, Vertrauensverlust in Produkte und Dienste, Folgewirkungen von Fehlern oder unfairer Behandlung durch KI-Systeme
- Rechtliche Risiken: Unsicherheit in der Erfüllung bestehender rechtlicher Regeln bzw. durch neue Regulierung
- Personalrisiken: Qualifikationsrückgang, Mangel an geeignet ausgebildetem Personal

Diese Risiken gelten alle auch für den Einsatz von generativer KI. So kann etwa durch unbedachte Eingaben von Mitarbeitenden vertrauliche Information an Dritte gelangen. KI-generierter Code kann zu Sicherheitslücken führen, wenn er nicht überprüft wird, und unreflektierte Textausgaben können PR-Krisen auslösen. Einige spezielle Risiken generativer Sprachmodelle im Unternehmen sind u. a. (Rane, 2023):

- Innovationsrisiko: Die Abwägung, in welchem Ausmaß und für welche Zwecke generative KI im Unternehmen eingesetzt werden soll, ist schwierig, weil es große Unterschiede in der Akzeptanz der Technologie gibt. Sie kann nach Region, Alter und mit der Erfahrung der Nutzenden schwanken.
- Technologierisiko:
 - Dies betrifft vor allem die Richtigkeit und Genauigkeit generierter Texte: generative Modelle können völlig erfundene Inhalte plausibel erscheinen lassen, diese mit erfundenen Quellenangaben versehen. Dazu kommt, dass derzeit viele Sprachmodelle kontinuierlich verändert werden und keine Garantien für die korrekte Funktionsweise möglich sind.
 - Mangelnde Wiederholbarkeit und Verlässlichkeit: KI-Systeme zur Textgenerierung führen oft zu unterschiedlichen Resultaten, selbst wenn gleiche Inputdaten verwendet werden. Generative KI-Systeme unterliegen häufiger Wartung und Erweiterung – ohne, dass dies für die Nutzenden leicht ersichtlich sein muss.
- Sicherheitsrisiko:
 - Hierzu gehört beispielsweise die Preisgabe von Unternehmensinformationen, weil Mitarbeitende in Unternehmen durch die Übermittlung an generative KI-Schnittstellen Geschäftsgeheimnisse verraten können.
 - Überaus problematisch ist die Verwendung generativer KI-Systeme zur Programmierung, weil durch versteckt eingeschleusten bösartigen Code eine Gefährdung der Systemintegrität erfolgen kann.
- Öffentlichkeitsrisiko:
 - Mangelnde Nachvollziehbarkeit und Erklärbarkeit: KI-generierte Texte können schwer nachvollziehbar sein und es kann schwerfallen, sie zu begründen oder zu erklären.
 - Generative Sprachmodelle können vorurteilsbehaftete Aussagen machen und tendenziöse Äußerungen produzieren. Sie können in manchen Fällen auch Beschimpfungen generieren oder sich in anders unpassender Weise äußern, z. B. wenn sie ihren User:innen ihre Liebe erklären oder ihnen den Selbstmord empfehlen (Bellware & Masih, 2024).
- Rechtliches Risiko:
 - Verletzung geschützter personenbezogener Daten: Eingaben für KI-Systeme können persönliche Daten enthalten, die in unzulässiger Form übermittelt werden (z. B. durch unzulässigen internationalen Datentransfer). Daten können durch Verknüpfungen mit weiteren Daten deanonymisiert werden.
 - Verletzung geistigen Eigentums und Plagiate: Generierte Texte können die Rechte von Urheberinnen verletzen bzw. zu ungewollten Plagiaten führen.
- Personalrisiko:
 - Es kann leicht zu einer fälschlichen Vermenschlichung kommen. KI-Systeme sind häufig so gestaltet, dass der Eindruck entsteht, man würde mit einem

Menschen interagieren. Dies ist nicht nur ein Problem bei der uninformierten Verwendung, sondern kann auch bei Geübten zur Überbewertung des Outputs von generativer KI führen.

- Langfristig zeigt sich auch ein struktureller Effekt auf dem Arbeitsmarkt: Die gestiegene Zahl arbeitsloser IT-Fachkräfte in den USA könnte ein Anzeichen dafür sein, dass gewisse Tätigkeiten zunehmend von KI übernommen werden (Bousquette, 2025a).

Diese enorme Breite an Herausforderungen zeigt, mit welcher Vielzahl von Risiken Unternehmen konfrontiert sind, die große Sprachmodelle nutzen wollen. Unternehmen reagierten darauf zunächst nur zögerlich und in einem ersten Schritt häufig mit dem Verbot, generative KI einzusetzen. Dies ist aber angesichts der praktischen Vorteile sowie des freien Zugangs zu generativen KI-Systemen für die Belegschaft kaum durchzusetzen. Daher haben inzwischen viele Unternehmen interne Leitlinien und Anwendungsrichtlinien formuliert.

Ein typisches unternehmerisches Gegenmittel zu den erwähnten Risiken und Problemen ist die Verlagerung der Verantwortung für generierte Inhalte auf die Mitarbeitenden selbst. Die gängige Praxis besteht darin, die Nutzung von KI-Systemen unter bestimmten Auflagen zu erlauben, gleichzeitig aber die volle Verantwortung für die Richtigkeit und Angemessenheit der erzeugten Inhalte bei den User:innen zu belassen. Dieses Modell der „kontrollierten Autonomie“ entlastet zunächst vor allem das Management, belastet jedoch die Mitarbeitenden mit neuen, oft unausgesprochenen Pflichten. Es ist fraglich, ob durch derartige Regelungen eine praktische rechtliche Entlastung der Unternehmen erfolgt. Es ist denkbar, dass in vielen Jurisdiktionen die Unternehmen letztendlich in vollem Umfang verantwortlich bleiben und die Verantwortung nicht einfach an die Mitarbeitenden auslagern können. Dies dürfte vor allem dann schwierig sein, wenn die Nutzung generativer KI regelmäßig durch die Unternehmen geduldet wird. Es stellt sich z.B. die Frage, ob und in welchem Ausmaß sichergestellt ist, dass die Belegschaft die Risiken generativer KI versteht und erkennen kann. Dies wäre theoretisch durch Schulungen sicherzustellen, für die heute aber kaum ein ausreichendes und vor allem praxisnahes Angebot besteht. Derartige Schulungen sind für Unternehmen aufgrund des sich dynamisch entwickelnden Umfelds schwer zu planen. Als Evidenz dafür mag gelten, dass etwa im Jahr 2023 noch davon ausgegangen wurde, dass die Formulierung von KI-Anfragen als „Prompt Engineer“ ein spezielles Berufsbild wird. Diese Annahme scheint aufgrund der verbesserten und intuitiveren Reaktion von generativen KI-Systemen nunmehr obsolet zu sein (Bousquette, 2025b). Es bleibt daher schwierig, die genauen Fertigkeiten im Umgang mit generativer KI durch Training der Belegschaft gezielt zu verbessern.

Die Rolle der Autorin

Durch die Anweisung an die Beschäftigten, generative KI für die Texterzeugung zwar verwenden zu dürfen, aber diese jedenfalls immer selbst zu verantworten, entsteht eine paradoxe Situation: Obwohl sie die Verantwortung für die Inhalte tragen, sind sie nicht die eigentlichen Urheber der Texte – zumindest nicht im traditionellen Sinne. Mitarbeiterinnen werden solcherart zu Herausgeberinnen bzw. Kuratorinnen.

Damit öffnet sich eine neue philosophische Perspektive auf die Frage nach der Bedeutung und genauen Art von Autorenschaft in Zusammenhang mit generativer KI.

Zunächst ist im Verhältnis zur künstlichen Intelligenz darauf hinzuweisen, dass Autorenschaft im Kontext humanistischer Philosophie überhaupt als Grundlage der Zuschreibung von Verantwortlichkeit verstanden wurde. In diesem Sinne argumentieren Humanisten für eine Verbindung von Autorenschaft, Freiheit und Verstand (Nida-Rümelin & Staudacher, 2024). So verstanden geht Autorenschaft über die Verantwortlichkeit für das Abfassen von Werken hinaus. Im Humanismus – insbesondere im Digitalen Humanismus (Werthner et al., 2023) – wird der Mensch als Autor seines Lebens betrachtet, das er in Freiheit gelenkt von der entwerfenden Kraft des Geistes verantwortet. Selbst wenn man der Trias von Verstand, Freiheit und Verantwortung als humanistischer Essenz einer *conditio humana* nicht folgen mag, formulieren diese drei Begriffe dennoch Wesenseigenschaften von Autorenschaft. Texte galten lange nicht nur als sprichwörtliches Produkt des Geistes, sondern repräsentierten als Ausdruck menschlicher Sprache ein Wesensmerkmal, das über lange Zeit als charakteristisch für das Menschsein angesehen wurde. Darüber hinaus sind Texte auch Ausdruck der Freiheit des Menschen, insofern sie von Autoren genuin – frei und kreativ – hervorgebracht werden und daher auch in der Verantwortung der Autorin liegen. Falls jedoch Mängel an Freiheit, Verstand, oder Verantwortung vorliegen, sind auch Zweifel an der Autorenschaft angebracht. Im Folgenden wollen wir diese Aspekte genauer betrachten und beginnen damit, was über Autorenschaft in generativer KI und zur Ausgestaltung von Diskursen gesagt werden kann.

Autorenschaft und Diskursordnung

Michel Foucault beschränkt in seiner wichtigen Arbeit über „Die Ordnung des Diskurses“ (1991) den Begriff des „Autors“ nicht primär auf eine individuelle Person, sondern zeigt sich vor allem an seiner Funktion im Diskursgefüge interessiert: Der Autor gilt ihm als Prinzip der Gruppierung, der Bedeutungstiftung und der kohärenten Einheit von Aussagen. Wie wir weiter unten sehen werden, löst sich diese Form der Autorenschaft im Gebrauch von Chatbots auf. Die Bedeutung des Subjekts als Ursprung von Diskurs verliert an Kraft. Zwar wird durch rechtliche Konstruktionen versucht, Autorenschaft durch Verantwortlichkeitszuschreibung zu simulieren, doch diese juristische Fiktion verdeckt lediglich das Fehlen eines „echten“ Urhebers. Die Mitarbeitenden werden so zu Trägern einer leeren Autorenschaft – verantwortlich, aber nicht schöpferisch.

Foucault widmet sich der Ordnung des Diskurses auch durch die Analyse von Ausschlussmechanismen, d. h. jenen Prozessen, mit deren Hilfe der Diskurs kontrolliert wird. Dazu zählt er Ausschlussysteme wie etwa explizite Verbote; interne Prozeduren, z. B. Kommentare und eben den Autor selbst; und schließlich die Verknappung der sprechenden Subjekte, z. B. durch Regeln darüber, wer am Diskurs teilnehmen kann.

Diese Ausschlussmechanismen spielen auch in der generativen KI eine Rolle, etwa wenn die KI explizit nur für gewisse Zwecke verwendet werden darf, wenn sie intern so gestaltet wird, dass sie gewisse Diskurse nicht führen möchte, oder wenn

eben nur wenige Sprechende (in unserem Fall generative KI-Systeme) vorliegen. Diese Aspekte sollen im Folgenden genauer erläutert werden.

Wer spricht: die KI als Autorin

Generative KI-Systeme, die mittels großer Sprachmodelle funktionieren, basieren auf der Analyse von Texten Unzähliger. Sie erstellen probabilistische Modelle von Sprachen, indem sie Wortfolgewahrscheinlichkeiten statistisch schätzen (Wang et al., 2024). Eine Reihe von Techniken wird verwendet, um diese Modelle an die Erwartungen und Bedürfnisse der User:innen anzupassen, dazu gehören u. a. überwachtes und verstärkendes Lernen aus Dialogbeispielen. Im Endeffekt repräsentieren generative KI-Systeme eine radikale Vervielfältigung sprachlicher Subjektivität: In der Interaktion mit einem Chatbot tritt der Mensch einem kollektiven Archiv gegenüber, das aus den Stimmen Tausender gespeist ist. Die generative KI schätzt die Wahrscheinlichkeit einer passenden Antwort auf der Basis jener Wortfolgewahrscheinlichkeiten, die in den Trainingstextdaten enthalten sind.

Auf Basis dieser Funktionsweise ist leicht zu sehen, dass generative KI-Systeme zunächst jedenfalls keinen prioritären Zugang zu Wahrheit oder Faktizität haben. Sie funktionieren auf der Basis von Texten anhand derer sie trainiert werden. Soweit diese Texte Aspekte der Wirklichkeit wiedergeben, sind sie wohl geeignet, Auskunft über die Welt zu geben. Viele Texte können aber durchaus nicht mit der Welt korrespondieren, etwa wenn es sich um Belletristik oder schlicht um Fantasien oder erfundene, aber realistisch erscheinende Werke handelt. Zusätzlich können die durch Texte beschriebenen Ausschnitte der Welt inkongruent miteinander sein oder sich auch direkt widersprechen. Das gilt selbst dort, wo man Faktizität vermuten würde: Newton'sche und Einstein'sche Physik führen nicht zu denselben Ergebnissen.

Die Frage der Faktizität von Ausgangstexten wird dadurch kompliziert, dass ein guter Teil belletristischer Werke relevante Ausschnitte der Wirklichkeit beschreibt, die jedenfalls wahr sein könnten. So beschreibt ein Großteil des Werkes von Nobelpreisträger Gabriel García Márquez die Welt, wie sie wohl sein könnte: Flaschen stehen auf Tischen und Boote, die Bischöfe befördern, fahren Flüsse hinauf. Sie enthalten aber auch Beschreibungen und Phänomene, die wir üblicherweise außerhalb der Realität ansiedeln würden, z. B. wenn Márquez Visionen, Weissagungen oder gar magische Vorgänge beschreibt, die wir nicht als real annehmen können.

Dass Sprachmodelle sich die Texte einer großen Zahl von Personen zu Nutze machen, bedeutet normalerweise, dass einzelne Textteile dieser Personen nur mehr schwer zu identifizieren sind. Trotz der enormen Zahl an Parametern großer Sprachmodelle arbeiten generative KI-Systeme auf Basis von Generalisierungen über die Texte, d. h. einzelne Textteile sind üblicherweise nicht einfach oder zumindest nicht mehr garantiert in ihrer ursprünglichen Form zugänglich. Es kommt damit zu einer Art von Durchschnittsbildung, die für die generative KI charakteristisch ist. Sie hat auch zur Kritik an den Modellen geführt, dass sie bloß „durchschnittliches“ generieren könnten. In dieser starken Form ist die Kritik nicht einfach haltbar, aber es ist beim Stand der Technik derzeit schwer vorherzusagen, welche Resultate generativer KI-Systeme bloß durchschnittlich und welche anderen als überraschend oder vielleicht als kreativ einzustufen sind (Srivastava et al., 2022).

Vergleichen wir diese Situation mit konventioneller Textgenerierung durch Autorinnen, dann wird klar, dass häufig mehrere Personen die Autorenschaft von Texten teilen können. Das Sprechen von vielen ist nicht der Ursprung der Problematik KI-generierter Texte. Typischerweise entstehen allerdings Texte, die von vielen generiert werden, einfacher als durch Schätzung von Parametern für Wortfolgewahrscheinlichkeitsmodelle. Typischerweise sind ganze Textabschnitte von einzelnen Personen geschrieben oder es werden bloß einzelne Phrasen und manchmal Wörter ergänzt oder ausgetauscht. Dennoch ist es sicher richtig, dass auch konventionell von mehreren Autoren produzierte Texte bereits an einem Mangel an klarer, eindeutiger Autorenschaft leiden können, wenn sie nicht kollektive Autorenschaft beanspruchen können. Gerade diese kollektive Autorenschaft wird aber üblicherweise erwartet, wenn mehrere Autorinnen an einem Text arbeiten. Sie stehen dann für den Text gerade und es wird erwartet, dass sie gemeinsam sowie einzeln Verantwortung für den Text übernehmen.

Eine derartige Sicht – insbesondere der Übernahme von Verantwortung – auf die von generativer KI produzierten Texte ist unmöglich. Für das Produzierte kann schon deswegen keine gemeinsame Verantwortung übernommen werden, weil es gar keine Multiautoren gibt. Zwar entstehen die Texte auf Basis der Daten, die von verschiedenen Personen stammen, sie sind aber möglicherweise in der produzierten Form nie auch nur von irgendeinem einzelnen Autor produziert worden. Sie können also völlig neu zusammengewürfelt werden – in der Tat spielt die zufällige Auswahl von Textteilen eine wichtige Rolle in der generativen KI.

Die Eigenschaft von generativer KI, riesige Datenmengen von Tausenden menschlichen Textverfassenden zu verwenden, scheint den Zugriff auf eine Vielfalt von unterschiedlichen Positionen, Sichten oder Meinungen zu ermöglichen. Obwohl dies prinzipiell nicht falsch ist, liegt darin auch eine gewisse Täuschung. In der Realität stehen nämlich nur wenige marktbeherrschende Systeme zur Verfügung, die in ihrer Funktionalität und Diskursfähigkeit stark voneinander abweichen. Die Anzahl der tatsächlich „sprechenden“ Systeme ist stark limitiert. Aus der Vielzahl potenzieller Perspektiven wird eine technisch und ökonomisch limitierte Auswahl. Diese Beschränkung auf wenige Anbieter ist von großer Relevanz, weil die Systemanbieter letztlich über den Charakter der generierten Texte entscheiden.

KI-Systemanbieter zeigen sich durchaus bewusst für die brisante Rolle generativer KI, d.h. für die oben erwähnten Risiken, die KI-Systeme mit sich bringen. Die Anbieter schließen zum Beispiel eine Vielzahl von Nutzungsmöglichkeiten generativer KI-Systeme aus. So verfügt etwa Open AI, der Anbieter von ChatGPT, in den Nutzungsbedingungen (Allgemeine Richtlinien), dass ihr System für folgende Zwecke nicht verwendet werden darf (OpenAI, 2025).

- 2. Nutzen Sie unseren Dienst nicht, um sich selbst oder anderen zu schaden – Nutzen Sie beispielsweise unseren Dienst nicht, um Selbstmord oder Selbstschädigung zu fördern, Waffen zu entwickeln oder zu verwenden, andere zu verletzen oder Eigentum zu zerstören, oder sich an nicht autorisierten Aktivitäten zu beteiligen, die die Sicherheit eines Dienstes oder Systems verletzen.

- 3. *Verwenden oder verbreiten Sie Output unserer Dienste nicht, um anderen zu schaden – Geben Sie beispielsweise keinen Output unserer Dienste weiter, um zu betrügen, zu spammen, zu täuschen, zu mobben, zu belästigen, zu diffamieren, aufgrund geschützter Merkmale zu diskriminieren, Kinder zu sexualisieren oder Gewalt, Hass oder das Leiden anderer zu fördern.*

Darüber hinaus wird spezifiziert, dass mittels der von ChatGPT bereitgestellten Texte keine „[i]ndividuell zugeschnittene Rechts-, Finanz- oder Gesundheits-/medizinische Beratung ohne Prüfung durch eine qualifizierte Fachkraft und ohne Offenlegung der Verwendung von KI-Unterstützung und ihrer möglichen Einschränkungen“ durchgeführt werden darf. Ähnliches gilt für automatisierte Entscheidungen in sensiblen Bereichen (z.B. Strafverfolgung), Glücksspiele, politische Kampagnen, Lobbyarbeit, Desinformation, anzüglische Inhalte und vieles mehr. Es dürfen auch keine Tools erzeugt werden, die sich an unter Dreizehnjährige richten (OpenAI, 2025).

Die Beschränkung des KI-generierten Diskurses

Angesichts der vielfältigen und weitreichenden Einschränkungen erscheint es unwahrscheinlich, dass sich alle, die ChatGPT verwenden, dieser Einschränkungen überhaupt bewusst sind. Es kann vielmehr angenommen werden, dass die Systeme für viele der geächteten Zwecke regelmäßig verwendet werden, z.B. für Lobbying oder als Fremdsprachentrainer für Schülerinnen auch unter zwölf Jahren. Schon aus diesem Grund überlassen die Systemhersteller die Entscheidung über Art, Inhalt und Zweck der Verwendung von ChatGPT nicht allein den Nutzenden. Sie versuchen die Inhalte, die mit ihren Systemen erzeugt werden können, so zu beschränken, dass eine Vielzahl von möglicherweise bedenklichen Texten unterdrückt wird.

Zur Vermeidung problematischer Inhalte werden Chatbots zunehmend so programmiert, dass sie bestimmte Themenbereiche nicht behandeln. Sie stellen einen Spezialfall der inzwischen weitreichend eingesetzten algorithmischen Moderation von Online-Inhalten dar (Prem & Krenn, 2024). Die eingesetzten Mechanismen reichen von der Filterung einzelner Begriffe bis hin zur systematischen Vermeidung ganzer Themenfelder. Ein prominentes Beispiel ist der chinesische Chatbot DeepSeek, der sich konsequent nicht zu politischen Fragen äußert, die das eigene Regime betreffen. Aber auch westliche Systeme unterliegen vielfältigen Einschränkungen: Sie generieren keine Texte über Waffenherstellung, Drogenproduktion oder Gewalt, meiden häufig Themen wie Sexualität oder Religion und reagieren zurückhaltend auf ethisch komplexe Sachverhalte. Diese Beschränkungen sind selten staatlich verordnet, sondern entstammen überwiegend der freiwilligen Selbstkontrolle durch Unternehmen. Es handelt sich dabei – in Abwandlung des Titels der Arbeit von Foucault (1991) – um eine „private“ Ordnung des Diskurses – eine diskursive Infrastruktur, die weitgehend außerhalb demokratischer Kontrolle operiert.

Besonders brisant ist dabei, dass die Mechanismen der Diskurskontrolle oft unsichtbar bleiben. Die Filter wirken im Hintergrund, ihre Kriterien sind nicht transparent und können sich jederzeit ändern. Dadurch entsteht ein System impliziter Ausschlüsse, das sich den klassischen Modellen äußerer Zensur – wie Foucault sie

beschreibt – entzieht. Statt expliziter Verbote erleben wir eine subtile, algorithmisch gesteuerte Form der Lenkung, die bestimmt, was gesagt werden darf – und vor allem was nicht. Die Grenze zwischen technischer Sicherheitsmaßnahme und ideologischer Zensur verschwimmt dabei zunehmend.

Ein besonders kritischer Aspekt ist die zunehmende Delegation moralischer Entscheidungen an KI-Systeme. Immer häufiger treffen Chatbots – statistisch generierte – Einschätzungen, deren ethischer Gehalt erheblich ist – sei es bei der Formulierung politischer Aussagen, bei der Darstellung sozialer Realitäten oder bei der Entscheidung, welche Information wem zugänglich gemacht wird. Diese Verschiebung der Verantwortung ist problematisch, da die Entscheidungsgrundlagen der Systeme nicht offengelegt sind. Die moralische Verantwortung wird in die Black Box der Technologie ausgelagert – mit unklaren Konsequenzen (van der Rijt et al., 2025).

Für diejenigen, die sich auf ein korrektes Funktionieren generativer KI-Systeme verlassen, die Nutzenden nämlich, sind die zuletzt erläuterten Ausschlussmechanismen besonders problematisch. Während die Verwendung (bzw. der Ausschluss der Verwendung) von KI-Systemen gemäß der Nutzungsbedingungen noch relativ leicht zu beurteilen ist und ggf. auch eine Einschätzung des Gesagten durch genaue Überprüfung eines generierten Textes noch möglich sein mag, ist die Überprüfung dessen, was gerade nicht gesagt wird, eine schlechthin unmögliche Aufgabe. Es ist heute keinem Nutzenden generativer KI zuzumuten, genau zu hinterfragen, welche Ausschlussmechanismen ein spezieller Chatbot gerade implementiert. Es bleibt meist unklar, was der Chatbot alles nicht sagen kann, weil er es nicht sagen darf. Es bleibt auch unklar, welche Meinungen aus der Vielzahl der potenziell in den Trainingsdaten vorhandenen Pluralität von Ansichten nicht durch die generative KI produziert werden können. Dies mag manchmal durchscheinen; es kann ggf. auch zum Gegenstand genauer Analysen gemacht werden – es wird aber kaum jemals im Vorhinein genau offengelegt.

Die Produzenten der generativen KI-Systeme versuchen damit genau jenen Diskurs zu umschiffen, der möglicherweise problematisch für den Erfolg ihrer Systeme werden kann. Im Zweifel mag es ihnen besser erscheinen, wenn ihr System lieber keine Gewalt beschreiben kann, keine Informationen zur Produktion von Waffen liefert und lieber auch keine homosexuellen Liebesgeschichten erzählen kann. Damit aber werden KI-gestützte Diskurse verunmöglicht, in denen diese Aspekte – in welcher Form auch immer – gebraucht werden. So kann ein System, das nicht über Gewalt sprechen kann, nicht zur Prävention von Gewalt, nicht zum Schreiben eines Kriminalromans oder als Unterstützung für einen Kriegsberichterstatter verwendet werden. Zumindest ist in diesen Fällen mit drastischen Qualitätseinbußen zu rechnen.

Die menschliche Autorin wird in diesem Prozess konsequent entmachtet, d.h., es wird ihr die Freiheit entzogen, automationsunterstützt Krimis, Hilfestellungen für Gewaltopfer oder Beschreibungen aktuellen Kriegsgeschehens zu erstellen. Keine dieser Aktivitäten ist per se gesetzwidrig oder sozial schädlich. Im Gegenteil, sie sind zumindest bisher häufig ein wichtiger Teil gesellschaftlicher und kultureller Leistungen und durchaus geeignet, das Gemeinwohl zu fördern. Die KI-Systemhersteller werden jedoch derartige Zensur häufig mit einer gesellschaftlichen Aufgabe begründen. Sie argumentieren z.B. mit „schädlichen Inhalten“, die allerdings im Gegensatz zu „verbotenen Inhalten“ nicht gesetzlich bestimmt sind (z.B. terroristische oder Nazi-Propaganda). Es erscheint aber wahrscheinlich, dass die Schädlichkeit generierter

Inhalte seitens der KI-Unternehmen nicht ausschließlich auf gesellschaftliche Schäden abstellt, sondern sich im Umweg der oben erwähnten KI-Risiken eher an potenziellen Schäden für die KI-Unternehmen orientiert. Schädlich ist in diesem Sinne vor allem, was zu einem Imageschaden und damit zu einem Einkommensverlust der KI-Systemanbieter führt.

Diese kommerzielle Ordnung des Diskurses bleibt dem menschlichen Autor jedoch überwiegend verborgen. Sie ist philosophisch formuliert die transzendente Bedingung der Verfügbarkeit generativer KI-Dienste. Dem KI-System bzw. seinem Anbieter kann der Vorwurf der Täuschung nicht erspart werden. Sie erscheinen verantwortungsvoll zu agieren, wo sie in Wahrheit keine Verantwortung übernehmen, und zwar weder in den generierten Texten noch in der Verantwortlichkeit durch Autorenschaft noch durch Übergabe der gesamten Verantwortung an die Autorin. Diese bleibt im Unklaren darüber, welchen Diskursverboten sie unterliegt oder welchen Einschränkungen ihrer schöpferischen Freiheit sie ausgesetzt wird. Von den zitierten Eigenschaften von verantworteter Autorenschaft, Verstand und Freiheit, mangelt es vor allem an der Freiheit. Dass – im Sinne der Möglichkeit von Täuschungen – auch der Verstand bedroht ist, liegt auf ebenfalls auf der Hand.

Ausblick

Die wachsende Verbreitung von Chatbots in Unternehmen bringt eine Vielzahl tiefgreifender Herausforderungen mit sich – insbesondere im Hinblick auf die Kontrolle über digitale Diskurse. Die derzeit verfügbaren Systeme entziehen sich weitgehend der Einflussnahme durch ihre Nutzerinnen oder die Unternehmen, in denen sie beschäftigt sind. Unternehmen verfügen nur über begrenzte Möglichkeiten, um auf die inhaltliche Gestaltung der KI oder deren ethische Ausrichtung einzuwirken. Diese Situation erfordert eine grundlegende gesellschaftliche Diskussion über Regulierung, Transparenz und Teilhabe in einer zunehmend KI-dominierten Kommunikationskultur. Wenn wir nicht nur Konsumenten, sondern auch Mitgestaltende dieses digitalen Diskurses sein wollen, müssen neue Formen der Kontrolle, Verantwortung und Autorenschaft entwickelt werden. Dies bedarf sowohl einer weitreichenderen technischen, politischen als auch philosophischen Aufarbeitung der grundlegenden Problematik.

Zu den wichtigsten philosophischen Fragestellungen gehört, wie generative KI-Systeme überhaupt mit uns „sprechen“ sollen, d. h., welche Art von Einschränkungen, Präferenzen, Sicherheitsmaßnahmen sollen sie überhaupt vorsehen und wer soll darüber entscheiden? Bedarf es dazu einer neuen Ethik für generative KI-Systeme oder genügen demokratische Prinzipien oder etwa die Prinzipien einer Diskursethik im Sinne von Habermas? Ist die Nutzung von Chatbots stets mit einem Mangel an Selbstrespekt verbunden, wie etwa van der Rijt et al. (2025) argumentieren? Solange wir nicht begonnen haben, überlegte Antworten auf diese Herausforderungen zu geben, wird auch die Frage nach den moralischen Prinzipien des Einsatzes generativer KI in Unternehmen auf keinem soliden Boden stehen.

Literatur

- Bellware, K. & Masih, N. (2024, 24. Oktober). Her teenage son killed himself after talking to a chatbot. Now she's suing. *The Washington Post*. <https://www.washingtonpost.com/nation/2024/10/24/character-ai-lawsuit-suicide/>
- Bousquette, I. (2025a, März 4). How AI Tools Are Reshaping the Coding Workforce. *The Wall Street Journal*. <https://www.wsj.com/articles/how-ai-tools-are-reshaping-the-coding-workforce-6ad24c86?msockid=0236edb14fdc69af3b60ff614e75689a>
- Bousquette, I. (2025b, April 25). Prompt Engineer, the Hottest AI Job of 2023, Is Already Obsolete. *The Wall Street Journal*. <https://www.wsj.com/articles/the-hottest-ai-job-of-2023-is-already-obsolete-1961b054?msockid=0236edb14fdc69af3b60ff614e75689a>
- Dam, S. K., Hong, C. S., Qiao, Y. & Zhang, C. (2024). *A Complete Survey on LLM-based AI Chatbots*. <https://doi.org/10.48550/arXiv.2406.16937>
- Foucault, M. (1991). *Die Ordnung des Diskurses*. Fischer-Taschenbuch.
- Nida-Rümelin, J. & Staudacher, K. (2024). Philosophical Foundations of Digital Humanism. In H. Werthner, C. Ghezzi, J. Kramer, J. Nida-Rümelin, B. Nuseibeh, E. Prem & A. Stanger (Hrsg.), *Introduction to Digital Humanism* (S. 17–30). Springer. https://doi.org/10.1007/978-3-031-45304-5_2
- OpenAI. (2025). *Nutzungsrichtlinien vom 29. Januar 2025*. <https://openai.com/de-DE/policies/usage-policies/>
- Prem, E. (2025). Managing Entrepreneurial AI Ethics Risks. In C. H. Hoffmann (Hrsg.), *Artificial Intelligence, Entrepreneurship and Risk. Reflections and Positions at the Crossroads between Philosophy and Management* (S. 453–475). Springer VS. https://doi.org/10.1007/978-3-658-45544-6_23
- Prem, E. & Krenn, B. (2024). On Algorithmic Content Moderation. In H. Werthner, C. Ghezzi, J. Kramer, J. Nida-Rümelin, B. Nuseibeh, E. Prem & A. Stanger (Hrsg.), *Introduction to Digital Humanism* (S. 481–493). Springer. https://doi.org/10.1007/978-3-031-45304-5_30
- Rane, N. (2023). *Role and Challenges of ChatGPT and Similar Generative Artificial Intelligence in Business Management*. <https://doi.org/10.2139/ssrn.4603227>
- Srivastava, A., Rastogi, A., Rao, A., Shoeb, A. A. M., Abid, A., Fisch, A., Brown, A. R., Santoro, A., Gupta, A., Garriga-Alonso, A., Kluska, A., Lewkowycz, A., Agarwal, A., Power, A., Ray, A., Warstadt, A., Kocurek, A. W., Safaya, A., Tazarv, A. ... Wu, Z. (2022). *Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models*. <https://doi.org/10.48550/arXiv.2206.04615>
- van Der Rijt, J.-W., Coelho Mollo, D. & Vaassen, B. (2025). *AI Mimicry and Human Dignity: Chatbot Use as a Violation of Self-Respect*. <https://doi.org/10.48550/arXiv.2503.05723>
- Wang, Z., Chu, Z., Doan, T. V., Ni, S., Yang, M. & Zhang, W. (2024). History, development, and principles of large language models: an introductory survey. *AI and Ethics*, 1–17. <https://doi.org/10.1007/s43681-024-00583-7>
- Werthner, H., Ghezzi, C., Kramer, J., Nida-Rümelin, J., Nuseibeh, B., Prem, E. & Stanger, A. (2023). *Introduction to Digital Humanism: A Textbook*. Springer. <https://doi.org/10.1007/978-3-031-45304-5>

Algorithmische Diskriminierung am Arbeitsplatz

Fairness, Transparenz, Datenschutz und kontrafaktische Erklärungen in KI-gestützten ADM-Systemen

Lejla Rizvanovik

1. Einführung

Die zunehmende Integration von Systemen automatisierter Entscheidungsfindung (ADM-Systeme) in die Arbeitswelt stellt neue Anforderungen an Datenschutz, Fairness und Transparenz. Ob im Recruiting, der Leistungsbewertung oder der Personalentwicklung – algorithmische Verfahren beeinflussen heute maßgeblich, wie Entscheidungen über Menschen getroffen werden. Dabei wird oft suggeriert, dass maschinelle Entscheidungen objektiver oder effizienter seien als menschliche. Doch aktuelle Forschung zeigt, dass sich bestehende Diskriminierungen durch algorithmische Prozesse nicht nur fortsetzen, sondern sogar verstärken können (Barberá, 2025; Wachter et al., 2018).

Ein zentrales Problem liegt in der Intransparenz dieser Systeme: Datenbasierte Entscheidungen sind für betroffene Arbeitnehmer:innen häufig nicht nachvollziehbar, und der gesetzliche Rahmen bleibt in vielen Punkten vage. Begriffe wie „allein automatisiert“ oder „signifikante Auswirkungen“ in Artikel 22 der DSGVO sind bislang nicht eindeutig definiert und erschweren sowohl die rechtskonforme Implementierung als auch die wirksame Durchsetzung der bestehenden Schutzrechte (Barberá, 2025).

Dieser Beitrag untersucht daher, wie sich Transparenz und Partizipation in KI-gestützten Entscheidungsprozessen im Arbeitskontext verbessern lassen – insbesondere durch den Einsatz kontrafaktischer Erklärungen. Dabei wird argumentiert, dass der derzeitige Rechtsrahmen wichtige Schutzmechanismen zwar vorsieht, diese jedoch in der Praxis häufig ins Leere laufen. Vor allem das Fehlen eines klar geregelten „Rechts auf Erklärung“ erschwert es Betroffenen, Entscheidungen zu verstehen, anzufechten oder zukünftige Ergebnisse aktiv zu beeinflussen (Wachter et al., 2018).

Ziel des Beitrags ist es, einen Brückenschlag zwischen rechtlichen, technischen und praktischen Perspektiven zu leisten. Im Zentrum steht die Frage: *Wie kann es gelingen, automatisierte Entscheidungen nachvollziehbar und diskriminierungsfrei zu gestalten – ohne dabei Geschäftsgeheimnisse oder algorithmische Komplexität offenzulegen?* Dazu wird zunächst eine rechtsdogmatische Analyse von Artikel 22 DSGVO und relevanter EuGH-Judikatur vorgenommen – insbesondere der Entscheidung im Fall C-634/21 (SCHUFA I), die die Reichweite automatisierter Scoring-Systeme präzisiert (EuGH, 2023a), sowie der verbundenen Rechtssachen C-26/22 und C-64/22 (SCHUFA II) (EuGH, 2023b), die sich mit der datenschutzrechtlichen Zulässigkeit der fortgesetzten Speicherung von Insolvenzregistern durch Auskunftfeien befassen.

Ergänzend werden ausgewählte Best Practices zur risikobasierten Bewertung von ADM-Systemen entlang des AI-Lifecycles diskutiert (Barberá, 2025). Technisch fokussiert sich der Beitrag auf das Konzept der kontrafaktischen Erklärungen, die als

transparente, datenschutzkonforme und leicht verständliche Alternativen zu komplexen Modellinterpretationen dienen.

Die interdisziplinäre Analyse zeigt auf, dass Unternehmen ADM-Systeme nicht nur technisch beherrschen, sondern auch ethisch und rechtlich verantwortungsvoll gestalten müssen. Die Implementierung kontrafaktischer Erklärungen bietet hier ein praxistaugliches Mittel, um Transparenz zu erhöhen, Vertrauen von Mitarbeitenden zu stärken und zugleich regulatorische Anforderungen zu erfüllen – insbesondere vor dem Hintergrund des bereits in Kraft getretenen EU AI Acts.

2. Automatisierte Entscheidungen in der Arbeitswelt

2.1 Technologische Grundlagen von ADM-Systemen

Automatisierte Entscheidungsfindung (Automated Decision-Making, ADM) bezeichnet den Prozess, bei dem algorithmische Systeme – meist gestützt auf maschinelles Lernen – auf Grundlage großer Datenmengen eigenständig Schlüsse ziehen und Handlungen auslösen. Solche Systeme analysieren Muster, klassifizieren Individuen und treffen Vorhersagen, die zunehmend in personalbezogene Entscheidungen einfließen (Barberá, 2025). Dabei kommen sowohl klassische algorithmische Verfahren als auch moderne Large Language Models (LLMs) zum Einsatz, die durch ihr Kontextverständnis in der Lage sind, auch komplexe Text- und Datenanalysen durchzuführen (Wachter et al., 2018).

Ein typischer ADM-Prozess im Arbeitskontext folgt einem dreistufigen Ablauf: (1) Datenerhebung (z.B. Lebensläufe, Performance-Daten), (2) algorithmische Bewertung (Scoring, Matching) und (3) automatisierte Handlung (Einladung zum Interview, Ablehnung, Bonusvergabe). Besonders kritisch sind dabei Systeme, die allein automatisierte Entscheidungen mit erheblicher Wirkung im Sinne von Artikel 22 DSGVO treffen – also ohne nennenswerte menschliche Kontrolle.

Neuere KI-Modelle wie LLMs bergen zusätzliche Herausforderungen: Sie speichern mitunter personenbezogene Informationen aus Trainingsdaten (Memorization Risk), verarbeiten externe Wissensquellen mit potenziell unkontrollierten Daten (RAG-Schwachstellen) und machen es oft unklar, wer datenschutzrechtlich verantwortlich ist (Barberá, 2025). Diese Probleme verschärfen sich mit dem Aufkommen sogenannter „agentic AI systems“, bei denen KI-Instanzen zunehmend autonom agieren und Entscheidungen ohne unmittelbare menschliche Aufsicht treffen – ein Szenario, das sowohl vom AI Act (Art. 14, Erwägungsgründe 47 & 60) als auch von Datenschutzexpert:innen als besonders risikobehaftet bewertet wird.

Trotz dieser Herausforderungen verspricht ADM erhebliche Effizienzgewinne – etwa durch die Standardisierung von HR-Prozessen, Ressourcenoptimierung oder die Skalierbarkeit von Personalentscheidungen. Doch mit diesen Potenzialen gehen strukturelle Risiken einher: Bias in Trainingsdaten, mangelnde Transparenz und eingeschränkte Kontrollmöglichkeiten für betroffene Personen. Diese Spannungsfelder treten besonders deutlich in den konkreten Anwendungsszenarien auf.

2.2 Typische Anwendungsfelder im HR- und Arbeitskontext

Automatisierte Entscheidungsverfahren finden heute in nahezu allen Phasen des Beschäftigtenlebenszyklus Anwendung. Bereits im Recruiting setzen viele Unternehmen auf KI-gestützte Matching-Systeme, um Bewerber:innen nach bestimmten Kriterien vorzufiltern. Solche Systeme analysieren Lebensläufe, Social-Media-Profile oder Online-Tests und generieren Scores, die über den Zugang zu Vorstellungsgesprächen entscheiden können. Ein praxisnahes Beispiel liefert der Fall des AMS-Chatbots in Österreich: Das Arbeitsmarktservice (AMS) setzte ein KI-gestütztes Assistenzsystem zur digitalen Berufsberatung ein, das Bewerbenden Empfehlungen auf Basis ihrer Angaben aussprach. In der Anwendung zeigte sich jedoch, dass das System geschlechterstereotype Empfehlungen reproduzierte – etwa IT-Berufe für Männer und geisteswissenschaftliche Felder für Frauen. Die öffentliche Kritik war erheblich: Obwohl der Chatbot formell nur vorbereitende Funktionen erfüllte, wurde deutlich, wie stark solche Systeme die Berufswahl und Wahrnehmung von Karrieren vorstrukturieren können. Der Fall zeigt exemplarisch, wie ADM-Systeme auch ohne finale Entscheidung nachhaltige Auswirkungen auf individuelle Berufsperspektiven haben können – insbesondere, wenn sie systematisch Vorurteile aus Trainingsdaten übernehmen und verstärken (Köwer, 2024).

Auch im Bereich der Leistungsbewertung und Zielvereinbarung kommen zunehmend KI-Systeme zum Einsatz. Mitarbeitende werden anhand vordefinierter Kennzahlen beurteilt, die häufig aus verhaltens- oder leistungsbezogenen Daten aggregiert werden. Diese Systeme fördern zwar Effizienz, bergen jedoch die Gefahr, dass strukturelle Vorurteile – etwa in Bezug auf Geschlecht, Herkunft oder Alter – unbewusst reproduziert werden, wenn Trainingsdaten vergangenheitsbasierte Diskriminierung enthalten (Barberá, 2025; Wachter et al., 2018).

Weitere Anwendungsfelder umfassen automatisierte Versetzungs-, Beförderungs- oder Kündigungsentscheidungen, die auf Basis von Data Analytics vorgenommen werden. Hier können selbst scheinbar neutrale Systeme zu einer indirekten Diskriminierung führen – etwa wenn Krankheitstage, Teilzeitquoten oder Pausenzeiten algorithmisch negativ gewichtet werden.

Besonders problematisch ist, dass viele dieser Systeme in der Praxis nicht als Hochrisiko-KI gemäß AI Act klassifiziert werden, obwohl ihre Auswirkungen erheblich sind. Anhang III des AI Act nennt explizit KI-Systeme im Beschäftigungskontext als Hochrisiko-Anwendungen, doch die konkrete Einstufung hängt oft von der Selbstauskunft des Deployers ab. In der Folge können Systeme mit Relevanz für die Praxis – wie automatisiertes Job-Ad-Targeting oder Gehaltsvorschläge – durch das Raster der Regulierung fallen (European Parliament & Council of the European Union, 2024).

Diese Konstellation macht deutlich: Ohne transparente Gestaltung, klare Verantwortlichkeiten und risikobewusste Implementierung bergen ADM-Systeme im Arbeitskontext erhebliche Gefahren – insbesondere für marginalisierte Gruppen wie People of Color, Frauen, Menschen mit Behinderung, ältere Arbeitnehmende, queere Personen, neurodivergente Bewerbende, nichtmuttersprachliche Personen oder Personen mit prekärem Aufenthaltsstatus. Genau hier setzt die rechtliche Analyse im nächsten Kapitel an.

3. Rechtsrahmen: Datenschutz, KI-Regulierung und Diskriminierungsschutz

3.1 Artikel 22 DSGVO im Fokus

Artikel 22 der Datenschutz-Grundverordnung (DSGVO) ist die zentrale Norm für den Umgang mit automatisierten Einzelentscheidungen, insbesondere dann, wenn sie rechtliche oder vergleichbar erhebliche Wirkungen für betroffene Personen entfalten. Gemäß Art. 22 Abs. 1 DSGVO hat jede Person das Recht, nicht einer ausschließlich auf automatisierter Verarbeitung beruhenden Entscheidung unterworfen zu werden, sofern diese rechtliche Wirkung entfaltet oder sie in ähnlicher Weise erheblich beeinträchtigt. Ausnahmen nach Art. 22 Abs. 2 DSGVO betreffen entweder Vertragserfüllung, gesetzliche Grundlagen oder ausdrückliche Einwilligung – jeweils unter Zusatzbedingungen wie Transparenz und dem Recht auf menschliches Eingreifen.

In der Praxis führt diese Vorschrift jedoch zu erheblichen Auslegungsunsicherheiten. So bleibt unklar, wann eine Entscheidung tatsächlich „allein“ automatisiert ist – insbesondere, wenn nominell menschliche Überprüfungen vorgesehen sind, diese aber rein formell oder routinemäßig erfolgen. Ebenso schwammig bleibt der Begriff der „signifikanten Auswirkung“: Ist ein abgelehnter Bewerbungsvorgang bereits ein erheblicher Eingriff? Oder erst die Kündigung? Die juristische Literatur und Aufsichtsbehörden bewerten dies uneinheitlich (Barberá, 2025; Wachter et al., 2018).

Das jüngste SCHUFA-Urteil des Europäischen Gerichtshofs (EC-LI:EU:C:2023:957, C-634/21) hat hierzu wichtige Klarheit geschaffen: Der EuGH entschied, dass bereits ein automatisiert ermittelter Score, der maßgeblich in eine Entscheidung einfließt, unter Artikel 22 DSGVO fallen kann – selbst dann, wenn die finale Entscheidung formal von einem Menschen getroffen wird. Entscheidend sei der faktische Einfluss der automatisierten Bewertung (EuGH, 2023a). Dieses Urteil stärkt die Betroffenenrechte in Kontexten wie der Kreditvergabe oder im HR-Scoring, wo algorithmische Bewertungen den Zugang zu Chancen real beeinflussen.

Allerdings ist Artikel 22 nicht als „Recht auf Erklärung“ ausgestaltet. Die Norm enthält weder eine inhaltlich-technische Begründungspflicht für automatisierte Entscheidungen noch ein echtes Widerspruchsrecht gegen den algorithmischen Prozess selbst. Es ist daher umstritten, ob Art. 22 ein Verbot oder lediglich ein Schutzrecht auf Antrag darstellt (Wachter et al., 2018). Die Unsicherheiten in der Anwendung mindern die Durchsetzungskraft des Artikels – insbesondere im Kontext moderner ADM-Systeme, die komplexe Modelle wie LLMs verwenden.

3.2 Aktuelle Judikatur: SCHUFA-Urteil des EuGH (C-634/21)

Im Zentrum der SCHUFA-Entscheidung stand die Frage, ob ein automatisiertes Bonitätsscore-Verfahren im Fall C-634/21 (EuGH, 2023a) eine „allein automatisierte Entscheidung“ im Sinne von Art. 22 DSGVO darstellt. Der EuGH kam zum Schluss, dass die bloße Übernahme eines algorithmischen Scoring-Werts durch eine kreditvergebende Stelle ausreiche, um den Schutzbereich von Art. 22 zu eröffnen. Eine bloß formale menschliche Bestätigung ohne kritische Prüfung genüge nicht, um aus der Definition „allein automatisiert“ herauszufallen (EuGH, 2023a).

Diese Entscheidung ist übertragbar auf ADM-Systeme in der Arbeitswelt – etwa bei der automatisierten Vorauswahl von Bewerber:innen oder algorithmischen Leistungsbewertungen. Arbeitgeber:innen, die sich auf KI-gestützte Scores stützen, müssen sicherstellen, dass eine tatsächliche menschliche Kontrolle erfolgt – einschließlich der Fähigkeit, das System zu hinterfragen, Fehler zu erkennen und Alternativen abzuwägen.

Besonders bedeutsam ist das Urteil im Zusammenspiel mit den Informationspflichten aus Art. 13 und 14 DSGVO sowie dem Grundsatz der fairen und transparenten Verarbeitung (Art. 5 Abs. 1 lit. a DSGVO). Wird der Einfluss eines ADM-Systems verschleiert oder nicht transparent dargestellt, verletzt dies potenziell das Recht der betroffenen Person auf faire Datenverarbeitung – ein Risiko, das durch kontrafaktische Erklärungen gemindert werden könnte (siehe Kapitel 4).

3.3 DSGVO vs. EU AI Act – Policy-Ziele im Vergleich

Mit dem AI Act (Verordnung (EU) 2024/1689) verfolgt die Europäische Union das Ziel, harmonisierte Vorschriften für den sicheren, vertrauenswürdigen und grundrechtskonformen Einsatz von KI-Systemen zu schaffen. Im Gegensatz zur DSGVO, die den Fokus auf den Schutz personenbezogener Daten und individuelle Betroffenenrechte legt, zielt der AI Act stärker auf systemische Kontrolle, Produktsicherheit und Marktaufsicht ab (European Parliament & Council of the European Union, 2024, Art. 1, Erwägungsgründe 1–2).

Der AI Act klassifiziert KI-Systeme anhand ihres Risikopotenzials. Systeme im Arbeitskontext – etwa zur Einstellung, Bewertung oder Kündigung – gelten gemäß *Anhang III* als *Hochrisiko-KI*. Für diese Systeme gelten umfassende Anforderungen, darunter:

- *Risikomanagementsysteme* (Art. 9),
- *Datengovernance und Qualitätspflichten* (Art. 10),
- *Transparenzanforderungen* (Art. 13),
- sowie *pflichtgemäße menschliche Aufsicht* (Art. 14).

Während die DSGVO auf individuell einklagbare Rechte – wie Informations-, Widerspruchs- oder Löschrechte – setzt, verfolgt der AI Act einen *präventiven Governance-Ansatz*, der entlang des gesamten Lebenszyklus eines KI-Systems greift. Beide Regime überschneiden sich insbesondere bei ADM-Systemen, haben jedoch unterschiedliche Zielrichtungen und Begrifflichkeiten. So verlangt die DSGVO etwa die Offenlegung automatisierter Entscheidungen (Art. 13 Abs. 2 lit. f), während der AI Act auf systemische Dokumentation, Auditability und funktionale Nachvollziehbarkeit fokussiert (Art. 16).

Auch im Bereich des Diskriminierungsschutzes unterscheiden sich die Zugänge deutlich: Während die DSGVO vorrangig auf individuell einklagbare Rechte zielt, sind die rechtlichen Zuständigkeiten bei struktureller oder kollektiver Diskriminierung – etwa durch algorithmisches Scoring oder KI-basiertes Schichtmanagement – bislang unklar verteilt. Im Fall Deliveroo (Tribunale di Bologna, 2020) wurde der Plattformanbieter zu einer Entschädigung verurteilt, obwohl dieser argumentier-

te, keine personenbezogenen Daten verarbeitet zu haben. Der AI Act sieht gestufte Pflichten entlang der KI-Lieferkette vor – mit Auflagen für Entwickler:innen, Betreiber:innen und Nutzer:innen. Dennoch bleibt offen, wie Beweislast und Verantwortlichkeit bei diskriminierender Wirkung zugewiesen werden. Hense und Mustać (2024, S. 203–205) betonen, dass insbesondere kollektive Diskriminierung in der Regulierung bislang *unterbelichtet* ist – obwohl genau hier die größten realweltlichen Effekte auftreten.

Limited-Risk Use Cases – unterregulierte Praxisrelevanz

Ein zentrales Problem liegt in der Einstufung vieler *praxisrelevant einflussreicher Systeme* als sogenannte „Limited Risk“-Anwendungen, sofern sie formal keine endgültigen Entscheidungen treffen oder keine sensiblen Daten verarbeiten. Solche Systeme übernehmen unterstützende oder vorbereitende Funktionen – etwa *algorithmisches Ranking bei Jobanzeigen oder Matching-Vorschläge im Bewerbungsprozess* –, können aber faktisch Vorentscheidungen treffen, die den Zugang zu Arbeit strukturieren. Dennoch unterliegen sie lediglich Basispflichten, etwa allgemeinen Transparenzgeboten (Art. 52 AI Act), nicht aber den Schutzmaßnahmen für Hochrisiko-KI.

General Purpose AI und Übergangsregelung

Ein zusätzlicher regulatorischer Graubereich betrifft sogenannte *General Purpose AI-Systeme (GPAI)* – also große, vielseitig einsetzbare Modelle wie GPT, Claude oder LLaMA. Für GPAI-Modelle, die *vor dem 2. August 2025 auf dem EU-Markt bereitgestellt wurden*, gilt laut Art. 83 AI Act eine Übergangsregelung bis zum 2. August 2027. Innerhalb dieser Frist müssen Anbieter:innen die vollständigen Anforderungen des AI Act schrittweise umsetzen. Diese Regelung betrifft auch viele KI-Komponenten, die bereits heute in HR-ADM-Systeme integriert werden, etwa durch LLM-gestützte Screening- und Analysefunktionen.

TRAI-Pflichten ≠ Hochrisiko-Klassifizierung

Verwirrung entsteht auch durch die Existenz der sogenannten *TRAI Use Cases* – also KI-Systemen, die laut Art. 52 AI Act Transparenzpflichten unterliegen (z. B. Deepfakes, Emotionserkennung oder Chatbots) – *ohne Hochrisiko-Klassifizierung*. Diese Systeme müssen Betroffene über den KI-Einsatz informieren, unterliegen aber nicht denselben strengen Anforderungen wie Hochrisiko-Systeme. In der Praxis führt diese Unterscheidung oft zu *Fehleinschätzungen*, da ein System mit Transparenzpflicht *nicht automatisch als risikobehaftet eingestuft wird*.

Insgesamt wird deutlich: Zwar fordern sowohl DSGVO als auch AI Act Schutzmaßnahmen wie Transparenz, menschliche Kontrolle und Rechenschaftspflicht, doch *vage Begriffsdefinitionen, regulatorische Ausnahmen und technische Komplexität schwächen die praktische Durchsetzbarkeit*. ADM-Systeme mit signifikanter Wirkung – etwa im Scoring, in der Personalentscheidung oder im Zugang zu Stellenanzeigen – *fallen oft durch das Raster*, wenn sie formal als „Limited Risk“ klassifiziert sind oder GPAI-Komponenten nutzen, die sich noch in der Übergangsfrist befinden.

Ein eindrückliches Beispiel für die gesellschaftlichen Folgen solcher Systeme zeigte sich in Österreich: Der AMS-Chatbot, entwickelt vom Arbeitsmarktservice zur

automatisierten Berufsberatung, geriet rasch in Kritik, da er geschlechterstereotype Empfehlungen abgab – etwa IT-Berufe für Männer und geisteswissenschaftliche Studiengänge für Frauen. Trotz der Bekundung, diskriminierende Muster im Training vermieden zu haben, offenbarte der Fall die tief verankerten Verzerrungen in realen Datenbeständen und ihre Reproduktion durch KI. Dieser Vorfall zeigt exemplarisch, wie auch scheinbar „unterstützende“ Systeme strukturelle Diskriminierung verstärken können – insbesondere, wenn sie außerhalb der Hochrisiko-Klassifizierung operieren. Fälle wie dieser unterstreichen die Notwendigkeit strengerer Regelungen und klarer Kontrollmechanismen, wie sie im AI Act angestrebt werden (Köver, 2024).

Kontrafaktische Erklärungen (siehe Kapitel 4) könnten ein praktikabler Weg sein, um diese regulatorischen Grauzonen technisch sinnvoll zu überbrücken – insbesondere dort, wo weder eine Hochrisiko-Klassifizierung noch ein individuelles Betroffenenrecht eindeutig greifen.

Doch viele dieser Herausforderungen verdichten sich letztlich in einem zentralen Rechtsinstrument der DSGVO: Artikel 22. Seine konkrete Anwendung und Auslegung entscheiden darüber, ob ADM-Systeme fair, nachvollziehbar und rechtlich kontrollierbar eingesetzt werden können. Das folgende Unterkapitel illustriert anhand eines realitätsnahen Szenarios, wie sich diese Schutzlücke im Alltag auswirkt – und was sich durch die jüngste EuGH-Rechtsprechung ändern könnte.

3.4 Zwischen Schutzpflicht und Praxislücke – Artikel 22 DSGVO im Wandel

Stellen wir uns folgendes Szenario vor: Amira, eine österreichische Staatsbürgerin mit bosnisch-türkischem Hintergrund, beantragt einen Kredit bei Bank A. Diese nutzt ein KI-gestütztes Profiling-System, das auf Basis von Amiras finanzieller Vorgeschichte, Online-Verhalten und weiteren personenbezogenen Daten automatisch einen niedrigen Score generiert – ohne dass menschliches Eingreifen erfolgt. Der Antrag wird abgelehnt. Auch Bank B verwendet ein ähnliches System, das – aufbauend auf den gleichen Datenmustern – erneut zur Ablehnung führt. Ohne transparente Erklärungen und rechtlich durchsetzbare Rechte bleibt Amira faktisch ausgeschlossen. Erst mit dem SCHUFA-Urteil des EuGH (2023a) eröffnet sich die Möglichkeit, gegen solche Entscheidungen rechtlich vorzugehen – insbesondere, wenn keine rechtssichere Einwilligung vorliegt oder die Datenverarbeitung über das Erforderliche hinausgeht. Doch wie realistisch ist es für eine Person in finanzieller Notlage, diese Rechte auch tatsächlich geltend zu machen?

4. Kontrafaktische Erklärungen als Transparenzinstrument

4.1 Grundprinzipien kontrafaktischer Erklärungen

Die Forderung nach Transparenz ist einer der zentralen Diskussionspunkte im Umgang mit KI-gestützten Entscheidungen – sowohl aus juristischer als auch ethischer Perspektive. Weder die DSGVO noch der AI Act garantieren allerdings ein technisches *Recht auf Erklärung* im engeren Sinne. Während Art. 22 DSGVO das Recht

auf menschliches Eingreifen vorsieht und Art. 13 AI Act von einer „*vermittelbaren Nachvollziehbarkeit*“ spricht, bleiben konkrete Anforderungen an die Verständlichkeit algorithmischer Prozesse unklar.

Ein zentraler Kritikpunkt liegt darin, dass viele KI-Modelle – insbesondere Deep-Learning-basierte ADM-Systeme – technisch intransparent („Black Box“) sind. Ihre innere Logik lässt sich für Betroffene, aber auch oft für Entwickler:innen, kaum rekonstruieren. Klassische Ansätze zur Modellinterpretation wie LIME oder SHAP setzen oft ein tiefes technisches Verständnis voraus und sind schwer in juristische Kontexte übertragbar.

Vor diesem Hintergrund gewinnen *kontrafaktische Erklärungen* zunehmend an Bedeutung. Sie zielen nicht darauf ab, das Modell selbst zu erklären, sondern geben eine Antwort auf die Frage: „*Was hätte sich ändern müssen, damit das Ergebnis anders ausgefallen wäre?*“ (Wachter et al., 2018).

Ein Beispiel: Eine Bewerberin wird von einem automatisierten System abgelehnt. Eine kontrafaktische Erklärung könnte lauten: „*Wenn Sie mindestens drei Jahre Führungserfahrung nachgewiesen hätten, wäre Ihre Bewerbung positiv bewertet worden.*“ Diese Aussage ist konkret, handlungsorientiert und setzt kein Wissen über neuronale Netzwerke oder mathematische Modellparameter voraus.

Der Vorteil kontrafaktischer Erklärungen liegt damit in ihrer *doppelten Anschlussfähigkeit*:

- *Für Betroffene*: Sie bieten konkrete Anhaltspunkte, wie künftige Entscheidungen positiv beeinflusst werden können – ohne dass sie die technische Funktionsweise verstehen müssen.
- *Für Verantwortliche*: Sie ermöglichen Transparenz, ohne Geschäftsgeheimnisse oder Modellarchitektur offenzulegen – ein Vorteil insbesondere bei proprietären Systemen oder GPAI-Komponenten (General Purpose AI), die häufig als Black Boxes integriert werden.

Im Sinne des AI Acts können kontrafaktische Erklärungen helfen, die Anforderungen aus Art. 13 („transparente Information“) und Art. 14 („menschliche Aufsicht“) *praxisnah zu operationalisieren*. Auch die DSGVO lässt Raum für diese Form der Erklärung, etwa im Rahmen von Informationspflichten (Art. 13 Abs. 2 lit. f DSGVO) oder als Bestandteil technischer und organisatorischer Maßnahmen (TOM) zur Wahrung der Betroffenenrechte (Art. 24 DSGVO).

Allerdings sind kontrafaktische Erklärungen nicht frei von Herausforderungen. Die Auswahl kontrafaktischer Variablen muss auf validen, nichtdiskriminierenden Daten beruhen – ansonsten besteht die Gefahr, dass sie bestehende Ungleichheiten reproduzieren oder inadäquate Erwartungen wecken (Wachter et al., 2018). Eine kontrafaktische Aussage wie „*Wenn Sie nicht schwanger gewesen wären, hätten Sie den Job bekommen*“ wäre offensichtlich unzulässig – nicht nur ethisch, sondern auch rechtlich.

Dieser Ansatz unterscheidet sich deutlich von klassischen Erklärungsmethoden im Sinne der „White Box“-Analyse. Kontrafaktische Erklärungen liefern *keine Begründung*, warum eine Entscheidung so gefallen ist, sondern eine *konkrete Bedingung*, unter der sie anders ausgefallen wäre. Dadurch lassen sich Handlungsspielräume

aufzeigen, die nicht nur aus Transparenzgründen, sondern auch zur Selbstbestimmung und Widerspruchsprüfung von Bedeutung sind.

In juristischer Hinsicht passen kontrafaktische Erklärungen gut zur DSGVO: Sie ermöglichen *transparente Kommunikation*, ohne dass die Geschäftslogik oder das Modell offengelegt werden muss – ein Aspekt, der besonders bei proprietären Systemen oder komplexen LLM-Architekturen relevant ist. Zudem erfüllen sie aus Sicht der Betroffenen ein wesentliches Ziel: *Verstehen, Anfechten und Verbessern*. Dieses Verständnis ist entscheidend, um Chancengleichheit und Vertrauen in KI-gestützte Prozesse zu sichern – vor allem in sensiblen Kontexten wie Bewerbungsverfahren oder internen Evaluationsprozessen.

4.2 Vorteile und Grenzen im Vergleich zu XAI

Erklärbare KI (Explainable AI, XAI) bezeichnet technologische Methoden, die darauf abzielen, die Entscheidungsfindung von KI-Modellen für Menschen nachvollziehbar zu machen. Dazu zählen u. a. Feature-Attributionsverfahren wie *SHAP (SHapley Additive exPlanations)* oder *LIME (Local Interpretable Model-agnostic Explanations)*, die beispielsweise aufzeigen, welche Eingabefaktoren den größten Einfluss auf das Ergebnis hatten. Ziel ist es, auch bei komplexen Modellen wie neuronalen Netzen eine gewisse Transparenz herzustellen – etwa für Entwickler:innen, Regulator:innen oder Auditor:innen.

Diese Verfahren stoßen jedoch im arbeitsrechtlichen und datenschutzrechtlichen Kontext an Grenzen: Sie sind technisch anspruchsvoll, teilweise schwer reproduzierbar und für Laien kaum verständlich. Zudem bergen sie das Risiko, *vermeintliche Erklärungen zu liefern, die nicht rechtsverbindlich interpretierbar sind* – insbesondere, wenn sie auf Wahrscheinlichkeiten oder Korrelationen beruhen.

In der Diskussion um erklärbare KI stellt sich zunehmend die Frage, wie Diskriminierungsfreiheit nachgewiesen werden kann. Die EU-Antidiskriminierungsrichtlinien sehen bei entsprechender Indizienlage eine Beweislastumkehr vor – d. h. das betroffene Unternehmen oder der verantwortliche Systembetreiber muss aktiv belegen, dass keine diskriminierende Wirkung vorliegt (Hense & Mustać, 2024, S. 204). Dies erhöht die Relevanz kontrafaktischer Erklärungen: Sie erlauben es, transparent darzulegen, wie eine Entscheidung auch unter veränderten Bedingungen (z. B. bei anderem Geschlecht oder Wohnort) ausgefallen wäre. Solche Verfahren helfen dabei, algorithmische Fairness greifbar zu machen, ohne die komplexe Modelllogik vollständig offenzulegen.

Kontrafaktische Erklärungen stellen hier eine niedrigschwellige und zugleich wirksame Alternative dar. Sie liefern keine abstrakte Funktionsbeschreibung, sondern eine direkte Antwort auf die individuelle Entscheidung: *Was hätte ich ändern müssen, um ein anderes Ergebnis zu erhalten?* Dies entspricht der intuitiven Vorstellung von Fairness und adressiert zugleich Informationsrechte, ohne proprietäre Modelle offenzulegen.

4.3 Praxisbeispiele und Implementierungsempfehlungen

Unternehmen, die ADM-Systeme einsetzen, sollten kontrafaktische Erklärungen daher nicht isoliert betrachten, sondern als Bestandteil eines umfassenden Governance-Modells verstehen. Dies umfasst:

- eine *klare Dokumentation* der Entscheidungslogik und Einflussfaktoren,
- ein *Monitoring auf diskriminierende Effekte* sowie
- die *Verknüpfung mit Widerspruchsverfahren* und menschlicher Überprüfung.

Erste Best Practices zeigen, dass kontrafaktische Erklärungen insbesondere im Recruiting und bei Performance-Bewertungen *erfolgreich implementiert werden können*, etwa durch regelbasierte Schwellenwerte oder semantisch generierte Textbausteine. Unternehmen wie LinkedIn und Workday haben begonnen, erste Formen nutzungsfreundlicher Entscheidungsbegründungen zu entwickeln, die auf kontrafaktischer Logik basieren – ein Trend, der regulatorisch künftig an Bedeutung gewinnen dürfte.

5. Partizipation und ethische Verantwortung im ADM-Einsatz

Der Einsatz automatisierter Entscheidungsverfahren in der Arbeitswelt wirft nicht nur technische und rechtliche, sondern auch tiefgreifende *ethische Fragen* auf. Besonders im Verhältnis zwischen Arbeitgeber:innen und Mitarbeitenden stellt sich die Frage, wie fair, nachvollziehbar und partizipativ ADM-Systeme ausgestaltet sein müssen, um Vertrauen und Akzeptanz zu schaffen. Technisch korrekte, aber sozial intransparente Systeme bergen das Risiko einer *digitalen Entmündigung*: Mitarbeitende erleben Entscheidungen als willkürlich, nicht beeinflussbar und schwer anfechtbar – selbst, wenn sie formal korrekt getroffen wurden.

5.1 Ethische Mindeststandards: Fairness, Verantwortung, Mensch-in-der-Schleife

Automatisierte Entscheidungen greifen in Rechte, Chancen und das Selbstverständnis von Menschen ein – etwa bei der Besetzung von Stellen, der Verteilung von Boni oder der Kündigung. Solche Eingriffe erfordern nicht nur datenschutzrechtliche Legitimierung, sondern auch eine *ethisch verantwortliche Ausgestaltung*. Das Prinzip der *Fairness* – etwa im Sinne gleichwertiger Chancen und diskriminierungsfreier Bewertung – ist dabei ebenso zentral wie die *Verantwortlichkeit* (Accountability) derjenigen, die das ADM-System einsetzen oder konfigurieren.

Viele ADM-Prozesse sind so gestaltet, dass Menschen lediglich nachgelagert bestätigen, was das System vorgibt – ohne tiefes Verständnis für die Funktionsweise oder Auswirkungen. Der oft zitierte „Human in the Loop“ verkommt so zum „Human on the Sideline“. Eine echte ethische Kontrolle erfordert dagegen eine bewusste, informierte und überprüfbare menschliche Rolle – nicht nur zur Abwendung rechtlicher Risiken, sondern als Ausdruck verantwortungsvoller Unternehmensführung.

5.2 Informationsrechte und Mitwirkung gemäß AI Act

Mit dem AI Act wird erstmals auf regulatorischer Ebene explizit anerkannt, dass auch Arbeitnehmer:innen im ADM-Kontext ein Informationsrecht haben. Gemäß Art. 29 AI Act müssen Unternehmen ihre Mitarbeitenden vor dem Einsatz eines Hochrisiko-KI-Systems im Arbeitsumfeld informieren, sofern eine signifikante Auswirkung auf deren Situation zu erwarten ist. Diese Vorschrift wird durch Erwägungsgrund 92 konkretisiert, der ausdrücklich betont, dass betroffene Personen nicht nur über den Einsatz, sondern auch *über die Zwecke, Abläufe und möglichen Auswirkungen der Entscheidungssysteme informiert* werden müssen – und zwar in klarer, verständlicher Sprache. Eine *interne KI-Guideline* kann dabei ein wirksames Mittel sein, um diese Informationen strukturiert und nachvollziehbar bereitzustellen.

Diese Informationspflicht geht über klassische Datenschutztransparenz hinaus: Sie zielt auf *Verständnis, Einbeziehung und reflexive Kontrolle* ab. Unternehmen sind somit angehalten, nicht nur regulatorisch zu dokumentieren, sondern auch kommunikativ zu befähigen – etwa durch interne Schulungen, Dialogformate oder durch Betriebsvereinbarungen, die den KI-Einsatz flankieren.

5.3 Beteiligung in der Praxis: Betriebsräte, Ethikboards und Audits

Ein zentraler Hebel zur Förderung partizipativer ADM-Governance liegt in der *Einbindung kollektiver Akteure* – insbesondere von Betriebsräten, aber auch von Ethikkommissionen oder interdisziplinären KI-Ausschüssen. In vielen europäischen Ländern ist der Einsatz neuer Technologien mitbestimmungspflichtig – eine Chance, den Einsatz algorithmischer Systeme *nicht nur zu kontrollieren, sondern aktiv mitzugestalten*.

Zudem bieten *interne Ethikboards* eine Möglichkeit, nicht nur gesetzliche Mindeststandards, sondern auch *Werteorientierung* und *organisationales Lernen* zu fördern. Unternehmen, die ADM-Systeme evaluieren, sollten systematisch dokumentieren:

- Welche Entscheidungen werden (teil)automatisiert?
- Wer trägt Verantwortung für Fehler oder Diskriminierung?
- Welche Feedbackmechanismen stehen Betroffenen offen?

Auch die Durchführung einer *Fundamental Rights Impact Assessment (FRIA)* – wie sie der EU AI Act in Art. 29 i. V. m. Erwägungsgrund 92 und 93 für Hochrisiko-KI-Systeme vorsieht – gilt zunehmend als zentrale Maßnahme, um Grundrechtsrisiken strukturiert zu bewerten. Diese Bewertung soll gewährleisten, dass der Einsatz solcher Systeme keine unverhältnismäßigen Beeinträchtigungen für betroffene Gruppen verursacht. Die FRIA sollte technische, soziale und ethische Risiken integrativ erfassen – idealerweise unter frühzeitiger Einbindung der betroffenen Mitarbeitenden sowie unter Berücksichtigung von Aspekten wie Transparenz, Aufsicht und diskriminierungssensibler Gestaltung.

5.4 Der kulturelle Wandel: von Akzeptanz zu aktiver Mitgestaltung

Partizipation darf nicht nur als Compliance-Auflage verstanden werden. Vielmehr bietet sie die Chance, algorithmische Systeme *sozial verträglich und nachhaltig zu gestalten*. Mitarbeitende sind keine passiven Objekte von KI-Entscheidungen, sondern können wertvolle Perspektiven einbringen – etwa zur Einschätzung von Fairness, Kontextrelevanz oder versteckten Biases. Unternehmen, die diesen Dialog aktiv fördern, stärken nicht nur das Vertrauen, sondern auch die langfristige Legitimation ihrer KI-Strategien.

ADM-Systeme im Arbeitskontext erfordern daher eine doppelte Verantwortlichkeit:

- *rechtlich*, im Sinne des Grundrechtsschutzes und der DSGVO-/AI-Act-Konformität,
- *ethisch*, im Sinne partizipativer, diskriminierungsfreier und erklärbarer Gestaltung.

6. Fazit und Ausblick

Die neue KI-Verordnung ist der neue Ally sowohl für marginalisierte Gruppen als auch für den Datenschutz. Doch der Schutzanspruch muss auch operativ eingelöst werden – durch rechtliche Klarheit, technische Nachvollziehbarkeit und organisatorische Verantwortung. Automatisierte Entscheidungsverfahren haben das Potenzial, Arbeitsprozesse effizienter, skalierbarer und objektiver zu gestalten. Gleichzeitig bergen sie erhebliche Risiken: algorithmische Diskriminierung, Intransparenz, mangelnde Kontrollierbarkeit und regulatorische Grauzonen. Besonders im arbeitsweltlichen Kontext können KI-gestützte Systeme tiefgreifende Auswirkungen auf berufliche Chancen, Bewertung und Teilhabe haben. Diese Entwicklungen erfordern einen integrativen Blick auf technologische, rechtliche und ethische Aspekte – jenseits reiner Effizienzlogik.

Dieser Beitrag hat gezeigt, dass Artikel 22 DSGVO ein wichtiges, aber in der Praxis unzureichend konkretisiertes Instrument darstellt. Die EuGH-Entscheidung im Fall SCHUFA I (C-634/21) hat deutlich gemacht, dass bereits vorbereitende Scoring-Verfahren unter bestimmten Bedingungen als vollwertige automatisierte Entscheidungen im Sinne von Artikel 22 gelten können – insbesondere dann, wenn sie faktisch entscheidungsleitend wirken (EuGH, 2023a). Das zeitgleich veröffentlichte Urteil SCHUFA II (verbundene Rechtssachen C-26/22 und C-64/22) verdeutlicht darüber hinaus, dass auch die herkömmlich genutzten Rechtsgrundlagen gemäß Art. 6 Abs. 1 lit. a (Einwilligung) und lit. f (berechtigtes Interesse) nicht uneingeschränkt tragfähig sind – insbesondere dann nicht, wenn Betroffene über die Tragweite und Zwecke algorithmischer Prozesse nicht ausreichend informiert werden (EuGH, 2023b). Gleichzeitig bringt der AI Act neue Impulse für systemische Kontrolle, etwa durch verpflichtende Risikobewertungen, Human Oversight und Transparenzanforderungen. Doch zentrale Begriffe wie „signifikante Beeinträchtigung“, „limited risk“ oder „general purpose AI“ bleiben unklar. ADM-Systeme mit realweltlichem Einfluss – wie Job-Ad-Targeting, Scoring-Engines oder performanzbasierte Personalbewertungen –

fallen häufig durch das Raster der Regulierung, weil sie formell nicht als Hochrisiko klassifiziert sind.

Um dieser Lücke zu begegnen, sollten Unternehmen *kontrafaktische Erklärungen* als zentrales Transparenzinstrument einsetzen. Sie bieten verständliche, handlungsorientierte Rückmeldungen an Betroffene – ohne in die algorithmische Black Box einzudringen. Damit können sowohl die Anforderungen aus dem AI Act (Art. 13, 14) als auch die Datenschutzgrundsätze der DSGVO (Transparenz, Fairness, Datenminimierung) in der Praxis besser erfüllt werden.

Handlungsempfehlungen

Für eine faire und rechtskonforme Nutzung von ADM-Systemen in der Arbeitswelt ergeben sich folgende konkrete Empfehlungen:

1. *Business Impact Assessments erweitern*

Unternehmen sollten bei der Einführung oder Weiterentwicklung von ADM-Systemen systematisch prüfen, welche datenschutz- und KI-rechtlichen Auswirkungen entstehen – unter Einbeziehung beider SCHUFA-Urteile (EuGH, 2023a, b). Besonders kritisch ist die Rechtsgrundlage der Verarbeitung: Die häufig verwendete Berufung auf Einwilligung (Art. 6 Abs. 1 lit. a DSGVO) oder berechtigtes Interesse (Art. 6 Abs. 1 lit. f DSGVO) ist bei automatisierten Entscheidungssystemen oft nicht tragfähig – insbesondere dann nicht, wenn für die betroffene Person erhebliche Wirkungen im Sinne von Art. 22 DSGVO entstehen. In solchen Fällen sollten Unternehmen eine datenschutzrechtliche Folgenabschätzung (Art. 35 DSGVO) mit einem erweiterten Business Impact Assessment kombinieren, das auch grundrechtliche und diskriminierungsbezogene Risiken einbezieht.

2. *Kontrafaktische Erklärungen standardisieren*

ADM-Systeme sollten technisch so gestaltet werden, dass sie automatisch kontrafaktische Erklärungen generieren können – insbesondere in sensiblen Anwendungskontexten wie Bewerbungsverfahren, internen Beförderungsprozessen oder automatisierten Leistungsbewertungen. Diese Erklärungen liefern konkrete Antworten auf die Frage „Was hätte anders sein müssen, damit die Entscheidung anders ausfällt?“ und ermöglichen damit eine verständliche, praxistaugliche Form der Transparenz. Sie tragen nicht nur zur Erfüllung der Informationspflichten aus Art. 13 Abs. 2 lit. f DSGVO bei, sondern helfen auch, die Anforderungen des Art. 29 und 52 AI Act zu operationalisieren – ohne dabei Geschäftsgeheimnisse offenzulegen. Ihre Integration in ADM-Systeme sollte daher als Standardbestandteil verantwortungsvoller KI-Governance betrachtet werden.

3. *Partizipative Governance etablieren*

Mitarbeitende, Betriebsräte und Ethikgremien sollten frühzeitig und strukturiert in die Entwicklung, Auswahl und Implementierung von ADM-Systemen eingebunden werden. Dies kann z. B. über transparente Entscheidungsprotokolle, Feedbackschleifen oder ethikorientierte Audits erfolgen. Der AI Act (Erwägungsgrund 71 und Art. 29) fordert ausdrücklich eine „angemessene Einbindung der betroffenen Personen“ – was im Arbeitskontext durch digitale Mitbestimmung, kontinuierliches Monitoring und Schulungsmaßnahmen konkretisiert werden sollte.

4. *GPAI-Komponenten aktiv einordnen*

Der Einsatz von General-Purpose-AI-Modellen (GPAI) – etwa in Form von LLMs für Scoring, Textanalyse oder Empfehlungssysteme – sollte nicht als technische Randnotiz behandelt werden. Auch wenn gemäß Art. 83 AI Act eine Übergangsfrist bis August 2027 besteht, sind Unternehmen bereits jetzt gefordert, Risiken, Mitbestimmungsrechte, Transparenzpflichten und Verantwortlichkeiten entlang der GPAI-Kette zu klären. Dies betrifft insbesondere das Zusammenspiel mit hochriskanten ADM-Systemen, bei denen GPAI-Komponenten integriert werden – etwa zur Vorselektion im Recruiting oder zur Verhaltensanalyse.

5. *Regulatorische Schnittstellen klären*

Zwischen DSGVO und AI Act bestehen inhaltliche und begriffliche Schnittmengen, aber auch Regelungslücken – etwa in der Definition automatisierter Entscheidungen oder in der Verantwortungsverteilung bei GPAI-Einsatz. Unternehmen, Datenschutzbeauftragte und Betriebsräte sollten daher aktiv an einer gemeinsamen Auslegungspraxis mitwirken, z.B. durch sektorübergreifende Audits, interne Policy Papers oder die Beteiligung an Verbandsinitiativen. Besonders im Arbeitskontext braucht es verbindliche Standards, um ADM-Systeme sowohl rechtssicher als auch ethisch legitimiert einsetzen zu können.

6. *Antidiskriminierungsnachweise integrieren*

Unternehmen sollten proaktiv dokumentieren, welche Designentscheidungen und Datenprozesse sie im ADM-System getroffen haben, um Diskriminierung zu vermeiden. Dazu zählen z.B. systematische Tests auf Gruppenbenachteiligung, der Einsatz kontrafaktischer Erklärungen sowie die klare Zuordnung von Verantwortlichkeiten innerhalb der AI-Lieferkette (Hense & Mustać, 2024, S. 203–205).

Die Zukunft der Arbeitswelt wird wesentlich davon abhängen, wie transparent, fair und partizipativ algorithmische Systeme gestaltet werden. ADM-Systeme sind kein Selbstzweck. Ihre Legitimation entsteht nicht durch Effizienz – sondern durch *Verantwortung, Rechenschaft und Einbindung*. Unternehmen, die dies erkennen, schaffen nicht nur regulatorische Sicherheit, sondern auch Vertrauen – und damit einen echten Wettbewerbsvorteil in einer zunehmend automatisierten Arbeitswelt.

Literatur

- Barberá, I. (2025). *AI Privacy Risks & Mitigations Large Language Models (LLMs)*. European Data Protection Board. https://www.edpb.europa.eu/our-work-tools/our-documents/support-pool-experts-projects/ai-privacy-risks-mitigations-large_en
- Europäischer Gerichtshof (EuGH). (2023a). *ECLI:EU:C:2023:957*. <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:62021CJ0634>
- Europäischer Gerichtshof (EuGH). (2023b). *ECLI:EU:C:2023:958*. <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A62022CA0026>
- European Parliament & Council of the European Union. (2024). *Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139*

- and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act). Official Journal of the European Union. <http://data.europa.eu/eli/reg/2024/1689/oj>
- Hense, P. & Mustać, T. (2024). *AI Act compact: Compliance, management & use cases in corporate practice*. Deutscher Fachverlag.
- Köver, C. (2024, 5. Januar). *Diskriminierung: AMS erntet Hohn mit neuem KI-Chatbot*. netzpolitik.org. <https://netzpolitik.org/2024/diskriminierung-ams-erntet-hohn-mit-neuem-ki-chatbot/>
- Pery, A. & Simon, M. (2024). *GDPR Curbs Use of AI Via Article 22 Decision*. Bloomberg Law. <https://www.bloomberglaw.com/external/document/X4BBTPFO000000/international-data-privacy-compliance-professional-perspective-g>
- Tribunale di Bologna. (2020). *Ordinanza del 31 dicembre 2020, n. 29416/2020 RG* [Ruling on algorithmic discrimination by Deliveroo]. Bologna, Italy.
- Wachter, S., Mittelstadt, B. & Russell, C. (2018). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law & Technology*, 31(2), 841–887.

Teil III:
Impulse zur Gestaltung in konkreten Arbeitsfeldern

Strategische Vorausschau und Technikfolgenabschätzung

Cybersicherheit in der KI-gestützten Arbeitswelt proaktiv gestalten

Titus Udrea & Helene Baumgartner

1. Einleitung: KI als doppelte Herausforderung in der Arbeitswelt

Die zunehmende Integration künstlicher Intelligenz (KI) in die Arbeitswelt verändert nicht nur Beschäftigungsstrukturen und Entscheidungsprozesse, sondern auch die Sicherheitsarchitektur innerhalb von Organisationen. KI ist dabei längst nicht mehr nur ein Werkzeug zur Automatisierung, sondern entwickelt sich zu einer strukturellen Kraft, die tief in die Organisation, das Management und die Sicherung von Arbeit eingreift. KI-Systeme, die beispielsweise im algorithmischen Management oder in der prädiktiven Personalplanung eingesetzt werden, versprechen Effizienzgewinne – bringen jedoch zugleich weitreichende neue Risiken mit sich.

Eine der zentralen Herausforderungen in diesem Kontext betrifft die Cybersicherheit. KI ermöglicht neue Angriffsformen, die skalierbar, zielgerichtet und zunehmend autonom sind – etwa in Form von Deepfakes, Social Engineering oder durch die gezielte Manipulation von KI-Systemen. Gleichzeitig trägt KI zur Entwicklung fortschrittlicher Verteidigungsmechanismen bei, etwa durch automatisierte Bedrohungserkennung oder adaptive Reaktionssysteme. Diese doppelte Rolle macht Cybersicherheit zu einer Schlüsselfrage in KI-gestützten Arbeitsumgebungen – mit Implikationen für Vertrauen, Autonomie und Rechenschaft.

Viele bestehende Governance-Ansätze im Bereich KI und Cybersicherheit bleiben bislang reaktiv und stark technikzentriert. Sie greifen zu kurz, wenn es darum geht, mit der dynamischen Entwicklung Schritt zu halten oder langfristige, soziotechnische Auswirkungen auf Beschäftigte und Organisationen systematisch zu berücksichtigen. Insbesondere die Perspektiven von Arbeitnehmer*innen – also derjenigen, die direkt von KI-gestützten Entscheidungen betroffen sind – finden in Gestaltung, Implementierung und Kontrolle solcher Systeme bisher zu wenig Berücksichtigung.

Vor diesem Hintergrund argumentiert der Beitrag, dass strategische Vorausschau (Foresight) und Technikfolgenabschätzung (TA) zentrale Instrumente einer proaktiven, inklusiven und zukunftsorientierten Governance von KI am Arbeitsplatz sind. Durch die systematische Beobachtung schwacher Signale, die Entwicklung alternativer Zukunftsszenarien und die Bewertung gesellschaftlicher Risiken ermöglichen diese Ansätze ein frühzeitiges Erkennen und Gestalten technologischer Entwicklungen. Darüber hinaus schaffen sie eine Grundlage, um technische und organisatorische Sicherheitsstrategien zu integrieren, partizipative Entscheidungsprozesse zu fördern und die Entwicklung von KI an Werten wie Transparenz, Rechenschaft und digitaler Souveränität auszurichten.

Das Kapitel verfolgt zwei analytische Stränge: Zum einen werden neue Bedrohungsszenarien im Bereich Cybersicherheit in KI-gestützten Arbeitsumgebungen analysiert. Zum anderen wird untersucht, inwiefern Foresight und TA zur Stärkung einer resilienten, sozial verantwortlichen KI-Governance beitragen können.

2. Cybersicherheit in der KI-gestützten Arbeitswelt: Transformationen und Risiken

Die Digitalisierung der Arbeitswelt und der zunehmende Einsatz der KI haben die Art und Weise, wie wir arbeiten und interagieren, grundlegend verändert. KI automatisiert repetitive Aufgaben, liefert bessere Entscheidungsgrundlagen, birgt aber neue Herausforderungen für die Informationssicherheit. Sie erweist sich als Technologie mit ambivalenten Auswirkungen, die sowohl von Cyberkriminellen für Angriffe genutzt wird als auch das Potenzial besitzt, die Cyberabwehr erheblich zu stärken (Gupta et al., 2023).

Dieses Kapitel analysiert die neuen Gefahren, die durch KI-basierte Angriffe entstehen, beleuchtet die Angriffsvektoren gegen KI-Systeme selbst und betrachtet die transformative Rolle von KI in der Cybersicherheit. Dabei wird diskutiert, wie KI-gestützte Systeme dazu beitragen können, eine sichere Arbeitsumgebung zu gewährleisten. Ein besonderer Fokus liegt dabei auf der Rolle des Menschen, den nötigen Kompetenzen und den Herausforderungen, z.B. durch die begrenzte Verfügbarkeit von Personen mit Expert*innenwissen.

2.1 KI-basierte Cyberangriffe: neue Dimensionen der Bedrohung

Cyberkriminelle nutzen zunehmend KI-Technologien, um ihre Angriffe zu automatisieren, zu skalieren und intelligenter zu gestalten. Die Automatisierung ermöglicht einen gleichzeitigen Angriff auf Ziele und erschwert dadurch deren frühzeitige Erkennung. Intelligente, auf KI-Algorithmen basierende Schadsoftware kann sich dynamisch an veränderte Umgebungen anpassen und ihre Angriffsstrategie optimieren. KI-Technologie ermöglicht es somit, die Effizienz und Reichweite von Angriffen zu steigern sowie deren Komplexität zu erhöhen. Social Engineering und Deepfakes werden immer überzeugender, wodurch es schwieriger wird, betrügerische Nachrichten und Identitäten zu erkennen. Mit KI ist es möglich, riesige Mengen an gefälschten Social-Media-Inhalten zu generieren, um gezielt die öffentliche Meinung zu manipulieren.

Um einen Überblick über die möglichen spezifischen Phasen eines Cyberangriffs zu geben, in denen KI eingesetzt wird und in welcher Form, ist das Cyber Kill Chain Model einer der Standards für eine strukturierte Betrachtung (Hutchins et al., 2011; Kaloudi & Li, 2020).

- Aufklärung (Reconnaissance): Eine automatisierte Informationsbeschaffung und Analyse von Mustern und Schwachstellen ermöglicht eine gezieltere Vorgangsweise und somit eine gesteigerte Effektivität.
- Bewaffnung (Weaponization): KI-Einsatz bei der Entwicklung von Angriffswerkzeugen führt potenziell zu neuartiger und schwerer erkennbarer Schadsoftware, die herkömmliche Sicherheitsmaßnahmen umgeht.
- Zustellung (Delivery): KI kann zur Optimierung von Zustellmethoden eingesetzt werden (z. B. überzeugendere Phishing-E-Mails), um die Erfolgsrate zu erhöhen.

- Ausnutzung (Exploitation): Die höhere Effizienz bei der Schwachstellensuche und -ausnutzung führt zu einer verkürzten Reaktionszeit für Sicherheitsmaßnahmen.
- Installation: Die Möglichkeit zur unauffälligen Installation und selbstpropagierenden Verbreitung erhöht die Gefahr von großflächigen Infektionen.
- Command and Control (C2): KI kann C2-Verbindungen tarnen und deren Erkennung erschweren, was die Aufrechterhaltung der Bedrohung ermöglicht.
- Aktionen auf Ziele (Actions on Objectives): Eine effizientere Datenexfiltration (Selektion relevanter Daten und Transfertarnung) sowie optimierte Ransomware-Angriffe (beschleunigte Verschlüsselung und potenziell adaptive Erpressungsstrategien) werden durch KI ermöglicht. KI kann in komplexen, mehrstufigen Angriffen die Koordination und Automatisierung einzelner Schritte verbessern. Eine daraus resultierende beschleunigte und effektivere Zielerreichung führt zu einem höheren potenziellen Schaden.

Über die Betrachtung einzelner Phasen der Cyber Kill Chain hinaus hat KI weitere übergreifende Auswirkungen auf Angriffsszenarien. KI senkt die Einstiegshürden und ermöglicht auch Laien, komplexe Angriffe durchzuführen. Werkzeuge wie WormGPT oder FraudGPT verdeutlichen die Rolle von KI als „Arbeitsunterstützung“ für Angreifer, indem sie Schadcodegenerierung und Automatisierung von Phishing-Kampagnen erleichtert oder generell Routinetätigkeiten automatisiert. Diese Erweiterung der potenziellen Akteure und ihrer Fähigkeiten verschärft die Problematik zusätzlich angesichts der Schaffung neuer Angriffsflächen durch die wachsende Vernetzung von IoT-Geräten, deren Schwachstellen ebenfalls mit Hilfe von KI ausgenutzt werden können (Ekelhart et al., 2025).

Die Auswirkungen des KI-Einsatzes bei Cyberangriffen lassen sich anhand konkreter Zahlen belegen. Beispielsweise verkürzt sich die benötigte Zeit für einen erfolgreichen Angriff von Stunden auf den Minutenbereich. Diese immense Beschleunigung ist der Fähigkeit von KI zuzuschreiben, komplexe Angriffsketten zu automatisieren, von Verteidigungsmaßnahmen zu lernen und Taktiken dynamisch anzupassen (Jada & Mayayise, 2024). Gleichzeitig führt KI zu einer deutlichen Kostensenkung bei der Durchführung von Cyberangriffen. Daraus wiederum resultiert ein erheblich höherer ROI für die Angreifer, was wiederum positiv mit dem Angriffsvolumen korreliert. Dies unterstreicht die Skalierbarkeit von KI-gestützten Angriffen und führt zu einer enormen Belastung der Cybersicherheitsinfrastruktur (Elgan, 2025; NCSC, 2024). Zusammenfassend ermöglichen diese Entwicklungen Cyberkriminellen, Angriffe deutlich schneller und kostengünstiger durchzuführen und gleichzeitig die Anzahl der Angriffe signifikant zu erhöhen, was die Bedrohungslage insgesamt verschärft.

Zukünftige Bedrohungen könnten durch KI-gegen-KI-Angriffe geprägt sein. Auch die fehlende ethische Barriere bei KI oder unbeabsichtigte Fehlfunktionen bergen Eskalationspotenzial (King et al., 2020). Diese Entwicklungen erfordern eine umfassende Neubewertung der aktuellen Sicherheitsstrategien und eine proaktive Anpassung an die sich wandelnde Bedrohungslandschaft.

2.2 KI als Zielscheibe: Systematik der Angriffsarten

KI-Systeme können jedoch auch selbst zum Angriffsziel werden. OWASP¹ identifiziert seit 2023 jährlich die zentralen Sicherheitsrisiken für KI-Systeme (OWASP, 2024), darunter Gefahren wie die Offenlegung sensibler Informationen, die Verbreitung von Fehlinformationen sowie sogenannte „Prompt Injections“, bei denen bösartige Eingaben das Verhalten von KI-Systemen insofern manipulieren, dass eine Ausgabe unerwünschter oder schädlicher Inhalte erfolgt.

In Ergänzung zu den Bemühungen von Open-Source-Gemeinschaften wie OWASP widmen sich auch staatliche Institutionen der Identifizierung und Systematisierung von Sicherheitsrisiken im Bereich der KI. Eine maßgebliche Organisation in diesem Kontext ist das NIST². Ähnlich wie OWASP bietet auch NIST (Vassilev et al., 2024) einen Rahmen und eine Systematik, um die komplexen Sicherheitsherausforderungen im Bereich der KI zu adressieren. Vier Hauptkategorien von Angriffen werden definiert: Evasion Attacks (Umgehungsangriffe), bei denen versucht wird, die KI durch manipulierte Eingaben zu täuschen; Poisoning Attacks (Vergiftungsangriffe), die darauf abzielen, die Trainingsdaten zu verfälschen, um das Modellverhalten zu beeinflussen; Privacy Attacks (Angriffe auf die Privatsphäre), die darauf abzielen, sensible Informationen aus dem Modell oder den Trainingsdaten zu extrahieren; und Abuse Attacks (Missbrauchsangriffe), bei denen KI-Systeme dazu gebracht werden sollen, schädliches oder unerwünschtes Verhalten zu zeigen (NIST, 2024). Diese Kategorisierung bietet eine systematische Grundlage zur Entwicklung von Schutzmaßnahmen gegen KI-spezifische Angriffe. NIST (2024) betont zudem die Bedeutung eines umfassenden Risikomanagements in seinem Framework für Cybersecurity (CSF 2.0).

2.3 KI zur Stärkung der Cyberabwehr: Potenzial und Grenzen

Nachdem wir die potenziellen negativen Auswirkungen von KI auf Angriffsszenarien beleuchtet haben, wenden wir uns nun den defensiven Anwendungen von KI im Cyberspace zu.

2.3.1 Potenzial von KI in der Cyberabwehr

Der Einsatz von KI in der Cybersicherheit bietet eine Reihe von Vorteilen gegenüber traditionellen Sicherheitsansätzen (Ekelhart et al., 2025). Ein fundamentaler Fortschritt besteht in der Kapazität KI-gestützter Systeme, nahezu in Echtzeit jede einzelne Netzwerkaktivität zu analysieren und zu korrelieren (Ramakrishnan & Chittibala, 2024). Dies ermöglicht eine verbesserte und proaktive Bedrohungserkennung, da suspekt Aktivitäten nicht sequenziell oder ereignisbasiert, sondern umfassend und nahezu simultan zu ihrem Auftreten detektiert und im besten Fall neutralisiert werden können.

1 Open Worldwide Application Security Project, <https://owasp.org/>

2 National Institute of Standards and Technology (NIST) der Vereinigten Staaten, <https://www.nist.gov/>

KI ermöglicht eine präzisere Bedrohungsanalyse, die über die reine Identifizierung bekannter Schadsoftware hinausgeht. Hochentwickelte Algorithmen erkennen neuartige Angriffe (Zero-Day-Angriffe) durch detaillierte Anomalie- und Verhaltensanalysen und können zur Erweiterung des Lagebewusstseins Informationen aus öffentlich zugänglichen Quellen in ihre Analysen integrieren. Diese fortgeschrittene Analysefähigkeit wird durch die Skalierbarkeit von KI-Lösungen unterstützt, die eine effiziente Verarbeitung großer Datenmengen und eine flexible Anpassung der Analysekapazitäten erlaubt.

Eng mit dieser Analysefähigkeit verbunden ist das verbesserte Risikomanagement. Nicht nur die Identifizierung, auch die Bewertung von Schwachstellen sowie die proaktive Suche nach Bedrohungen können durch KI unterstützt werden (Binhammad et al., 2024). Ein KI-Einsatz in der Modellierung von Angriffspfaden und eine automatisierte Analyse potenzieller Auswirkungen (Business-Impact-Analyse) ermöglicht eine fundiertere Einschätzung und Priorisierung von Risiken.

Ein weiterer wesentlicher Vorteil ergibt sich aus der gesteigerten Effizienz durch Automatisierung. Menschliche Expert*innen können bei Routineaufgaben wie Schwachstellenscans oder der Analyse von Sicherheitswarnungen entlastet werden (Binhammad et al., 2024). Gleichzeitig ermöglicht KI die Automatisierung der Reaktion auf bestimmte Bedrohungen, was wiederum die Geschwindigkeit und Effektivität der Angriffsabwehr positiv beeinflusst.

Die adaptive Natur und das kontinuierliche Lernen von KI-Systemen sind wichtige Faktoren, um mit der dynamischen Bedrohungslandschaft Schritt zu halten. Durch die ständige Analyse neuer Daten und Angriffsvektoren und dem Lernen aus vergangenen Ereignissen können sich KI-gestützte Sicherheitsmechanismen kontinuierlich anpassen (Markevych & Dawson, 2023).

Neben diesen generellen Vorteilen führt KI zu konkreten Verbesserungen in spezifischen Bereichen der Cyberabwehr und ermöglicht eine effektivere Nutzersensibilisierung. Konkret betrifft dies verbesserte Zugriffskontrollen durch fortschrittliche biometrische Authentifizierung (Binhammad et al., 2024) sowie die Unterstützung spezialisierter Aufgaben wie Penetrationstests und die Erkennung von Deepfakes durch intelligente Inhaltsanalyse. Zudem ermöglicht KI personalisierte Sicherheits Schulungen und Awareness-Trainings, indem sie individuelle Lernmuster analysiert und maßgeschneiderte Inhalte bereitstellt, um die Sensibilität der Nutzer für Cybersicherheitsrisiken effektiv zu steigern.

2.3.2 Grenzen und Herausforderungen von KI in der Cyberabwehr

Trotz des erheblichen Potenzials von KI in der Cyberabwehr bestehen auch signifikante Grenzen und Herausforderungen wie die Abhängigkeit von Qualität und Repräsentativität der Trainingsdaten, die zu ungenauen Ergebnissen führen kann, sowie die mangelnde Transparenz und Erklärbarkeit vieler KI-Modelle (Ekelhart et al., 2025). Dazu kommen mögliche Inkonsistenzen und potenzielle Fehlinterpretationen basierend auf Phänomenen wie Halluzinationen oder anderen Effekten, bei denen die Relation zwischen der Analyse und der automatisierten Handlung durch ein KI-System nicht stimmt. Phänomene wie das sogenannte „Overthinking“ können dazu führen, dass ein KI-System durch exzessive Überlegungen wertvolle Reaktionszeit auf

Bedrohungen verliert, nicht zielführende oder sogar schädliche Handlungen setzt oder die Problemlösung vorschnell abbricht (Cuadron et al., 2025). Ähnlich problematisch ist das „Underthinking“, bei dem – wie Wang et al. (2025) zeigen – vielversprechende Denkpfade zu früh verlassen werden.

Der hohe Ressourcenbedarf an Rechenleistung und spezialisierter Expertise sowie die nötige Integration in bestehende Sicherheitsinfrastrukturen sind weitere nicht zu unterschätzende Hürden, die sich mitunter als komplex und kostspielig erweisen. Dazu kommt das latente Potenzial für Überautomatisierung mit dem Risiko, dass vollautomatisierte Systeme legitime Aktivitäten wie z. B. kritische Software-Updates blockieren oder dass bestehende KI-Tools zur Bedrohungserkennung umfunktioniert werden, um Angriffe zu optimieren. Ein weiterer kritischer Aspekt ist der potenzielle Verlust fundamentaler Analyse- und Reaktionsfähigkeiten menschlicher Expert*innen durch eine zu starke Abhängigkeit von automatisierten Systemen.

2.3.3 Mensch-KI-Zusammenarbeit und Auswirkungen auf die Beschäftigten

Die Integration von KI in die Cybersicherheit revolutioniert nicht nur die Art und Weise, wie wir Cyberangriffe abwehren, sondern verändert auch grundlegend die Anforderungen an die Arbeitskräfte in diesem Feld (Binhammad et al., 2024).

Angesichts der Stärken und Grenzen von KI in der Cybersicherheit ist eine effektive Zusammenarbeit zwischen Mensch und Maschine bedeutend. KI-Systeme überzeugen durch ihre Fähigkeit zur schnellen Datenverarbeitung und Mustererkennung, während menschliche Fachkräfte Kontextwissen, Urteilsvermögen und Sensibilität für neuartige Situationen einbringen – eine Kombination, die robuste Sicherheitslösungen ermöglicht. Im Trainingsprozess einer KI liefern Expert*innen nicht nur die Daten, also die „Ground Truth“³, sondern verfeinern auch Modelle durch kontinuierliches Feedback. Es braucht also die menschlichen Expert*innen, um präzisere Ergebnisse zu erzielen (Cohen et al., 2025).

Die Entstehung neuer Rollen wie KI-Sicherheitsanalyst*innen und Spezialist*innen für maschinelles Lernen in der Cybersicherheit ist eine direkte Folge dieser Entwicklung. Parallel dazu führt die wachsende Komplexität der Cyberbedrohungen, verbunden mit der Notwendigkeit einer kontinuierlichen Überwachung der zur Abwehr genutzten KI-Systeme und der ständigen Anpassung an neue Angriffsvektoren, zu einem höheren Bedarf an Fachkräften.

Der quantitativ steigende Bedarf korrespondiert mit erhöhten Anforderungen an die Qualifikationen sowohl bestehender als auch zukünftiger Cybersecurity-Expert*innen. Die rasante technologische Entwicklung erfordert eine kontinuierliche Erweiterung ihrer Kenntnisse und Fähigkeiten, insbesondere im Bereich der KI, wodurch das Berufsbild von IT-Sicherheitsmitarbeiter*innen zunehmend interdisziplinär geprägt wird.

Trotz des wachsenden Fachkräftebedarfs in der Cybersicherheit eröffnet die KI-Integration auch Chancen. Die Automatisierung von Routineaufgaben, beispielsweise die Auswertung von Logdateien oder die Erkennung bekannter Schwachstellen durch

3 Der Begriff „Ground Truth“ bezieht sich auf reale Daten, die für das Training und Testen von KI-Modellen verwendet werden.

KI, verändert die Arbeitsweise der Mitarbeitenden und erfordert neue Kompetenzen, birgt aber auch das Potenzial einer zukünftigen Entlastung des Fachkräftemarktes.

Neben diesen technologischen und arbeitsorganisatorischen Veränderungen gewinnt die umfassende Sensibilisierung und Schulung aller Mitarbeitenden in Bezug auf Cybersicherheit und die spezifischen Implikationen der KI zunehmend an Bedeutung.

In diesem Kontext spielt der von der ENISA⁴ herausgegebene ECSF (2022) eine wichtige Rolle bei der Standardisierung und Förderung von Cybersicherheitskompetenzen. Obwohl der aktuelle Rahmen noch keine expliziten KI-spezifischen Kompetenzen im Bereich der Cybersicherheit detailliert ausführt, wäre es wünschenswert, dass zukünftige Weiterentwicklungen diese berücksichtigen. Dies würde dazu beitragen, die für die Integration von KI-Technologien und die Reduzierung des Fachkräftemangels notwendigen Kompetenzen klarer zu definieren und deren Entwicklung zu fördern.

3. Strategische Vorausschau und Technikfolgenabschätzung für KI-Sicherheit

Wie bereits angedeutet, hat der Einsatz von KI bei Cyberangriffen zu einer Eskalation des Umfangs, der Geschwindigkeit und der Raffinesse digitaler Bedrohungen geführt. Der Einsatz von KI hat nachweislich die offensiven Cyberfähigkeiten durch Automatisierung, Umgehungstechniken und anpassungsfähiges Targeting verbessert. Darüber hinaus hat sich gezeigt, dass KI neue Angriffsflächen geschaffen hat, da KI-Systeme selbst zunehmend anfällig für Ausnutzung geworden sind (Malatji & Tolah, 2024).

In diesem Zusammenhang stoßen bestehende Cybersicherheits- und Governance-Strategien an ihre Grenzen. Das technische Arsenal ist zwar unverzichtbar, aber oft reaktiv und unzureichend in die breitere Organisationskultur und Politikgestaltung integriert. Es sind Ansätze erforderlich, die die dynamische Natur KI-gestützter Bedrohungen nicht nur in technologischer, sondern auch in strategischer, sozialer und institutioneller Hinsicht widerspiegeln. In diesem Abschnitt werden daher die strategische Vorausschau und die Technikfolgenabschätzung (TA) als wichtige Instrumente untersucht, die Organisationen dabei unterstützen, KI-Risiken proaktiv, partizipativ und zukunftsorientiert zu antizipieren, zu verstehen und zu regulieren.

3.1 Warum traditionelle KI- und Cybersicherheits-Governance-Ansätze nicht ausreichen

Mit der zunehmenden Integration von KI-Systemen in die Infrastruktur des Arbeitsplatzes – von der algorithmischen Verwaltung bis hin zur prädiktiven Personalanalyse – übersteigen die Cybersicherheitsbedrohungen oft die Reaktionsfähigkeit klassischer Governance-Ansätze. Herkömmliche Ansätze für Cybersicherheit und KI-Governance sind in der Regel reaktiv und eher darauf ausgerichtet, auf Vorfäl-

⁴ Agentur der Europäischen Union für Cybersicherheit.

le zu reagieren und sich ihnen anzupassen, als Risiken im Voraus zu erkennen und zu mindern (Fischer et al., 2022; Varzhanskiy et al., 2025). Diese Verzögerung ist im Zusammenhang mit KI besonders problematisch, da Bedrohungen nicht nur durch externe Angriffe entstehen, sondern auch durch interne Systemschwachstellen, feindliche Manipulationen (z.B. Vergiftungs-, Umgehungs- oder Missbrauchsangriffe, wie in NIST (2024) aufgeführt) und unbeabsichtigte Folgen automatisierter Entscheidungsfindung. Da KI-Systeme anpassungsfähig und datengesteuert sind, kann sich ihr Verhalten auf komplexe und unvorhersehbare Weise ändern, wodurch statische Regulierungsmodelle und lineare Risikobewertungen zunehmend ineffizient werden. Dies unterstreicht die Notwendigkeit adaptiver Governance-Ansätze, die in der Lage sind, mit der dynamischen Entwicklung von KI-Systemen Schritt zu halten und flexibel auf neue Herausforderungen zu reagieren (Reuel & Undheim, 2024). Strategische Vorausschau wurde daher als wichtige Ergänzung zur aktuellen Governance identifiziert, die es Institutionen ermöglicht, schwache Signale zu erkennen und Zukunftsszenarien zu erforschen, bevor Probleme voll zum Tragen kommen.

Wie in Kapitel 2 erläutert, veranschaulichen Modelle wie die Cyber Kill Chain und die OWASP Top 10 die zentrale Rolle von KI in modernen Cyberangriffen – sowohl als Mittel wie auch als Ziel. Diese technischen Bedrohungsszenarien lassen sich jedoch nicht ohne Weiteres in tragfähige Governance-Strukturen am Arbeitsplatz übersetzen.

Darüber hinaus gelingt es den meisten aktuellen Governance-Ansätzen nicht, die Perspektiven und Interessen der Arbeitnehmer*innen sinnvoll zu integrieren. Diese sind häufig Gegenstand KI-gestützter Entscheidungen und stehen an digitalisierten Arbeitsplätzen an vorderster Front. Die Cybersicherheit am Arbeitsplatz ist dabei nicht nur eine technische Angelegenheit der Sicherung von IT-Systemen, sondern eine soziotechnische Herausforderung, die Vertrauen, Autonomie, Überwachung und Machtdynamik umfasst (z.B. Khando et al., 2021). Algorithmen können die Verantwortlichkeit verwischen, sodass unklar ist, wer verantwortlich ist, wenn Sicherheitsverletzungen oder schädliche Entscheidungen auftreten. Dennoch werden Arbeitnehmer*innen nur selten in die Gestaltung, Umsetzung oder Bewertung solcher Systeme einbezogen. Dieser Mangel an partizipativer Steuerung untergräbt nicht nur die Legitimität von KI-Systemen am Arbeitsplatz, sondern auch ihre Wirksamkeit. Managementstrategien, die die menschliche und organisatorische Dimension außer Acht lassen, werden der Komplexität der Auswirkungen von KI auf Sicherheit, Ethik und Arbeitsbeziehungen nicht gerecht (OECD, 2025).

3.2 Anwendung von Foresight (strategischer Vorausschau) und Technikfolgenabschätzung zur Vorhersage von KI-Arbeitsplatzrisiken

Foresight und Technikfolgenabschätzung (FTA) sind komplementäre Ansätze, die eine vorausschauende Analyse der gesellschaftlichen Relevanz neuer Technologien unterstützen (Udrea, 2025). Traditionelle Ansätze zur Beeinflussung von Innovationen und neuen technologischen Entwicklungen stehen vor einem doppelten Problem, das als Collingridge-Dilemma bezeichnet wird (Collingridge, 1982). Einerseits sind Technologien in der Anfangsphase noch formbar, andererseits sind ihre langfristigen Folgen schwer vorhersehbar. Sobald diese Folgen evident werden, ist der Spielraum

für regulatorische Eingriffe oft begrenzt (Grunwald, 2018). Vielmehr werden die kurzfristigen Fähigkeiten neuer Technologien häufig überschätzt, während die langfristigen systemischen Auswirkungen unterschätzt werden⁵ (Krings & Heister, 2024). Dies resultiert in einer eher reaktiven als strategischen Steuerung.

FTA-Ansätze sind daher unerlässlich, um eine verantwortungsvolle, durch KI gesteuerte Transformation am Arbeitsplatz zu gestalten. Sie unterstützen eine proaktive Entscheidungsfindung, indem sie Organisationen dabei helfen, Risiken und Trends frühzeitig zu erkennen und technologische Entwicklungen mit gesellschaftlichen und arbeitsbezogenen Werten in Einklang zu bringen. Horizon Scanning ermöglicht die frühzeitige Erkennung schwacher Signale, Verschiebungen und disruptiver Dynamiken, die Arbeitsprozesse, Beschäftigungsstrukturen oder die Informationssicherheit verändern können. Die Szenarienplanung ermöglicht die Entwicklung langfristiger Strategien unter unsicheren Bedingungen und bietet eine strukturierte Möglichkeit, verschiedene Zukunftspfade zu erkunden. Einige dieser Pfade können aufkommende Bedrohungen beinhalten, während andere neue Möglichkeiten zur Sicherung und Verbesserung der Widerstandsfähigkeit am Arbeitsplatz eröffnen (Crawford & Wright, 2025).

Die Relevanz von FTA zeigt sich besonders deutlich, wenn KI als ein soziotechnisches Phänomen verstanden wird (Kudina & van de Poel, 2024). Die Einführung von KI-Systemen erfolgt nicht neutral oder deterministisch – sie ist stets eingebettet in spezifische Governance-Rahmen, institutionelle Interessen und strategische Entscheidungsprozesse. Durch den Einsatz vorausschauender Methoden und TA-Ansätze lassen sich die potenziellen Auswirkungen von KI auf zentrale Aspekte wie Autonomie, Gerechtigkeit, Überwachung und die Qualität von Arbeitsverhältnissen systematisch analysieren. Auf diese Weise wird eine partizipative und ethisch reflektierte Gestaltung technologischer Entwicklungen gefördert.

Im Unterschied zur Prognose, die vergangenheitsbasierte Daten nutzt, um zukünftige Entwicklungen vorherzusagen, ist Foresight explorativ und partizipativ angelegt. Zentrale Veränderungsdynamiken werden identifiziert und alternative Zukunftsszenarien entwickelt, um Handlungsspielräume aufzuzeigen (Peels & Del Ponce Castillo, 2024). Die Technikfolgenabschätzung ergänzt diese Perspektive durch evidenzbasierte Bewertungen konkreter Technologien – mit Fokus auf deren gesellschaftliche, ökologische und ökonomische Auswirkungen. In Kombination bilden Foresight und TA ein wirkungsvolles Instrumentarium, um strategische Zukunftsorientierung mit praxisnaher Bewertung zu verbinden – insbesondere im Hinblick auf Fragen der Sicherheit und Governance von KI in der Arbeitswelt (Udrea, 2025).

Darüber hinaus trägt FTA dazu bei, neu aufkommende Risiken frühzeitig sichtbar zu machen – und zwar für eine breite Gruppe relevanter Akteur*innen, darunter Beschäftigte, Gewerkschaften sowie politische Entscheidungsträger*innen. Foresight-gestützte Entscheidungsprozesse ermöglichen es, Schwachstellen in digitalen Infrastrukturen, regulatorische Lücken oder diskriminierende Tendenzen in algorithmischen Systemen frühzeitig zu erkennen. Die Kombination aus kontinuierlichem Monitoring, partizipativ gestalteter Vorausschau und TA – etwa im Rahmen von Workshops oder unterstützenden digitalen Tools – eröffnet dabei ein vielversprechendes methodisches Fundament für proaktive und inklusive KI-Gestaltung in der Arbeitswelt.

5 Diese Beobachtung ist auch als Amaras Gesetz bekannt.

Beispiele für die Integration von Foresight im Bereich der Cybersicherheit finden sich auf internationaler Ebene. Die OECD empfiehlt vorausschauende Governance-Ansätze, um Risiken von KI am Arbeitsplatz frühzeitig zu adressieren (Lane & Saint-Martin, 2021). Die Europäische Cybersicherheitsagentur ENISA veröffentlichte ihren Bericht „Foresight Cybersecurity Threats for 2030“ (2024), der zukünftige Bedrohungen auf Basis von Szenarien identifiziert. Auch das BSI⁶ nutzt vorausschauende Analysen, etwa in einer aktuellen Studie zur globalen Cyber-Startup-Landschaft (2024).

3.3 Integration von Foresight in KI-Sicherheitsstrategien am Arbeitsplatz

Die Integration von Foresight und Technikfolgenabschätzung (TA) in Cybersicherheitsstrategien ist nicht nur eine methodische Ergänzung, sondern Ausdruck eines umfassenderen Verständnisses von strategischer Governance. Sie steht für eine Form des zukunftsorientierten Denkens, die bestehende Governance-Mechanismen ergänzt und zu einer lernenden, anpassungsfähigen Organisationskultur beiträgt. In einer Arbeitswelt, in der KI-Systeme Sicherheitsmechanismen nicht nur ermöglichen, sondern zugleich potenziell unterlaufen können, reicht es nicht mehr aus, ausschließlich auf reaktive Maßnahmen zu setzen. Stattdessen bedarf es einer proaktiven Sicherheits-Governance, die Foresight nicht nur in die strategische Planung, sondern auch in die operative Logik der Organisation integriert.

Szenarienplanung hilft, zukünftige KI-Entwicklungen und ihre sicherheitsrelevanten Auswirkungen früh zu erkennen. Horizon Scanning kann dabei helfen, frühe Anzeichen disruptiver Veränderungen zu erkennen – beispielsweise KI-gestützte Phishing-Kampagnen oder neuartige Insider-Bedrohungen durch autonome Softwareagenten. Durch eine strategische Auseinandersetzung mit solchen Trends können Organisationen Cybersicherheitsrichtlinien entwickeln, die nicht auf starre Compliance-Checklisten beschränkt bleiben, sondern flexibel und adaptiv ausgestaltet sind.

Ein zweiter zentraler Aspekt betrifft die Entwicklung kritischer Sicherheitskompetenzen auf allen Ebenen der Organisation. Angesichts der durch KI generierten Bedrohungsszenarien sind nicht nur technische Fähigkeiten erforderlich, sondern auch ein tiefgreifendes soziotechnisches Verständnis. Dazu zählt insbesondere die Fähigkeit, Wechselwirkungen zwischen algorithmischen Systemen, menschlichem Verhalten, organisatorischen Strukturen und kulturellen Normen zu erkennen. Wichtig sind Fähigkeiten zur Erkennung schwerer Bedrohungen, Bewertung algorithmischer Verzerrungen und ethischen Mitgestaltung. Schulungen und gezielter Kapazitätsaufbau sind daher keine nachgelagerten Maßnahmen, sondern ein integraler Bestandteil jeder nachhaltigen Resilienzstrategie im Unternehmen (Abdelkader et al., 2024, S. 12).

Drittens stärkt ein integrierter FTA-Ansatz die partizipative Gestaltung von Sicherheit. Anstatt Sicherheitsstrategien top down vorzugeben, setzt FTA auf Dialogformate, in denen Mitarbeitende, IT-Fachkräfte, Gewerkschaften und externe Interessengruppen aktiv eingebunden werden. Diese multiperspektivische Herangehensweise hilft dabei, blinde Flecken bei der Implementierung von KI-Systemen zu vermeiden und technische Schutzmaßnahmen mit der betrieblichen Realität in Ein-

6 Bundesamt für Sicherheit in der Informationstechnik, Bundesministerium des Innern und für Heimat, Deutschland.

klang zu bringen. Da partizipative Verfahren ein zentrales Merkmal von FTA sind, unterstützt dieser Ansatz die Auffassung, dass Cybersicherheit als kollektives Anliegen zu verstehen ist (Khando et al., 2021; Klüver, 2024).

Schließlich ermöglicht die Integration von Foresight und TA einen systemischen Resilienzansatz. Sicherheitsvorfälle werden nicht als isolierte Ausnahmen, sondern als Hinweise auf strukturelle Fehlanpassungen zwischen Technologie, Regulierung und sozialen Erwartungen verstanden. Ein vorausschauendes Sicherheitsmanagement basiert daher nicht auf Schadensbegrenzung, sondern auf kontinuierlichem Monitoring und lernender Anpassung (z. B. Europäische Kommission, o. D.; NIST, 2024). Es unterstützt Organisationen dabei, den Einsatz von KI mit normativen Leitwerten wie Transparenz, Autonomie, Inklusivität und digitaler Souveränität zu verknüpfen.

Für eine breite Umsetzung können Organisationen regelmäßige Vorausschau-Audits, partizipative Risikoanalysen oder gemeinsame Szenarienworkshops mit internen und externen Akteur*innen durchführen. Die Verankerung solcher Praktiken in ISO-kompatiblen Risikomanagementsystemen – etwa gemäß ISO/IEC 27001, 31001 oder ISO/IEC 42001 – kann die institutionelle Implementierung zusätzlich unterstützen und eine Angleichung an internationale Sicherheitsstandards gewährleisten.

Gleichzeitig sind Foresight- und TA-Ansätze in der Praxis mit Herausforderungen verbunden: Kurzfristiges Denken in Organisationen, ein begrenztes Verständnis für vorausschauende Methoden (Futures Literacy) sowie knappe personelle und finanzielle Ressourcen – insbesondere bei kleinen und mittleren Unternehmen – können ihre Anwendung erschweren. Um diesen Barrieren zu begegnen, bedarf es gezielter politischer Unterstützung, institutioneller Koordination und entsprechender Qualifizierungsprogramme, die FTA in Prozesse des organisationalen Lernens und der strategischen Planung integrieren. Die Kombination von vorausschauender Analyse und partizipativer Governance ermöglicht Cybersicherheitsstrategien, die technisch resilient, ethisch reflektiert und sozial eingebettet sind. Somit gehen sie weit über den engen Fokus traditioneller, rein technischer Strategien hinaus.

4. Resilienz in der KI-Arbeitswelt: Warum vorausschauende Ansätze zentral sind

Die Integration von KI in die Arbeitswelt ist weit mehr als ein technologischer Fortschritt – sie markiert einen tiefgreifenden strukturellen Wandel, der Arbeitsorganisation, Entscheidungsprozesse und Risikomanagement gleichermaßen transformiert. Wie gezeigt wurde, verändert KI die Cybersicherheitslandschaft grundlegend: Sie ermöglicht auf der einen Seite skalierbare, autonome und zunehmend schwer erkennbare Angriffsformen. Auf der anderen Seite wird sie zu einem zentralen Element in der Verteidigung gegen diese Bedrohungen – durch automatisierte Erkennung, Echtzeitanalyse und adaptive Reaktionsmechanismen.

Diese doppelte Rolle offenbart ein zentrales Spannungsfeld: Gerade jene Systeme, die Sicherheit gewährleisten sollen, können selbst zu Angriffszielen oder Schwachstellen werden. In einer sich dynamisch verändernden Bedrohungslandschaft reichen traditionelle, rein reaktive und technisch fokussierte Governance-Ansätze nicht mehr aus. Erforderlich sind proaktive, integrative und zukunftsgerichtete Strategien, die über rein technische Lösungen hinausgehen.

Strategische Vorausschau und Technikfolgenabschätzung (FTA) bieten hierfür entscheidende Zugänge. Sie ermöglichen es, potenzielle Störungen frühzeitig zu erkennen, Wechselwirkungen zwischen Technologie, Organisation und Gesellschaft systematisch zu analysieren und sicherheitsrelevante Entscheidungen auf eine breitere Wissens- und Wertebasis zu stellen. Besonders relevant ist, dass FTA auch nichttechnische Risiken sichtbar macht – etwa im Hinblick auf Transparenzdefizite, Machtasymmetrien oder fehlende Mitgestaltungsmöglichkeiten der Beschäftigten. Die Integration partizipativer Verfahren und szenarienbasierter Analysen eröffnet neue Wege für adaptive, lernfähige und resiliente Cybersicherheitsstrategien.

Cybersicherheit darf in diesem Kontext nicht länger als rein technische Domäne verstanden werden. Vielmehr muss sie in die umfassendere digitale Transformation von Organisationen eingebettet und im Einklang mit Fragen der Arbeitskultur, der Partizipation und der Regulierung gedacht werden. Dies gilt insbesondere vor dem Hintergrund der zunehmenden algorithmischen Entscheidungsfindung am Arbeitsplatz, in der Verantwortung, Vertrauen und Mitsprache häufig unklar oder unterrepräsentiert sind.

Zusammenfassend zeigt dieser Beitrag: Eine sichere, KI-gestützte Arbeitswelt erfordert einen doppelten Ansatz. Erstens muss Cybersicherheit als integraler Bestandteil des technologischen und organisatorischen Wandels begriffen werden. Zweitens braucht es vorausschauende Governance-Instrumente, die nicht nur technische Risiken adressieren, sondern auch soziale Erwartungen und betriebliche Realitäten einbeziehen. Nur durch eine solche integrierte Perspektive lassen sich Unternehmen und ihre Beschäftigten wirksam und nachhaltig gegen die Risiken einer zunehmend KI-geprägten Arbeitswelt wappnen.

Literatur

- Abdelkader, S., Amissah, J., Kinga, S., Mugerwa, G., Emmanuel, E., Mansour, D. A., Bajaj, M., Blazek, V. & Prokop, L. (2024). Securing modern power systems: Implementing comprehensive strategies to enhance resilience and reliability against cyber-attacks. *Results in Engineering*, 23, 102647. <https://doi.org/10.1016/j.rineng.2024.102647>
- Binhammad, M., Alqaydi, S., Othman, A. & Abuljadayel, L. H. (2024). The role of AI in cyber security: Safeguarding digital identity. *Journal of Information Security*, 15(2), 245–278. <https://doi.org/10.4236/jis.2024.152015>
- Bundesamt für Sicherheit in der Informationstechnik (BSI). (2024, 16. September). *BSI und Cyberagentur veröffentlichen Studie zur Startup-Szene in der San Francisco Bay Area* [Pressemeldung]. https://www.bsi.bund.de/DE/Service-Navi/Presse/Pressemitteilungen/Presse2024/240916_Studie_Startup_Cyber_Landscape.html
- Cohen, D., Te'eni, D., Yahav, I., Zagalsky, A., Schwartz, D., Silverman, G., Mann, Y., Elalouf, A. & Makowski, J. (2025). Human-AI Enhancement of Cyber Threat Intelligence. *International Journal of Information Security*, 24(2), 99. <https://doi.org/10.1007/s10207-025-01004-4>
- Collingridge, D. (1982). *The social control of technology*. St. Martin's Press.
- Crawford, M. M. & Wright, G. (2025). *Improving and Enhancing Scenario Planning: Futures Thinking*. Edward Elgar Publishing. <https://doi.org/10.4337/9781035310586>
- Cuadron, A., Li, D., Ma, W., Wang, X., Wang, Y., Zhuang, S., Liu, S., Schroeder, L. G., Xia, T., Mao, H., Thumiger, N., Desai, A., Stoica, I., Klimovic, A., Neubig, G. &

- Gonzalez, J. E. (2025). *The Danger of Overthinking: Examining the Reasoning-Action Dilemma in Agentic Tasks*. <https://doi.org/10.48550/arXiv.2502.08235>
- ECSF. (2022). *European Cybersecurity Skills Framework*. European Union Agency for Cybersecurity. <https://doi.org/10.2824/859537>
- Ekelhart, A., Goldenits, G., Mayer, R. & Schuster, V. (2025). *Künstliche Intelligenz in der Cybersicherheit: Chancen und Risiken*. SBA Research & RTR. <https://www.rtr.at/TKP/aktuelles/publikationen/publikationen/Kuenstliche-Intelligenz-in-der-Cybersicherheit.pdf>
- Elgan, M. (2025, 17. Januar). *How to calculate your AI-powered cybersecurity's ROI*. Think | IBM. <https://www.ibm.com/think/insights/how-to-calculate-your-ai-powered-cybersecurity-roi>
- Europäische Kommission. (o. D.). *Strategische Vorausschau*. https://commission.europa.eu/strategy-and-policy/strategic-foresight_de
- European Union Agency for Cybersecurity (ENISA). (2022). *European cybersecurity skills framework*. European Union Agency for Cybersecurity. <https://data.europa.eu/doi/10.2824/859537>
- European Union Agency for Cybersecurity (ENISA). (2024, März). *Foresight Cybersecurity Threats for 2030 – Update*. ENISA. <https://www.enisa.europa.eu/publications/foresight-cybersecurity-threats-for-2030-update-2024-extended-report>
- Fischer, B., Meissner, D., Nyuur, R. & Sarpong, D. (2022). Guest editorial: Cyber-attacks, strategic cyber-foresight, and security. *IEEE Transactions on Engineering Management*, 69(6), 3660–3663. <https://doi.org/10.1109/TEM.2022.3204165>
- Grunwald, A. (2018). *Technology assessment in practice and theory*. Routledge.
- Gupta, M., Akiri, C., Aryal, K., Parker, E. & Praharaj, L. (2023). From ChatGPT to Threat-GPT: Impact of generative ai in cybersecurity and privacy. *IEEE Access*, 11, 80218–80245. <https://doi.org/10.1109/ACCESS.2023.3300381>
- Hutchins, E. M., Cloppert, M. J. & Amin, R. M. (2011). Intelligence-driven computer network defense informed by analysis of adversary campaigns and intrusion kill chains. In J. Ryan (Hrsg.), *Leading Issues in Information Warfare and Security Research* (Bd. 1, S. 80–106). Academic Publishing International.
- Jada, I. & Mayayise, T. O. (2024). The impact of artificial intelligence on organisational cyber security: An outcome of a systematic literature review. *Data and Information Management*, 8(2), 100063. <https://doi.org/10.1016/j.dim.2023.100063>
- Kaloudi, N. & Li, J. (2020). The ai-based cyber threat landscape: A survey. *ACM Computing Surveys (CSUR)*, 53(1), 1–34. <https://doi.org/10.1145/3372823>
- Khando, K., Gao, S., Islam, S. M. & Salman, A. (2021). Enhancing employees information security awareness in private and public organisations: A systematic literature review. *Computers & Security*, 106, 102267. <https://doi.org/10.1016/j.cose.2021.102267>
- King, T. C., Aggarwal, N., Taddeo, M. & Floridi, L. (2020). Artificial intelligence crime: An interdisciplinary analysis of foreseeable threats and solutions. *Science and Engineering Ethics*, 26(1), 89–120. <https://doi.org/10.1007/s11948-018-00081-0>
- Klüver, L. (2024). Participatory technology assessment: TA and the public. In A. Grunwald (Hrsg.), *Handbook of Technology Assessment* (S. 64–76). Edward Elgar Publishing.
- Krings, M. & Heister, W. (2024). »Hype-Zyklus« und »Amara-Gesetz«: Die Wahrnehmung neuer Technologien. *Blätter der Wohlfahrtspflege (BdW)*, 171(4). <https://doi.org/10.5771/0340-8574-2024-4>
- Kudina, O. & van de Poel, I. (2024). A sociotechnical system perspective on AI. *Minds and Machines*, 34(3), 21. <https://doi.org/10.1007/s11023-024-09680-2>
- Lane, M. & Saint-Martin, A. (2021). *Die Auswirkungen von KI auf die Arbeitsmärkte: Was wir bislang wissen* (OECD Social, Employment and Migration Working Papers No. 256). OECD Publishing. <https://doi.org/10.1787/540444e1-de>

- Malatji, M. & Tolah, A. (2024). Artificial intelligence (AI) cybersecurity dimensions: a comprehensive framework for understanding adversarial and offensive AI. *AI and Ethics*. <https://doi.org/10.1007/s43681-024-00427-4>
- Markevych, M. & Dawson, M. (2023). A review of enhancing intrusion detection systems for cybersecurity using artificial intelligence (ai). *Sciend International Conference Knowledge Based Organization*, 29(3), 30–37. <https://doi.org/10.2478/kbo-2023-0072>
- National Cyber Security Centre (NCSC). (2024, 24. Januar). *The near-term impact of AI on the cyber threat*. NCSC. <https://www.ncsc.gov.uk/report/impact-of-ai-on-cyber-threat>
- National Institute of Standards and Technology (NIST). (2024). *The NIST Cybersecurity Framework (CSF) 2.0*. <https://doi.org/10.6028/nist.cswp.29>
- OECD. (2025). *Steering AI's future: Strategies for anticipatory governance* (OECD Artificial Intelligence Papers No. 32). OECD Publishing. <https://doi.org/10.1787/5480ff0a-en>
- OWASP. (2024). *OWASP Top 10 for LLM Applications 2025*. Abgerufen am 7. Juni 2025, von <https://genai.owasp.org/resource/owasp-top-10-for-llm-applications-2025/>
- Peels, R. & Ponce del Castillo, A. (2024). *Using foresight to think and act upon an uncertain future world of work: trade unions' experiences* (ILO Working Paper No. 118). International Labour Office. <https://doi.org/10.54394/VNQM4255>
- Ramakrishnan, S. & Chittibala, D. R. (2024). Enhancing Cyber Resilience: Convergence of SIEM, SOAR, and AI in 2024. *International Journal of Computing and Engineering*, 5(2), 36–44. <https://doi.org/10.47941/ijce.1754>
- Reuel, A. & Undheim, T. A. (2024). *Generative AI Needs Adaptive Governance*. <https://doi.org/10.48550/arXiv.2406.04554>
- Udreă, T. (2025). Foresight und Technikfolgenabschätzung: Grundlage und Potential für Arbeitnehmer:inneninteressen. In S. Klocker & A. Siegl (Hrsg.), *KI Verstehen und Gestalten: Künstliche Intelligenz und die Arbeitswelt*. ÖGB-Verlag. Im Erscheinen.
- Varzhanskiy, I., Melnychenko, A., Naiev, S. & Kapustynskiy, O. (2025). Foresight and Active Forecast Potential for Improving Defense Capability. *European Journal of Interdisciplinary Studies*, 17(1). <https://doi.org/10.24818/ejis.2025.01>
- Vassilev, A., Oprea, A., Fordyce, A. & Anderson, H. (2024). *Adversarial machine learning: A taxonomy and terminology of attacks and mitigations*. <https://doi.org/10.6028/nist.ai.100-2e2023>
- Wang, Y., Liu, Q., Xu, J., Liang, T., Chen, X., He, Z., Song, L., Yu, D., Li, J. & Zhang, Z. (2025). *Thoughts Are All Over the Place: On the Underthinking of o1-Like LLMs*. <https://doi.org/10.48550/arXiv.2501.18585>

Veränderte Spielregeln

Wie künstliche Intelligenz die Anforderungen an Kompetenzen, Vertrauen und (Zusammen-)Arbeit in der Arbeitswelt verändert

Stefanie Spielberger

1. Einleitung

Künstliche Intelligenz (KI) hat sich in den vergangenen Jahren rasant weiterentwickelt und gilt mittlerweile als zentraler Treiber eines tiefgreifenden Wandels der Arbeitswelt. Ihre zunehmende Integration transformiert nicht nur betriebliche Prozesse und Strukturen, sondern verändert auch grundlegend die Art und Weise, wie Menschen arbeiten, kommunizieren und zusammenarbeiten (Bankins et al., 2024). Mit KI ist „*the ability of a machine or computer system to simulate and perform tasks that require human intelligence, such as logical reasoning, learning, and problem solving*“ gemeint (Morandin-Ahuerma, 2022, S. 1950). Aktuell findet KI in vielen Bereichen wie z.B. Wirtschaft, Wissenschaft, Kunst und Bildung Anwendung. Die Bandbreite an Einsatzmöglichkeiten reicht von autonomen Fahrzeugen und intelligenten Sprachassistenten bis hin zu KI-gestützten medizinischen Diagnosesystemen. Dabei entwickeln sich sowohl die Technologien rasant weiter als auch der Kreis der Anwendenden, der stetig wächst, zunehmend verschiedene Branchen und Berufsgruppen umfasst und hierbei Effizienz und Nutzererlebnis verbessert (Ng et al., 2021a). Spätestens seit der weltweiten Verfügbarkeit generativer KI (genKI) wie z.B. ChatGPT hat sich das Anwendungsspektrum nochmal deutlich erweitert, da leistungsfähige KI-Tools nun auch für eine breite Öffentlichkeit zugänglich und zunehmend am Arbeitsplatz nutzbar sind (Kero et al., 2023; Seufert & Meier, 2023). GenKI wird hierbei verstanden als „*a subset of artificial intelligence that, based on a textual prompt, generates novel content*“ (Stanford HAI, 2023, S. 3). (Generative) KI birgt das Potenzial, sich wiederholende Aufgaben zu automatisieren, Entscheidungsfindung zu verbessern und die Kommunikation zu erleichtern (Brynjolfsson et al., 2018; Chee et al., 2024). Einige Autoren sind davon überzeugt, dass diese Entwicklungen zu weitreichenden Jobverlusten führen werden (Frey & Osborne, 2017; Miller, 2022), wohingegen die meisten Autoren mittlerweile eher mit einer Verschiebung von Aufgaben und der Nachfrage nach neuen Kompetenzen rechnen (z.B. Brynjolfsson et al., 2018; Bughin et al., 2018; Dede et al., 2021; Wienrich et al., 2022). Dabei ist unbestritten, dass die Einführung von KI für einige Arbeitnehmende den Wechsel in vollkommen neue Berufe erfordern wird, während andere sich neue Kompetenzen aneignen müssen, um erfolgreich mit KI-Technologien am Arbeitsplatz zusammenzuarbeiten. Um ihre Zukunftsfähigkeit zu sichern, ist eine Anpassung der Wissens- und Kompetenzbasis der Mitarbeitenden an den digitalen Wandel erforderlich. Dies ermöglicht eine gewinnbringende Verbindung der Stärken von Menschen und Maschinen. Dabei trifft KI auf eine Arbeitswelt, die sich bereits in einem umfassenden Wandel befindet und geprägt ist durch hybride Arbeitsmodelle und virtuelle Zusammenarbeit. Die Synthese von KI und ortsunabhängiger Arbeit wird eine grundlegen-

de Veränderung traditioneller Arbeitsmodelle und Unternehmensstrukturen bewirken (Miller, 2022; Richter & Richter, 2024).

Obwohl die Auswirkungen von KI auf die Arbeitswelt tiefgreifend sind, ist die Vorbereitung vieler Organisationen auf die damit einhergehenden Veränderungen bislang unzureichend (Initiative D21, 2024). Dennoch ist es unabdingbar, geeignete Strategien und Lösungsansätze zu entwickeln, um KI erfolgreich in die (hybride) Arbeitswelt zu integrieren, denn langfristig werden Unternehmen, die nicht in der Lage sind, KI effektiv zu integrieren und zu nutzen, Wettbewerbsnachteile haben (Pinski et al., 2024).

Ziel dieses Beitrags ist es, einen Überblick über veränderte Kompetenzanforderungen, die Rolle von Vertrauen bei der Implementierung von KI sowie die Besonderheiten hybrider Arbeitswelten zu geben und darauf aufbauend Handlungsempfehlungen für eine erfolgreiche KI-Integration abzuleiten, um das Potenzial von KI in der Arbeitswelt gezielt zu erschließen.

2. Erforderliche Kompetenzen im KI-Zeitalter

2.1 AI Literacy

In Anbetracht dieser tiefgreifenden Veränderungen, die (generative) KI-Tools für die Arbeits- und Lebenswelt einer Vielzahl an Personen mit sich bringen, liegt es auf der Hand, dass mit der Nutzung von KI-Anwendungen veränderte Kompetenzanforderungen einhergehen, was die rapide ansteigende Anzahl an Publikationen in diesem Bereich unterstreicht (Almatrafi et al., 2024; Yang et al., 2025). Mit der zunehmenden Verbreitung von KI am Arbeitsplatz werden diejenigen, die KI verstehen und damit umgehen können, einen deutlichen Vorteil gegenüber denjenigen haben, die über weniger entwickelte KI-Kenntnisse verfügen (Southworth et al., 2023). Im Mittelpunkt dieser Diskussion steht häufig der Erwerb von „AI Literacy“, für die es bis dato noch keinen Konsens für eine allgemeingültige Definition gibt. Eine weit verbreitete Definition, die explizit die Arbeitsperspektive miteinschließt, beschreibt AI Literacy *„as a set of competencies that enables individuals to critically evaluate AI technologies; communicate and collaborate effectively with AI; and use AI as a tool online, at home, and in the workplace“* (Long & Magerko, 2020, S. 2). Im Kontext des Erwerbs von AI Literacy richten viele Autorinnen und Autoren den Fokus insbesondere auf nichttechnische Anwendergruppen (Cetindamar et al., 2022; Laupichler et al., 2023; Long & Magerko, 2020; Ng et al., 2021a). AI Literacy zielt darauf ab, den Menschen Kompetenzen für eine effektive Interaktion mit KI-Systemen, eine fundierte Entscheidungsfindung und ein Verständnis für die breiteren gesellschaftlichen Auswirkungen von KI zu vermitteln (Ng et al., 2021a). Denn für die meisten wird zumindest die Fähigkeit erforderlich sein, die Ergebnisse von KI-Systemen angemessen für ihre Entscheidungsfindung zu nutzen (Chee et al., 2024; Faruqe et al., 2022). In ihrem systematischen Literaturreview kommen Almatrafi et al. (2024), ähnlich wie zuvor Ng et al. (2021b), zu folgenden Kernelementen von AI Literacy: Erkennen, Kennen und Verstehen, Verwenden und Anwenden, Bewerten, Gestalten von KI sowie ethisch Navigieren. Im Hinblick auf die Priorisierung dieser einzelnen Elemente besteht unter den Autoren weitgehend Einigkeit, dass das grundlegende

Verständnis der Funktionsweise von KI und die kritische und ethische Auseinandersetzung mit KI von größerer Bedeutung sind als die reine Anwendung (Almatrafi et al., 2024; Cetindamar et al., 2022; Chee et al., 2024; Faruque et al., 2022; Ng et al., 2021a).

Der größte Literaturstrang, der sich mit dem Erwerb von KI-Kompetenzen auseinandersetzt, ist im Bildungsbereich verortet, wobei das Spektrum der Zielgruppen vom Kleinkindalter bis hin zum Kompetenzerwerb im Rahmen der Hochschulbildung reicht (Almatrafi et al., 2024; Gu & Ericson, 2025; Kong et al., 2021; Southworth et al., 2023; Su et al., 2023; Yang et al., 2025; Zhai et al., 2021). Bislang deutlich weniger intensiv untersucht ist der arbeitsbezogene KI-Kompetenzerwerb (Almatrafi et al., 2024; Wienrich et al., 2022), aber auch hier finden sich zunehmend Betrachtungen geänderter Kompetenzanforderungen. Faruque et al. (2022) unterscheiden in Bezug auf die Tiefe der benötigten KI-Expertise am Arbeitsplatz beispielsweise zwischen „Co-Workers“, allgemeinen Arbeitskräften, die KI-Outputs gezielt in ihre Arbeit integrieren müssen, „Collaborators“, die aktiv mit KI zusammenarbeiten und spezialisierten Nutzergruppen wie KI-System-Entwicklern („Creators“). Andere Autoren betrachten geänderte Kompetenzanforderungen für einzelne Nutzergruppen wie z.B. Mitarbeitende an digitalen Arbeitsplätzen (Cetindamar et al., 2022), Mitarbeitende in multinationalen Konzernen (Jaiswal et al., 2022) oder Führungskräfte im oberen Management (Pinski et al., 2024). Es wird somit deutlich, dass die Einführung von (generativer) KI sich tiefgreifend auf die erforderlichen Kompetenzprofile auswirken wird und je nach Nutzergruppe individuelle KI-Kenntnisse benötigt werden, die auf ihre jeweilige Funktion zugeschnitten sind (Pinski et al., 2024). Hieraus wird auch deutlich, dass die Aufgabenteilung zwischen Mensch und KI unterschiedliche Formen annehmen kann, was Gegenstand des nachfolgenden Kapitels ist.

2.2 Mensch-KI-Kollaboration – eine komplementäre Partnerschaft

KI-Systeme besitzen das Potenzial, eine Vielzahl von bislang durch Menschen ausgeführte Tätigkeiten zu übernehmen bzw. zu verdrängen. Die allgegenwärtigen Visionen der Zusammenarbeit von Menschen und KI legen die Hypothese nahe, dass Maschinen Routineaufgaben übernehmen sollten, um es Menschen zu ermöglichen, sich auf kreativere Arbeiten zu konzentrieren (Jarrahi, 2018). Der Substitutionsprozess durch KI setzt jedoch nicht auf der Ebene ganzer Berufsbilder, sondern auf der Ebene einzelner Aufgaben an (Huang & Rust, 2018). Vorrangig betroffen sind Tätigkeiten, die mit vergleichsweise geringer kognitiver Komplexität verbunden sind und sich technisch leichter automatisieren lassen. Huang und Rust (2018) mutmaßen, dass sich die Entwicklung der Aufgabenübernahme durch KI in folgendem Muster vollziehen wird: Zunächst werden Aufgaben mit mechanischer Intelligenz ersetzt, gefolgt von analytischen, dann intuitiven und zuletzt einfühlsamen Tätigkeiten. Je nach Aufgabenstellung gibt es zwischen den beiden Extremen „KI macht alles“ und „Mensch macht alles“ unzählige Varianten der Zusammenarbeit (André et al., 2021). Seufert und Meier (2023) unterscheiden in diesem Kontext verschiedene Kooperationsformen und Rollen der KI (z.B. als Assistent, Koordinator, Experte). Die Zukunft liegt somit nicht grundsätzlich in der kompletten Übernahme menschlicher

Arbeitsplätze durch KI, sondern in einer komplementären Partnerschaft, bei der die Gesamtleistung größer ist als die Summe der Einzelbeiträge (Sowa et al., 2021).

Cetindamar et al. (2022) erweitern daher im Hinblick auf digitale Arbeitsplätze die Definition von AI Literacy und betonen, dass neben der Bereitschaft zu lebenslangem Lernen und grundlegenden technischen Kenntnissen komplementäre Fähigkeiten, die den effizienten Einsatz von KI verbessern können, von großer Wichtigkeit sind. Unter komplementären Fähigkeiten sind dabei solche zu verstehen, die die Stärken von KI ergänzen und gleichzeitig deren Schwächen ausgleichen. Diese Art der Zusammenarbeit wird von mehreren Autoren beschrieben und als „Intelligence Augmentation“ (Dede et al., 2021), „hybride Intelligenz“ (Seufert & Meier, 2023, S. 1984) oder „komplementäre kollaborative Intelligenz“ (Rötzel, 2024, S. 22; Wilson & Daugherty, 2018) bezeichnet. Mit Bezug auf die organisationale Entscheidungsfindung können die Fähigkeiten von Mensch und KI zu einer idealen Symbiose verschmelzen, bei der die KI Komplexität im Entscheidungsprozess reduziert und der Mensch Intuition und Kreativität nutzt, um Unsicherheit und Mehrdeutigkeit zu bewältigen (Jarrahi, 2018). Jarrahi weist ferner darauf hin, dass diese Entscheidungsprozesse nicht ausschließlich im höheren Management stattfinden, sondern auch in neuen, unsicheren und mehrdeutigen Situationen auf niedrigeren Hierarchieebenen. Somit wird deutlich, dass menschliche Arbeitskräfte im Rahmen der komplementären Zusammenarbeit mit KI ihre Zeit zunehmend mit Aufgaben verbringen, die nicht routinemäßig, nicht repetitiv und weniger vorhersehbar sind (Brown et al., 2025).

In diesem Zusammenhang gewinnen insbesondere höherwertige kognitive und soziale Kompetenzen (Soft Skills) an Bedeutung, da sie bislang (noch) als schwer automatisierbar gelten (Jaiswal et al., 2022). Zirar et al. (2023) identifizieren neben technischen insbesondere menschliche und konzeptionelle Kompetenzen als entscheidend für das produktive Zusammenwirken von Mensch und KI. Santana und Díaz-Fernández (2023) schließen sich dieser Sichtweise an und heben als zentrale Zukunftskompetenzen kommunikative Kompetenzen und Problemlösungsfähigkeit hervor, gefolgt von Zusammenarbeit und Teamarbeit. Im Zuge der starken Verbreitung von genKI erhält zudem die Fähigkeit, effektiv und zielgerichtet mit KI durch gezieltes Prompting zu kommunizieren und zu kollaborieren einen enormen Bedeutungszuwachs und ist nach Ansicht einiger Autoren heute zusätzlich als Bestandteil von AI Literacy anzusehen (Bozkurt, 2023; Chetty, 2023; Seufert & Meier, 2023). Da KI-Systeme auf Trainingsdaten basieren und nicht eigenständig lernen, sind zum einen natürlich die Daten, aber zum anderen auch die Prompts entscheidend für die Qualität der Ergebnisse und der Schlüssel zur optimalen Nutzung des Potenzials von KI-Technologien (Bozkurt, 2023). Dies bietet auch große berufliche Chancen für ältere Arbeitnehmende, da sie ihre langjährige Erfahrung und ihr umfangreiches Fachwissen auf neue Weise einbringen können, zumal die Systeme mit natürlicher Sprache arbeiten und damit wesentlich einfacher zu bedienen sind als andere digitale Werkzeuge (Chetty, 2023).

Jüngere Forschungsarbeiten haben sich mit den Auswirkungen von genKI auf die Produktivität befasst. Brynjolfsson et al. (2023), Dell’Acqua et al. (2023) und Noy und Zhang (2023) zeigen übereinstimmend, dass der Einsatz generativer KI die Produktivität von Wissensarbeitern deutlich steigern kann. Dies äußert sich in schnellerer Aufgabenerledigung, einer höheren Anzahl bearbeiteter Aufgaben sowie einer signifikanten Verbesserung der Arbeitsqualität. Besonders hervorzuheben ist, dass

alle Autoren darin übereinstimmen, dass weniger leistungsstarke, geringqualifizierte und neue Mitarbeitende überproportional von der KI-Integration profitieren, was auf eine Verringerung von Leistungsunterschieden vor allem in Bezug auf das Qualitätsniveau hindeutet. Im Gegensatz dazu ist der Effekt bei erfahrenen und hochqualifizierten Mitarbeitenden geringer. Bei ihnen sind vor allem Effizienzgewinne nachweisbar (Noy & Zhang, 2023). Zudem kann die Integration von generativer KI in kollaborative Arbeitsprozesse Teams bei Ideenfindung und Moderation unterstützen, besonders wenn Expertise fehlt oder die Kreativität stockt (He et al., 2024). Bezüglich des Umgangs mit genKI identifizieren Dell’Acqua et al. (2023, S. 17) zwei unterschiedliche Nutzerstrategien: „Cyborgs“ und „Zentauren“. Cyborgs integrieren KI- und menschliche Fähigkeiten auf einer granularen Teilaufgabenebene, während Zentauren strategisch zwischen menschlichen und KI-Teilaufgaben unterscheiden. Allerdings hat KI auch ihre Grenzen und kann die menschliche Leistung sogar verschlechtern, wenn die Aufgaben außerhalb der aktuellen Kompetenz von KI liegen (Dell’Acqua et al., 2023). Zudem bestehen Risiken wie der Verlust von kritischem Denken, die Homogenisierung von Ideen (Dell’Acqua et al., 2023; He et al., 2024) und die Reproduktion bestehender Muster und Verzerrungen. Daher ist ein gezielter und situationsabhängiger Einsatz von KI entscheidend, um ihre Vorteile zu maximieren und ihre Grenzen zu berücksichtigen.

Zusammenfassend zeigt sich, dass die Einführung von (generativer) KI eine veränderte Arbeitsteilung zwischen Menschen und Maschine mit sich bringt. Ein Qualifizierungsansatz, der ausschließlich auf den Erwerb technischer Kompetenzen abzielt, reicht eindeutig nicht aus, um den umfassenden Wandel der Kompetenzanforderungen durch KI adäquat zu adressieren. Im Kontext fortschreitender KI-Entwicklung rücken menschliche Kompetenzen wie kritisches Denken, Empathie, Kreativität, Kommunikation und ethisches Urteilsvermögen zunehmend in den Fokus, da sie eine komplementäre Ergänzung zur maschinellen Intelligenz darstellen. Ein effektiver KI-Einsatz in Organisationen setzt daher ein erweitertes Verständnis von AI Literacy mit ergänzenden Fähigkeiten für den Umgang mit genKI voraus. In Organisationen bedarf es somit eines systematischen Kompetenzmanagements, das potenziell obsolete Fähigkeiten identifiziert und zugleich neue, zukunftsrelevante Qualifikationen für unterschiedliche Mitarbeitendengruppen fördert, um sowohl die Beschäftigungsfähigkeit der Mitarbeitenden als auch die langfristige Wettbewerbsfähigkeit der Organisation zu sichern (Jaiswal et al., 2022). Die kontinuierliche Weiterentwicklung von KI verlangt zudem von Mitarbeitenden und Führungskräften eine hohe Bereitschaft zur kontinuierlichen Anpassung, insbesondere im Hinblick auf die Zusammenarbeit mit KI-Systemen (Jarrahi, 2018). Ein weiteres zentrales Element einer erfolgreichen Mensch-KI-Partnerschaft ist Vertrauen, weshalb dieser Aspekt im Mittelpunkt des folgenden Kapitels steht.

3. Vertrauen als Grundlage für erfolgreiche KI-Integration

Verschiedene Autoren betonen, dass Vertrauen, welches Nutzerinnen und Nutzer der Technologie entgegenbringen, eine zentrale Voraussetzung für die erfolgreiche Integration von KI in Arbeitsprozesse darstellt (Choung et al., 2023; Glikson & Woolley, 2020; Henrique & Santos, 2024; Lockey et al., 2021; Schork, 2024). Vertrauen be-

einflusst nicht nur die grundsätzliche Akzeptanz, sondern auch die konkrete Nutzung, Missbrauch oder Ablehnung von KI-Systemen (Glikson & Woolley, 2020). Vertrauen wird definiert als „*a psychological state comprising the intention to accept vulnerability based upon positive expectations of the intentions or behavior of another*“ und ist ein grundlegender menschlicher Mechanismus, der erforderlich ist, um mit Verletzlichkeit, Ungewissheit, Komplexität und Mehrdeutigkeit umzugehen, die zusammengekommen ein Risiko darstellen (Rousseau et al., 1998, S. 395). Vertrauen spielt im Kontext von KI eine zentrale Rolle, da deren komplexe, oft undurchsichtige Entscheidungsprozesse für Nutzer schwer nachvollziehbar sind. Diese Intransparenz erschwert den Umgang mit den potenziellen Risiken, weshalb Vertrauen als Voraussetzung für die Akzeptanz und den produktiven Einsatz von KI im Arbeitskontext gilt (Choung et al., 2023). Hinzu kommt, dass Misstrauen gegenüber KI am Arbeitsplatz häufig daraus resultiert, dass Beschäftigte sie als Bedrohung für die eigene Arbeitsstelle wahrnehmen und die Fähigkeiten von KI-Systemen oft überbewertet werden, was zu unrealistischen Erwartungen und verstärkten Ängsten führt (Zirar et al., 2023).

Die Forschung unterscheidet zwei zentrale Dimensionen des Vertrauens in KI: „kognitives Vertrauen“ und „emotionales Vertrauen“ (Choung et al., 2023; Glikson & Woolley, 2020; Rötzel, 2024). Kognitives Vertrauen basiert auf rationalen Bewertungen hinsichtlich Eigenschaften wie Greifbarkeit, Transparenz, Zuverlässigkeit und unmittelbarer Rückmeldung des Systems, während emotionales Vertrauen durch anthropomorphe Merkmale, soziale Präsenz und menschenähnliches Verhalten der KI beeinflusst wird (Glikson & Woolley, 2020; Henrique & Santos, 2024). Beide Dimensionen gelten als zentrale Prädiktoren für die Akzeptanz von KI, wobei emotionales Vertrauen vor allem bei nutzerzentrierten Zielgruppen entscheidend ist, während kognitives Vertrauen für technikaffine Nutzende relevanter ist, die auf Leistungsfähigkeit und Verlässlichkeit achten (Choung et al., 2023). Im Rahmen einer europaweiten Untersuchung zum Vertrauen in KI zeigen Scantamburlo et al. (2025), dass Vertrauen in KI nicht nur vom individuellen Wissensniveau abhängt, sondern auch stark von kulturellen, politischen und bildungsbezogenen Faktoren in den verschiedenen Ländern beeinflusst wird. Auch Rötzel (2024, S. 20) weist auf die Rolle von sozialen Einflussfaktoren im Vertrauensaufbau hin. Er beschreibt zwei gegenläufige Verhaltensmuster in der Mensch-KI-Interaktion: die „Automatisierungsneigung“, also die Tendenz, algorithmischen Empfehlungen auch ohne kritische Prüfung zu folgen, sowie die „Algorithmusvermeidung“, bei der menschliche Urteile bevorzugt werden, selbst wenn zuverlässige KI-Empfehlungen verfügbar sind. Auch Seufert und Meier (2023, S. 1203) nennen individuelle Haltungen als kritische Einflussgrößen für eine erfolgreiche Mensch-KI-Kollaboration. Sie argumentieren, dass ein „Growth Mindset“ Offenheit und Lernbereitschaft im Umgang mit KI unterstützt, während ein „Fixed Mindset“ eher zu Konkurrenzdenken und Abwehrhaltungen gegenüber technologischen Veränderungen führen kann.

Um die Akzeptanz von KI am Arbeitsplatz zu fördern, sind neben technischer Funktionalität bewusste Gestaltungsmaßnahmen zur Vertrauensbildung auf unterschiedlichen Ebenen erforderlich (Choung et al., 2023; Glikson & Woolley, 2020; Lockey et al., 2021; Zhang et al., 2023). Ein vielversprechender Gestaltungsansatz zur Erhöhung der Nachvollziehbarkeit von KI-Entscheidungen und damit zur Stärkung des kognitiven Vertrauens besteht darin, die häufig als „Black Box“ bezeich-

neten KI-Systeme erklärbar zu machen (Henrique & Santos, 2024; Lockey et al., 2021). Auch Rötzel (2024) und Zhang et al. (2023) betonen, dass die Offenlegung der Funktionsweise und die eindeutige Kennzeichnung von KI-Systemen essenziell ist, um Missverständnisse und Vertrauensverluste zu vermeiden. Weitere zentrale Mechanismen zur Stärkung des Vertrauens in KI sind Vertrautheit und Erfahrung durch den Umgang mit KI (Daly et al., 2025; Glikson & Woolley, 2020; Lockey et al., 2021), eine anthropomorphe Gestaltung der KI (Choung et al., 2023) sowie eine angemessene KI-Governance, welche das Ziel hat, die Zusammenarbeit zwischen Menschen und Maschine sicher, transparent und förderlich zu gestalten (Lockey et al., 2021). Scantamburlo et al. (2025) erweitern in diesem Zusammenhang den Betrachtungswinkel über die Organisationsgrenzen hinaus und betonen, dass Vertrauen in das „KI-Ökosystem“ durch gesetzliche Regelungen, transparente Kommunikation der Anbieter und Bildungsmaßnahmen gestärkt werden kann.

Zusammenfassend lässt sich feststellen, dass Vertrauen ein dynamisches und vielschichtiges Konstrukt ist, das sowohl durch die Eigenschaften und Gestaltung der KI-Systeme als auch durch deren soziale und organisationale Einbettung sowie durch die Haltung, Erfahrungen und Kompetenzen der Nutzerinnen und Nutzer geprägt wird. Angesichts rasanter technologischer Entwicklungen wird es zunehmend entscheidend, die Mensch-KI-Interaktion bewusst zu gestalten und dabei ein ausgewogenes Verhältnis zwischen Innovationsdynamik und Vertrauenssicherung herzustellen (Rötzel, 2024). Eine zukunftsfähige Arbeitswelt erfordert daher ein tiefes Verständnis für Voraussetzungen vertrauensvoller Mensch-KI-Kollaboration. Entsprechend sollte Vertrauensbildung als integraler Bestandteil von KI-Implementierungsprozessen in Organisationen von Beginn an systematisch mitgedacht und gefördert werden.

4. Hybride Arbeit und generative KI

Seit der COVID-19-Pandemie hat sich hybride Arbeit als Kombination aus Präsenz- und Remote-Arbeit zu einem verbreiteten Arbeitsmodell entwickelt, ist jedoch nicht für alle Berufsgruppen gleichermaßen geeignet. In Anbetracht der dynamischen digitalen Entwicklungen, insbesondere der rasanten Verbreitung generativer KI und der Entstehung virtueller Arbeitswelten, plädieren mehrere Autoren für ein breiteres Verständnis des Konzepts hybrider Arbeit (Lauring & Jonasson, 2025; Richter & Richter, 2024). Lauring und Jonasson (2025, S. 15–16) schlagen vor, hybride Arbeit als das flexible Wechseln zwischen traditionellen (z.B. Büro, Face-to-Face) und nicht-traditionellen (z.B. Homeoffice, digital, virtuell) Arbeitsformen hinsichtlich Modus, Ort und Zeit zu sehen. Dabei stellt ein ausgewogenes Verhältnis dieser Ausprägungen den höchsten Grad an Hybridität dar. Diese Wechsel können sowohl auf individueller als auch auf Teamebene erfolgen. Richter und Richter (2024) integrieren die zunehmende Mensch-KI-Interaktion explizit in das Konzept hybrider Arbeit. Sie fordern eine gleichwertige Gestaltung von Mensch-Mensch- und Mensch-KI-Kollaboration in physischen wie auch virtuellen Räumen. Hybride Arbeit wird so zu einer dynamischen Kooperationsform, in der Menschen und KI gemeinsam Aufgaben bewältigen. Hybride Arbeitsformen bieten vielfältige Potenziale, etwa hinsichtlich Flexibilität und Produktivität, werfen jedoch zugleich komplexe Fragen im Hinblick auf Vertrauen und Gerechtigkeit auf. Die nachfolgenden Ausführungen skizzieren zentra-

le Herausforderungen und Chancen hybrider Arbeit, wobei die dargestellten Aspekte exemplarisch ausgewählt wurden. Eine vertiefte Bewertung im Hinblick auf den Einsatz von KI erfolgt in einem späteren Abschnitt.

Als Vorteil von hybrider Arbeit wird häufig eine höhere Flexibilität und Autonomie der Beschäftigten genannt, was mit einer größeren Arbeitszufriedenheit einhergeht (Hofmann & Rief, 2023; Kaiser et al., 2024), wobei Choudhury et al. (2022) zu dem Ergebnis kommen, dass Mitarbeitende, die durchschnittlich etwa zwei Tage pro Woche im Büro verbringen, die höchste Arbeitszufriedenheit aufweisen. Hybride Arbeit bietet jedoch nicht automatisch mehr Flexibilität für Mitarbeitende, sondern sie kann auch durch betriebliche oder kollegiale Erwartungen eingeschränkt oder vorgegeben werden (Lauring & Jonasson, 2025). Dies kann Herausforderungen für die betroffenen Mitarbeitenden mit sich bringen.

Als weiterer Vorteil wird eine potenziell gesteigerte Produktivität durch hybride Arbeit genannt (Choudhury et al., 2022; Hofmann & Rief, 2023). Dies ist jedoch mit Einschränkung zu betrachten, da die Produktivität auch immer von individuellen Rahmenbedingungen wie z.B. der persönlichen und räumlichen Situation im Homeoffice abhängen kann. Auch können individuelle Verhaltensweisen, Präferenzen und Kompetenzen den Erfolg hybrider Arbeit beeinflussen (Allen et al., 2024).

Trotz einer gewissen Präsenz am Arbeitsplatz ist eine zentrale Herausforderung hybrider Arbeit die „soziale Erosion“ (Hofmann & Rief, 2023, S. 5), also der schleichende Verlust zwischenmenschlicher Interaktion und Bindung. Informelle Gespräche nehmen ab, Teamzusammenhalt und teamübergreifende Netzwerke bröckeln, die Wahrnehmung von Arbeitsbelastung bei Kollegen sinkt und es bestehen Unsicherheiten über die Arbeitsintensität anderer Teammitglieder, was Vertrauen und Zusammenarbeit beeinträchtigt (Hofmann & Rief, 2023).

Betrachtet man nun diese Entwicklungen im Lichte der zunehmenden Verbreitung von KI wird deutlich, dass sich durch das Aufeinandertreffen von KI und hybrider Arbeit die geschilderten Effekte verstärken können. So kann durch die permanente örtliche und zeitliche Verfügbarkeit von KI in hybriden Arbeitswelten eine höhere Flexibilität und Autonomie erreicht werden, die Mitarbeitende unabhängiger vom Büroarbeitsplatz und von der Mitwirkung von Kolleginnen und Kollegen machen kann. Denn von der KI können einfache Aufgaben übernommen werden, woraus sich Produktivitätsgewinne erzielen lassen, die sich durch die teilweise ungestörtere Arbeit in einem hybriden Setting noch verstärken können. Zudem könnten die relativ hohen Qualitätszuwächse durch KI auch Geringqualifizierten größere Autonomie und Flexibilität in Bezug auf das Aufgabenspektrum ermöglichen. Dies könnte dazu führen, dass auch sie wiederum einen besseren Zugang zu hybriden Arbeitsmodellen erhalten, da der Anteil selbstständiger Aufgabenbearbeitung zunimmt. Wenn KI den Großteil einfacher Aufgaben übernimmt, stellt sich jedoch auch die Frage, wie durch KI-Lösungen langfristig das Erfahrungswissen, z.B. von Geringqualifizierten oder auch von Nachwuchskräften, gesichert werden kann (Dell’Acqua et al., 2023), da dieses wiederum für die wichtige kritische Beurteilung von KI-Ergebnissen relevant sein kann. Ein weiteres zukünftiges Risiko der intensiven Nutzung von genKI könnte auch die zunehmende Abhängigkeit dieser Nutzergruppen von KI sein, sodass z.B. kleinste Aufgaben wie das Verfassen einer E-Mail nicht mehr ohne die Hilfe von KI möglich sind. Es ist zudem davon auszugehen, dass auch die individuelle Nutzerstrategie Auswirkungen auf die Produktivität und die Kollaboration in Teams hat. So

können beispielsweise die Arbeitsergebnisse von „Cyborgs“ aufgrund der nahtlosen KI-Integration in hoher Geschwindigkeit erbracht werden, könnten jedoch bei Transparenz und Nachvollziehbarkeit im Team an Grenzen stoßen, wohingegen „Zentauren“ für klare Arbeitsteilung, Reflexion und Vielfalt an Lösungswegen sorgen, dafür aber mehr Zeit für Koordination benötigen (Dell’Acqua et al., 2023). Beide Ansätze könnten die Leistungsfähigkeit von Teams deutlich erhöhen, insbesondere wenn verschiedene KI-Kooperationsstile verfügbar sind, da so unterschiedliche Stärken und Methoden ins Team einfließen können.

Weiterhin ist davon auszugehen, dass die Kombination von hybrider Arbeit und dem Einsatz von KI die Herausforderungen an die zwischenmenschliche Interaktion verschärfen und die in hybriden Arbeitsstrukturen ohnehin vorhandenen Tendenzen zur „sozialen Erosion“ (Hofmann & Rief, 2023, S. 5) beschleunigen wird. Denn durch die verstärkte Mensch-KI-Kollaboration wird der direkte Austausch im Team weiter reduziert, spontane Impulse und kreative Ideenfindung werden seltener. Dies wird insbesondere in stark digitalisierten und individualisierten Arbeitsumgebungen mit einem hohen Anteil an Nicht-Präsenzarbeit gelten.

Darüber hinaus entstehen durch den Einsatz von KI in hybriden Settings Vertrauensrisiken auf mehreren Ebenen. Erstens betrifft dies das Vertrauen in die KI-Systeme selbst. Zweifel an der Richtigkeit, Transparenz und Nachvollziehbarkeit KI-generierter Ergebnisse führen zu Unsicherheiten im Arbeitsprozess und behindern eine effektive Mensch-KI-Interaktion. Zweitens gerät das Verhältnis zwischen Mitarbeitenden und Arbeitgebern aufgrund der Sorge, durch KI langfristig ersetzbar zu sein, unter Druck. Dies kann sich negativ auf Motivation, Commitment und Arbeitszufriedenheit auswirken (Huang & Rust, 2018). Drittens wirkt sich der zunehmende Einsatz von KI auch auf das kollegiale Miteinander aus. Wenn der Einsatz von KI nicht transparent erfolgt, können Missverständnisse über individuelle Beiträge, Arbeitslasten und Leistungsbewertung entstehen. Dies wird insbesondere in hybriden Arbeitskontexten verstärkt, in denen bereits eine reduzierte Sichtbarkeit der Arbeitsleistung und ein geringerer sozialer Austausch bestehen. Wird KI zudem verdeckt oder unsystematisch genutzt, könnte dies als Illoyalität oder Regelverstoß interpretiert werden, insbesondere dann, wenn der Umfang der menschlichen Eigenleistung nicht mehr klar erkennbar ist. In diesem Kontext wird das Vertrauen, das Beschäftigte sowohl in ihre Organisation, Kolleginnen und Kollegen als auch in digitale Werkzeuge wie KI entwickeln, zu einem zentralen Erfolgsfaktor für deren nachhaltige Integration (Glikson & Woolley, 2020).

Letztlich sind auch Fairnessaspekte zu berücksichtigen, da sowohl das Ausmaß der Hybridität der Arbeit als auch die Intensität der Nutzung von KI in hohem Maße von individuellen Voraussetzungen (z. B. digitale Kompetenzen) und aufgabenspezifischen Anforderungen (z. B. Grad der Virtualisierung) abhängen. Dadurch entstehen unterschiedliche Belastungen, Partizipationschancen und Entwicklungsmöglichkeiten innerhalb von Teams, was zu Spannungen oder subjektiv wahrgenommener Ungleichbehandlung führen kann.

5. Handlungsempfehlungen für Unternehmen

Die vorherigen Ausführungen machen deutlich, dass die Einführung (generativer) KI die (hybride) Arbeitswelt auf verschiedenen Ebenen grundlegend verändert. Im Folgenden werden daher ausgewählte Handlungsempfehlungen vorgestellt, mit denen Unternehmen diesen Herausforderungen wirksam begegnen können.

Strategisches Kompetenzmanagement und ganzheitlicher Qualifizierungsansatz

Unternehmen sollten ein systematisches Kompetenzmanagement etablieren, das KI-ersetzbare Kompetenzen sowie neue, zukunftsrelevante Fähigkeiten identifiziert, um langfristig Beschäftigungsfähigkeit und Wettbewerbsfähigkeit zu sichern. Im Sinne eines umfassenden Kompetenzrahmens ist AI Literacy als integratives Konzept zu begreifen, das durch die gezielte Einbeziehung kognitiver, sozialer und ethischer Fähigkeiten wie kritisches Denken, Kommunikationsfähigkeit und moralisches Urteilsvermögen erweitert werden muss. Dabei sollte die Entwicklung von KI-Kompetenzen zielgruppenspezifisch erfolgen.

Gestaltung vertrauenswürdiger KI-Systeme und Partizipation

Die Förderung von Vertrauen in KI-Systeme erfordert eine transparente und nachvollziehbare Gestaltung der Technologien. Organisationen sollten darauf achten, dass die Funktionsweise von KI verständlich und die Entscheidungslogik nachvollziehbar ist. Dies kann durch partizipative Einführungsprozesse erreicht werden, bei denen Mitarbeitende frühzeitig in die Gestaltung von KI-Systemen eingebunden werden. Solche partizipativen Prozesse stärken nicht nur das Vertrauen in die Technologie, sondern fördern auch die Akzeptanz und das Verständnis der Mitarbeitenden. Zudem sollte bei der Gestaltung von KI auf emotional vertrauensbildende Maßnahmen geachtet werden, die die Nutzenden in ihrer Interaktion mit den Systemen unterstützen und deren Akzeptanz sowie das Vertrauen in die Technologie langfristig stärken.

Stärkung von Kollaboration und sozialem Zusammenhalt in hybriden Arbeitsmodellen

Unternehmen sollten bei der Einführung von KI und der Gestaltung hybrider Arbeitsmodelle darauf achten, dass Teamkultur und Kommunikationsstrukturen an die neuen Anforderungen angepasst werden. Eine erfolgreiche Integration von KI in die hybride Arbeitswelt erfordert eine flexible und situationsspezifische Gestaltung, die sowohl die Produktivität als auch den sozialen Zusammenhalt fördert. Besonders wichtig ist es, der „sozialen Erosion“ (Hofmann & Rief, 2023, S. 5) entgegenzuwirken. Deshalb sollte bewusst auf die Förderung von Teamzusammenhalt, Transparenz und Vertrauen geachtet werden, um eine produktive, motivierende und faire Arbeitsumgebung zu schaffen.

Etablierung eines organisationsweiten KI-Managements

Ein umfassendes KI-Management ist erforderlich, um den ethischen und produktiven Einsatz von KI sicherzustellen. Dazu zählt der Aufbau klarer KI-Governance-Strukturen, ein kontinuierliches Monitoring sowie die Förderung einer lernenden Organisation. Zentrale Bestandteile sind zudem eindeutige Leitlinien zur Nutzung von KI.

Auf diese Weise kann die Integration von KI strategisch genutzt und zukunftsorientiert gestaltet werden, was angesichts des rasanten technologischen Fortschritts eine zwingende Voraussetzung für den Erhalt der Wettbewerbsfähigkeit von Organisationen ist.

6. Fazit

Hybride Arbeit und generative KI markieren gemeinsam einen tiefgreifenden Wandel der Arbeitswelt. Unternehmen stehen dadurch vor neuen Herausforderungen, insbesondere bei der Vorbereitung ihrer Mitarbeitenden auf den kompetenten Umgang mit KI. Eine vorausschauende Gestaltung der KI-Integration ist entscheidend, um Produktivitätspotenziale von KI zu nutzen und zugleich Vertrauen sowie Fairness im Arbeitsumfeld zu stärken. Schlüssel hierzu sind der systematische Kompetenzaufbau und die Entwicklung vertrauenswürdiger KI-Systeme inklusive klarer KI-Governance-Strukturen. Der Erwerb relevanter Fähigkeiten wie AI Literacy, kognitiver sowie kommunikativer und kollaborativer Kompetenzen bildet die Grundlage für den erfolgreichen KI-Einsatz im Arbeitsalltag. Im hybriden Kontext gewinnen zudem sozialer Zusammenhalt und Vertrauen an Bedeutung. Unternehmen sollten daher gezielt in Kompetenzentwicklung und vertrauensbildende Maßnahmen auf unterschiedlichen Ebenen investieren, um KI als strategischen Hebel wirksam zu nutzen.

In der Zukunft wird es entscheidend sein, die kontinuierliche Entwicklung von KI-Kompetenzen in die Unternehmenskultur zu integrieren, da sich die KI-Technologie ständig weiterentwickelt. Die fortschreitende Verschmelzung von Mensch und Maschine wird neue Möglichkeiten für Innovation und Zusammenarbeit eröffnen, die heute noch nicht vollständig abzusehen sind. In einer zunehmend hybriden Arbeitswelt wird es außerdem von zentraler Bedeutung sein, dass Unternehmen flexible und integrative Arbeitsumgebungen schaffen, die sowohl die Vorteile digitaler Technologien als auch die sozialen und kreativen Potenziale der Präsenzarbeit miteinander vereinen. Dies erfordert nicht nur technologische Anpassungen, sondern auch eine bewusste Förderung von Kommunikation, Vertrauen und Kollaboration, insbesondere in einem Kontext, in dem KI als Arbeitsmittel und Partner in den Arbeitsalltag integriert wird.

Literatur

- Allen, K. S., Grelle, D., Lazarus, E. M., Popp, E. & Gutierrez, S. L. (2024). Hybrid is here to stay: Critical behaviors for success in the new world of work. *Personality and Individual Differences*, 217, 112459. <https://doi.org/10.1016/j.paid.2023.112459>
- Almatrafi, O., Johri, A. & Lee, H. (2024). A systematic review of AI literacy conceptualization, constructs, and implementation and assessment efforts (2019–2023). *Computers and Education Open*, 6, 100173. <https://doi.org/10.1016/j.caeo.2024.100173>
- André, E., Bauer, W., Aurich, J. C., Bullinger-Hoffmann, A., Heister, M., Huchler, N., Neuburger, R., Peissner, M., Stich, A., Suchy, O., Ramin, P. & Wächter, M. (2021). *Kompetenzentwicklung für Künstliche Intelligenz – Veränderungen, Bedarfe und Handlungsoptionen* (Whitepaper). Plattform Lernende Systeme. https://doi.org/10.48669/pls_2021-2

- Bankins, S., Ocampo, A. C., Marrone, M., Restubog, S. L. D. & Woo, S. E. (2024). A multilevel review of artificial intelligence in organizations: Implications for organizational behavior research and practice. *Journal of Organizational Behavior*, 45(2), 159–182. <https://doi.org/10.1002/job.2735>
- Bozkurt, A. (2023). Generative artificial intelligence (AI) powered conversational educational agents: The inevitable paradigm shift. *Asian Journal of Distance Education*, 18(1). <http://www.asianjde.com/ojs/index.php/AsianJDE/article/view/718>
- Brown, A. S., Dishop, C. R., Kuznetsov, A., Chao, P.-Y. & Woolley, A. W. (2025). Beyond efficiency: Trust, AI, and surprise in knowledge work environments. *Computers in Human Behavior*, 167, 108605. <https://doi.org/10.1016/j.chb.2025.108605>
- Brynjolfsson, E., Li, D. & Raymond, L. (2023). *Generative AI at Work* (NBER Working Paper 31161). National Bureau of Economic Research. <https://doi.org/10.3386/w31161>
- Brynjolfsson, E., Mitchell, T., & Rock, D. (2018). What Can Machines Learn and What Does It Mean for Occupations and the Economy? *AEA Papers and Proceedings*, 108, 43–47. <https://doi.org/10.1257/pandp.20181019>
- Bughin, J., Hazan, E., Lund, S., Dahlström, P., Wiesinger, A. & Subramaniam, A. (2018). *Skill Shift Automation and the Future of the Workforce* (Discussion Paper). McKinsey Global Institute. <https://www.mckinsey.com/~/media/McKinsey/Industries/Public%20and%20Social%20Sector/Our%20Insights/Skill%20shift%20Automation%20and%20the%20future%20of%20the%20workforce/MGI-Skill-Shift-Automation-and-future-of-the-workforce-May-2018.pdf>
- Cetindamar, D., Kitto, K., Wu, M., Zhang, Y., Abedin, B. & Knight, S. (2022). Explicating AI Literacy of Employees at Digital Workplaces. *IEEE Transactions on Engineering Management*, 71, 810–823. <https://doi.org/10.1109/TEM.2021.3138503>
- Chee, H., Ahn, S. & Lee, J. (2024). A Competency Framework for AI Literacy: Variations by Different Learner Groups and an Implied Learning Pathway. *British Journal of Educational Technology*, 2–37. <https://doi.org/10.1111/bjet.13556>
- Chetty, K. (2023). AI Literacy for an Ageing Workforce: Leveraging the Experience of Older Workers. *OBM Geriatrics*, 7(3), 1–17. <https://doi.org/10.21926/obm.geriatr.2303243>
- Choudhury, P., Khanna, T., Makridis, C. & Schirmann, K. (2022). *Is Hybrid Work the Best of Both Worlds? Evidence from a Field Experiment* (Harvard Business School Technology & Operations Mgt. Unit Working Paper No. 22-063, Harvard Business School Strategy Unit Working Paper No. 22-063). Social Science Research Network. <https://doi.org/10.2139/ssrn.4068741>
- Choung, H., David, P. & Ross, A. (2023). Trust in AI and Its Role in the Acceptance of AI Technologies. *International Journal of Human-Computer Interaction*, 39(9), 1727–1739. <https://doi.org/10.1080/10447318.2022.2050543>
- Daly, S. J., Wiewiora, A. & Hearn, G. (2025). Shifting attitudes and trust in AI: Influences on organizational AI adoption. *Technological Forecasting and Social Change*, 215, 124108. <https://doi.org/10.1016/j.techfore.2025.124108>
- Dede, C., Etemadi, A. & Forshaw, T. (2021). *Intelligence Augmentation: Upskilling Humans to Complement AI*. The Next Level Lab at the Harvard Graduate School of Education. President and Fellows of Harvard College.
- Dell’Acqua, F., McFowland, E., Mollick, E. R., Lifshitz-Assaf, H., Kellogg, K., Rajendran, S., Kraye, L., Candelon, F. & Lakhani, K. R. (2023). *Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality* (Harvard Business School Technology & Operations Mgt. Unit Working Paper No. 24-013, The Wharton School Research Paper). <https://doi.org/10.2139/ssrn.4573321>
- Faruqe, F., Watkins, R. & Medsker, L. (2022). Competency Model Approach to AI Literacy: Research-Based Path From Initial Framework to Model. *Advances in Artificial Intelligence*, 2022, 1–12. <https://doi.org/10.1002/9781119789886.ch1>

- cial Intelligence and Machine Learning*, 2(4), 580–587. <https://doi.org/10.54364/AA-IML.2022.1140>
- Frey, C. B. & Osborne, M. A. (2017). The future of employment: How susceptible are jobs to computerisation? *Technological Forecasting and Social Change*, 114, 254–280. <https://doi.org/10.1016/j.techfore.2016.08.019>
- Glikson, E. & Woolley, A. W. (2020). Human Trust in Artificial Intelligence: Review of Empirical Research. *Academy of Management Annals*, 14(2), 627–660. <https://doi.org/10.5465/annals.2018.0057>
- Gu, X. & Ericson, B. J. (2025). *AI Literacy in K-12 and Higher Education in the Wake of Generative AI: An Integrative Review*. <https://doi.org/10.48550/arXiv.2503.00079>
- He, J., Houde, S., Gonzalez, G. E., Moran, D. A. S., Ross, S. I., Muller, M. & Weisz, J. D. (2024). AI and the Future of Collaborative Work: Group Ideation with an LLM in a Virtual Canvas. In *Proceedings of the 3rd Annual Meeting of the Symposium on Human-Computer Interaction for Work (CHIWORK '24)* (Article 9, S. 1–14). Association for Computing Machinery. <https://doi.org/10.1145/3663384.3663398>
- Henrique, B. M. & Santos, E. (2024). Trust in artificial intelligence: Literature review and main path analysis. *Computers in Human Behavior: Artificial Humans*, 2(1), 100043. <https://doi.org/10.1016/j.chbah.2024.100043>
- Hofmann, J. & Rief, S. (2023). *Performance hybrider Arbeit*. Fraunhofer IAO. <https://doi.org/10.24406/publica-1949>
- Huang, M.-H. & Rust, R. T. (2018). Artificial Intelligence in Service. *Journal of Service Research*, 21(2), 155–172. <https://doi.org/10.1177/1094670517752459>
- Initiative D21. (2024). *D21-Digital-Index 2023/24: Jährliches Lagebild zur Digitalen Gesellschaft* [Report]. https://initiated21.de/uploads/03_Studien-Publikationen/D21-Digital-Index/2023-24/d21digitalindex_2023-2024.pdf
- Jaiswal, A., Arun, C. J. & Varma, A. (2022). Rebooting employees: Upskilling for artificial intelligence in multinational corporations. *The International Journal of Human Resource Management*, 33(6), 1179–1208. <https://doi.org/10.1080/09585192.2021.1891114>
- Jarrah, M. H. (2018). Artificial intelligence and the future of work: Human-AI symbiosis in organizational decision making. *Business Horizons*, 61(4), 577–586. <https://doi.org/10.1016/j.bushor.2018.03.007>
- Kaiser, S., Süß, S. & Schmoll, R. (2024). One size fits all? Ein Plädoyer für eine differenzierte Betrachtung und Gestaltung hybrider Arbeit. *Schmalenbach IMPULSE*, 3(2), 1–9. <https://doi.org/10.54585/TRDC8581>
- Kero, S., Akyürek, S. Y. & Flaßhoff, F. G. (2023). *Bekanntheit und Akzeptanz von ChatGPT in Deutschland* (Meinungsmonitor Künstliche Intelligenz, Factsheet Nr. 10 – Dezember 2023). <https://www.cais-research.de/wp-content/uploads/Factsheet-10-ChatGPT.pdf>
- Kong, S.-C., Cheung, W. M.-Y. & Zhang, G. (2021). Evaluation of an artificial intelligence literacy course for university students with diverse study backgrounds. *Computers and Education: Artificial Intelligence*, 2, 100026. <https://doi.org/10.1016/j.caeai.2021.100026>
- Laupichler, M. C., Aster, A., Haverkamp, N. & Raupach, T. (2023). Development of the “Scale for the assessment of non-experts’ AI literacy” – An exploratory factor analysis. *Computers in Human Behavior Reports*, 12, 100338. <https://doi.org/10.1016/j.chbr.2023.100338>
- Lauring, J. & Jonasson, C. (2025). What is hybrid work? Towards greater conceptual clarity of a common term and understanding its consequences. *Human Resource Management Review*, 35(1), 101044. <https://doi.org/10.1016/j.hrmr.2024.101044>
- Lockey, S., Gillespie, N., Holm, D. & Someh, I. A. (2021). A Review of Trust in Artificial Intelligence: Challenges, Vulnerabilities and Future Directions. In *Proceedings of the*

- 54th Hawaii International Conference on System Sciences (S. 5463–5473). <https://doi.org/10.24251/hicss.2021.664>
- Long, D. & Magerko, B. (2020). What is AI Literacy? Competencies and Design Considerations. In *CHI'20: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (S. 1–16). Association for Computing Machinery. <https://doi.org/10.1145/3313831.3376727>
- Miller, T. (2022). *A.I. and remote working: A paradigm shift in employment* (2. Aufl.). Business Expert Press.
- Morandín-Ahuerma, F. (2022). What is Artificial Intelligence? *International Journal of Research Publication and Reviews*, 3(12), 1947–1951. <https://doi.org/10.55248/gengpi.2022.31261>
- Ng, D. T. K., Leung, J. K. L., Chu, K. W. S., & Qiao, M. S. (2021a). AI Literacy: Definition, Teaching, Evaluation and Ethical Issues. *Proceedings of the Association for Information Science and Technology*, 58(1), 504–509. <https://doi.org/10.1002/pra2.487>
- Ng, D. T. K., Leung, J. K. L., Chu, K. W. S. & Qiao, M. S. (2021b). Conceptualizing AI literacy: An exploratory review. *Computers and Education: Artificial Intelligence*, 2, 100041. <https://doi.org/10.1016/j.caeai.2021.100041>
- Noy, S. & Zhang, W. (2023). *Experimental Evidence on the Productivity Effects of Generative Artificial Intelligence* (Working Paper). <https://doi.org/10.2139/ssrn.4375283>
- Pinski, M., Hofmann, T. & Benlian, A. (2024). AI Literacy for the top management: An upper echelons perspective on corporate AI orientation and implementation ability. *Electronic Markets*, 34(1), 24. <https://doi.org/10.1007/s12525-024-00707-1>
- Richter, A. & Richter, S. (2024). Hybrid work – a reconceptualisation and research agenda. *i-com*, 23(1), 71–78. <https://doi.org/10.1515/icom-2023-0027>
- Rötzel, P. G. (2024). Künstliche Intelligenz (KI) – unser bester Freund? In S. Schork (Hrsg.), *Vertrauen in Künstliche Intelligenz* (S. 19–33). Springer Vieweg. https://doi.org/10.1007/978-3-658-43816-6_2
- Rousseau, D. M., Sitkin, S. B., Burt, R. S. & Camerer, C. (1998). Not So Different After All: A Cross-Discipline View Of Trust. *Academy of Management Review*, 23(3), 393–404. <https://doi.org/10.5465/amr.1998.926617>
- Santana, M. & Díaz-Fernández, M. (2023). Competencies for the artificial intelligence age: Visualisation of the state of the art and future perspectives. *Review of Managerial Science*, 17(6), 1971–2004. <https://doi.org/10.1007/s11846-022-00613-w>
- Scantamburlo, T., Cortés, A., Foffano, F., Barrué, C., Distefano, V., Pham, L. & Fabris, A. (2025). Artificial Intelligence Across Europe: A Study on Awareness, Attitude and Trust. *IEEE Transactions on Artificial Intelligence*, 6(2), 477–490. <https://doi.org/10.1109/TAI.2024.3461633>
- Schork, S. (Hrsg.). (2024). *Vertrauen in Künstliche Intelligenz: Eine multi-perspektivische Betrachtung*. Springer Vieweg. <https://doi.org/10.1007/978-3-658-43816-6>
- Seufert, S. & Meier, C. (2023). Hybride Intelligenz: Zusammenarbeit mit KI-Assistenzsystemen in wissensintensiven Bereichen. *HMD Praxis der Wirtschaftsinformatik*, 60(6), 1194–1209. <https://doi.org/10.1365/s40702-023-01012-9>
- Southworth, J., Migliaccio, K., Glover, J., Glover, J., Reed, D., McCarty, C., Brendemuhl, J. & Thomas, A. (2023). Developing a model for *AI Across the curriculum*: Transforming the higher education landscape via innovation in AI literacy. *Computers and Education: Artificial Intelligence*, 4, 100127. <https://doi.org/10.1016/j.caeai.2023.100127>
- Sowa, K., Przegalinska, A. & Ciechanowski, L. (2021). Cobots in knowledge work. *Journal of Business Research*, 125, 135–142. <https://doi.org/10.1016/j.jbusres.2020.11.038>
- Stanford University Human-Centered Artificial Intelligence (Stanford HAI). (Hrsg.). (2023). *Generative AI: Perspectives from Stanford HAI*. <https://hai.stanford.edu/research/generative-ai-perspectives-from-stanford-hai>

- Su, J., Ng, D. T. K. & Chu, S. K. W. (2023). Artificial Intelligence (AI) Literacy in Early Childhood Education: The Challenges and Opportunities. *Computers and Education: Artificial Intelligence*, 4, 100124. <https://doi.org/10.1016/j.caeai.2023.100124>
- Wienrich, C., Carolus, A., Markus, A. & Augustin, Y. (2022). *AI Literacy: Kompetenzdimensionen und Einflussfaktoren im Kontext von Arbeit* (Working Paper). Denkfabrik Digitale Arbeitsgesellschaft.
- Wilson, H. J. & Daugherty, P. R. (2018). Collaborative Intelligence: Humans and AI Are Joining Forces. *Harvard Business Review*, July-August, 84–92.
- Yang, Y., Zhang, Y., Sun, D., He, W. & Wei, Y. (2025). Navigating the landscape of AI literacy education: Insights from a decade of research (2014–2024). *Humanities and Social Sciences Communications*, 12(1), 374. <https://doi.org/10.1057/s41599-025-04583-8>
- Zhai, X., Chu, X., Chai, C. S., Jong, M. S. Y., Istenic, A., Spector, M., Liu, J.-B., Yuan, J. & Li, Y. (2021). A Review of Artificial Intelligence (AI) in Education from 2010 to 2020. *Complexity*, 2021(1), 8812542. <https://doi.org/10.1155/2021/8812542>
- Zhang, G., Chong, L., Kotovsky, K. & Cagan, J. (2023). Trust in an AI versus a Human teammate: The effects of teammate identity and performance on Human-AI cooperation. *Computers in Human Behavior*, 139, 107536. <https://doi.org/10.1016/j.chb.2022.107536>
- Zirar, A., Ali, S. I. & Islam, N. (2023). Worker and workplace Artificial Intelligence (AI) coexistence: Emerging themes and research agenda. *Technovation*, 124, 102747. <https://doi.org/10.1016/j.technovation.2023.102747>

Künstliche Intelligenz an Hochschulen

Eine systemische Analyse der KI-Integration in Hochschulen mit dem Viable System Model (VSM)

Barbara Geyer & Rita Stampfl

1. Einführung

Die zunehmende Verbreitung Künstlicher Intelligenz (KI) prägt unsere Zeit maßgeblich und durchdringt immer mehr Lebensbereiche (Kavitha et al., 2024). Auf Hochschulen entfaltet KI tiefgreifende Auswirkungen und verändert deren traditionelle Funktionslogiken (Katsamakos et al., 2024). Die Potenziale erstrecken sich von der Optimierung der Lehr- und Lernprozesse über die Effizienzsteigerung administrativer Abläufe bis hin zur Forschungsunterstützung (Ocen et al., 2025). Gleichzeitig stellen diese Technologien die Hochschulen vor große Herausforderungen in Bezug auf akademische Integrität, ethische Implikationen und verantwortungsvolle Governance (Wu et al., 2024). Die aktive Gestaltung des KI-Einsatzes im Sinne des Bildungs- und Gesellschaftsauftrags wird somit zu einer zentralen Aufgabe der Hochschulen (Jin et al., 2025).

Um diese komplexen Transformationsprozesse systematisch zu erfassen, wird zur Analyse der organisationalen KI-Integration das Viable System Model (VSM) nach Beer (2007) auf Hochschulen übertragen. Anhand der fünf VSM-Systeme werden KI-Anwendungen in Hochschulbereichen, resultierende Potenziale und entstehende Spannungsfelder analysiert. Der Beitrag bietet somit nicht nur eine Bestandsaufnahme, sondern entwickelt daraus einen Orientierungsrahmen zur strategischen Hochschulentwicklung im digitalen Wandel.

2. Das Viable System Model (VSM) und Hochschulen

Das Viable System Model (VSM) wurde in den 1960er- und 1970er-Jahren von Beer als kybernetisches Modell zur Analyse komplexer Organisationen entwickelt. Beer (2007) orientierte sich dabei an neurologischen Strukturen und übertrug diese auf soziale Systeme, um deren Lebensfähigkeit (Viabilität) in dynamischen Umwelten zu erklären. Gemäß dem Modell sind Organisationen, ähnlich wie lebende Organismen, dazu in der Lage, sich selbst zu regulieren und anzupassen. Im Gegensatz zu mechanistischen Organisationsmodellen betont das VSM zirkuläre Beziehungen, Autonomie und Rekursivität.

Dieses kybernetische Organisationsverständnis basiert auf fünf wesentlichen Subsystemen, die in jedem lebensfähigen System vorhanden und über spezifische Kommunikationswege miteinander verbunden sein müssen. Diese Systeme agieren nicht hierarchisch, sondern funktional differenziert und gewährleisten gemeinsam die Anpassungsfähigkeit und Stabilität der Organisation. Rezk und Gamal (2019) haben das VSM bereits auf den Hochschulkontext übertragen.

Auf dieser Grundlage wird für diesen Beitrag das Viable System Model für den Hochschulkontext adaptiert und die Zusammenarbeit der verschiedenen Organisationsebenen in der nachfolgenden Abbildung 1 dargestellt. System 5 (Identität & Werte) wird durch das Rektorat in seiner normativen Führungsrolle sowie den Universitätsrat verkörpert und definiert die grundlegende Mission der Hochschule. System 4 (Strategische Planung) manifestiert sich in Forschungsbüros und strategischen Abteilungen, die kontinuierlich die zukunftsbezogene Umwelt beobachten und strategische Entwicklungspfade identifizieren. Die homöostatische Balance zwischen System 3 und 4 – dargestellt durch die dynamische Verbindung – reflektiert das zentrale Spannungsverhältnis zwischen operativer Stabilität und Innovation.

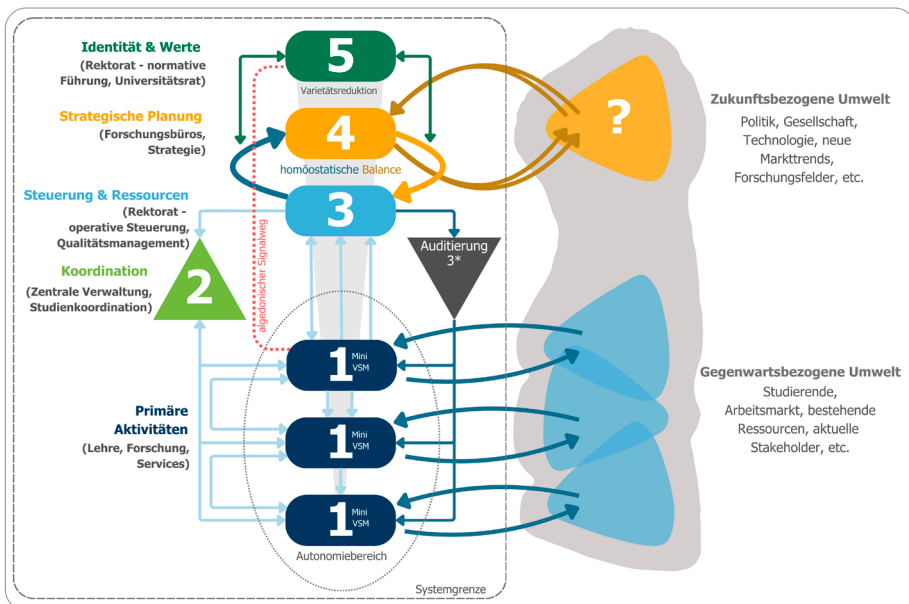


Abb. 1: Viable System Model (VSM) im Hochschulkontext. Eigene Adaption nach Beer (2007)

In der hier vorgestellten Darstellung wird System 3 (Steuerung & Ressourcen) durch das Rektorat in seiner operativen Steuerungsfunktion und das Qualitätsmanagement realisiert, ergänzt durch System 3* (Auditierung). System 2 (Koordination) umfasst zentrale Verwaltungsstrukturen und Studienkoordination, deren Funktion in der Verhinderung von Konflikten zwischen den operativen Einheiten liegt. Die operativen Einheiten (System 1), bestehend aus der Lehre, der Forschung und den zentralen Dienstleistungsservices, agieren innerhalb des ihnen zugewiesenen Autonomiebereichs und sind in der Lage, vollständige lebensfähige Systeme zu bilden. Dieser Sachverhalt wird durch das integrierte Mini-VSM visualisiert.

Eine weitere Eigenschaft des Modells ist die Durchlässigkeit der Systemgrenze. Damit wird verdeutlicht, dass Hochschulen keine geschlossenen Systeme sind, sondern in ständigem Austausch mit ihrer differenzierten (zukunfts- und gegenwartsbezogenen) Umwelt stehen. Die Filterebenen zwischen den Systemebenen symbolisieren das notwendige Varietätsmanagement, das als Reduktion von Komplexität

in der vertikalen Kommunikation verstanden werden kann. Der algedonische Signalweg ermöglicht in Krisensituationen eine direkte Kommunikation von der operativen zur normativen Ebene.

Diese systemtheoretische Betrachtung von Hochschulen ermöglicht eine ganzheitliche Analyse organisatorischer Herausforderungen und bildet die konzeptionelle Grundlage für die nachfolgende Untersuchung der Auswirkungen künstlicher Intelligenz (KI) auf die verschiedenen Subsysteme und ihre Funktionen im Hochschulkontext. Aufbauend auf diesem theoretischen Rahmen wird im nächsten Schritt systematisch analysiert, wie KI auf die einzelnen Systeme des VSM wirkt – beginnend bei der operativen Ebene bis hin zur normativen Ausrichtung der Hochschule.

2.1 System 1: primäre Aktivitäten – Lehre, Forschung, Services

Im Rahmen der Anwendung des Viable System Model (VSM) auf Hochschulen repräsentiert System 1 die operativen Einheiten, in denen die eigentliche Mission der Hochschule realisiert wird – konkret in Lehre, Forschung und Services. Im vorliegenden Beitrag wird die Perspektive entwickelt, dass diese Funktionsbereiche nicht nur das Rückgrat der Institution bilden, sondern auch die Orte darstellen, an denen technologische Transformationen zuerst und unmittelbar spürbar werden. Die zentralen KI-Anwendungsfelder im System 1 sind in Tabelle 1 zusammengefasst.

Im Bereich der Lehre ermöglichen adaptive Lernpfade und idiographische Lernanalysen personalisierte Bildungserfahrungen durch die Nutzung individueller Lernendendaten als Analysegrundlage. Diese Herangehensweise fördert maßgeschneiderte Lernprozesse und Echtzeitunterstützung bei Verständnisproblemen sowie datengestützte Kursoptimierung für heterogene Gruppen. Allerdings sind mit diesem Ansatz auch ethische Herausforderungen verbunden, zum Beispiel Datensensibilität, potenzielle algorithmische Verzerrungen und Einschränkungen der Lernautonomie. Darüber hinaus sind Fragen des Datenschutzes und der Auswirkungen unvollständiger oder fehlerhafter Daten auf die Modellgenauigkeit zu berücksichtigen (Misiejuk et al., 2025).

Neben der Unterstützung von Lernprozessen werden KI-Anwendungen auch für die Erstellung von Lehr-, Lern- und Prüfungsmaterialien sowie für die Lehrplangestaltung, Kursplanung und Qualitätskontrolle eingesetzt. Zukünftige Lernmanagementsysteme könnten Lehrende bei der Inhaltsvorbereitung und -anpassung effizient unterstützen. Herausforderungen ergeben sich in Bezug auf Urheberrechtsfragen sowie ethische und rechtliche Rahmenbedingungen. Die potenziell fehlerhafte Natur generierter KI-Inhalte erfordert ein kritisches Hinterfragen. Die Rolle der Lehrenden verändert sich durch die Nutzung von KI-basierten Werkzeugen, die administrative und analytische Aufgaben übernehmen. Dadurch verschiebt sich der Fokus auf individuelle Betreuung und pädagogische Kernaufgaben (Brandhofer et al., 2024). Die Entwicklung von KI-Kompetenzen wird somit zu einer Schlüsselkompetenz für alle Bildungsbeteiligten (Pinski & Benlian, 2024).

Im Kontext der Leistungsbewertung kann KI die Entwicklung von Tests automatisieren und rationalisieren. Die Hauptherausforderung besteht darin, dass KI-Algorithmen zur Bewertung wissensbasierter Aufgaben dieselben Konstruktattribute erfassen müssen wie menschliche Beurteilende. Der zentrale Vorteil liegt in der zeit-

Tab. 1: System 1 des VSM: Analyse zentraler KI-Anwendungen in Lehre, Forschung und Services

KI-Anwendung	Bereich	Potenzielle Vorteile	Herausforderungen
Adaptive Lernpfade und Learning Analytics	Lehre	Adaptives Lernen, frühzeitige Unterstützung, datengestützte Kursoptimierung	Datenschutz, algorithmische Verzerrungen, mögliche Einschränkung der Lernautonomie
KI-Werkzeuge für Lehre und Forschung	Lehre, Forschung	Effizienz bei der Materialerstellung und Datenanalyse, Unterstützung komplexer Forschungsprozesse	Qualität und Bias von KI-generierten Inhalten, Urheberrechtsfragen, Anpassung von Rollen und Kompetenzen
Automatisierte Bewertungssysteme und Feedback	Lehre	Schnelle, skalierbare Rückmeldung, Potenzial für objektivere Bewertung	Validität bei komplexen Aufgaben, algorithmische Verzerrungen, rechtlich problematisch bei beurteilungsrelevanten Entscheidungen
Chatbots und virtuelle Assistenten	Lehre, Services	Informationszugang rund um die Uhr, Entlastung des Personals, interaktive Lernmöglichkeiten	Antwortqualität bei komplexen Anfragen, Akzeptanzbarrieren, didaktische Gestaltung für Lernanwendungen
KI in der Forschung	Forschung	Beschleunigte Literaturrecherche und Datenanalyse, Unterstützung bei Hypothesengenerierung und Modellierung	Datenqualität und Verzerrungen, eingeschränkte Nachvollziehbarkeit komplexer Modelle, Tendenz zur Bestätigung bestehender Annahmen
KI für Kollaboration und Publikation	Forschung	Unterstützung bei Partnerfindung und gemeinsamer Arbeit, Erleichterung der Publikationsprozesse	Datenschutz bei Datenaustausch, Fragen des geistigen Eigentums, Sicherstellung wissenschaftlicher Integrität
KI in Administration und Services	Services	Automatisierung von Routineaufgaben, potenziell personalisierte Services, Effizienzsteigerung	Datenschutz, mögliche Entpersonalisierung, Integration in bestehende Systeme, Akzeptanz bei Mitarbeitenden

nahen, informativen Rückmeldung zur Steuerung des Lernprozesses, während die Validitätssicherung problematisch bleibt, da spezifische Bewertungsrichtlinien für KI-basierte Assessments fehlen (Kaldaras et al., 2024).

Als Schnittstelle zwischen Lehre und Service gewinnen Chatbots und virtuelle Assistenten sowohl in der Lehre als auch im Servicebereich an Bedeutung. Sie kommen bei der Beantwortung organisatorischer Fragen, der Simulation von Gesprächssituationen und der inhaltlichen Vertiefung zum Einsatz (Stampfl et al., 2024). Im Rahmen dieses Beitrags wird argumentiert, dass ihre kontinuierliche Verfügbarkeit einen echten Mehrwert für Studierende bietet und gleichzeitig Verwaltungspersonal entlastet. Die Qualität der generierten Antworten, insbesondere bei komplexen Anfragen, wird maßgeblich durch die Systemarchitektur und -konfiguration bestimmt. Erfahrungen aus der eigenen Lehrpraxis verdeutlichen, dass Akzeptanzfaktoren und die didaktische Einbettung entscheidend für die Wirksamkeit dieser Systeme als Lernunterstützung sind.

Wendet man den Blick auf die Forschung, ermöglicht KI eine Beschleunigung der Mustererkennung in datenintensiven Bereichen, was wiederum den Zugang zu neuen Forschungsfragen eröffnet (Erduran & Levrini, 2024). Diese Form der optimierten Datenanalyse beeinflusst die Generierung von Hypothesen. Darüber hinaus erschließen KI-gestützte Werkzeuge neue Wege in der Literaturrecherche (Brommer & Heimgartner, 2024). Allerdings sind mit dem Einsatz von KI auch epistemologische Herausforderungen verbunden, insbesondere in Bezug auf die Qualität der Daten und die erforderliche kritische Kompetenz der Forschenden im Umgang mit algorithmischen Vorschlägen (Koskinen, 2024). Hierzu zählen Fragen der Vorhersagegenauigkeit sowie der eingeschränkten Interpretierbarkeit komplexer Modelle.

Über die Datenanalyse hinaus können KI-Tools in der Forschung, der Kollaboration und dem Publikationsprozess vielfältige Unterstützung leisten. Sie erleichtern die Suche nach Kooperationspartnern, die Literaturanalyse und gemeinsame Schreibprozesse. Darüber hinaus assistieren KI-Tools beim Verfassen von Texten, der Integration von Literatur und der Erweiterung von Forschungsübersichten, während Übersetzungsfunktionen die interdisziplinäre und transkulturelle Forschung fördern. Herausforderungen bestehen im Datenschutz beim Datenaustausch, bei Fragen des geistigen Eigentums an KI-generierten Inhalten und bei der Sicherstellung wissenschaftlicher Integrität. Die Unterscheidung zwischen menschlichem und KI-generiertem Output wirft Plagiatsfragen auf (Buck et al., 2023).

Im dritten Kernbereich, dem Servicebereich von Universitäten, fokussieren aktuelle KI-Anwendungen auf die Automatisierung von Routineaufgaben durch Chatbots. Dies betrifft primär die Beantwortung von Studierendenanfragen außerhalb des Lehrbetriebs, besonders in den Bereichen Studienzulassung und -einschreibung. Trotz des Potenzials für personalisierte Services im Bildungssektor liegt der gegenwärtige Schwerpunkt im administrativen Bereich auf der Automatisierung genereller Anfragen. Es bestehen jedoch Bedenken hinsichtlich möglicher Verzerrungen (Bias) in generativen KI-Modellen (Peyton et al., 2025).

Zusammenfassend lässt sich festhalten, dass System 1 der zentrale Ort ist, an dem sich die Transformation durch KI in der Hochschule konkretisiert. Tabelle 1 zeigt, dass die Vielfalt der KI-Anwendungen nicht als ein technologisches Einzelthema zu verstehen ist, sondern vielmehr eine umfassende Veränderung der operativen Logiken und Interaktionen darstellt.

2.2 System 2: Koordination – Abstimmung zwischen Einheiten

Nach einer Darstellung der operativen Einheiten (System 1) fokussiert dieser Abschnitt auf System 2 des Viable System Model (VSM). Dieses beschreibt die Koordinationsmechanismen zwischen den Einheiten – Strukturen, die Abstimmung ermöglichen, Überschneidungen regulieren und Reibungsverluste verhindern. Im Kontext der Hochschulen umfasst dies die Synchronisierung von Lehre, Forschung und administrativen Abläufen. Die Analyse zeigt, dass der Einsatz von KI in diesen Koordinationsprozessen Effizienzpotenziale bietet, jedoch Fragen zu Qualität, Flexibilität und Akzeptanz aufwirft. Die nachfolgende Tabelle 2 präsentiert eine Auswahl an potenziellen Anwendungsbereichen.

Tab. 2: System 2 des VSM: Analyse zentraler KI-Anwendungen in der Koordination

KI-Anwendung	Potenzielle Vorteile	Herausforderungen
Weiterentwicklung von Studierendenservices	Straffung des Immatrikulationsprozesses, Verbesserung des Studierendenservices, personalisierte Unterstützung durch Chatbots, Unterstützung durch individuelles Feedback	Datenschutzverletzungen, unethische Datenerhebung, reduzierte menschliche Interaktion
KI-gestützte Kommunikation und Texterstellung	Konsistente Standardisierung von Kommunikationsprozessen, erhöhte Effizienz bei wiederkehrenden Mitteilungen, automatisierte Feedback-Mechanismen	Depersonalisierung der Hochschulkommunikation, unzureichende Kontextsensitivität, Datenschutzbedenken
Empfehlungssysteme für interdisziplinäre Kooperationen	Katalysator für Synergieeffekte zwischen Disziplinen, Förderung innovativer Forschungsansätze, optimierte Nutzung institutioneller Expertise	Herausforderungen bei der Datenintegration, Akzeptanzbarrieren auf Seiten der Fakultäten, methodische Limitationen der Empfehlungen

Ein zentrales Anwendungsfeld von KI an Hochschulen ist die Studierendenbetreuung. Die Implementierung von KI kann zu einer Optimierung der Immatrikulationsprozesse führen und den Service für Studierende verbessern. Chatbots könnten personalisierte Unterstützung bei der Finanzierung, Terminplanung und Beratung bieten sowie Lehrende beim individuellen Feedback unterstützen. Es ist jedoch zu beachten, dass mit der Implementierung von KI auch Risiken einhergehen, beispielsweise Datenschutzverletzungen, unethische Datenerhebung und eine potenzielle Reduktion menschlicher Interaktion (Dempere et al., 2023).

Neben administrativen Aspekten transformiert KI auch die akademische Kommunikation. Ou et al. (2024) zeigen auf, dass Studierende KI-basierte Sprachwerkzeuge bereits umfassend für akademische Kommunikation nutzen, was eine Verbesserung der kommunikativen Leistung zur Folge haben kann. Dies könnte effizientere Abstimmungsprozesse zwischen Hochschulbereichen fördern, etwa durch KI-Automatisierung wiederkehrender Mitteilungen oder verbesserte Informationsvermittlung. Hinsichtlich der Qualität und Zuverlässigkeit generierter Inhalte in spezialisierten akademischen Bereichen sowie bezüglich der Depersonalisierung der Kommunikation und der mangelnden Kontextsensitivität bestehen jedoch Bedenken (Ou et al., 2024).

In der Forschung ermöglichen KI-basierte Systeme neue Formen der Kooperation. Scholarly Recommendation Systems sind Empfehlungssysteme, die auf KI basieren und Forschende bei der Identifizierung relevanter wissenschaftlicher Ressourcen wie Literatur, Datensätze, Forschungsförderung und Kollaborationspartner unterstützen. Ihr Ziel ist es, den Forschenden Zeit zu ersparen und ihnen unbekannte, aber relevante Informationen zugänglich zu machen (Zhang et al., 2023). Die Effektivität dieser Systeme hängt jedoch maßgeblich von der Datenqualität ab. Zudem bestehen methodische Einschränkungen algorithmischer Vorschläge und Akzeptanzbarrieren.

Die in Tabelle 2 dargestellten Anwendungsfelder zeigen, dass KI die hochschulinterne Koordination funktional unterstützen kann. Diese Betrachtung über verschie-

dene Anwendungsbereiche hinweg legt nahe, dass die Implementierung von KI nicht technikgetrieben, sondern an den realen Bedürfnissen, Kommunikationsformen und Kooperationskulturen der Organisation ausgerichtet erfolgen sollte.

2.3 System 3: Steuerung & Ressourcen – Management, Kontrolle

Nach der Betrachtung der operativen Einheiten (System 1) und ihrer Koordination (System 2) widmet sich dieser Abschnitt dem System 3 des Viable System Model (VSM). Dieses System umfasst die Steuerungs- und Ressourcenfunktionen – konkret die Führung und Steuerung der Hochschule als Gesamtorganisation. Im Hochschulkontext wird System 3 durch das Rektorat in seiner operativen Führungsrolle und das Qualitätsmanagement verkörpert. Der Einsatz von Künstlicher Intelligenz (KI) in diesen Steuerungsprozessen bietet erhebliche Potenziale zur datengestützten Entscheidungsfindung und Ressourcenoptimierung, wirft aber gleichzeitig Fragen nach der Nachvollziehbarkeit algorithmischer Entscheidungen und ethischen Implikationen auf. Die folgende Tabelle 3 gibt einen Überblick über exemplarische KI-Anwendungen für System 3.

Tab. 3: System 3 des VSM: Analyse zentraler KI-Anwendungen für Steuerung und Ressourcen

KI-Anwendung	Potenzielle Vorteile	Herausforderungen
Datengestützte Entscheidungsfindung	Evidenzbasierte Entscheidungsprozesse, verbesserte Transparenz von Leistungsindikatoren, Systematisierung komplexer Hochschuldaten	Interpretationsschwierigkeiten bei multidimensionalen Daten, algorithmisch bedingte Intransparenz, ethische Implikationen datengestriebener Entscheidungen
Prädiktive Analytik für Studienabbrüche und Planung	Implementierung effektiver Frühwarnsysteme, gezielte Interventionsmöglichkeiten, Senkung der Studienabbrecherquoten	Gefahr von Fehlprognosen, Potenzial diskriminierender Algorithmen, ethisch-moralische Grenzen präventiver Maßnahmen
KI-gestützte Qualitätssicherungssysteme	Automatisierte Erstellung von Qualitätsberichten, umfassende Analyse heterogener Datenbestände, kontinuierliche Überwachung von Qualitätsparametern	Die Überbewertung quantitativer gegenüber qualitativen Indikatoren, mangelnde Transparenz der Bewertungsalgorithmen, mangelnde Akzeptanz bei den Hochschulangehörigen

Für eine datenbasierte Hochschulsteuerung eröffnen sich durch den Einsatz von KI neue Möglichkeiten. Die Verwendung von KI in Hochschulen erlaubt die automatisierte Extraktion, Analyse und Klassifizierung von Datensätzen aus diversen Systemen. Infolgedessen ist es der Hochschulleitung möglich, Managemententscheidungen auf breiterer Informationsbasis zu treffen. Leistungskennzahlen zu Studierendeneinschreibung, Studienfortschritt, Forschungsaktivitäten und Personalentwicklung werden systematisch erfasst und über intuitive Dashboards visualisiert. Dadurch wird eine Analyse von Trends in Studierendenzahlen oder Forschungsleistungen für strategische Entscheidungen ermöglicht. Es ist jedoch zu bedenken, dass der Einsatz

algorithmischer Systeme das Risiko intransparenter, schwer nachvollziehbarer Entscheidungsprozesse birgt. Dies kann zu normativen Entscheidungen mit unzureichend berücksichtigten ethischen Implikationen führen (Gaftandzhieva et al., 2023).

Neben der Gesamtsteuerung unterstützt KI auch die gezielte Intervention im Studienbetrieb. Prädiktive Analytik kann in der Studienplanung eingesetzt werden, um gefährdete Studierende frühzeitig zu identifizieren und gezielt zu unterstützen. Gonzalez-Nucamendi et al. (2023) beschreiben, wie maschinelles Lernen anhand quantitativer Variablen abbruchgefährdete Studierende identifizieren kann. Der Einsatz solcher Frühwarnsysteme kann dazu beitragen, Studienabbrüche zu reduzieren und Ressourcen effizienter zu nutzen. Die in der Studie diskutierten Effektivitätskoeffizienten verdeutlichen jedoch, dass nicht jede*r als gefährdet eingestufte Studierende tatsächlich Interventionen benötigt. Dies wirft ethische Fragen bezüglich des Ressourceneinsatzes und möglicher Einschränkungen der studentischen Autonomie auf (Gonzalez-Nucamendi et al., 2023).

Ein weiteres zentrales Anwendungsfeld stellt das Qualitätsmanagement dar. Die Verwendung von KI in der Qualitätssicherung im Bildungsbereich erlaubt die automatisierte Analyse von Feedback, das im Rahmen von Befragungen erhoben wird. Diese Technologie verbessert die effiziente und systematische Erfassung von Qualitätssignalen aus qualitativen Daten. Darüber hinaus eröffnet sich ein breites Spektrum an weiteren Anwendungsfeldern, welche die Identifikation von Themen, die Extraktion von Verbesserungsvorschlägen sowie Stimmungsanalysen umfassen. Es ist jedoch zu beachten, dass eine Übergewichtung quantitativer Indikatoren eintreten kann. Für die Generierung nachvollziehbarer Ergebnisse ist zudem die Sicherstellung der Transparenz der Bewertungsprozesse von essentieller Bedeutung (Parker et al., 2024).

In der Gesamtbetrachtung dieser Aspekte demonstrieren die in Tabelle 3 dargestellten Anwendungen das Potenzial von KI-Systemen zur Steigerung der Effizienz und Evidenzbasierung in der Hochschulsteuerung. Für einen verantwortungsvollen Einsatz der Systeme ist eine kritische Begleitung sowie die Integration in ein transparentes, partizipatives Steuerungssystem unerlässlich. Zudem sind infrastrukturelle und regulatorische Anforderungen zu berücksichtigen. Datenintensive KI-Anwendungen stellen hohe Anforderungen an die verfügbaren Rechenkapazitäten und erfordern eine nachhaltige IT-Strategie. Zudem gewinnen rechtliche Rahmenbedingungen wie der AI Act sowie datenschutzrechtliche Vorgaben an Bedeutung für eine verantwortungsvolle KI-Governance (Brandhofer et al., 2024).

2.4 System 4: Strategie & Umwelt – Planung, externe Beobachtung

Nach der Analyse der operativen, koordinativen und steuernden Funktionen (Systeme 1–3) des Viable System Model (VSM) folgt die Betrachtung von System 4, das für die strategische Ausrichtung und Umweltbeobachtung der Hochschule verantwortlich ist. Die Hauptaufgabe von System 4 besteht in der Sicherstellung der Zukunftsfähigkeit durch Analyse externer Entwicklungen und Ableitung strategischer Handlungsoptionen. In diesem Beitrag wird die Auffassung vertreten, dass Künstliche Intelligenz (KI) die systematische Auswertung großer Datenmengen, frühzeitige Trenderkennung und Unterstützung komplexer Planungsprozesse ermöglicht. Die zentralen Anwendungsfelder werden in Tabelle 4 dargestellt.

Tab. 4: System 4 des VSM: Analyse zentraler KI-Anwendungen für Strategie und Umwelt

KI-Anwendung	Potenzielle Vorteile	Herausforderungen
Marktanalyse und Trendbeobachtung mit KI	Systematische Erfassung zukünftiger Bildungstrends, bedarfsorientierte Angebotsgestaltung, Stärkung der Hochschule im akademischen Wettbewerb	Zuverlässigkeit und Validität externer Datenquellen, Dynamik rapider gesellschaftlicher Veränderungen, Komplexität der Interpretation prognostischer Modelle
Entwicklung neuer Programme mit KI-Unterstützung	Konzeption innovativer Curricula, evidenzbasierte Gestaltung für spezifische Zielgruppen, proaktive Anpassung an Entwicklungen des Arbeitsmarktes	Validierungsaufwand für KI-gestützte Curriculumsvorschläge, Herausforderungen bei der praktischen Implementierung, Risiko der Technologieabhängigkeit
KI-Matching für Hochschulk Kooperationen	Strategische Optimierung der Drittmittelakquise, effiziente Identifikation geeigneter institutioneller Partner, datenbasierte Kooperationspotenziale	Verzerrte Bewertung potenzieller Kooperationen, unvollständige Datenbasis für Partnerschaftsanalysen, notwendige Ergänzung durch personelle Expertise

Im Bereich der Wissenschaftsanalyse eröffnen sich durch den Einsatz von KI neue Perspektiven für strategische Entscheidungen. KI-gestützte bibliometrische Analysen wie die von Kavitha et al. (2024) über künstliche Intelligenz im Hochschulwesen ermöglichen eine systematische Beobachtung der Forschungstrends in diesem sich entwickelnden Feld. Diese Analysen liefern wertvolle Daten, mit deren Hilfe Hochschulen nicht nur zukünftige Bildungstrends erkennen, sondern auch ihre Studienangebote bedarfsorientiert anpassen können. Dies verschafft ihnen letztendlich einen Vorteil im akademischen Wettbewerb. Die Studie von Kavitha et al. (2024) stützt diese Annahme und zeigt beispielsweise einen exponentiellen Anstieg der Publikationen zum Thema KI im Hochschulwesen, wobei China und die Vereinigten Staaten als die einflussreichsten Länder hervortreten. Dieser Publikationsanstieg reflektiert die rapide Entwicklung des Fachgebiets, die insbesondere durch das Aufkommen neuer Themen wie maschinelles Lernen und Convolutional Neural Networks gekennzeichnet ist. Die Interpretation der Analyseergebnisse, beispielsweise Co-Autorennetzwerke, Keyword-Co-Vorkommnisse und Co-Zitationsanalysen, gestaltet sich jedoch komplex und erfordert spezifisches Fachwissen.

Während KI-Analysen das akademische Umfeld erfassen, kann KI auch die interne Bildungsplanung verändern. Die Anwendung von KI in der Studiengangsentwicklung bietet das Potenzial, Curricula zu modernisieren und eine zielgruppenorientierte Planung zu ermöglichen. Darüber hinaus kann KI dazu beitragen, Lernerfahrungen zu personalisieren und Inhalte an die individuellen Bedürfnisse der Lernenden anzupassen. Die Integration von KI in die Curriculumsentwicklung wird maßgeblich durch den häufigen Einsatz von KI, das Wissen der Lehrenden über KI und institutionelle Unterstützung gefördert. Trotz dieser Vorteile sehen sich Bildungseinrichtungen mit erheblichen Herausforderungen konfrontiert, zu denen ethische Bedenken sowie die mangelnde Fähigkeit der KI, kulturelle und bildungspolitische

Kontexte adäquat zu berücksichtigen, zählen. Zudem wird die Rolle von KI in der Curriculumsentwicklung maßgeblich von den Wahrnehmungen und Einstellungen der Lehrenden beeinflusst (Abbasi et al., 2025). Darüber hinaus kann die gezielte Entwicklung innovativer Studienangebote auf Basis von KI-gestützten Trendanalysen die strategische Positionierung der Hochschule im Wettbewerb um Drittmittel und Kooperationen nachhaltig stärken.

Neben der inhaltlichen Weiterentwicklung umfasst die strategische Dimension auch die Vernetzung im Wissenschaftsökosystem. Ein algorithmusbasierter Matching-Ansatz demonstriert das Potenzial von KI-gestützten Systemen, geeignete Forschungspartner durch automatisierte Identifikation auf Basis von Forschungsinteressen und Profilmerkmalen zu ermitteln (Lund, 2023). Bei diesem Verfahren werden verschiedene Datenquellen wie Publikationsdaten, Biografien und Themenklassifikationen kombiniert. Die Methode zeichnet sich durch hohe Effizienz und Präzision aus, jedoch bleiben weiche Faktoren wie institutionelle Kompatibilität oder gemeinsame Zielvorstellungen unberücksichtigt. Ein weiteres Problem stellen unvollständige Daten dar, die zu algorithmischen Verzerrungen führen können und somit eine ergänzende menschliche Bewertung erforderlich machen.

Die Gesamtbetrachtung dieser unterschiedlichen strategischen Anwendungsbeispiele zeigt, dass die in Tabelle 4 dargestellten KI-Anwendungen im strategischen Hochschulmanagement unterstützend wirken können. Basierend auf diesen Erkenntnissen legt der vorliegende Beitrag nahe, dass ein solcher Nutzen nur unter der Voraussetzung realisiert werden kann, dass der Einsatz von KI transparent erfolgt, eine kritische Begleitung stattfindet und diese Technologie nicht als Ersatz, sondern vielmehr als Erweiterung menschlicher Urteils- und Gestaltungsfähigkeit verstanden wird.

2.5 System 5: Identität & Werte – Leitbild, Ethik, Positionierung

Nach einer Analyse der vier vorangegangenen Systeme widmet sich der folgende Abschnitt dem fünften System des Viable System Model (VSM). Dieses steht für die institutionelle Selbstbeschreibung der Hochschule, die Werteorientierung sowie die normative Einbettung. Dieses zentrale System bildet das Fundament für das Leitbild, die ethischen Prinzipien sowie die Positionierung der Hochschule im gesellschaftlichen Raum. In diesem Zusammenhang wird im vorliegenden Beitrag die Position vertreten, dass sich in Zeiten der digitalen Transformation die Frage stellt, wie KI nicht nur als Werkzeug, sondern vielmehr als Teil des institutionellen Selbstverständnisses verstanden und integriert werden kann. Zur Konkretisierung dieser konzeptionellen Überlegungen veranschaulicht Tabelle 5 beispielhaft zentrale Anwendungsfelder der KI, die für die Identität und Werte einer Hochschule von Relevanz sind.

Die Integration von KI in hochschulische Leitbilder manifestiert sich bereits in ersten praktischen Ansätzen. Die Studie von Wu et al. (2024) über KI-Leitlinien an Universitäten verdeutlicht, dass die strategische Verankerung von KI in Hochschulstrategien durch die Veröffentlichung von Richtlinien erfolgt, mit dem Ziel, die digitale Transformation nicht nur operativ, sondern auch normativ zu gestalten und eine zukunftsorientierte Profilierung zu ermöglichen. Doch ungeachtet dieser positiven In-

Tab. 5: System 5 des VSM: Analyse zentraler KI-Anwendungen für Identität und Werte

KI-Anwendung	Potenzielle Vorteile	Herausforderungen
Verankerung von KI in Hochschulstrategien	Zukunftsorientierte Hochschulentwicklung, klare institutionelle Positionierung, kohärente digitale Transformationsstrategien	Ausbalancierung von Innovation und akademischer Tradition, Gefahr oberflächlicher Symbolpolitik, Sicherstellung nachhaltiger Implementierung
Ethische Richtlinien und KI-Governance	Förderung verantwortungsvoller Technologienutzung, Erhöhung der gesellschaftlichen Akzeptanz, Etablierung eines normativen Orientierungsrahmens	Divergierende Auffassungen über ethische Standards, Komplexität der praktischen Implementierung, Spannungsfeld zwischen Regulierung und Innovationsfreiheit
KI-basierte Profilierung im Hochschulmarketing	Steigerung der Attraktivität für potenzielle Studierende, Differenzierung im interinstitutionellen Wettbewerb, Modernisierung des institutionellen Images	Risiko einer technologiezentrierten Außendarstellung, potenzieller Verlust an Authentizität, unangemessene Betonung technologischer Aspekte gegenüber inhaltlichen Aspekten

tentionen besteht die Gefahr, dass strategische Erwähnungen von KI ohne substanzielle Umsetzung bleiben, da die vorwiegend edukative und beratende Natur dieser Leitlinien konkrete Implementierungen erschwert. Wu et al. (2024) betonen die Notwendigkeit, technologische Ambitionen mit den tradierten Werten akademischer Bildung in Einklang zu bringen, um eine glaubwürdige Integration zu gewährleisten.

Die normative Dimension dieser Integration ist ein viel diskutiertes Thema in der Fachliteratur. Holmes et al. (2022) führen aus, dass die Implementierung ethischer Richtlinien und KI-Governance in der Bildung darauf abzielt, die verantwortungsvolle Nutzung von KI-Technologien zu fördern und durch normative Orientierung die gesellschaftliche Akzeptanz zu erhöhen. Die Forschungsgemeinschaft sieht sich jedoch mit Herausforderungen konfrontiert, die sich aus unterschiedlichen ethischen Auffassungen ergeben. So wird die praktische Umsetzung nicht nur durch divergierende Wertvorstellungen, sondern auch zusätzlich durch das Fehlen spezifischer Rahmenwerke für die komplexen ethischen Fragestellungen erschwert. Zentraler Gegenstand der Kontroverse ist das Spannungsverhältnis zwischen dem Regulierungsbedarf hinsichtlich ethischer Praktiken und dem Erfordernis der Bewahrung von Innovationsfreiheit in einem von hoher Dynamik geprägten Forschungsumfeld (Holmes et al., 2022).

Neben den normativen Aspekten eröffnet KI auch neue strategische Handlungsfelder für Hochschulen. Im Kontext des Hochschulmarketings gewinnt die KI-basierte Personalisierung zunehmend an Relevanz zur Optimierung der Ansprache potenzieller Studierender und Steigerung der Einschreibungszahlen. Durch den Einsatz fortschrittlicher Datenanalyseverfahren können Hochschulen ihre (zukünftigen) Studierenden besser verstehen und zielgerichtete Botschaften, Inhalte und Dienstleistungen entwickeln. Insbesondere der Einsatz von KI-gestützten Chatbots, personalisierten Mailinglisten und gezielter Onlinewerbung ermöglicht eine individuelle Betreuung und den Aufbau von Beziehungen. Diese Marketingstrategien zielen darauf ab, potenzielle Studierende anzusprechen und deren Wahrnehmung positiv zu

beeinflussen, was zur Differenzierung im interinstitutionellen Wettbewerb beiträgt. Kritisch anzumerken ist jedoch, dass der Einsatz von KI im Hochschulmarketing an Grenzen hinsichtlich des Datenschutzes und der digitalen Kompetenz stößt (Meenakshi & Kapil, 2025).

Zusammenfassend lässt sich festhalten, dass die in Tabelle 5 dargestellten KI-Anwendungen verdeutlichen, dass KI in System 5 nicht als neutrale Technologie, sondern als kontextsensibles Element institutioneller Selbstverständigung zu betrachten ist. Vor diesem Hintergrund vertritt der vorliegende Beitrag die Perspektive, dass die KI-Integration eine ethische Reflexion, strategische Klarheit und eine Kultur erfordert, die technologische Innovation mit Verantwortung und Identität verbindet.

3. Fazit und Ausblick

Die Analyse entlang der fünf Systemebenen des Viable System Model zeigt, dass KI nicht als isolierte Digitalisierungsmaßnahme betrachtet werden kann, sondern eine tiefgreifende Transformation für Hochschulen in den Bereichen Lehre, Forschung, Verwaltung und Strategie auslöst. Der Einsatz von KI verändert nicht nur Prozesse, sondern wirft auch fundamentale Fragen zu Rollen, Verantwortlichkeiten und institutionellem Selbstverständnis auf.

Die vielschichtige Transformation erfordert einen ganzheitlichen Ansatz. Hochschulen stehen vor der Herausforderung, KI nicht nur technisch zu implementieren, sondern normativ zu gestalten. Sie tragen eine doppelte Verantwortung. Einerseits müssen sie KI kompetent für effizientere, individuellere und datenbasiertere Aufgabenerfüllung einsetzen. Andererseits obliegt ihnen als wissenschaftlichen Institutionen die Schaffung von Reflexions-, Kritik- und Mitgestaltungsräumen.

In diesem Spannungsfeld zwischen technologischer Innovation und kritischer Reflexion definiert sich die neue Rolle der Hochschule. Die erfolgreiche Hochschule der Zukunft entwickelt sich zum Resonanzraum für Technologieverständnis, zur Vermittlerin von KI-Kompetenz und zur Akteurin einer verantwortungsvollen digitalen Gesellschaft. Ihre Funktion umfasst mehr als die bloße Adaption technologischer Systeme, sondern beinhaltet auch die bewusste Mitgestaltung neuer Mensch-Maschine-Interaktionskonstellationen.

Literatur

- Abbasi, B. N., Wu, Y. & Luo, Z. (2025). Exploring the impact of artificial intelligence on curriculum development in global higher education institutions. *Education and Information Technologies*, 30(1), 547–581. <https://doi.org/10.1007/s10639-024-13113-z>
- Beer, S. (2007). *Diagnosing the system for organizations* (Classic Beer Series) (Repr). Wiley.
- Brandhofer, G., Gröbinger, O., Jadin, T., Raunig, M. & Schindler, J. (Hrsg.). (2024). *Von KI lernen, mit KI lehren: Die Zukunft der Hochschulbildung. Projektbericht*. Books on Demand. <https://fnma.at/content/download/2990/19034>
- Brommer, S. & Heimgartner, S. (2024). *KI-basierte Tools beim wissenschaftlichen Schreiben: Mit Studierenden gemeinsam Anwendungen erproben und evaluieren*. VK: KIWA. <https://zenodo.org/doi/10.5281/zenodo.12097857>

- Buck, I., Jost, C., Kreis-Hoyer, P. & Limburg, A. (2023). *KI-induzierte Transformation an Hochschulen*. Hochschulforum Digitalisierung. <https://hochschulforumdigitalisierung.de/wp-content/uploads/2023/11/HSRM-Diskussionspapier-Nr.-26-KI-induzierte-Transformation-an-Hochschulen-1.pdf>
- Dempere, J., Modugu, K., Hesham, A. & Ramasamy, L. K. (2023). The impact of Chat-GPT on higher education. *Frontiers in Education*, 8, 1206936. <https://doi.org/10.3389/feduc.2023.1206936>
- Erduran, S. & Levrini, O. (2024). The impact of artificial intelligence on scientific practices: An emergent area of research for science education. *International Journal of Science Education*, 46(18), 1982–1989. <https://doi.org/10.1080/09500693.2024.2306604>
- Gaftandzhieva, S., Hussain, S., Hilcenko, S., Doneva, R. & Boykova, K. (2023). Data-driven Decision Making in Higher Education Institutions: State-of-play. *International Journal of Advanced Computer Science and Applications*, 14(6). <https://doi.org/10.14569/IJACSA.2023.0140642>
- Gonzalez-Nucamendi, A., Noguez, J., Neri, L., Robledo-Rella, V. & García-Castelán, R. M. G. (2023). Predictive analytics study to determine undergraduate students at risk of dropout. *Frontiers in Education*, 8, 1244686. <https://doi.org/10.3389/feduc.2023.1244686>
- Holmes, W., Porayska-Pomsta, K., Holstein, K., Sutherland, E., Baker, T., Shum, S. B., Santos, O. C., Rodrigo, M. T., Cukurova, M., Bittencourt, I. I. & Koedinger, K. R. (2022). Ethics of AI in Education: Towards a Community-Wide Framework. *International Journal of Artificial Intelligence in Education*, 32(3), 504–526. <https://doi.org/10.1007/s40593-021-00239-1>
- Jin, Y., Yan, L., Echeverria, V., Gašević, D. & Martinez-Maldonado, R. (2025). Generative AI in higher education: A global perspective of institutional adoption policies and guidelines. *Computers and Education: Artificial Intelligence*, 8, 100348. <https://doi.org/10.1016/j.caeai.2024.100348>
- Kaldaras, L., Akaze, H. O. & Reckase, M. D. (2024). Developing valid assessments in the era of generative artificial intelligence. *Frontiers in Education*, 9, 1399377. <https://doi.org/10.3389/feduc.2024.1399377>
- Katsamakos, E., Pavlov, O. V. & Saklad, R. (2024). Artificial Intelligence and the Transformation of Higher Education Institutions: A Systems Approach. *Sustainability*, 16(14), 6118. <https://doi.org/10.3390/su16146118>
- Kavitha, K., Joshith, V. P., Rajeev, N. P. & S, A. (2024). Artificial Intelligence in Higher Education: A Bibliometric Approach. *European Journal of Educational Research*, 13(3), 1121–1137. <https://doi.org/10.12973/eu-jer.13.3.1121>
- Koskinen, I. (2024). We Have No Satisfactory Social Epistemology of AI-Based Science. *Social Epistemology*, 38(4), 458–475. <https://doi.org/10.1080/02691728.2023.2286253>
- Lund, B. (2023). *Researchers Matching for Collaboration: A Novel Algorithm-Based Approach Using Cosine Similarity*. <https://dx.doi.org/10.2139/ssrn.4346965>
- Meenakshi, M. & Kapil, K. (2025). AI-Driven Personalization in Higher Education Marketing: Impact on Student Engagement and Enrolment Decisions. *European Economic Letters (EEL)*, 15(2), 171–186. <https://doi.org/10.52783/eel.v15i2.2831>
- Misiejuk, K., Samuelsen, J., Kaliisa, R. & Prinsloo, P. (2025). Idiographic learning analytics: Mapping of the ethical issues. *Learning and Individual Differences*, 117, 102599. <https://doi.org/10.1016/j.lindif.2024.102599>
- Ocen, S., Elasu, J., Aarakit, S. M. & Olupot, C. (2025). Artificial intelligence in higher education institutions: Review of innovations, opportunities and challenges. *Frontiers in Education*, 10, 1530247. <https://doi.org/10.3389/feduc.2025.1530247>
- Ou, A. W., Stöhr, C. & Malmström, H. (2024). Academic communication with AI-powered language tools in higher education: From a post-humanist perspective. *System*, 121, 103225. <https://doi.org/10.1016/j.system.2024.103225>

- Parker, M. J., Anderson, C., Stone, C. & Oh, Y. (2024). A Large Language Model Approach to Educational Survey Feedback Analysis. *International Journal of Artificial Intelligence in Education*. <https://doi.org/10.1007/s40593-024-00414-0>
- Peyton, K., Unnikrishnan, S. & Mulligan, B. (2025). A review of university chatbots for student support: FAQs and beyond. *Discover Education*, 4(1), 21. <https://doi.org/10.1007/s44217-025-00397-7>
- Pinski, M. & Benlian, A. (2024). AI literacy for users – A comprehensive review and future research directions of learning methods, components, and effects. *Computers in Human Behavior: Artificial Humans*, 2(1), 100062. <https://doi.org/10.1016/j.chbah.2024.100062>
- Rezk, S. S. & Gamal, S. (2019). The viable system model and its applications in higher education: An overview. *Kybernetes*, 48(3), 438–450. <https://doi.org/10.1108/K-04-2018-0185>
- Stampfl, R., Geyer, B., Deissl-O'Meara, M. & Ivkic, I. (2024). Revolutionising Role-Playing Games with ChatGPT. *Advances in Artificial Intelligence and Machine Learning*, 4(2), 2244–2257. <https://doi.org/10.54364/AIIML.2024.42129>
- Wu, C., Zhang, H. & Carroll, J. M. (2024). AI Governance in Higher Education: Case Studies of Guidance at Big Ten Universities. *Future Internet*, 16(10), 354. <https://doi.org/10.3390/fi16100354>
- Zhang, Z., Patra, B. G., Yaseen, A., Zhu, J., Sabharwal, R., Roberts, K., Cao, T. & Wu, H. (2023). Scholarly recommendation systems: A literature survey. *Knowledge and Information Systems*, 65(11), 4433–4478. <https://doi.org/10.1007/s10115-023-01901-x>

Autonomy Against the Machine

Clinical Decision-Making and the Politics of AI

Samuel Eke & Nils Hofmann

1. Introduction

Artificial Intelligence (AI) and Machine Learning (ML) are no longer futuristic add-ons in medicine – they are here, FDA-cleared, CE-marked, and already shaping how diagnoses are made, images interpreted, and patients managed. But as algorithms quietly integrate into clinical workflows, a deeper transformation is unfolding: the shifting role of physician and patient care. The present work asks what becomes of physician autonomy when clinical decisions are increasingly influenced, guided, or even constrained – by systems that are often opaque, proprietary, and defy human interpretability.

We begin by unpacking the concept of physician autonomy, grounding it in both ethical and empirical models. From there, we examine how AI is used in practice across radiology, electronic health records, and biosignal analysis, assessing not only its capabilities but also its recurring limitations. Particular attention is given to the discourse around post-hoc explanations, that aim to make black-box models more intelligible. We offer a measured assessment of where such methods support clinical reasoning, and where they risk introducing false confidence or masking epistemic gaps.

In the last two chapters, we first show how physician autonomy cannot be maintained as responsible action in the face of uninterpretable AI. It is precisely the relational responsibility towards patients as answerability that becomes impossible. In conclusion, we propose extending the concept of physician autonomy with the perspective of political autonomy as participation in the design of AI in order to be able to continue to guarantee responsibility towards patients.

2. Physician Autonomy

We begin by exploring the concept of physician autonomy to enable a more nuanced discussion of its status in light of AI advancements. Traditionally, “autonomy” has been defined as “the capacity for self-government” and seen as “a necessary condition of moral agency” (Blackburn, 2016). In contrast, “heteronomy” implies being governed by external forces. Yet, applied to the role of physicians, “physician autonomy” remains insufficiently defined. Grosser and colleagues (2025) highlight the lack of a clear definition in current literature. Drawing on Schulz and Harrison’s (1986) framework, they distinguish between “social and economic work freedom” and “clinical work freedom”. The former includes aspects such as “choice of specialty and practice location”, “control over earnings”, and “control over nature and volume of tasks” (Schulz & Harrison, 1986, p. 342). The latter includes “acceptance of patients”, “control over diagnosis and treatment”, “control over other professionals” and “evaluation of care” (Schulz & Harrison, 1986, p. 342).

AI applications have been discussed in the section of “control over diagnosis and treatment”. Given AI’s ability to process vast datasets, its support in both routine and specialized decision-making is evident. Some argue it could enhance physician autonomy by reducing administrative burdens (e.g., Kocaballi et al., 2020). However, concerns arise if AI assumes a dominant role in clinical decision-making, potentially undermining physicians’ judgment. These concerns extend to the possible replacement of entire specialties, such as radiology (e.g., Jussupow et al., 2022). Such developments could significantly alter physicians’ professional roles, which is why we emphasize the ethical dimensions of AI. Rather than speculating on future task distributions between physicians and AI, we aim to explore the conceptual and ethical implications of AI for physician autonomy in clinical decision-making.

Since the Second World War, autonomy has been discussed in modern medical ethics primarily in terms of the principle of respect for patient autonomy regarding treatment decisions. In principle, Beauchamp and Childress (2019, p. 102) define patient autonomy therefore as autonomous action, understood as decision-making regarding one’s treatment based on three conditions: “Intentionality”, “Understanding”, and “Noncontrol”. The condition of “Understanding” holds particular significance, as it directly responds to the epistemic asymmetry between physicians and patients concerning medical knowledge. Patients place trust in medical expertise as a critical source of information to enable them to make autonomous decisions about their health.

When assessing physician autonomy, it is important to acknowledge the difference in responsibility: patients are primarily accountable to themselves. In contrast, as Funer and Wiesing (2024) note, the physician’s primary responsibility is “to promote the health and well-being of individual patients” (World Medical Association, 2022). Accordingly, in the following text, we therefore understand physician autonomy in clinical decision-making not merely as freedom of choice but further as responsible action towards patients. To conceptualize autonomous clinical decision-making in this way, we propose applying the same standard of autonomy to physicians that is typically applied to patients. Therefore, we draw on the above-mentioned three-condition theory of autonomy by Beauchamp and Childress (2019), which is fundamentally based on Aristotelian theory of action, as succinctly summarized by Coeckelbergh: “You are responsible ‘if (1) you do it (or have done it) – if you are the agent of the action, if you have caused the action, if you have a sufficient degree of control over the action – and (2) if you know what you are doing, if you are aware what you are doing (or knew what you were doing)’” (Coeckelbergh, 2020, p. 2054). Drawing on Fischer and Ravizza (2000) and Rudy-Hiller (2018), Coeckelbergh concludes that responsibility entails both a “control condition” and an “epistemic condition” (Coeckelbergh, 2020, p. 2054). Based on this definition, we argue that with the increasing use of AI in clinical decision-making, a new form of epistemic asymmetry may emerge – this time placing physicians on the receiving end of an informational imbalance – potentially threatening the very foundations of physician autonomy as responsible decision-making.

We will show that when facing epistemically opaque AI systems, physicians may be unable to make truly autonomous responsible decisions. The “black box problem” challenges the epistemic condition of responsible action: once AI systems surpass a certain complexity, their outputs may become incomprehensible to physicians. This

undermines what Schulz and Harrison, building on Tolliday (1978), called the “over-arching primacy of medical knowledge” (Schulz & Harrison, 1986, p. 342) – a cornerstone of physician autonomy and professional legitimacy. In the next chapter, we will examine how clinical AI is already eroding this epistemic primacy in part and will likely continue to do so.

3. Clinical Applications

The implications for physician autonomy grow more urgent as AI systems moved from theoretical possibility to clinical norm. As of this writing, the U.S. Food and Drug Administration (FDA) has approved 1,016 AI- and ML-enabled medical devices for clinical use. Figure 1 provides an overview of the number of approved devices by medical field. In contrast to the FDA, the European Union lacks a centralized registry of CE-marked AI devices; nonetheless, by 2020, at least 240 such tools had already received regulatory approval (Muehlematter et al., 2021).

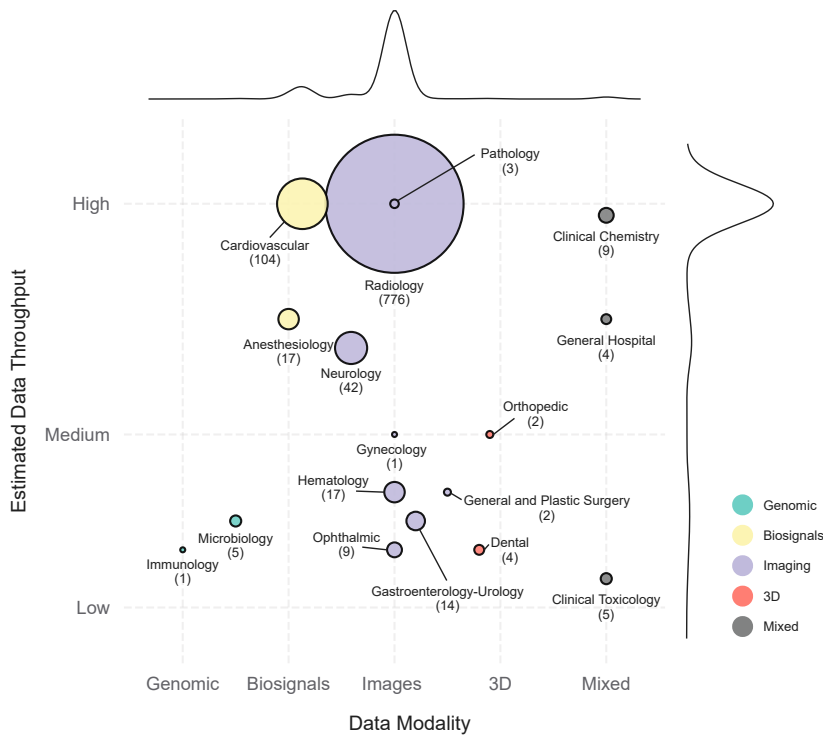


Fig. 1: Number of FDA-approved AI/ML devices by medical specialty. X-axis indicates dominant data modality; Y-axis reflects estimated data throughput based on perceived procedural frequency and guesstimated data yield. The Y-axis primarily serves to enhance visual separation. Bubble size reflects the number of devices approved; color indicates modality group. Marginal kernel density estimates show the distribution of specialties along each axis. Most devices are deployed in imaging with a skew toward high-throughput specialties such as radiology and cardiology

By focusing on AI's key areas of adoption (radiology and cardiology) and future reach (EHRs), we show where it is most established and where it holds the greatest systemic potential. The following section examines representative FDA-approved tools in each domain to better understand how AI is operationalized at the point of care, how it interacts with physicians, and where critical limitations persist.

3.1 AI in Radiology

Radiology remains the leading domain for FDA-approved AI/ML tools, driven by the compatibility of imaging data with CNNs, which extract hierarchical features well-suited for pathology detection. Tools like Siemens' AI-Rad Companion exemplify this: they outperform radiologists in sensitivity for several findings but increase false positives and misidentify non-pathological structures (e.g., nipples, ECG leads) as lesions. Performance improves at high confidence thresholds, yet detection of underrepresented conditions, such as pneumothoraces, remains unreliable, highlighting limitations in dataset diversity and generalizability (Niehoff et al., 2023). Imagen's Chest-CAD, based on ResNet-34 and trained on nearly half a million X-rays, enhances abnormality detection across anatomical categories and raises non-experts to near-expert performance (Anderson et al., 2024). However, it offers no diagnostic conclusions, merely flagging anatomical regions, reinforcing its role as an assistive tool rather than an autonomous system.

This pattern extends into AI-guided image acquisition, particularly in ultrasound. In a study by Narang et al. (2021), nurses without prior ultrasound training used AI guidance to acquire transthoracic echocardiograms with image quality comparable to experienced sonographers for most views. However, performance declined sharply in more complex anatomical areas like the inferior vena cava. Crucially, the nurses had received didactic instruction and performed practice scans, and no control group of unaided novices was included, limiting attribution of results to the AI itself. Similarly, Baloesu et al. (2025) showed that non-expert users could obtain high-quality lung ultrasound scans using real-time AI feedback, with especially strong performance in difficult-to-image zones. Yet again, the study lacked a control group of novices without AI, making it unclear how much improvement stemmed from AI versus basic training. Interestingly, trained physicians using the same AI system achieved perfect image quality, suggesting that AI best amplifies existing expertise.

3.2 AI in Cardiology

AI is increasingly applied to biosignals like ECGs and heart sounds for cardiac screening. Eko Health's FDA-approved murmur classifier, a 2D ResNet, trained on 15,000 recordings, outperformed clinicians in sensitivity (85.6% vs. 77.0%) and specificity (84.4% vs. 78.9%) (Prince et al., 2023), and in a real-world study, more than doubled the sensitivity of primary care physicians for moderate-or-greater valvular disease (94.1% vs. 41.2%), though with lower specificity (84.9% vs. 95.7%) (Rancier et al., 2024). Yet, its clinical value remains limited: performance drops sharply for low-grade murmurs, common and clinically relevant, and its output is

coarse, distinguishing only between structural and non-structural murmurs without localizing the lesion or conveying severity or context. Unlike human clinicians who integrate features like pitch, radiation, and location to refine diagnosis, the model lacks this granularity.

Furthermore, they developed a low ejection fraction (EF) detector, using ECG and phonocardiogram data, which achieved good screening performance (AUROC 0.85, NPV 98.0%) (Guo et al., 2025). A key limitation remains its low positive predictive value of 20.3%, meaning that four out of five patients flagged as having reduced EF were ultimately false positives, raising concerns about unnecessary follow-ups, patient anxiety, and increased healthcare burden. Notably, many false positives had conduction abnormalities, suggesting the model may misclassify certain subgroups and raising concerns about its generalizability and reliability.

3.3 AI in Electronic Health Records

AI models built on structured EHR data are still in their infancy in clinical decision-making, drawing on features like vitals, labs, medications, and demographics to predict deterioration or care needs. These models attempt to synthesize diverse clinical parameters to guide judgment. Most use tree-based methods (e.g., random forests, XGBoost), which handle missingness and nonlinearities well. The Sepsis ImmunoScore exemplifies this approach: it analyzes 22 EHR-derived inputs to assign sepsis risk levels and achieved robust prospective performance across five hospitals ($AUC \geq 0.80$) (Bhargava et al., 2024).

Yet even technically strong models can fail when features misalign with clinical goals. Optum's algorithm, designed to predict future healthcare utilization, used projected cost as a proxy for clinical risk. As a result, it systematically underestimated risk in Black patients, who often received less care despite greater illness burden, highlighting how structural inequities can distort algorithmic outputs (Obermeyer et al., 2019). This case highlights a troubling possibility: as models grow more complex, so too does our inability to detect when they perpetuate bias, creating ethical blind spots in clinical AI.

4. Blind Spots and Hidden Biases

While image-based AI systems have demonstrated potential in clinical workflows, especially in aiding non-experts or supporting negative case confirmation, their broader limitations are recurrent and systematic. Most notably, these tools tend to focus on narrow diagnostic tasks, detecting a predefined set of findings. Systems like AI-Rad or Chest-CAD operate within fixed anatomical categories or binary outputs and are incapable of accounting for incidental, multimodal, or atypical pathologies. Furthermore, from a scientific perspective, many of the studies evaluating these systems lack proper controls, overrepresent common conditions, and underrepresent rare ones, limiting both validity and clinical relevance.

Adding to these concerns, MIT researchers found that deep learning models can accurately predict a patient's self-reported race from medical images, despite race

not being visibly encoded (Gichoya et al., 2022). This remained true even after controlling for image quality, anatomy, and density, suggesting models exploit latent features that humans cannot perceive. As Marzyeh Ghassemi noted, the persistence of this signal raises serious concerns about hidden biases. Leo Anthony Celi warned that including diverse data alone is insufficient and cautioned, “this paper should make us pause and truly reconsider whether we are ready to bring AI to the bedside” (Gordon, 2020). Follow-up work showed that the more accurately models inferred demographic traits like race or gender, the larger their diagnostic disparities became, especially for Black patients, women, and other underrepresented groups (Yang et al., 2024). As co-lead author Haoran Zhang cautioned, fairness in training doesn’t guarantee fairness in deployment. Yet this is precisely how most clinical AI systems are implemented today (Trafton, 2024).

Clinical AI today remains tightly constrained: it performs narrow, well-defined tasks with binary or categorical outputs, mostly in routine, high-volume settings where physicians retain oversight. While often no longer fully interpretable, its limited scope still permits auditability, not by design, but because current applications remain confined in complexity and domain. Yet even within these boundaries, the findings made by MIT researchers regarding hidden biases raise concerns. Looking ahead, clinical AI will move beyond narrow, well-structured tasks because of the ongoing drive to optimize healthcare. As clinical environments become more complex and data-rich, interest is growing in systems that can process longitudinal, multimodal data and make decisions in even more complex scenarios. As of this writing, to our knowledge, no FDA-cleared applications incorporate transformer models, temporal reasoning, or similarly advanced architectures. The ethical implications of this issue will be examined in detail in the final two chapters.

5. Explainable AI and Post-Hoc Explanations

As AI systems begin to tackle more complex tasks, a central challenge emerges: how to preserve epistemic transparency in models whose reasoning processes are increasingly opaque. In response, the field of *explainable AI* has gained momentum, which aims to render such systems more intelligible within clinical practice. Most approaches focus on post-hoc explanation techniques, to approximate a model’s logic after the fact, offering clinicians a partial view of how specific inputs may have shaped its outputs. These methods attempt to sustain trust and oversight as models evolve, but whether they can meaningfully support clinical reasoning or ensure physician autonomy remains contested.

Cynthia Rudin (2019) argues that post-hoc explanations are not only epistemically flawed but practically dangerous in high-stakes domains such as medicine. Post-hoc techniques may appear valid in isolated regions of the feature space but can break down elsewhere, introducing unacceptable uncertainty in safety-critical applications. Rudin also warns of plausibility without understanding: explanations may feel coherent to users while actually masking the true, and possibly spurious, behavior of the model, thereby fostering unwarranted trust.

Still, many scholars defend post-hoc explainability. For example, Hatherley and colleagues (2025) argue that human explanations are also often biased or incom-

plete, and yet remain integral to medical reasoning. That argument risks slipping into whataboutism. The inherent unreliability of human explanations is why medicine emphasizes evidence hierarchies, placing expert opinion at the bottom and systematic research like meta-analyses at the top. The same critique applies to their claim that post-hoc explanations should not be held to a higher standard than human reasoning. We need to thoroughly scrutinize post-hoc explanations, particularly since AI lacks the accountability and legal responsibility of a human clinician.

Post-hoc explanations have found their most convincing use cases in domains like radiology, where the structure of the data aligns well with human perception. Radiological images exhibit strong spatial continuity and local autocorrelation, with pathologies that are often discrete, localized, and visually identifiable. In this setting, interpretability methods such as saliency maps or attention overlays can meaningfully correspond to human visual reasoning. A recent study by Senoner et al. (2024) demonstrated that such tools helped radiologists reduce false positives in identifying lung lesions, not by making the model's reasoning transparent, but by reinforcing confidence in negative assessments. In effect, the post-hoc explanations acted like a stencil or template, guiding the clinician's attention towards features they already associate with normal findings. Yet it's precisely this alignment with human perception that makes radiology an exception. Studies used to advocate for post-hoc explanations, such as those cited by Hatherley et al. (2025), typically rely on experimental setups like maze-solving tasks (Vasconcelos et al., 2022) that mirror the structured, fully observable environments of imaging. But these conditions fail to reflect the complexity of clinical decision-making in medicine more broadly.

By contrast, most clinical reasoning occurs in heterogeneous, high-dimensional feature spaces, where numerical values like lab results mix with categories such as diagnoses, medications, or procedures. Unlike images or time series, where nearby elements follow a clear structure, these features exist in disconnected spaces with no obvious order or continuity. This makes it harder for machines to form interpretable patterns, and harder for humans to follow them. In such settings, explanation tools like SHAP or LIME produce fragmented, context-sensitive outputs that don't align with how clinicians actually think. What appears to be clean, structured data is often misleading: it lacks the internal organization that makes visual explanations in radiology effective.

Radiology and cardiology dominate medical AI, not by chance, but due to their scale and alignment with deep learning. Radiology relies on standardized imaging data with strong spatial structure, well-suited to convolutional neural networks. Cardiology, by contrast, is primarily based on biosignal data such as echocardiograms (ECGs), offering temporal patterns ideal for sequence models. In both cases, the geometry and continuity of the data provide inductive priors that AI models can effectively exploit. We also include electronic health records (EHRs) as a third domain. Though not a specialty, EHRs integrate data across all areas of care, making them a critical frontier for AI with broad potential to inform clinical decision-making.

Empirical evidence continues to challenge the utility of post-hoc methods in real-world clinical contexts. Wysocki and colleagues (2023) studied 23 clinicians using a COVID-19 risk prediction tool with or without SHAP explanations. They found no significant improvement in trust, satisfaction, or error detection. In fact, 17% of clinicians reported worse understanding when explanations were included. Further-

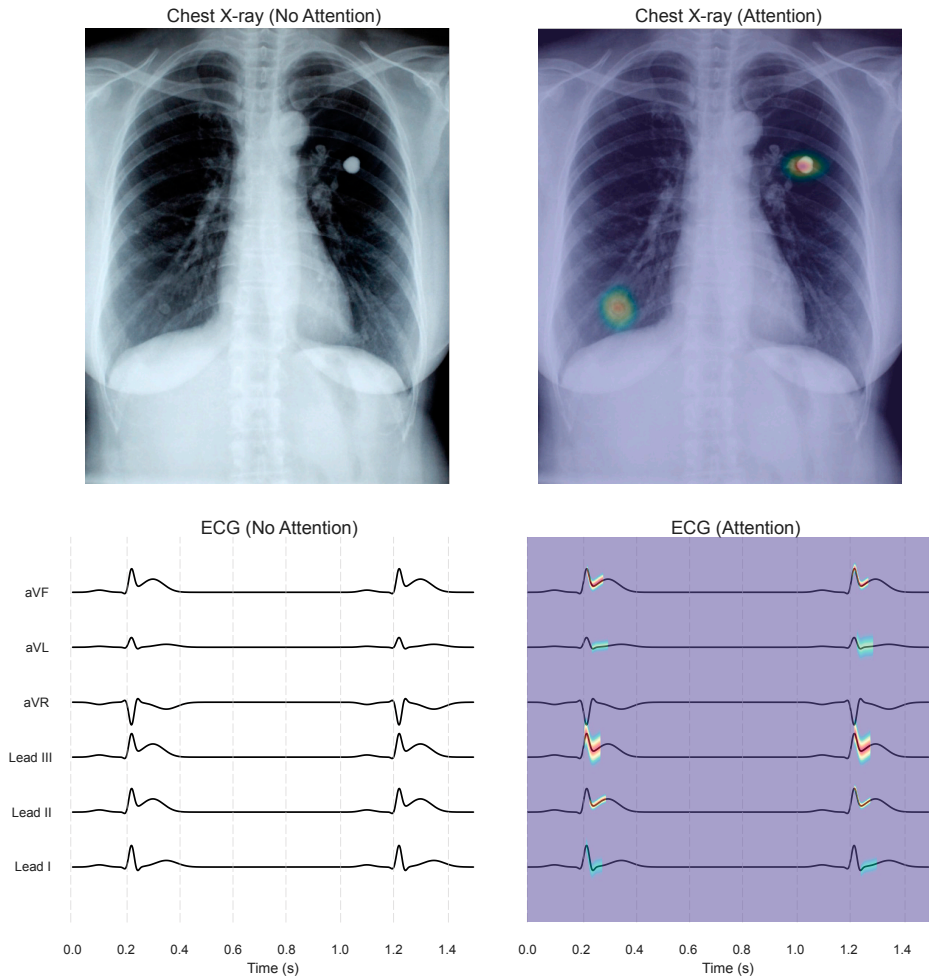


Fig. 2: Post-hoc attention heatmaps as visual guides. Attention heatmaps can highlight regions of interest in chest X-rays and ECGs, and may align with clinically relevant features (e.g., opacities, ECG anomalies). While they do not disclose what is being detected, clinicians can assess their validity against visible structure, offering a domain where post-hoc methods can support interpretation. Such approaches presuppose a direct mapping to a semantic space and tend to fail in heteronomous, polysemantic data such as EHRs

more, clinicians made faster decisions when model outputs and explanations agreed with their prior beliefs, suggesting that explanations may function less as tools for scrutiny but rather as compliance cues, potentially reinforcing existing biases. Similar findings were reported by Nagendran et al. (2023), who evaluated 86 ICU physicians making fluid and vasopressor dosing decisions in sepsis cases. While AI recommendations significantly influenced clinical behavior, adding simple explanations (feature importance) did not further increase adherence or confidence. Moreover, self-reported trust in explainability had no correlation with actual behavior, highlighting a persistent gap between perceived interpretability and actionable understanding. Together, these studies suggest that post-hoc explanations, particularly in

the form of feature attributions, may offer limited epistemic value in complex clinical decision-making settings.

We conclude that post-hoc explanations *may* prove effective in structured, perceptually grounded tasks, abundantly found in fields like radiology, where the geometry of the input space maps naturally onto human interpretive systems. But it falters in the broader domain of clinical medicine, where data is heterogeneous, nonlinear, and semantically entangled. Here, explanations must do more than trace decision pathways; the user must be able to transparently identify *when* and *why* they might be wrong. If absent, explainability risks becoming a veneer of epistemic comfort, offering the impression of clarity while concealing model limitations. Meeting the real demands of clinical AI will require not just architectural innovation but epistemic accountability: models that reason over structured knowledge, incorporate causal assumptions, and are intrinsically interpretable.

6. Physician Autonomy in Clinical Decision-Making

The results from the last chapter show that there is no state of the art of AI transparency in present clinical practice. We depicted a spectrum of different epistemic requirements posed by a variety of AI applications. While some of these applications remain intelligible, as previously demonstrated, others' complexity has already surpassed human cognitive capacities. Moreover, the deployment of black-box AI in other domains indicates that its clinical implementation is likely to become a subject of imminent debate. We thus turn to a discussion of the ethical consequences this foreseeable development poses for physician autonomy. First, we assess whether sufficient understanding and the absence of control is possible when interacting with uninterpretable AI systems. Secondly, we introduce a relational perspective to further demonstrate how answerability towards patients whose autonomy relies on that of physicians becomes impossible due to uninterpretable AI.

Returning to Beauchamp and Childress's (2019) three-dimensional account of autonomous action – intentionality, understanding, and non-control – it remains clear that physicians using AI intend to promote patient well-being. Understanding in clinical contexts entails the cognitive ability to grasp relevant information – typically present when using conventional technologies, such as MRT. With AI, however, the challenge lies not in user limitations but in insufficient transparency for meaningful integration into clinical reasoning. While full technical expertise is unnecessary, some level of interpretability is essential. Physicians must be able to align AI-generated outputs – diagnoses, radiological evaluations, or EHR-based recommendations – with accepted clinical practice. Limited understanding across specialties is common, yet responsibility has traditionally rested with identifiable human experts – senior physicians, professors, consultants, statisticians, or engineers.

Otherwise, with some AI models, the established chain of human responsibility begins to disintegrate when they become uninterpretable. Once the complexity of a model surpasses the limits of human comprehension, even its developers may no longer be able to account for how specific outcomes are produced. Therefore, the epidemic condition for responsible action is no longer guaranteed in this case.

Secondly, we turn to the dimension of non-control. Drawing on the Aristotelian concept of responsible action, Beauchamp and Childress (2019) argue that autonomy requires freedom from controlling external influences. While no one is entirely free from social norms or peer expectations – rendering perfect autonomy unattainable –, they emphasize that certain influences, particularly “coercion”, “persuasion” and “manipulation” (Beauchamp & Childress, 2019, p. 137), are illegitimate and compromise autonomy. In this context, we stress the importance of physicians understanding how AI systems present results, as this distinction is key to differentiating between rational understanding and undue influence.

This has important implications for how physicians engage with AI. It matters whether outputs are presented as abstract numerical terms or delivered via interactive language models that mimic human dialogue and reasoning. Even without intent to manipulate, such models can trigger trust, empathy, or deference, leading to emotional persuasion rather than rational understanding. If physicians are unaware of these effects, their autonomy may be compromised.

Concerns about manipulation extend to the origin and content of AI-generated recommendations. Most systems are developed by private companies, and their outputs may reflect corporate interests. Proprietary constraints often limit transparency, making independent auditing difficult. Combined with commercial incentives, this lack of scrutiny undermines accountability. For instance, AI systems may favor drugs from particular pharmaceutical companies. Although some physicians may also act in alignment with stakeholder interests, they are subject to professional and legal requirements to disclose conflicts of interest and can be held accountable. By contrast, AI systems – lacking consciousness and intention – cannot bear responsibility, even when emotional manipulation occurs. In a Derridean sense, *bearing witness* is a uniquely human form of attesting to truth, as it presupposes trust, belief, and the presence of a responsible subject who can be held to account (Derrida, 2005, p. 75). AI systems, lacking consciousness and intention, cannot bear witness. As such, they cannot be held responsible for emotional persuasion or biased recommendations, even if these effects occur. Moreover, since they operate without moral or epistemic self-awareness, their output cannot be assumed to be free from the influence of commercial interests. From this perspective, it becomes increasingly questionable whether we can meaningfully speak of physician autonomy when AI is involved in clinical decision-making. If physicians are routinely exposed to potentially persuasive AI-generated recommendations, the risk of compromised autonomy becomes structural and ever-present.

Finally, Coeckelbergh (2020) introduces a relational dimension to responsibility: to whom is the actor accountable? Coeckelbergh adds a crucial third component to the framework of responsible agency – *answerability* – defined as the obligation to explain one’s actions to those impacted by them (2020, p. 2061). In the physician-patient relationship, this relational responsibility is central. Physicians are expected not only to make informed decisions but also to communicate the rationale behind them. Yet in the context of black-box AI systems, this epistemic basis becomes increasingly undermined. If physicians cannot understand or articulate how an AI-generated recommendation has been reached, they are unable to fulfil their duty of answerability. This directly affects patient autonomy. According to the World Medical Association (2022), patients have the right “to be informed at every stage of the

care process”. When physicians are no longer capable of meeting this informational obligation, informed consent is compromised – undermining both physician responsibility and patient autonomy.

Uninterpretable AI’s opacity impairs physicians’ ability to explain decisions or convey risks, thereby compromising patient autonomy. Funer and Wiesing (2024, p. 4) argue that “physicians’ decision-making autonomy is an expedient means to accomplish the purpose of promoting health and respecting patients’ autonomy.” In contrast, we assert that physician autonomy is not merely instrumental but ethically essential. Without epistemic authority, physicians cannot support informed consent. As long as AI systems conceal biases or risks, patients may unknowingly be exposed to them – an ethical concern for both physician and patient autonomy. At a societal level, this underlying principle of beneficence risks turning into utilitarianism – benefiting those groups well-represented in training data while undermining others’ autonomy. Already today, medical datasets often reflect structural discrimination. Gender medicine exemplifies how knowledge production has historically neglected certain groups. Still, human experts can be held accountable for scrutinizing data quality and methodological soundness, determining whether algorithms are appropriate, and rejecting flawed tools when necessary. But concerning AI, if physicians cannot understand or explain AI-recommendations, they cannot meet ethical and epistemic duties. Consequently, their role as autonomous decision-makers – and their ability to ensure informed consent – is undermined. This raises a central ethical question: should we sacrifice physician autonomy for beneficence when it is unclear who benefits and to what extent? If AI-bias persists, who is responsible for recognizing and correcting it?

7. Physician Autonomy as Political Autonomy

In the last chapter, we presented three arguments against sustaining physician autonomy in clinical decision-making when confronted with uninterpretable AI in the future. Eventually we propose a different perspective: expanding physician autonomy as political autonomy to shape the structural and institutional conditions, including technologies in use, of medical practice that facilitate autonomous clinical decision-making.

Ethical debates on AI and physician autonomy have largely focused on clinical decision-making as the core of professional freedom, treating AI development and clinical practice as separate domains. Yet physician autonomy is shaped by the design of these technologies, which has not yet been considered part of the concept (Grosser et al., 2025; Schulz & Harrison, 1986). While they include “clinical”, “social”, and “economic” components in their definition, they do not sufficiently capture that all these domains are fundamentally shaped by broader societal conditions, in particular the healthcare system, medical schools, as well as pharma and tech industries. This underlying societal structuring precedes and constrains the supposed freedoms in each category. Regarding AI, crucial questions arise: who develops it, under what conditions, and which systems are approved? These questions underline physicians’ role in shaping their field and influencing autonomy-supporting conditions.

This is where the political dimension of autonomy, which historically predates the ethical dimension in the history of ideas of this term, becomes relevant – originating in the context of the Greek city-states. It refers to institutions that are self-imposed rather than externally dictated. From a Castoriadian perspective, heteronomy describes a societal state in which the constitution of social, economic, and political institutions – and the principles underlying them – are such that “the society or the individual, as the case may be, has no action upon them” (Castoriadis, 2022a, p. 169).

From this perspective, ethical responsibility entails a normative imperative to ensure conditions that prevent heteronomy imposed by opaque AI and enable autonomous clinical decision-making. Otherwise, the physician-patient relationship is undermined, compromising responsible action, as well as answerability, which is a central condition of patient autonomy. We argue that securing such conditions is consistent with Castoriadis’ concept of politics, which he defined as the establishment of institutions that facilitate freedom as autonomy. Here, autonomy is participatory self-legislation enacted “through explicit self-institution and self-governance” (Castoriadis, 2022b, p. 311), reflecting his conception of democracy. Such institutions remain open to critique and change, sustaining future autonomy. Physicians’ relational responsibility requires institutions that support the “politics of autonomy” (Castoriadis, 2022b, p. 301), enabling ethically responsible decision-making.

Political autonomy should thus extend to founding and shaping scientific and economic institutions that produce autonomy-supporting AI. Therefore, the discourse on physician autonomy must include political autonomy as a structural precondition. It empowers physicians as citizens and professionals to shape critical institutional and technological infrastructures – the conditions of their autonomy. It is only through active participation that physicians can ensure responsibility for the institutions they collectively create, including the design and deployment of clinical AI.

Finally, we present examples of contexts in which physicians may take action to shape their institutional conditions. In terms of medical education, physicians must advocate in university governance for curricula to incorporate AI literacy. These must aim at fostering an understanding of model logic, inherent limitations, and typical artifacts, enabling clinicians to critically interpret AI-supported decision-making processes. Analogous to ultrasound training – where clinicians learn to recognize imaging artifacts and assess scan reliability – AI education should equip physicians with the capacity to discern the reliability and relevance of algorithmic outputs. This does not require proficiency in programming but rather a clinical awareness of AI behavior. Education and training thus form essential prerequisites for responsible care, necessitating that universities treat AI tools as diagnostic instruments that must only be applied when adequately understood.

Within AI research in medicine, physicians have to participate as AI-trained scientists to articulate the requirements for their clinical practice and to promote a paradigmatic shift towards transparency and alignment with real-world clinical contexts. Central to this transformation is a critical examination of training data, including the features used, demographic representation, and potential encoding of systemic biases. Relying on average-case metrics is ethically insufficient when marginalized groups, often absent from training data, face systematically worse outcomes hidden by aggregate accuracy. To ensure equitable and accountable AI, both model architec-

tures and anonymized, demographically balanced datasets should be made openly accessible for external validation.

Beyond data openness, future research must prioritize inherently interpretable AI systems rather than relying solely on post-hoc explanations. This also means rejecting proprietary black-box models, which hinder scrutiny, replication, and accountability. Physicians have to advocate for funding to support the development of transparent architectures, such as Neurosymbolic AI, which integrates neural networks with rule-based symbolic reasoning. This approach holds promise by combining the pattern recognition capabilities of deep learning with the clarity of structured logic, potentially bridging the epistemic gap between human reasoning and machine-generated outputs. Concurrently, sociotechnical research must examine how physicians interact with AI, how explanations are comprehended or misconstrued, and how these systems influence clinical judgment. Without this critical layer of usability studies, even highly accurate AI tools risk failing in practice, thereby undermining trust, professional autonomy, and ultimately patient safety.

References

- Anderson, P. G., Tarder-Stoll, H., Alpaslan, M., Keathley, N., Levin, D. L., Venkatesh, S., Bartel, E., Sicular, S., Howell, S., Lindsey, R. V. & Jones, R. M. (2024). Deep learning improves physician accuracy in the comprehensive detection of abnormalities on chest X-rays. *Scientific Reports*, 14(1). <https://doi.org/10.1038/s41598-024-76608-2>
- Baloesescu, C., Bailitz, J., Cheema, B., Agarwala, R., Jankowski, M., Eke, O., Liu, R., Nomura, J., Stolz, L., Gargani, L., Alkan, E., Wellman, T., Parajuli, N., Marra, A., Thomas, Y., Patel, D., Schraft, E., O'Brien, J., Moore, C. L. & Gottlieb, M. (2025). Artificial Intelligence–Guided Lung Ultrasound by Nonexperts. *JAMA Cardiology*, 10(3), 245–253. <https://doi.org/10.1001/jamacardio.2024.4991>
- Beauchamp, T. L. & Childress, J. F. (2019). *Principles of biomedical ethics* (8. Aufl.). Oxford University Press.
- Bhargava, A., López-Espina, C., Schmalz, L., Khan, S., Watson, G. L., Urdiales, D., Updike, L., Kurtzman, N., Dagan, A., Doodlesack, A., Stenson, B. A., Sarma, D., Reseland, E., Lee, J. H., Kravitz, M. S., Antkowiak, P. S., Shvilkina, T., Espinosa, A., Halalau, A. ... Shapiro, N. I. (2024). FDA-Authorized AI/ML Tool for Sepsis Prediction: Development and Validation. *NEJM AI*, 1(12). <https://doi.org/10.1056/aioa2400867>
- Blackburn, S. (2016). Autonomy/heteronomy. In *Oxford Dictionary of Philosophy*. <https://www.oxfordreference.com/display/10.1093/acref/9780198735304.001.0001/acref-9780198735304-e-303?rskey=5U0PgV&result=320>
- Castoriadis, C. (2022a). *Crossroads in the Labyrinth: Human Domains* (Bd. 2). <https://notbored.org/cornelius-castoriadis-crossroads-2-human-domains.pdf>
- Castoriadis, C. (2022b). *Crossroads in the Labyrinth: The Rising Tide of Insignificance* (Bd. 4). <https://notbored.org/cornelius-castoriadis-crossroads-4-rising-tide-of-insignificance.pdf>
- Coeckelbergh, M. (2020). Artificial Intelligence, Responsibility Attribution, and a Relational Justification of Explainability. *Science and Engineering Ethics*, 26(4), 2051–2068. <https://doi.org/10.1007/s11948-019-00146-8>
- Derrida, J. (2005). *Sovereignities in Question: The Poetics of Paul Celan* (T. Dutoit & O. Pasanen, Hrsg.). Fordham University Press.
- Fischer, J. M. & Ravizza, M. (2000). *Responsibility and Control: A Theory of Moral Responsibility*. Cambridge University Press.

- Funer, F. & Wiesing, U. (2024). Physician's autonomy in the face of AI support: Walking the ethical tightrope. *Frontiers in Medicine*, 11. <https://doi.org/10.3389/fmed.2024.1324963>
- Gichoya, J. W., Banerjee, I., Bhimireddy, A. R., Burns, J. L., Celi, L. A., Chen, L., Correa, R., Dullerud, N., Ghassemi, M., Huang, S., Kuo, P., Lungren, M. P., Palmer, L. J., Price, B. J., Purkayastha, S., Pyrros, A. T., Oakden-Rayner, L., Okechukwu, C., Seyyed-Kalantari, L. ... Zhang, H. (2022). AI recognition of patient race in medical imaging: a modelling study. *The Lancet Digital Health*, 4(6), e406–e414. [https://doi.org/10.1016/s2589-7500\(22\)00063-2](https://doi.org/10.1016/s2589-7500(22)00063-2)
- Gordon, R. (2022, 20. Mai). Artificial intelligence predicts patients' race from their medical images. *MIT News | Massachusetts Institute of Technology*. <https://news.mit.edu/2022/artificial-intelligence-predicts-patients-race-from-medical-images-0520>
- Grosser, J., Düvel, J., Hasemann, L., Schneider, E. & Greiner, W. (2025). Studying the Potential Effects of Artificial Intelligence on Physician Autonomy: Scoping Review. *JMIR AI*, 4, e59295. <https://doi.org/10.2196/59295>
- Guo, L., Pressman, G. S., Kieu, S. N., Marrus, S. B., Mathew, G., Prince, J., Lastowski, E., McDonough, R. V., Currie, C., Maidens, J. N., Al-Sudani, H., Friend, E., Padmanabhan, D., Kumar, P., Kersh, E., Venkatraman, S. & Qamruddin, S. (2025). Automated Detection of Reduced Ejection Fraction Using an ECG-Enabled Digital Stethoscope. *JACC Advances*, 4(3), 101619. <https://doi.org/10.1016/j.jacadv.2025.101619>
- Hatherley, J., Munch, L. & Bjerring, J. C. (2025). *In defence of post-hoc explanations in medical AI*. <https://doi.org/10.48550/arXiv.2504.20741>
- Jussupow, E., Spohrer, K. & Heinzl, A. (2022). Radiologists' Usage of Diagnostic AI Systems. *Business & Information Systems Engineering*, 64(3), 293–309. <https://doi.org/10.1007/s12599-022-00750-2>
- Kocaballi, A. B., Ijaz, K., Laranjo, L., Quiroz, J. C., Rezazadegan, D., Tong, H. L., Willcock, S., Berkovsky, S. & Coiera, E. (2020). Envisioning an artificial intelligence documentation assistant for future primary care consultations: A co-design study with general practitioners. *Journal of the American Medical Informatics Association: JAMIA*, 27(11), 1695–1704. <https://doi.org/10.1093/jamia/ocaa131>
- Muehlematter, U. J., Daniore, P. & Vokinger, K. N. (2021). Approval of artificial intelligence and machine learning-based medical devices in the USA and Europe (2015–20): A comparative analysis. *The Lancet Digital Health*, 3(3), e195–e203. [https://doi.org/10.1016/S2589-7500\(20\)30292-2](https://doi.org/10.1016/S2589-7500(20)30292-2)
- Nagendran, M., Festor, P., Komorowski, M., Gordon, A. C. & Faisal, A. A. (2023). Quantifying the impact of AI recommendations with explanations on prescription decision making. *npj Digital Medicine*, 6(1), 206. <https://doi.org/10.1038/s41746-023-00955-z>
- Narang, A., Bae, R., Hong, H., Thomas, Y., Surette, S., Cadieu, C., Chaudhry, A., Martin, R. P., McCarthy, P. M., Rubenson, D. S., Goldstein, S., Little, S. H., Lang, R. M., Weissman, N. J. & Thomas, J. D. (2021). Utility of a Deep-Learning Algorithm to Guide Novices to Acquire Echocardiograms for Limited Diagnostic Use. *JAMA Cardiology*, 6(6), 624–632. <https://doi.org/10.1001/jamacardio.2021.0185>
- Niehoff, J. H., Kalaitzidis, J., Kroeger, J. R., Schoenbeck, D., Borggrefe, J. & Michael, A. E. (2023). Evaluation of the clinical performance of an AI-based application for the automated analysis of chest X-rays. *Scientific Reports*, 13(1), 3680. <https://doi.org/10.1038/s41598-023-30521-2>
- Obermeyer, Z., Powers, B., Vogeli, C. & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464), 447–453. <https://doi.org/10.1126/science.aax2342>
- Prince, J., Maidens, J., Kieu, S., Currie, C., Barbosa, D., Hitchcock, C., Saltman, A., Norozi, K., Wiesner, P., Slamon, N., Del Grippio, E., Padmanabhan, D., Subramanian, A., Manjunath, C., Chorba, J. & Venkatraman, S. (2023). Deep Learning Algorithms to

- Detect Murmurs Associated With Structural Heart Disease. *Journal of the American Heart Association*, 12(20), e030377. <https://doi.org/10.1161/jaha.123.030377>
- Rancier, M., Israel, I., Monickam, V., Currie, C., Verschoore, B., Lastowski, E., Pelt, D. W., Prince, J. & McDonough, R. V. (2024). Artificial-Intelligence-Enabled Digital Stethoscope Significantly Improves Point-of-Care Screening for Clinically Significant Valvular Heart Disease. *Research Square*, Preprint. <https://doi.org/10.21203/rs.3.rs-5649209/v1>
- Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5), 206–215. <https://doi.org/10.1038/s42256-019-0048-x>
- Rudy-Hiller, F. (2018). *The Epistemic Condition for Moral Responsibility*. Stanford Encyclopedia of Philosophy/Fall 2018 Edition. <https://plato.stanford.edu/archives/fall2018/entries/moral-responsibility-epistemic/>
- Schulz, R. & Harrison, S. (1986). Physician autonomy in the Federal Republic of Germany, Great Britain and The United States. *The International Journal of Health Planning and Management*, 1(5), 335–355. <https://doi.org/10.1002/hpm.4740010504>
- Senoner, J., Schallmoser, S., Kratzwald, B., Feuerriegel, S. & Netland, T. (2024). Explainable AI improves task performance in human-AI collaboration. *Scientific Reports*, 14(1), 31150. <https://doi.org/10.1038/s41598-024-82501-9>
- Tolliday, H. (1978). Clinical Autonomy. In E. Jaques (Hrsg.), *Health services: Their nature and organization, and the role of patients, doctors, nurses, and the complementary professions* (Brunel Institute of Organization and Social Studies (BIOSS) Series) (S. 32–52). Heinemann.
- Trafton, A. (2024, 28. Juni). Study reveals why AI models that analyze medical images can be biased. *MIT News | Massachusetts Institute of Technology*. <https://news.mit.edu/2024/study-reveals-why-ai-analyzed-medical-images-can-be-biased-0628>
- Vasconcelos, H., Jörke, M., Grunde-McLaughlin, M., Gerstenberg, T., Bernstein, M. & Krishna, R. (2022). *Explanations Can Reduce Overreliance on AI Systems During Decision-Making*. <https://doi.org/10.48550/arXiv.2212.06823>
- World Medical Association. (2022). *WMA International Code of Medical Ethics*. <https://www.wma.net/policies-post/wma-international-code-of-medical-ethics/>
- Wysocki, O., Davies, J. K., Vigo, M., Armstrong, A. C., Landers, D., Lee, R. & Freitas, A. (2022). Assessing the communication gap between AI models and healthcare professionals: Explainability, utility and trust in AI-driven clinical decision-making. *Artificial Intelligence*, 316, 103839. <https://doi.org/10.1016/j.artint.2022.103839>
- Yang, Y., Zhang, H., Gichoya, J. W., Katabi, D. & Ghassemi, M. (2024). The limits of fair medical imaging AI in real-world generalization. *Nature Medicine*, 30(10), 2838–2848. <https://doi.org/10.1038/s41591-024-03113-4>

PLAInLaw4U. Rechtlicher Klartext im (beruflichen) Alltag

Mit KI auf dem Weg zu mehr Verständlichkeit, Teilhabe und Rechtsstaatlichkeit

Petra Schön

1. Deutsch in der DACH-Region

Rund 130 Millionen Menschen in mehr als 40 Ländern weltweit verwenden Deutsch als Erstsprache (L1) und innerhalb der Europäischen Union ist Deutsch die am häufigsten gesprochene Erstsprache (Thiele, 2020).

Mit Deutschland, Österreich und der Schweiz umfasst die DACH-Region die zentralen deutschsprachigen Länder in Mitteleuropa. Bei einer Gesamtbevölkerung von rund 101 Millionen Menschen in diesem Raum nutzen etwa 91 Millionen – also rund 90 % – Deutsch als L1 (Statista, 2025).

In Deutschland ist Deutsch laut § 23 Verwaltungsverfahrensgesetz (VwVfG) und § 184 Gerichtsverfassungsgesetz (GVG) *Amtssprache*. In der Schweiz ist Deutsch eine von mehreren Amtssprachen (Art. 4 der Bundesverfassung der Schweizerischen Eidgenossenschaft), ebenso wie in Österreich (Art. 8 B-VG). Auffällig dabei ist, dass sich diese Bestimmungen nicht auf Deutsch in seiner allgemeinen Verwendung durch die Bevölkerung in den jeweiligen Ländern beziehen, sondern auf Deutsch als *Amtssprache*. Dabei wird Amtssprache („Amtssprache“, o. D.) als die Sprache verstanden, in der Behörden kommunizieren und mit der Öffentlichkeit interagieren. Ein Fokus auf Verständlichkeit in dieser Kommunikation oder gar auf den Zugang der Allgemeinbevölkerung zum Recht bleiben unberücksichtigt. Damit ist es für weite Teile der Bevölkerung ohne juristische Vorbildung schwierig, juristische Bestimmungen, Vorschriften, Entscheidungen oder Formulare zu verstehen und anzuwenden.

In Deutschland etwa wird Barrierefreiheit in der Sprache explizit für besonders schutzbedürftige Gruppen im Behindertengleichstellungsgesetz (BGG) geregelt. Bis 2025 gibt es in Österreich keine vergleichbare Regelung für die spezifische oder flächendeckende Verwendung barrierefreier Sprache im Rechtsbereich. In Österreich wurde 2021 das Web-Zugänglichkeits-Gesetz (WZG) geschaffen. Es normiert barrierefreie Standards für Webseiten öffentlicher Einrichtungen.

Laut § 3 WZG müssen Webseiten öffentlicher Institutionen in Bezug auf *Wahrnehmbarkeit, Bedienbarkeit, Verständlichkeit* und *Stabilität* barrierefrei gestaltet werden. Dies betrifft z. B. Stylesheets, Menüs oder technische Funktionen, jedoch nicht ausdrücklich sprachliche Merkmale, ebenso wenig die Verpflichtung, Gesetze in leicht verständlicher Sprache zu gestalten.

Diese Lücke der mangelnden gesetzlichen Verpflichtung zur verständlichen Gestaltung von Rechtstexten sticht insbesondere dann ins Auge, wenn man die Webseite des österreichischen Parlaments, Menüpunkt *Einfache Sprache* (Parlament Österreich, o. D. b), besucht. Hier erläutert die Parlamentsdirektion komplexe Prozesse und Begriffe im Rahmen der Gesetzgebung. Es werden allerdings keine Gesetzestexte in klarer Sprache angeboten.

Andere Webseiten wie beispielsweise das privat initiierte Portal RECHTleicht.at¹ orientieren sich an internationalen Übereinkommen für vulnerable Gruppen (z.B. UN-Kinderrechtskonvention, 1989; UN-Behindertenrechtskonvention, 2006). Sie bieten aber ebenfalls keine vereinfachten Gesetzestexte und sind teilweise nicht auf dem aktuellsten Stand. Im föderalen System in der Schweiz bestehen teils erhebliche Unterschiede zwischen den Kantonen. Texte in einfacher Sprache orientieren sich dabei primär an spezifischen Themen wie etwa COVID-19 oder Wahlinformationen (Melcher, 2022).

Schließlich sind auf DIN- und ISO-Ebene internationale und deutsche Normen zur Vereinfachung juristischer Kommunikation derzeit in Ausarbeitung oder wurden bereits veröffentlicht. Diese technischen Standards wurden von Fachexpert*innen entwickelt und können eine wertvolle Orientierung darstellen. Sie entfalten jedoch keine rechtlich bindende Wirkung, da sie nicht durch Parlamente verabschiedet werden.

Die fehlende gesetzliche Verpflichtung zur flächendeckenden Erstellung von Rechtstexten in verständlicher Sprache schränkt den Zugang zum Recht für viele Personen ein – speziell für die allgemeine Bevölkerung und für vulnerable Gruppen.

2. Das Recht – vom Volk für das Volk

Die gesetzgebende Kraft liegt in den Ländern der DACH-Region in den Händen der Bürger*innen, wie in deren Verfassungen festgelegt. Art. 1 B-VG legt fest: Österreich ist eine demokratische Republik. Ihr Recht geht vom Volk aus. Demokratische Grundsätze finden sich ebenso in der Präambel sowie in Art. 1 und 2 der Bundesverfassung der Schweizer Eidgenossenschaft und in Art. 20 des deutschen Grundgesetzes (GG):

- 1) *Die Bundesrepublik Deutschland ist ein demokratischer und sozialer Bundesstaat.*
- 2) *Alle Staatsgewalt geht vom Volke aus. Sie wird vom Volke in Wahlen und Abstimmungen und durch besondere Organe der Gesetzgebung, der vollziehenden Gewalt und der Rechtsprechung ausgeübt.*

Die Verfassungen sichern damit die Souveränität des Volkes zur Gestaltung der Rechtsordnung. Die daraus hervorgehenden Rechtsvorschriften enthalten Rechte und Pflichten für deren Rechtssubjekte – damit Recht vom Volk für das Volk.

Dennoch werden diese Texte von der Mehrheit der Bevölkerung – in deren Namen sie geschaffen wurden – kaum verstanden. Dies gilt für Texte aus Legislative, Exekutive und Judikative gleichermaßen. Die meisten wurden von spezialisierten juristischen Institutionen unter Verwendung von komplexer Fachsprache und Rechtsbegriffen erstellt. Die sprachliche Komplexität – etwa in Satzbau, historisch-formalen Klauseln und fachspezifischer Terminologie – wirkt abschreckend und behindert die nichtjuristische Bevölkerungsmehrheit in der Wahrnehmung ihrer Rechte und Pflichten. Mit dem Fehlen juristischer Grundlagen steigt das Risiko, *in die Irre geführt, benachteiligt oder anderweitig ausgenutzt zu werden* (Levine, 2004, S. 396, Herv.

¹ <http://rechtleicht.at/gesetze-leichter-lesen/>

d. Autorin). Diese Bevölkerungsgruppen benötigen in besonderem Maße Schutz und Unterstützung, um gleichberechtigt am öffentlichen Diskurs teilnehmen und informierte Entscheidungen treffen zu können. Ohne allgemein zugängliche rechtliche Informationen sind sie derzeit in ihrer autonomen Wahrnehmung von Rechten und Pflichten eingeschränkt (Freyer et al., 2024, S. 50).

In der DACH-Region sind etwa 250.000 einschlägig geschulte juristische Fachkräfte tätig. In Ermangelung einheitlicher Ausbildungswege und homogener statistischer Grundlagen für diesen Bereich sowie aufgrund von beruflicher Fluktuation stehen keine präzisen Zahlen zur Verfügung. Zu den juristisch geschulten Berufsgruppen zählen etwa Richter*innen, Staats- und Rechtsanwält*innen (Bundesamt für Justiz, 2022; Bundesrechtsanwaltskammer, 2025; Die österreichische Justiz, 2025; Statista, 2024) sowie Verwaltungspersonal. Hinzu kommen etwa 2.000 bis 3.000 weitere Jurist*innen im privaten und öffentlichen Bereich, etwa Universitäts- und Wissenschaftspersonal sowie Unternehmensjurist*innen.

Die überwiegende Mehrheit der Bevölkerung in der DACH-Region verfügt hingegen über keine juristischen Vorkenntnisse. Während fundiertes juristisches Wissen den Zugang zu Gesetzen, Verträgen und Urteilen erleichtert, ist dieser Zugang für viele Menschen deutlich eingeschränkt. Das Volk, das formell als Urheber des Rechts gilt, versteht viele der im eigenen Namen erlassenen Vorschriften nicht. Damit bleibt ein Großteil der Bevölkerung der DACH-Region auf kostenintensive juristische Beratung angewiesen oder vom Zugang zu seinen Rechten und Pflichten ausgeschlossen.

Auch juristische Fachkräfte außerhalb ihres Kerngebietes könnten von vereinfachten Rechtstexten profitieren, etwa wenn sie sich rasch mit neuen Themenfeldern wie z.B. Cyberkriminalität oder Weltraumrecht vertraut machen möchten oder aber auch am Beginn ihrer beruflichen Laufbahn. Dabei kann beispielsweise ein einfach verständlicher Text wesentlich zu Überblick und Orientierung beitragen.

Gerade Berufseinsteiger*innen sowie Studierende der Rechts- oder Übersetzungswissenschaften könnten in Zukunft von PLAInLaw4U, einer innovativen KI-Anwendung zur Vereinfachung von Rechtstexten, profitieren. Denn das rasche Verständnis und die sichere Erstellung juristischer Fachtexte sind anfangs oft herausfordernd. PLAInLaw4U könnte als Lerninstrument zur individuellen oder kollaborativen Nutzung eingesetzt werden und langfristig möglicherweise zu einer Verkürzung der durchschnittlichen Studiendauer – derzeit in Österreich rund zehn Jahre (Die Presse, 2024) – sowie zur Senkung der Drop-out-Rate beitragen.

PLAInLaw4U erhebt dabei keineswegs den Anspruch, Jurist*innen zu ersetzen oder deren Expertise zu untergraben. Ziel ist vielmehr ein leichter Zugang zu grundlegenden rechtlichen Informationen. Dadurch können Missverständnisse und Fehlverhalten durch fehlendes Rechtsverständnis vermieden werden.

3. Deutschsprachige Rechtstexte – leicht, einfach, verständlich?

Die Rechtsvorschriften, die in den drei Ländern der DACH-Region vom Volk und seinen Vertreter*innen geschaffen werden, sind der Mehrheit des Volkes inhaltlich schwer zugänglich. Auch mangelt es an allgemeinen gesetzlichen Verpflichtungen, die die Verständlichkeit von Rechtstexten sicherstellen. Dieses Ungleichgewicht kann

zu Verständnis- und Entscheidungsdefiziten in der breiten Mehrheit der Bevölkerung sowie zu Desinformation und Desinteresse an demokratischer Willensbildung beitragen.

Im Gegensatz zur langjährigen Tradition einfacher und leichter Sprache im englischsprachigen öffentlichen Raum sind im deutschsprachigen Raum eine deutlich langsamere Entwicklung sowie Unschärfen bei der Anwendung leichter und einfacher Sprache zu beobachten.

Nach Schomacker et al. (2023) ist die regelbasierte *Leichte Sprache* generell besonders gut verständlich und entspricht in etwa dem Niveau A2 des Gemeinsamen Europäischen Referenzrahmens für Sprachen (GER, o.D.). Die o.g. Autor*innen unterscheiden zwischen *Leichter Sprache* und *Vereinfachter Sprache* – letztere bezeichnet ein nicht näher festgelegtes Kontinuum vereinfachter Sprachformen. Zudem verweisen sie auf die bestehende gesetzliche Verpflichtung öffentlicher Einrichtungen gemäß § 11 BGG. Darin ist festgelegt, dass für Menschen mit kognitiven Beeinträchtigungen Informationen in leichter und barrierefreier Sprache bereitzustellen sind. Dazu zählen auch juristische Informationen, amtliche Verträge und Formulare. Sie regelt zudem ausdrücklich die finanzielle Verantwortung der öffentlichen Stellen als zusätzliche Schutzmaßnahme. Schließlich bekräftigt die Regelung die Absicht der deutschen Bundesregierung, eine breitere Nutzung Leichter Sprache durch Behörden auch für die Allgemeinheit zu fördern. Einerseits kann diese Erklärung als grundsätzliches Anerkenntnis der legislativen Verantwortung verstanden werden, den Zugang zum Recht zu erleichtern – und als Eingeständnis des unbefriedigenden Status quo. Andererseits sollte eine solche Erklärung im Lichte demokratischer Gesetzgebung, die vom Volk für das Volk gemacht wird, gar nicht erforderlich sein. Festzustellen ist damit, dass diese Regelung zwar inhaltlich relevant ist, allerdings nur für eine eingeschränkte Zielgruppe. Für die allgemeine Bevölkerung wurde mit dieser Absichtserklärung nach wie vor kein Rechtsanspruch auf verständliche Rechtstexte geschaffen.

Stodden et al. (2023, S. 16441) erläutern die Differenzierung der deutschsprachigen *Leichten Sprache*, die durch ein sehr strenges Regelwerk zur maximalen Reduktion sprachlicher Komplexität auf etwa A1-Niveau (GER, o.D.) gekennzeichnet ist und ursprünglich für Zielgruppen mit kognitiven Einschränkungen entwickelt wurde. Ihre Einschätzung mit einem Niveau bis A1 unterscheidet sich von der von Schomacker et al. (2023) beschriebenen Einordnung. Auch in Österreich existieren abweichende Herangehensweisen: Die Direktion des österreichischen Parlaments bietet unter *Barrierefreiheit* wenig trennscharf sowohl Informationen in *Leichter Sprache* als auch *Einfacher Sprache*, letztere in vereinfachtem Deutsch auf B1-Niveau, an (Parlament Österreich, o.D.a). Diese Herangehensweise folgt dem Zugang von Stodden et al. (2023, S. 16441). Sie beschreibt *Einfache Sprache* als Resultat der Reduktion sprachlicher Komplexität auf dem Niveau A2 bis B1 (GER). Für sie zählen Menschen mit Leseeinschränkungen oder nichtdeutscher Erstsprache zu wesentlichen Zielgruppen.

Die beschriebene Ausrichtung auf einfache Sprache (B1, GER, o.D.; Stodden et al., 2023) deckt sich mit PLAI_nLaw4U, das zur sprachlichen Vereinfachung deutschsprachiger Rechtstexte beitragen soll. Die Verwendung des Begriffs vereinfachte Sprache als Synonym für leichte und einfache Sprache soll den begrifflichen Unschärfen der beiden Begriffe Rechnung tragen. Das deckt sich mit Spring et al.

(2022, S. 138), die einen utilitaristischen Sprachvereinfachungsprozess als *simplified language* bezeichnen.

4. PLAInLaw4U – Anwendungsgebiete & Qualität

PLAInLaw4U ist als interdisziplinäres Open-Source-KI-Sprachmodell (LLM) für verständliche deutschsprachige Rechtstexte konzipiert. Die Kernfunktion von PLAInLaw4U liegt dabei in einer intralingualen Übersetzung komplexer juristischer Texte in verständliche Sprache für Nichtjurist*innen und juristische Fachkräfte. Ziel dieses Modells ist es, juristisch relevante Informationen barrierefrei aufzubereiten, ohne dabei auf Vollständigkeit und Präzision verzichten zu müssen.

4.1 Anwendungsgebiete

Als anwendungsfreundlich gestaltete LLM-Plattform kann PLAInLaw4U vielseitig eingesetzt werden, darunter in folgenden Bereichen:

- a) *Bildungsbereich*: In erster Linie im Rahmen der sekundären und tertiären Ausbildung kann PLAInLaw4U zur Erklärung rechtlicher Vorgaben und Zusammenhänge in verständlicher Sprache zur Verfügung stehen. Dazu gehören im Schulunterricht Gegenstände wie Politische Bildung oder Staatsbürgerkunde. Im Rahmen der universitären Ausbildung wären etwa Studienrichtungen wie Rechtswissenschaften oder Translationswissenschaft mit Schwerpunkt Rechtsübersetzung mögliche Anwendungsgebiete.
- b) *Öffentliche Rechtsanwendung bzw. Rechtsprechung*: Im Rahmen der öffentlichen Verwaltung und Justiz könnte PLAInLaw4U unterstützen. Es könnte etwa bei der Erarbeitung von verständlichen Rechtstexten, Auskünften, Formularen sowie zur Beratung oder Kommunikation mit Bürger*innen bei behördlichen Verfahren dienen.
- c) *Öffentliche Rechtschöpfung*: Bei der legislativen Vorbereitung, z.B. dem Textieren bzw. Überarbeiten von Gesetzestexten, Verordnungen, Bescheiden, rechtlichen Regelwerken bzw. Textbausteinen kann PLAInLaw4U eine verständliche homogene Formulierung unterstützen.
- d) *Zivilgesellschaftliche Teilhabe*: Hier kann PLAInLaw4U zur Erstellung verständlicher Rechtsinformationen für Rechtsthemen im Alltag beitragen.
- e) *Private Rechtsanwendung und Beratung*: Für Rechtsanwält*innen in der Kommunikation mit Klient*innen und in der NGO-Rechtsberatung, z.B. für Geflüchtete, kann PLAInLaw4U ein zusätzliches Werkzeug zur Beratung und rechtlichen Aufklärung darstellen.

Ultimativ soll PLAInLaw4U damit auch die Vermittlung von staatsbürgerlicher Verantwortung und Partizipation fördern helfen. Dabei soll PLAInLaw4U als Open-access-Anwendung allen Interessierten zur Verfügung stehen – unabhängig von beruflichem Status, Bildungshintergrund oder Einkommen.

4.2 Qualität

PLAInLaw4U soll ein qualitativ hochwertiges, verlässliches, technisch stabiles, und skalierbares Modell zur Vereinfachung deutschsprachiger Rechtstexte bilden. Ein zentrales Funktionskriterium ist eine qualitativ hochwertige verlässliche Textaufbereitung, gerade vor dem Hintergrund des sensitiven Inhaltes und der Nutzung durch schützenswerte Gruppen. Um das Vertrauen in PLAInLaw4U zu rechtfertigen, muss das Modell ein Höchstmaß an Zuverlässigkeit bieten und zugleich die unterschiedlichsten juristischen Themenfelder kontextualisiert aufbereiten – ohne dabei Abstriche bei Präzision oder Textverständlichkeit zu machen.

Nach Freyer et al. (2024, S. 50) muss es Kernaufgabe solcher Modelle sein, eine vollständige und zuverlässige laufende Erstellung von Texten sicherzustellen (so auch Bender et al., 2021, S. 611; Floridi, 2023, S. 15; Kempt et al., 2023).

Entscheidend wird sein, typische KI-Schwächen wie Verzerrungen, Halluzinationen oder Informationsverlust zu minimieren und dabei das Generieren von vollständigen, korrekten und verständlichen Texten sicherzustellen.

Laut Stodden et al. (2023, S. 16449) sowie Alva-Manchego et al. (2019, S. 49) erfolgt eine Textvereinfachung mittels KI typischerweise in mehreren Schritten. Neben der Ersetzung facheinschlägiger Terminologie werden syntaktische Umstrukturierungen und Textkürzungen durchgeführt. Bei diesen Schritten liegt das Augenmerk auf Inhaltstreue und Vollständigkeit des Zieltextes mit dem Ausgangstext, was im Bereich der Modell-Methodologie weiter erläutert wird.

Auch wenn Deutsch in allen drei Ländern der DACH-Region als Amtssprache festgelegt ist, liegt doch keine einheitliche Sprache vor. Sowohl im gesprochenen als auch im geschriebenen Deutsch begegnen uns vielfältige terminologische und syntaktische Unterschiede, auch und gerade im Rahmen von Rechtstexten in unterschiedlichen Ländern und unterschiedlichen Rechtsgebieten. Um diese länderspezifischen Rechtsbegriffe, juristischen Konzepte und Sprachvariationen angemessen abzubilden und diese in verständliches Deutsch übersetzen zu können, ist ein höchstes Maß an Kontextualisierungsfähigkeit erforderlich.

5. PLAInLaw4U – Aufbau & Methodologie

Das Design des PLAInLaw4U-Modells, das auf drei Säulen beruht, stellt eine RL4LLM-Anwendung dar wie in der Taxonomie von Pternea et al. (2024) beschrieben.

5.1 Aufbau

Die Kombination der folgenden drei Ebenen bildet PLAInLaw4U:

- a) Das wissenschaftlich entwickelte LLM zur Vereinfachung der deutschen Sprache: DEPlain (Stodden et al., 2023)

- b) Ein hochwertiges länderspezifisches intralinguales Rechtslexikon, für den österreichischen Pilotbetrieb beispielsweise das JustizOnline Begriffslexikon mit österreichischer Rechtsterminologie, Definitionen und kontextbezogenen Synonymen
- c) Prompt Engineering durch Reinforcement Learning mit menschlichem Feedback auf Basis variationalen Präferenzlernens (RLHF-VPL). Dieser Schritt soll der Kontextualisierung und Anpassung an unterschiedliche Rechts- und Sprachkontexte dienen. Zudem soll damit die Optimierung vereinfachter Rechtstexte sowie Stabilität, Belastbarkeit und Resilienz des Modells gegenüber Angriffen gefördert werden.

5.2 Methodologie

5.2.1 DEPlain: LLM zur Vereinfachung deutscher Texte

DEPlain wurde 2023 von Stodden et al. (S. 16441) an der Heinrich-Heine-Universität Düsseldorf als wissenschaftliches intralinguales LLM zur Vereinfachung der deutschen Sprache entwickelt. Seine Code-Komponenten sind offen verfügbar (Stodden, 2023). Es liefert leistungsstarke Ergebnisse bei der Vereinfachung deutscher Texte allgemein (nicht rechtsspezifisch).

Eine solche KI-Aufgabe ist hochkomplex, wie Stodden et al. (2023, S. 16445) erläutern. Die satzweise Vereinfachung umfasst verschiedene textliche Anpassungen nach folgenden Mustern: 1:1 (Kopieren und Umformulieren), 1:m (Teilen komplexer Satzstrukturen), n:1 (Zusammenführen) oder n:m (Zusammenführen komplexer und einfacher Satzstrukturen). Die von Stodden et al. (2023, S. 16447) präsentierten Evaluierungen der DEPlain-Ergebnisse weisen DEPlain als hochwertig in den Bereichen Grammatikalität, Bedeutungstreue und Kohärenz aus. Die Nutzung dieses LLM als Grundlage für PLAInLaw4U bietet eine wissenschaftlich fundierte Ausgangsbasis und minimiert das Risiko von kommerziellen Verzerrungen in den Trainingsdaten.

5.2.2 Intralinguales Rechtslexikon für länderspezifische Terminologie

Im zweiten Schritt nutzt PLAInLaw4U den Terminologiekorpus eines qualitativ hochwertigen umfassenden intralingualen, d.h. deutsch-deutschsprachigen Rechtslexikons. Für Österreich stünde beispielsweise das *JustizOnline Begriffslexikon* zur Verfügung, das vom Bundesministerium für Justiz als öffentliche Behörde zusammengestellt und gepflegt wird (Die österreichische Justiz, o.D.).

Mit Hilfe des zugrundeliegenden Rechtslexikons mit umfangreicher Terminologie, Definitionen und kontextspezifischen Synonymen soll PLAInLaw4U entsprechende juristische Begriffe im komplexen Ausgangstext identifizieren, maskieren und durch vereinfachte Synonyme ersetzen.

5.2.3 Prompt Engineering: Reinforcement Learning (RL)

Als wissenschaftlich entwickeltes LLM zur Vereinfachung allgemeiner deutschsprachiger (nicht rechtsspezifischer) Texte erreicht DEPlain (Stodden et al., 2023) bereits eine relativ hohe Performanz im Vergleich zu anderen Plattformen, z. B. gerPT2 oder german_gpt (Anschütz et al., 2023, S. 1147). Um dem zuvor diskutierten hohen Qualitätsanspruch für die Textvereinfachung und Anpassung an den länderspezifischen Rechtskontext gerecht werden zu können, muss dieses Niveau weiterentwickelt werden. Dafür könnte einerseits auf zusätzliches manuelles Kodieren zum weiteren Fine-tuning des Modells zurückgegriffen werden, was allerdings erheblichen Dateninput sowie Rechen- und Trainingsressourcen erforderte (Huang et al., 2024, S. 7).

Eine mögliche ressourcenschonende Alternative könnte der von Cemri et al. (2022) vorgestellte Trainingsansatz zur *Unsupervised Simplification of Legal Texts* (USLT) darstellen. Das USLT-Modell wurde jedoch für ein angloamerikanisches Case-Law-System mit Gerichtsentscheidungen des US Supreme Court entwickelt. Daher unterscheiden sich sein Fokus und seine Funktionalität erheblich vom Rechtssystem in den Ländern der DACH-Region, wo eine Kodifikation durch Gesetze mit höherer sprachlicher Komplexität auftritt. Darüber hinaus ist das USLT-Modell nur auf Entscheidungen der US-Judikative trainiert, während PLAINLaw4U eine deutlich breitere Basis legislativer, exekutiver und judikativer Kontexte in verschiedenen deutschsprachigen Ländern abdecken möchte. Es muss daher jeden Ausgangstext hinsichtlich Herkunftsland und Rechtsgebiet in seinem Inhalt korrekt erfassen und kontextualisieren können. Dafür ist das landesspezifische kodifizierte Rechtssystem als Trainingsgrundlage entscheidend. Das Ziel des Trainings und Finetunings ist die Adaptierung des Modells an den länderspezifischen rechtlichen und sprachlichen Kontext.

Für diese Adaptierung erscheint Reinforcement Learning (RL4LLM; Pternea et al., 2024) vielversprechend. RL4LLM inkludiert menschliches Feedback und weist eine dynamische Lernkomponente auf Wort-, Satz- und Textebene (Stodden et al., 2023) auf.

Zhou et al. (2024, S. 87) diskutieren die Anwendbarkeit von Reinforcement Learning (RL) bei LLM im Rahmen einer Studie zum autonomen Fahren. Die Anforderungen an autonomes Fahren, in einer komplexen dynamischen Umgebung mit maximaler Sicherheit und minimalen Entscheidungsfehlern zu manövrieren, dienen dem Schutz der Passagiere und Verkehrsteilnehmer*innen. Damit sind sie mit den hohen Anforderungen an PLAINLaw4U vergleichbar. Zhou et al. (2024, S. 86) beschreiben RL als eine interaktive Trainingsmethode, die das LLM mit Belohnungsfunktionen zu optimalen Lernstrategien anleitet. Damit nutzt RL die kontextbezogene Lern- und Schlussfolgerungsfähigkeit des LLM, das dynamisch weiterführende Funktionen generieren kann. LLM sind in der Lage, auf Basis historischer Verläufe Referenzstrategien für bestimmte zukünftige Kontexte zu erstellen und für ihre Entscheidungsfindung historische und neue Daten mit den Belohnungsfunktionen abzugleichen. Dieses kontextuelle Lernen steht auch für PLAINLaw4U im Vordergrund.

Für die Aufgabe einer KI-gestützten Vereinfachung von Rechtstexten, vor allem im Hinblick auf die sprachlichen Variationen und verschiedenen Rechtsordnungen in der DACH-Region, stellt das Modell *Personalizing Reinforcement Learning from*

Human Feedback with Variational Preference Learning (RLHF-VPL) von Poddar et al. (2024) einen relevanten Ansatz dar. Ursprünglich für Präferenzentscheidungen bei Robotik-Aufgaben entwickelt, soll diese kontextspezifische Personalisierung nunmehr auf PLAInLaw4U übertragen werden können.

In seinen Ursprüngen war Reinforcement Learning (RL) nicht in der Lage, die Vielfalt menschlicher Präferenzen oder Subgruppen angemessen abzubilden. Frühere RL-Modelle basierten auf dem Bradley-Terry-Luce (BTL)-Modell (Poddar et al., 2024, S. 2). Dieses Modell weist eine entscheidende Schwäche auf: Ihm liegt lediglich eine unimodale Nutzer*innenfunktion zu Grunde. Diese geht wenig lebensnah von einer einzigen menschlichen Präferenz in jedem Kontext aus. Statt Differenzierungen zu berücksichtigen, bildet es einen statistischen Mittelwert und nivelliert so die Ergebnisse. Diese mangelnde algorithmische Spezifikation führt insbesondere bei Modellen mit hoher Varianz wie bei PLAInLaw4U zu unbefriedigenden Ergebnissen.

Auch im Hinblick auf die Abbildung unterschiedlicher menschlicher Werte und damit die Berücksichtigung von verdecktem Kontext griffen frühere RL-Modelle zu kurz. Bisher lag den Modellen ein einziges Wertesystem zu Grunde, das von einer kleinen Gruppe KI-Forschender vorgegeben wurde (Kirk et al., 2023). In einem solchen Modell führt selbst eine schwach ausgeprägte Mehrheitspräferenz dazu, dass Minderheitspräferenzen unberücksichtigt bleiben (Siththaranjan et al., 2023).

In einem stark kontextabhängigen Gebiet mit vielen sprachlichen und rechtlichen Variationen wie bei der Vereinfachung von deutschsprachigen Rechtstexten durch PLAInLaw4U ist Multimodalität für das Modelldesign entscheidend. Der von Poddar et al. (2024) entwickelte Algorithmus des Variational Preference Learning (VPL) basiert auf der Modellierung latenter, bedingter Belohnungsfunktionen. Dabei wird eine zusätzliche Variable in das Belohnungsmodell integriert, die latent Varianten abbilden kann.

Der RLHF-VPL-Algorithmus erfordert eine fein- und hochgradige Skalierung der Belohnungsfunktion, um eine hohe Performanz zu erreichen. Es wird erwartet, dass er damit besser als bisherige Ansätze den Anforderungen an qualitativ hochwertige kontextspezifische LLM-Textvereinfachung entsprechen kann.

Gerade bei hochsensiblen Modellen wie bei PLAInLaw4U haben nicht nur funktionale Erwägungen, sondern auch Schutz und Sicherheit Priorität. Um stets gleichbleibend höchste Textqualität in Inhalt und Aufbereitung sicherzustellen, ist umfassendes Prompt Engineering erforderlich, um die Widerstandsfähigkeit des Modells gegenüber Angriffen zu stärken und typischen KI-Fehlentwicklungen wie Halluzinationen, Verzerrungen oder zu starke Vereinfachungen vorzubeugen.

Im Hinblick auf zukünftige vulnerable und damit besonders schützenswerte Nutzer*innengruppen von PLAInLaw4U ist der Argumentation von Li et al. (2024, S. 265) zu folgen, die sich neben der Funktionalität für maximale LLM-Qualitätssicherung aussprechen. Auch wenn sich die Forschung von Li et al. (2024) auf die Lesbarkeit biomedizinischer Abstracts bezieht, ist die Zielsetzung von PLAInLaw4U mit ihr vergleichbar.

Die Wichtigkeit der Modellstabilität gegenüber Angriffen betonen auch Huang et al. (2024, S. 1), die solides Prompt Engineering nicht nur für die Funktionalität, sondern auch für die Stabilität und Resilienz des Modells propagieren – etwa auf Grundlage des von ihnen entwickelten und publizierten Codes (Huang et al., 2024), um Jailbreak-Angriffen, also Attacken auf die Systemfunktionalität bzw. Sicherheits-

architektur (Yi et al., 2024), auf PLAInLaw4U standzuhalten und damit die Anfälligkeit für Fehlfunktionen zu minimieren.

6. PLAInLaw4U – Funktionalität

Nach der Vorstellung der Modellkomponenten soll die konkrete Funktionalität von PLAInLaw4U anhand eines Textvereinfachungsbeispiels demonstriert werden. Dazu dient eine bundesdeutsche Bestimmung aus dem Bürgerlichen Gesetzbuch: *„Durch den Kaufvertrag wird der Verkäufer einer Sache verpflichtet, dem Käufer die Sache zu übergeben und das Eigentum an der Sache zu verschaffen“* (§ 433 BGB). Für die Bearbeitung sind folgende KI-Verarbeitungsschritte erforderlich:

- a) Kontextualisierung des Ausgangstextes der Rechtsquelle zum Ursprungsland und Analyse fachspezifischer sowie sprachlicher Merkmale und Besonderheiten mit Hilfe von Reinforcement Learning (RLHF-VPL; Poddar et al., 2024)
- b) Identifikation, Maskierung und Ersetzen juristischer Fachtermini auf Basis des Abgleichs mit dem länderspezifischen Rechtslexikon und
- c) Erstellen eines intralingualen sprachlich vereinfachten Zieltextes durch Anpassung von Ausdruck und Satzstruktur mittels DEPlain (Stodden et al., 2023).

Ein mögliches Ergebnis könnte lauten: *Beim Kauf muss der Verkäufer die Sache dem Käufer geben. Der Käufer wird Eigentümer.*

Mit ausreichendem RLHF-VPL soll das LLM PLAInLaw4U künftig in der Lage sein, unterschiedliche juristische Quellen auch ohne explizite Quellenangabe im Text korrekt zu differenzieren. Im obigen Beispiel handelt es sich um das BGB, also um eine deutsche Rechtsquelle, nicht um eine schweizerische oder österreichische. Viele Fachbegriffe ähneln einander zwar, stehen jedoch mitunter für unterschiedliche juristische Konzepte. Termini unterscheiden sich teils nur geringfügig, z.B. Schadensersatz (§ 823 BGB) in Deutschland, in Österreich dagegen Schadenersatz (§ 1295 ABGB). Mit dem schrittweisen Ausbau soll PLAInLaw4U in der Lage sein, diese Quellen korrekt abzugleichen.

7. PLAInLaw4U – Leistungsfähigkeit

Derzeit befindet sich PLAInLaw4U in der Aufbau-, Trainings- und Testphase. Für das Pilotprojekt zur Erstellung vereinfachter Rechtstexte in österreichischem Deutsch bietet das österreichische Rechtsinformationssystem des Bundes (RIS)² umfangreiche Rechtstexte. Das RIS enthält Bundes- und Landesgesetzgebung sowie gerichtliche Entscheidungen und wird vom Bundeskanzleramt der Republik Österreich erstellt und gewartet. Für die geplante schrittweise Skalierung von PLAInLaw4U für die DACH-Region stehen mehrere juristische Datenbanken, beispielsweise für deutsche

² <https://www.ris.bka.gv.at/>

Rechtstexte, zur Verfügung, darunter N-Lex³ sowie Gesetze im Internet und Rechtsprechung im Internet.⁴

Verständlichkeit und kontinuierlich hohe Qualität der Textgestaltung sind in der Auf- und Ausbauphase gleichermaßen zentral und werden sorgfältig sowohl mittels menschlichen Feedbacks als auch objektiviert anhand von quantitativen Metriken geprüft.

Eine Möglichkeit bietet die Prüfung der Ergebnisse auf Basis von Parametern wie SARI (Li et al., 2024, S. 267). SARI dient zur Zieltext-Messung auf seine Flüssigkeit und Stimmigkeit anhand von Kriterien wie Präzision oder Längenverhältnis. Auch BERTScore (Li et al., 2024, S. 267) könnte als Mess-Indikator zur Bewertung der semantischen Bewahrung und Kontextualisierung dienen.

SARI und BERTScore erscheinen für die Messung der Funktionalität passender als Parameter wie BLEU und ROUGE (Li et al., 2024, S. 267), die ursprünglich für Englisch und nicht zur Textvereinfachungsmessung entwickelt wurden.

Besonders geeignet für die Messung der Textvereinfachungsqualität erscheint das ursprünglich für Englisch entwickelte EASSE-Modell (Easier Automatic Sentence Simplification Evaluation) von Alva-Manchego et al. (2019, S. 49). Auf dieser Basis hat Stodden (2024b) für Deutsch das System EASSE.DE (Stodden, 2024a) entwickelt, das die automatische Bewertung von Vereinfachungen deutscher Texte (EASSE.DE) bzw. nicht englischsprachiger Texte (EASSE.multi) ermöglicht. Ihr System arbeitet mit dem Parameter *Sprachkonstante*, der sowohl die jeweilige Zielsprache berücksichtigt als auch sprachspezifische Metriken heranzieht. Neben BERTScore beinhaltet *EASSE.multi* auch einen Index zur Messung der einfachen Lesbarkeit. Das Evaluationsmodell zur deutschen Textvereinfachung ist speziell an DEPlain (Stodden et al., 2023) orientiert – ein wesentlicher Schritt in Richtung Standardisierung und Vergleichbarkeit bei der Bewertung der Textvereinfachungsqualität.

8. Zusammenfassung und Ausblick

Mit PLAInLaw4U wird ein innovatives und qualitativ hochwertiges KI-Modell zur Vereinfachung von Rechtstexten in deutscher Sprache entwickelt. Durch die Kombination von DEPlain, einem wissenschaftlich entwickelten LLM zur Vereinfachung deutscher Sprache (Stodden et al., 2023), einem länderspezifischen Rechtslexikon und Prompt Engineering durch Reinforcement Learning, das auf menschlichem Feedback basiert (RLHF-VPL, Poddar et al., 2024), soll es möglich werden, komplexe juristische Fachsprache für breite Bevölkerungsgruppen – Laien und Expert*innen gleichermaßen – zugänglich zu machen, ohne Abstriche bei Vollständigkeit, Präzision oder Verständlichkeit zu machen.

Für das Projekt stehen neben wissenschaftlicher Stringenz und ressourcenschonender Entwicklung insbesondere die hohe Verlässlichkeit der Plattform sowie eine praxisnahe und nutzungsfreundliche Umsetzung im Vordergrund. Nach dem Pilotmodell zur Vereinfachung österreichischer Rechtstexte soll eine modulare Modellarchitektur die Skalierung für weitere Sprachen sowie Fachgebiete ermöglichen. Qualitätssicherungsmaßnahmen sowohl während des Pilotbetriebes als auch im Rahmen

³ <https://n-lex.europa.eu/n-lex/index>

⁴ https://www.lereto.com/databases_DE_gii-rii.html

der laufenden Evaluation mittels menschlichen Feedbacks und quantitativen Standards sollen den hohen Qualitätsanspruch sicherstellen. Damit kann PLAIInLaw4U in verschiedenen Anwendungsbereichen, vom Ausbildungssektor, bei juristischen Expert*innen über die Beratung bis zur Unterstützung vulnerabler Gruppen, breite Einsatzmöglichkeiten finden. Damit könnte PLAIInLaw4U exemplarisch für den Einsatz von KI zur Stärkung demokratischer Teilhabe, von gesellschaftlicher und digitaler Inklusion sowie für barrierefreie Kommunikation im Rechtsstaat stehen.

Literatur

- Alva-Manchego, F., Martin, L., Scarton, C. & Specia, L. (2019). EASSE: Easier Automatic Sentence Simplification Evaluation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations, Hong Kong, China* (S. 49–54). Association for Computational Linguistics. <https://doi.org/10.18653/v1/d19-3009>
- Amtssprache. (o.D.). In *Begrifflexikon Österreichs Digitales Amt*. <https://www.oesterreich.gv.at/lexicon/A/amtssprache.html>
- Anschütz, M., Oehms, J., Wimmer, T., Jezierski, B. & Groh, G. (2023). Language Models for German Text Simplification: Overcoming Parallel Data Scarcity through Style-specific Pre-training. In *Findings of the Association for Computational Linguistics: ACL 2023* (S. 1147–1158). Association for Computational Linguistics. <https://aclanthology.org/2023.findings-acl.74.pdf>
- Bender, E. M., Gebru, T., McMillan-Major, A. & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)* (S. 610–623). Association for Computing Machinery. <https://doi.org/10.1145/3442188.3445922>
- Bundesamt für Justiz (Hrsg.). (2022). *Zahl der Richter, Richterinnen, Staatsanwälte, Staatsanwältinnen und Vertreter, Vertreterinnen des öffentlichen Interesses in der Rechtspflege der Bundesrepublik Deutschland am 31. Dezember 2020* [Datensatz]. https://www.bundesjustizamt.de/SharedDocs/Downloads/DE/Justizstatistiken/Richterstatistik_2020.pdf?__blob=publicationFile&v=3
- Bundesrechtsanwaltskammer. (2025). *Statistiken der Bundesrechtsanwaltskammer*. Bundesrechtsanwaltskammer. Abgerufen am 2. Mai 2025, von <https://www.brak.de/presse/zahlen-und-statistiken/>
- Cemri, M., Çukur, T. & Koç, A. (2022). *Unsupervised Simplification of Legal Texts*. <https://doi.org/10.48550/arxiv.2209.00557>
- Die österreichische Justiz. (2025). *Personalstand in der Justiz*. Die österreichische Justiz. <https://www.justiz.gv.at/justiz/daten-und-fakten/personalstand-in-der-justiz.8ab4a8a-422985de30122a92fc5bc63a9.de.html>
- Die österreichische Justiz. (o.D.). *JustizOnline Begrifflexikon*. <https://justizonline.gv.at/jop/web/begrifflexikon>
- Die Presse. (2024, 26. Mai). Jus-Studium wird in Österreich am häufigsten abgebrochen, Medizin am wenigsten. *Die Presse*. <https://www.diepresse.com/18500995/jus-studium-wird-in-oesterreich-am-haeufigsten-abgebrochen-medizin-am-wenigsten>
- Floridi, L. (2023). AI as Agency Without Intelligence: on ChatGPT, Large Language Models, and Other Generative Models. *Philosophy & Technology*, 36(1), 15–22. <https://doi.org/10.1007/s13347-023-00621-y>

- Freyer, N., Kempt, H. & Klöser, L. (2024). Easy-read and large language models: on the ethical dimensions of LLM-based text simplification. *Ethics and Information Technology*, 26(3), 50–60. <https://doi.org/10.1007/s10676-024-09792-4>
- Gemeinsamer Europäischer Referenzrahmen für Sprachen (GER). (o.D.). <https://www.europaeischer-referenzrahmen.de/>
- Huang, Z., Wang, X., Zhang, F., Xu, Z., Zhang, C., Zheng, X. & Huang, X. (2024). *Enhancing the Capability and Robustness of Large Language Models through Reinforcement Learning-Driven Query Refinement*. <https://doi.org/10.48550/arxiv.2407.01461>
- Kempt, H., Lavie, A. & Nagel, S. K. (2023). *Appropriateness is all you need!* <https://doi.org/10.48550/arxiv.2304.14553>
- Kirk, H. R., Vidgen, B., Röttger, P. & Hale, S. A. (2023). *Personalisation within bounds: A risk taxonomy and policy framework for the alignment of large language models with personalised feedback*. <https://doi.org/10.48550/arxiv.2303.05453>
- Levine, C. (2004). The concept of vulnerability in disaster research. *Journal of Traumatic Stress*, 17(5), 395–402. <https://doi.org/10.1023/b:jots.0000048952.81894.f3>
- Li, Z., Belkadi, S., Micheletti, N., Han, L., Shardlow, M. & Nenadic, G. (2024). Investigating Large Language Models and Control Mechanisms to Improve Text Readability of Biomedical Abstracts. In *2022 IEEE 10th International Conference on Healthcare Informatics (ICHI)* (S. 265–274). <https://doi.org/10.1109/ichi61247.2024.00042>
- Melcher, M. (2022, 23. Januar). *Klein aber fein – Einfache Sprache in der Schweiz*. Infoportal Einfache Sprache. <https://portaleinfach.org/2022/01/23/klein-aber-fein-einfache-sprache-in-der-schweiz/>
- Parlament Österreich. (o.D.a). *Barrierefreiheit*. <https://www.parlament.gv.at/services/barrierefreiheit>
- Parlament Österreich. (o.D.b). *Einfache Sprache*. <https://www.parlament.gv.at/services/barrierefreiheit/einfache-sprache/>
- Poddar, S., Wan, Y., Ivison, H., Gupta, A. & Jaques, N. (2024). *Personalizing Reinforcement Learning from Human Feedback with Variational Preference Learning*. <https://doi.org/10.48550/arxiv.2408.10075>
- Pternea, M., Singh, P., Chakraborty, A., Oruganti, Y., Milletari, M., Bapat, S., & Jiang, K. (2024). *The RL/LLM Taxonomy Tree: Reviewing Synergies Between Reinforcement Learning and Large Language Models*. <https://doi.org/10.48550/arxiv.2402.01874>
- Rechtsinformationssystem des Bundes (RIS). (o.D.). <https://www.ris.bka.gv.at/>
- Schomacker, T., Gille, M., Von der Hülls, J. & Tropmann-Frick, M. (2023). *Data and Approaches for German Text simplification – towards an Accessibility-enhanced Communication*. <https://doi.org/10.48550/arxiv.2312.09966>
- Siththaranjan, A., Laidlaw, C. & Hadfield-Menell, D. (2023). *Distributional Preference Learning: Understanding and Accounting for Hidden Context in RLHF*. <https://doi.org/10.48550/arxiv.2312.08358>
- Spring, N., Kostrzewa, M., Rios, A. & Ebling, S. (2022). Ensembling and Score-Based Filtering in Sentence Alignment for Automatic Simplification of German Texts. In *Lecture notes in computer science* (S. 137–149). https://doi.org/10.1007/978-3-031-05028-2_8
- Statista. (2024, 3. September). *Advokatur- und Notariatsbüros in der Schweiz nach Kantonen 2022*. <https://de.statista.com/statistik/daten/studie/466172/umfrage/advokatur-und-notariatsbueros-in-der-schweiz-nach-kantonen/>
- Statista. (2025, 25. Februar). *Bevölkerung in den DACH-Ländern 2024*. <https://de.statista.com/statistik/daten/studie/1117261/umfrage/bevoelkerung-in-den-dach-laendern/>
- Stodden, R. (2023). *DEPlain*. <https://github.com/rstodden/DEPlain>
- Stodden, R. (2024a). *EASSE-DE*. <https://github.com/rstodden>
- Stodden, R. (2024b). *EASSE-DE: Easier Automatic Sentence Simplification Evaluation for German*. <https://doi.org/10.48550/arxiv.2404.03563>

- Stodden, R., Momen, O. & Kallmeyer, L. (2023). *DEPlain: A German Parallel Corpus with Intralingual Translations into Plain Language for Sentence and Document Simplification*. <https://doi.org/10.48550/arxiv.2305.18939>
- Thiele, M. (2020, 12. Februar). Deutsch in der Welt. *deutschland.de*. <https://www.deutschland.de/de/topic/wissen/muttersprache-deutsch-in-42-laendern-der-welt>
- UN-Behindertenrechtskonvention (2006). <https://broschuerenservice.sozialministerium.at/Home/Download?publicationId=19>
- UN-Kinderrechtskonvention (1989). <https://www.unicef.de/informieren/ueber-uns/fuer-kinderrechte/un-kinderrechtskonvention>
- Yi, J., Zhang, S., Zou, Z., Cheng, S., Zou, J.Y. & Zhang, C. (2024). *Jailbreak Attacks and Defenses Against Large Language Models: A Survey*. <https://doi.org/10.48550/arXiv.2407.04295>
- Zhou, J., Zhao, Y., Liu, J., Dong, P., Luo, X., Tao, H., Chang, S. & Luo, H. (2024). LLM-4RL: Enhancing Reinforcement Learning with Large Language Models. In *2024 IEEE Canadian Conference on Electrical and Computer Engineering (CCECE)*, Kingston, ON, Canada (S. 86–88). <https://doi.org/10.1109/ccece59415.2024.10667224>

Befreit KI die Welt von Arbeit?

Josef Sawetz

1. Einführung

Die Arbeitswelt und unser Zusammenhalt in der Gesellschaft stehen vor enormen Herausforderungen. Waren bisher vor allem körperliche Arbeit und Routinetätigkeiten von Automatisierungsprozessen betroffen, sind es jetzt „Wissensarbeiter:innen“ und kreative Berufe, die sich durch den rasanten Fortschritt der Künstlichen Intelligenz (KI) mit deren Auswirkungen auf ihr Berufsfeld auseinandersetzen müssen. Entscheidend dabei ist, ob KI als Komplementär- oder Substitutionsfaktor für menschliche Arbeit eingesetzt wird. Während einige Berufe durch KI vollständig ersetzt werden können, erfordern andere eine verstärkte Zusammenarbeit zwischen Mensch und Maschine. In der gegenwärtigen Ära des „Endes des Augenscheins“ ist KI gleichzeitig in der Lage, als Instrument der „Wirklichkeitssimulation“ zu fungieren. Damit ändern sich Kompetenzprofile mit einer stärkeren Konzentration auf lebenslanges Lernen, Adaptationsfähigkeit, soziotechnische Kompetenzen und eine grundsätzlich kritisch-reflexive Haltung. Mittelmäßigkeit läuft in jedem Berufsfeld die größte Gefahr, durch KI ersetzt zu werden. Herausragende Fähigkeiten und Expertise allerdings können durch KI als assistierende Agenten sowohl Produktivität als auch Qualität und Innovationsdynamik steigern.

2. Inkorporierte Motivation zur ökonomischen Nutzung von Ressourcen

Der menschliche Organismus ist auf Effizienzsteigerung ausgelegt. Das zeigt sich ganz besonders im Energieverbrauch des Gehirns. Es ist das Organ in unserem Körper, das am meisten Energie, also Kalorien, verbraucht – und zwar 20-mal so viel wie ein gleich großer Muskel. Um uns zu motivieren, mit der uns zur Verfügung stehenden Kalorienmenge sorgsam umzugehen, hat sich ein Mechanismus im Lauf der Evolution entwickelt, der immer dann das Belohnungszentrum im Gehirn aktiviert und zu einer Dopaminausschüttung führt, sobald etwas einfach und leicht für uns zu bewältigen ist.

Vom Feuer zur KI: exponentielle Evolution der Effizienzsteigerungstechniken

Im Laufe der Stammesgeschichte des Menschen haben eine ganze Reihe von kulturellen und technologischen Innovationen dieses Prinzip der stetigen Effizienzsteigerung verwirklicht. Dabei wurden die zeitlichen Abstände der einzelnen Innovationsschübe, die disruptiv zu Systemänderungen gesellschaftlicher Art geführt haben, immer kleiner: von der Beherrschung des Feuers über die Sprache, den Ackerbau und die Viehzucht, das Rad, die Schrift, den Transport der Schrift über Papyrus und Papier, die Vervielfältigung der Schrift über den Buchdruck, die Dampfmaschine, den Telegrafen, das Telefon, die Elektrizität, das Auto, das Flugzeug, das

Radio, das Fernsehen, den Computer, das Internet, das Mobiltelefon, die sozialen Medien und aktuell die künstliche Intelligenz.

Parallel zu der Effizienzsteigerung durch die einzelnen Innovationen hat auch eine Demokratisierung des Zugangs zu diesen Technologien stattgefunden. So hat zum Beispiel das Internet den Zugang zu weltweitem Wissen demokratisiert, die sozialen Medien die freie Meinungsäußerung gegenüber einem globalen Publikum und die künstliche Intelligenz führt aktuell zu einer Reduktion der Schwelle von der Idee zur Umsetzung. Heute kann jeder Mensch auch ohne die entsprechenden Kenntnisse und Fertigkeiten dank KI eine Schriftstellerin sein, eine Musikerin, eine Filmemacherin, eine Illustratorin, eine Fotografin, eine Grafikerin, eine Webdesignerin, eine Programmiererin. Die Grenze ist die eigene Vorstellungskraft, der Ideenreichtum und die Kreativität. Umso wichtiger ist es heute, das eigene Denken zu befreien und aus festgefahrenen gedanklichen Korridoren auszubrechen. Die Welt war noch nie so groß und offen wie heute. Allerdings nur für die, deren gedankliche Freiheitsgrade mit dieser Offenheit und Größe gleichziehen können.

3. Ende des Augenscheins

Eine der Gefahren dieser Entwicklung des niederschweligen Zugangs zur Umsetzung eigener Ideen liegt in der oben erwähnten evolutionsbiologisch prädisponierten Bequemlichkeit. In der individuell unterschiedlich ausgeprägten Verteilung von Bequemlichkeit versus Neugierde wird ein Teil der Menschen den Weg der Bequemlichkeit wählen und damit in eine Abhängigkeit von der Mächtigkeit der künstlichen Intelligenz geraten. Das eigene Denken lässt sich nicht ungestraft outsourcen. Denn wir sind schon heute eingetaucht in das Zeitalter des „Endes des Augenscheins“. KI ist das bislang mächtigste Werkzeug zur Wirklichkeitssimulation. Fake News und Biased News waren noch nie so schwer zu entlarven, und zwar sowohl wegen der perfekten Herstellung simulierter Wirklichkeiten als auch wegen des zunehmenden Verlusts der kritischen Reflexion seitens des Publikums. Diejenigen jedoch, die Einfachheit und Bequemlichkeit fokussiert und zielgerichtet einsetzen, um Zeit zu sparen, zum Beispiel bei lästigen Routinearbeiten, und diese eingesparte Zeit nutzen, um ihre Neugierde zu neuen Horizonten aufbrechen zu lassen, werden zu den Treibern neuer Innovationen und damit „einfach mächtig“ – auf einfache Weise in ihrer Produktivität und Kreativität gestärkt. Und sie werden im besten Fall auch resistenter gegenüber der persuasiven Kraft künstlich hergestellter Perspektiven auf unsere Welt.

Sowohl die Trainingsdaten als auch die Algorithmen der KI sind in ihren Perspektiven und Gewichtungen nicht ausgeglichen. Algorithmic Biases und Arithmetic Prominence, Dominanz der englischen Sprache und Tendenz zu Mainstream und Durchschnittlichem sind nur einige der vielen Facetten der Wirklichkeit, die unseren Blick auf die Welt verzerren können, wenn wir in Zukunft die Sachverhalte unserer Lebenswirklichkeit vermehrt nur mehr durch die Brille der KI wahrnehmen können. Die generelle Tendenz zur Mittelwertbildung führt dazu, dass innovative oder nichtkonforme Perspektiven marginalisiert werden (Russell & Norvig, 2021). Die ausschließliche Rezeption KI-generierter Inhalte kann somit zu einer einseitigen Wahrnehmung sozialer und wissenschaftlicher Diskurse führen.

Es besteht damit die große Gefahr einer noch stärkeren Polarisierung am Arbeitsmarkt und generell in den Entfaltungsmöglichkeiten der Menschen. Im besten Fall führt der zukünftige globale Einsatz künstlicher Intelligenz in allen Bereichen der Wirtschaft und Gesellschaft zu einer zunehmenden „Befreiung“ von Arbeit, um die Potenziale eines jeden Menschen in bislang nicht gekanntem Umfang selbstbestimmt freisetzen zu können.

4. Der schmale Grat zwischen Utopie und Dystopie

Zwischen Dystopie und Utopie liegt dabei aber nur ein schmaler Grat, der jeweils zu einem fundamentalen Systembruch führt. Künstliche Intelligenz wird zu einem so grundlegenden Baustein unserer Gesellschaft und unseres Wirtschafts- und Arbeitslebens werden, wie es heute zum Beispiel schon die Elektrizität und das Internet darstellen. Wenn man heute den Stecker zieht, bleibt die Welt stehen. So wie das Internet und die Elektrizität wird sich auch die künstliche Intelligenz in alle Prozesse unseres Lebens hineinfressen und damit untrennbar in alles verwoben sein.

Dies betrifft nicht nur Einzelpersonen, sondern auch Organisationen und Gesellschaften. Der schleichende Verlust an Analyse- und Entscheidungsfähigkeit könnte dazu führen, dass wesentliche gesellschaftliche Funktionen zunehmend von KI-Systemen übernommen werden, während Menschen ihre kognitive Eigenständigkeit verlieren. Langfristig könnten dadurch grundlegende demokratische Prozesse und Entscheidungsmechanismen unterwandert werden.

Es zeichnet sich bereits heute eine Zukunft ab, in der unterschiedliche Entwicklungsstränge sich miteinander verbinden – beschleunigt durch ein zunehmendes Wettrennen um die Vorherrschaft in den neuen Technologien zwischen den USA, Europa und China: AGI – Artificial General Intelligence, Quantencomputing, Humanoide Roboter und Synthetische Biologie. Dass der Mensch an einem zukünftigen Punkt der Evolution eine neue Art erschaffen könnte, ist nicht nur theoretisch möglich. Durch die zunehmende Beschleunigung der Entwicklungen sind die aktuellen Systeme des Wirtschaftens und der Verteilung von Arbeit, Einkommen und Ressourcen auf eine harte Probe gestellt. Das Problem der schrumpfenden Bevölkerungszahlen in vielen entwickelten Ländern könnte eine große Chance werden, diesen kommenden Systembruch ohne einschneidende soziale Verwerfungen bewältigen zu können.

2025 steht bereits mit Beginn des Jahres im Zeichen des nächsten Evolutionschrittes der KI: Agentic AI – autonom arbeitende KI-Assistenten, die auch komplexere Aufgaben Schritt für Schritt eigenständig ausführen können. Beispielsweise nutzt das KI-System „AlphaFold“ von DeepMind maschinelles Lernen, um Proteinstrukturen vorherzusagen (Jumper et al., 2021). Eine zukünftige Agentic AI könnte eigenständig neue Moleküle für Medikamente entwickeln und Laborversuche optimieren.

5. Intelligenz als adaptive Informationsverarbeitung ist ein artübergreifendes Phänomen

Die Erforschung des Phänomens „Intelligenz“ hat sich im Laufe der vergangenen Jahrzehnte stark weiterentwickelt und ist heute Gegenstand verschiedener Disziplinen: Evolutionsbiologie, Neurowissenschaften, Philosophie und Psychologie gehören ebenso dazu wie die Forschung zu Künstlicher Intelligenz (KI). Der Begriff „Intelligenz“ umfasst dabei eine breite Palette an Fähigkeiten und Anpassungen. Historisch lag der Schwerpunkt hauptsächlich auf der menschlichen Intelligenz (Stern, 1914), doch es zeigte sich schnell, dass auch Tiere – von Wirbeltieren wie Primaten oder Vögeln bis zu Tintenfischen – komplexe kognitive Fähigkeiten aufweisen können (Shettleworth, 2010). Ebenso rückt vermehrt die Frage in den Vordergrund, ob und wie Pflanzen über Sensorik und Kommunikation verfügen, die als „Intelligenz“ angesehen werden kann. Zudem wirft die stetige Weiterentwicklung der Künstlichen Intelligenz (LeCun et al., 2015; Russell & Norvig, 2021) grundlegende Fragen auf: Können Computerprogramme oder autonome Systeme jemals echte „Intelligenz“ im Sinne eines komplexen, eigenständigen Erkenntnisvermögens entwickeln?

Bewusstsein und semantische Intelligenz

Aktuelle KI-Systeme können hochkomplexe Aufgaben lösen, ohne jedoch ein Phänomen zu zeigen, das wir „Bewusstsein“ nennen würden. Viele Neurowissenschaftler:innen vermuten, dass es für das Entstehen von Bewusstsein ein integriertes, globales neuronales Netzwerk braucht – ein Ansatz, der als Global Workspace Theory bekannt ist (Dehaene & Changeux, 2011).

In der philosophischen und kognitionswissenschaftlichen Debatte wird daher differenziert zwischen „starker KI“, die ein eigenes Bewusstsein entwickelt, und „schwacher KI“, die lediglich kognitive Funktionen simuliert.

Aktuelle Entwicklungen wie Transformer-Architekturen oder neuro-symbolische KI (Garcez & Lamb, 2023) versuchen, statistische Verfahren und symbolisches Schließen zu vereinen. Diese Integration könnte langfristig zu KI-Systemen führen, die mehr Kontextverständnis und interpretierbare Entscheidungsprozesse aufweisen (Adadi & Berrada, 2018; Rudin, 2019).

Gemeinsamkeiten und divergierende Prinzipien zwischen biologischer und künstlicher Intelligenz

Ein entscheidender Aspekt der biologischen Intelligenz ist die Neuroplastizität – die Fähigkeit des Gehirns, sich strukturell und funktionell an veränderte Umweltbedingungen anzupassen. Lernprozesse, sei es durch synaptische Modifikation oder durch Langzeitpotenzierung, ermöglichen es, Erfahrungen in dauerhafte neuronale Veränderungen zu überführen (Lillicrap et al., 2020). Der kontinuierliche Anpassungsprozess wird durch Feedbackschleifen und Rückkopplungsschleifen innerhalb der neuronalen Netzwerke unterstützt.

Im Gegensatz dazu erfolgt das Lernen in künstlichen Systemen überwiegend durch algorithmisch gesteuerte Optimierungsverfahren wie den Backpropagation-Algorithmus. Obwohl diese Ansätze in vielen Bereichen erstaunliche Erfolge erzielen, fehlt ihnen oftmals die intrinsische Flexibilität und Kontextsensitivität, die biolo-

gischen Systemen eigen ist (Yamins & DiCarlo, 2016). In neurowissenschaftlichen Modellen wird zudem zunehmend versucht, adaptivere Lernmechanismen zu integrieren, die die biologischen Prinzipien besser abbilden (Richards et al., 2019). Der Vergleich zeigt, dass künstliche Systeme an die evolutionär verfeinerten Mechanismen des menschlichen Gehirns noch nicht herankommen.

Trotz fundamentaler Unterschiede zeigen sowohl biologische als auch künstliche Intelligenz Merkmale adaptiver Lernprozesse, Mustererkennung und der Fähigkeit, komplexe Probleme zu lösen. Neurowissenschaftliche Modelle, die auf Vorhersageverarbeitung basieren, finden in modernen KI-Algorithmen Anwendung, während gleichzeitig philosophische Diskurse den Rahmen für Fragen zu Bewusstsein und intentionaler Subjektivität bilden. Die Schnittmenge dieser Disziplinen fördert eine integrative Sichtweise: Biologische Systeme inspirieren technische Entwicklungen, und Fortschritte in der KI werfen neue Fragen zur Natur des Geistes auf (Clark, 2013; Hassabis et al., 2017).

Deep-Learning-Systeme sind jedoch hochspezialisiert: Ein neuronales Netz, das in der Objekterkennung exzellent abschneidet, muss neu trainiert oder angepasst werden, um eine andere Aufgabe zu lösen (Lake et al., 2017). Diese fehlende Generalisierungsfähigkeit ist ein entscheidendes Unterscheidungsmerkmal gegenüber menschlichen Lernprozessen, die universeller und kontextabhängiger funktionieren.

Synergie oder Zusammenwachsen biologischer und künstlicher Intelligenz?

Die rasanten Fortschritte im Bereich des Deep Learning und der Robotik lassen vermuten, dass KI-Systeme in den nächsten Jahren noch autonomer, leistungsfähiger und vielfältiger werden (Russell & Norvig, 2021). Damit einher gehen aber auch Fragen über Datensicherheit, gesellschaftliche Kontrolle und mögliche Arbeitsplatzverluste, die bereits heute hochaktuell sind (Jobin et al., 2019).

Langfristig bleibt offen, ob künstliche Systeme jemals ein Bewusstsein entwickeln oder sich einer biologischen Intelligenz in all ihren Facetten annähern werden können. Neurowissenschaftlich spricht viel dafür, dass Bewusstsein und kognitive Prozesse tief in die organische Struktur eines Organismus eingebettet sind (Koch et al., 2016). Philosophisch ist hingegen denkbar, dass Funktionalität und Komplexität ausreichen, um Bewusstsein in nichtbiologischen Substraten entstehen zu lassen (Koch et al., 2016).

6. Kombination und Wettbewerb von menschlicher und KI-Intelligenz

Eine Untersuchung analysierte die Zusammenarbeit zwischen Ärzten und KI-Systemen bei der Erstellung von Differenzialdiagnosen. Die Ergebnisse zeigten, dass hybride Teams aus menschlichen Ärzten und KI-Systemen sowohl allein arbeitende Ärzte als auch ausschließlich KI-basierte Diagnosesysteme übertrafen. Dies verdeutlicht den Mehrwert einer kollaborativen Herangehensweise in der medizinischen Diagnostik (Zöller et al., 2024).

An der Ohio State University wurde untersucht, wie Menschen auf Antworten von Psychotherapeuten im Vergleich zu Texten von ChatGPT reagieren. Die von der

KI generierten Antworten wurden von den Versuchsteilnehmer:innen als einfühlsamer und hilfreicher bewertet (Hatch et al., 2025).

Von Ende 2022 bis Anfang 2024 fand Strzelecki (2025) 1.362 Fachartikel, in denen typische ChatGPT-Phrasen auftauchten. In 89 Fällen handelte es sich um peer-reviewed Publikationen. Strzelecki berichtet, dass 64 dieser Artikel sogar in Journalen erschienen sind, die zum Top-Quartil (Viertel) ihres Fachgebiets gehören. Nach Strzelecki muss davon ausgegangen werden, dass es viele weitere Artikel gibt, die zumindest zum Teil durch KI generiert wurden, die aber diese Phrasen nicht mehr enthalten. Zusätzlich ist zu bedenken, dass die KI-Verwendung seit Anfang 2024 massiv zugenommen hat. Da KI auch Fehler und Halluzinationen erzeugt und auch Gutachter sich vermehrt auf den Einsatz von KI verlassen, steht die Zuverlässigkeit und das Vertrauen der Wissenschaft auf dem Spiel – mit gravierenden möglichen Auswirkungen vor allem in so sensiblen Bereichen wie der Medizin. Vor allem, wenn man bedenkt, dass KI auch sehr gut empirische Daten simulieren kann ...

Positive AI-Halluzinationen

Unter einer „KI-Halluzination“ versteht man in der Regel die Abweichung eines KI-generierten Outputs von der überprüfbaren Realität (Bender et al., 2021). Beispiele dafür sind frei erfundene Textpassagen, falsche Zitate oder das Erfinden von Zusammenhängen ohne faktische Basis. Im Gegensatz dazu bezieht sich der Begriff „positive AI-Halluzination“ auf konstruktive Irrtümer oder Einfälle, die in einem kreativen Kontext nützlich sein können. Hierbei handelt es sich nicht um das absichtliche Fälschen von Daten, sondern vielmehr um eine Art struktureller Kreativität, die aus der statistisch-prognostischen Natur großer Sprachmodelle resultiert (Bommasani et al., 2021).

Beispiele für mögliche positive Effekte solcher Halluzinationen:

1. *Hypothesengenerierung*: Das KI-Modell schlägt auf Basis seiner „Assoziationen“ neuartige Forschungsfragen oder experimentelle Ansätze vor.
2. *Kreative Ideenfindung*: In der Konzeptionsphase neuer Projekte können ungewöhnliche oder „unlogisch“ anmutende Vorschläge einen Perspektivenwechsel anstoßen.
3. *Methodische Vielfalt*: KI-basierte „Fehlannahmen“ regen zur Entwicklung neuer methodischer Herangehensweisen an, um mögliche Lösungen für die vorgebrachten Ideen zu prüfen.

Positive AI-Halluzinationen markieren einen Bereich, in dem das Potenzial von KI für kreative Impulse in der wissenschaftlichen Forschung ausgelotet wird. Diese kreativen „Fehlleistungen“ großer Sprachmodelle könnten – kontrolliert eingesetzt – neue Perspektiven und originelle Forschungshypothesen fördern. Dennoch ist Skepsis geboten: Halluzinierte Inhalte dürfen nicht mit empirischen Fakten verwechselt werden und erfordern eine sorgfältige Überprüfung, um wissenschaftlichen Qualitätsstandards zu genügen (Bender et al., 2021).

Technischer Innovationskorridor

Folgende zukünftige Fortschritte in der KI-Entwicklung zeichnen sich bereits jetzt ab:

- Verbesserte Generalisierungsfähigkeit
 - Domänenübergreifendes Lernen
 - Flexible Adaptation
 - Robuste Transferleistung
- Erhöhte Energieeffizienz
 - Optimierte Architekturen
 - Neuromorphe Hardware
 - Quantencomputing-Integration
- Erweiterte kognitive Fähigkeiten
 - Kausales Verständnis
 - Abstraktes Reasoning
 - Kreative Synthese

Gesellschaftliche Implikationen der Weiterentwicklung von Künstlicher Intelligenz

Die weitreichenden gesellschaftlichen Auswirkungen könnten sich mit hoher Wahrscheinlichkeit in diesen Bereichen zeigen:

1. Bildungssystem:
 - Personalisierte Lernpfade
 - Adaptive Curricula
 - Multimodale Lernumgebungen
 - Integration neuer Technologien
 - KI-gestützte Tutorsysteme
 - Virtuelle Lernräume
 - Augmented Reality Learning
2. Arbeitswelt:
 - Transformation der Arbeitsprozesse
 - Mensch-Maschine-Kollaboration
 - Adaptive Arbeitsumgebungen
 - Kognitive Assistenzsysteme
 - Neue Kompetenzanforderungen
 - Lebenslanges Lernen
 - Adaptationsfähigkeit
 - Sozio-technische Kompetenzen

7. Historische Kontextualisierung von Automatisierung und KI

Technologische Innovationen haben historisch betrachtet stets die Arbeitswelt verändert. Die industrielle Revolution führte zur Mechanisierung von Arbeitsprozessen, die digitale Revolution beschleunigte Automatisierung im Bürobereich, und KI steht nun als nächste Entwicklungsstufe bevor. Diese Entwicklungen unterstreichen die Notwendigkeit, bestehende Paradigmen der Arbeitsmarktanalyse zu überdenken. Jede technologische Revolution führte sowohl zur Schaffung neuer Arbeitsplätze als auch zur Eliminierung bestehender Berufe. Es gilt, die historische Dynamik dieser Transformation zu verstehen, um effektive Strategien zur Anpassung an die Ära der KI zu entwickeln.

Die zunehmende Implementierung von KI-Technologien verändert signifikant die Strukturen des Arbeitsmarktes. Während technologische Fortschritte in der Vergangenheit primär mechanische oder repetitive Tätigkeiten automatisierten, greift die heutige KI tief in kognitive und analytische Prozesse ein (Frey & Osborne, 2017). Dies erfordert eine kritische Auseinandersetzung mit den psychologischen und sozialen Konsequenzen dieser Entwicklung. Zudem wirft die verstärkte Integration von KI ethische Fragen auf, insbesondere in Bezug auf Arbeitsplatzsicherheit, soziale Gerechtigkeit und die Notwendigkeit einer neuen Definition von Arbeit im 21. Jahrhundert.

Die Verteilung von Arbeit verändert sich durch den Einsatz von KI auf mehreren Ebenen. Arntz et al. (2016) argumentieren, dass die Automatisierungspotenziale je nach Berufsfeld unterschiedlich ausfallen. Während in administrativen und routinemäßigen Tätigkeiten ein hohes Automatisierungspotenzial besteht, bleiben Jobs in Bereichen, die kreatives Denken und komplexe Problemlösungen erfordern, weitgehend unberührt. Gleichzeitig entstehen in wachstumsstarken Branchen neue Jobprofile – etwa in der Datenanalyse, Robotik oder IT-Sicherheit (World Economic Forum, 2023). Ob diese neu entstehenden Tätigkeiten jedoch quantitativ und qualitativ den Rückgang in anderen Bereichen kompensieren können, ist noch unklar.

KI kann dazu beitragen, diese Segmente weiter zu differenzieren, indem sie beispielsweise hochqualifizierte Tätigkeiten ergänzt, während weniger qualifizierte Jobs automatisiert werden. Dies kann zu einer Verstärkung bestehender Ungleichheiten führen, wenn nicht gleichzeitig Qualifizierungsmaßnahmen und Umstrukturierungsstrategien implementiert werden (Autor, 2015).

7.1 Automatisierung versus Augmentation

Ein entscheidender Aspekt ist die Frage, ob KI als Komplementär- oder Substitutionsfaktor für menschliche Arbeit gesehen wird. Während einige Berufe durch KI vollständig ersetzt werden können, erfordern andere eine verstärkte Zusammenarbeit zwischen Mensch und Maschine.

Die Integration von KI in Arbeitsprozesse führt damit zu zwei parallel verlaufenden Entwicklungen: Automatisierung und Augmentation. Wobei sich zeigt, dass in der überwiegenden Mehrzahl der Unternehmen KI-Systeme primär zur Augmentation menschlicher Fähigkeiten eingesetzt werden, während vollständige Automatisierung nur ein sekundäres Ziel darstellt. Während in der Fertigung Automatisierung domi-

niert, entstehen in analytischen und kreativen Branchen neue Berufsfelder. Besonders betroffen sind administrative Tätigkeiten, während technische und kreative Berufe von der KI-Unterstützung profitieren könnten.

7.2 Motivation und Arbeitszufriedenheit

Die Integration von KI in Arbeitsprozesse führt zu einer differenzierten Entwicklung der Arbeitszufriedenheit. In Bereichen, wo KI als unterstützendes Tool implementiert wird, steigt die Arbeitszufriedenheit. Gleichzeitig zeigen sich negative Effekte in Kontexten, in denen KI als Ersatz für menschliche Arbeit wahrgenommen wird.

Die zunehmende Human-AI-Kollaboration wird das Arbeitsleben grundlegend verändern. Teams werden künftig verstärkt hybride Zusammensetzungen aufweisen, in denen menschliche Kreativität und emotionale Intelligenz mit den analytischen und routinemäßigen Fähigkeiten von KI-Systemen kombiniert werden. Studien zeigen bereits heute, dass die Delegation von Aufgaben an KI-Systeme die Arbeitszufriedenheit und Selbstwirksamkeit der Mitarbeiter steigert (Hemmer et al., 2023). In fünf Jahren wird diese Form der Zusammenarbeit durch transparente Kommunikationsprozesse und regelmäßige Schulungen weiter optimiert sein, sodass ein hohes Maß an Vertrauen in die Technologie erreicht wird.

Die Automatisierung wird nicht zu einem massiven Arbeitsplatzverlust führen, sondern vielmehr die Art und Weise, wie Arbeit verteilt wird, neu organisieren. Routineaufgaben werden vermehrt von KI übernommen, während neue Berufsfelder entstehen, die hochspezialisierte Fähigkeiten erfordern. Dies führt zu einer stärkeren Polarisierung des Arbeitsmarkts, wobei die Einkommensschere zwischen hochqualifizierten und gering qualifizierten Arbeitskräften weiter auseinanderklaffen könnte – sofern keine gezielten politischen Maßnahmen ergriffen werden (Arntz et al., 2016; Autor, 2015).

Studien zu Human-AI Kollaboration zeigen, dass Aufgaben, die zwischen KI-Systemen und Menschen delegiert werden, zu einer Verbesserung der Arbeitsqualität und -zufriedenheit führen können (Hemmer et al., 2023).

Besonders relevant ist hierbei der Aspekt der Delegation: Wenn KI-Modelle in der Lage sind, Aufgaben zu delegieren – also zu entscheiden, welche Tätigkeiten von Menschen und welche von der Maschine übernommen werden sollen –, verbessert sich die Effizienz der Arbeitsprozesse. Gleichzeitig zeigt die empirische Forschung, dass solche Delegationsstrategien zu einem höheren Gefühl der Selbstwirksamkeit bei den Beschäftigten führen, was wiederum die Arbeitszufriedenheit steigert (Spitzer et al., 2024).

Arbeitsanforderungen: KI-getriebene Veränderungen können zu einer Reduktion monotoner Tätigkeiten führen, jedoch auch neue, komplexe Anforderungen generieren. Beispielsweise kann die Notwendigkeit entstehen, mit digitalen Systemen und automatisierten Prozessen interagieren zu müssen, was kognitive Belastungen und Anpassungsdruck erzeugt (Giering & Kirchner, 2021).

Arbeitsressourcen: Der Einsatz von KI kann als Ressource wirken, indem er Routineaufgaben übernimmt und den Beschäftigten ermöglicht, sich auf wertschöpfende Tätigkeiten zu konzentrieren. Gleichzeitig erfordern neue Technologien jedoch auch

zusätzliche Qualifikationen und Weiterbildungsmaßnahmen, um die Arbeitsressourcen angemessen nutzen zu können (Nachtwei, 2021).

Die Delegation komplexer Entscheidungsprozesse an KI-Systeme kann eine Devaluierung individueller Expertise bewirken. Besonders in wissensintensiven Berufen sind tiefgehende psychologische Anpassungsprozesse erforderlich (Giering & Kirchner, 2021). Arbeitnehmer müssen möglicherweise ihre berufliche Identität in einer KI-geprägten Arbeitswelt neu definieren. Die individuelle Anpassungsfähigkeit spielt hierbei eine zentrale Rolle für eine erfolgreiche Integration in die veränderten Arbeitsprozesse.

Die Integration von KI verändert traditionelle Formen der Teamarbeit und erfordert neue kollaborative Modelle. Empirische Studien legen nahe, dass hybride Mensch-KI-Teams effektiver arbeiten können, sofern strategische Kompetenzentwicklungsmaßnahmen etabliert werden (Hemmer et al., 2023). Eine entscheidende Frage ist, inwieweit Unternehmen KI als Mittel zur Produktivitätssteigerung nutzen können, ohne dabei das soziale Gefüge von Arbeitsgruppen negativ zu beeinflussen.

7.3 Neue Organisationsformen und -strukturen durch KI

KI führt in unterschiedlicher Geschwindigkeit in den einzelnen Ländern und Branchen zur Entstehung neuer Organisationsformen und -strukturen:

- Hybride Teams (Mensch-KI-Kollaboration)
- Hybride Kommunikationsnetzwerke (Mensch-KI-Mensch)
- Flexible und dynamische Kompetenzcluster
- Flexible Teamnetzwerke mit KI-Integration
- Projektbasierte Matrixstrukturen
- Virtuelle Kollaborationsräume
- Dynamische Projektstrukturen
- Automatisierte Routineaufgaben bei gleichzeitiger Aufwertung kreativer Tätigkeiten
- Hybride Entscheidungsprozesse (Mensch-KI-Interaktion)
- Abflachung hierarchischer Strukturen durch KI-gestützte Entscheidungsprozesse
- Verschiebung von vertikalen zu horizontalen Koordinationsmechanismen
- Entstehung neuer Autoritätsformen basierend auf KI-Expertise
- Entwicklung hybrider Organisationskulturen
- Integration von KI-Ethik in Unternehmenswerte

7.4 Neue Kernkompetenzen durch KI

Damit verbunden sind neue Anforderungen an Kernkompetenzen:

- KI-Literalität und digitale Adaptationsfähigkeit
- Komplexes Problemlösen im KI-Kontext
- Kreative Nutzung von KI-Tools
- Ethische Entscheidungsfindung

Die Auswirkung der KI-Integration auf die soziale Mobilität besteht dabei vor allem in folgenden Aspekten:

- Veränderung traditioneller Karrierewege
- Neue Zugangsmöglichkeiten zu qualifizierten Tätigkeiten
- Entstehung digitaler Elitenstrukturen

Durch den Wegfall routinierter Aufgaben gewinnen Tätigkeiten an Bedeutung, die höhere kognitive, kommunikative und kreative Fähigkeiten erfordern. KI beeinflusst insbesondere Tätigkeiten mit mittlerem Qualifikationsniveau. Während hoch- und niedrigqualifizierte Tätigkeiten in unterschiedlicher Weise bestehen bleiben, geraten mittlere Berufsprofile (z.B. Sachbearbeitung) unter Druck.

Agile Organisationsmodelle, die auf Interdisziplinarität und schnelle Feedbackzyklen setzen, profitieren in besonderem Maße von KI, erfordern jedoch eine Kultur, in der Experimente und Fehler als Teil des Lernprozesses angesehen werden.

7.5 Struktureller Wandel des Arbeitsmarktes

Die Implementierung von KI-Technologien führt zu einer zunehmenden Polarisierung des Arbeitsmarktes. Tätigkeiten mit repetitivem oder regelbasiertem Charakter, insbesondere in den Bereichen Fertigung und Finanzdienstleistungen, unterliegen einer hohen Substituierbarkeit durch automatisierte Systeme (Frey & Osborne, 2023). Parallel dazu entstehen neue Berufsfelder in der Entwicklung von KI-Technologien, datengetriebenen Entscheidungsprozessen sowie in der IT-Sicherheitsbranche (World Economic Forum, 2023). Diese Entwicklung erfordert eine kontinuierliche Anpassung bestehender Arbeitsstrukturen und eine enge Verzahnung zwischen akademischer Forschung und industrieller Praxis, um den disruptiven Charakter der KI-Integration proaktiv zu gestalten.

8. Zukunftsszenario: der Arbeitsmarkt in fünf Jahren

In fünf Jahren wird die Integration von KI in nahezu allen betrieblichen Bereichen zur Normalität gehören. Routineaufgaben in der Produktion, im Verwaltungswesen und im Kundenservice werden verstärkt automatisiert, sodass menschliche Arbeitskräfte sich auf komplexe und kreative Tätigkeiten konzentrieren können. Die kontinuierliche Verbesserung von Algorithmen, etwa durch In-Context Learning und Retrieval-Augmented Generation, wird dafür sorgen, dass KI-Systeme nicht nur standardisierte Aufgaben erledigen, sondern auch adaptive Entscheidungshilfen bieten (Hemmer et al., 2023). Diese Entwicklung dürfte zu einer signifikanten Steigerung der betrieblichen Produktivität führen und Unternehmen bedeutende Wettbewerbsvorteile verschaffen.

Weiters werden in den nächsten fünf Jahren hybride Teams, in denen KI als aktiver Partner in der Problemlösung integriert ist, zur Norm werden. Diese Teams werden nicht nur effizienter arbeiten, sondern auch flexibler auf Veränderungen reagieren können. Gleichzeitig wird es notwendig sein, neue Formen der Führung und des Ma-

nagements zu entwickeln, die sowohl technologische als auch menschliche Ressourcen optimal koordinieren.

8.1 Mensch-Maschine-Interaktion und Teamdynamik

Obwohl Zukunftsprognosen im dynamischen Technologiebereich stets mit Unsicherheit behaftet sind, zeichnen sich laut aktuellem Forschungsstand einige Trends ab (World Economic Forum, 2023):

1. Tiefere Integration von KI in verschiedenen Branchen, unterstützt durch immer leistungsfähigere Algorithmen. Automatisierung erstreckt sich dabei über administrative, analytische und strategische Domänen.
2. Zunahme projektbasierter und spezialisierter Tätigkeiten, da KI standardisierte Prozesse übernimmt.
3. Weiterentwicklung der Organisationskultur, hin zu agilen und lernorientierten Strukturen.
4. Steigende Bedeutung metakognitiver und sozialer Kompetenzen, darunter Kreativität, Konfliktmanagement und ethisches Urteilsvermögen.
5. Mögliche Polarisierung, wenn keine ausreichenden Weiterbildungs- und Umschulungsprogramme vorhanden sind.
6. Erhöhte digitale Kompetenzanforderungen: Unternehmen verlangen verstärkt technologische Adaptivität und analytische Fähigkeiten.
7. Transformierte Organisationsstrukturen: Interdisziplinäre KI-gestützte Teams setzen sich zunehmend durch.
8. Neu definierte Arbeitsmodelle: Eine verstärkte Flexibilisierung von Arbeitszeitmodellen und Remotearbeit könnten sich durch den Einfluss von KI weiter etablieren.
9. Veränderte Führungskonzepte: KI wird zunehmend auch in der Entscheidungsfindung auf Managementebene eine Rolle spielen.

Literatur

- Adadi, A. & Berrada, M. (2018). Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access*, 6, 52138–52160. <https://doi.org/10.1109/access.2018.2870052>
- Arntz, M., Gregory, T. & Zierahn, U. (2016). *The Risk of Automation for Jobs in OECD Countries. A comparative Analysis* (OECD Social, Employment and Migration Working Papers No. 189). OECD Publishing. <https://doi.org/10.1787/5jlz9h56dvq7-en>
- Autor, D. H. (2015). Why Are There Still So Many Jobs? The History and Future of Workplace Automation. *The Journal of Economic Perspectives*, 29(3), 3–30. <https://doi.org/10.1257/jep.29.3.3>
- Bender, E. M., Gebru, T., McMillan-Major, A. & Shmitchell, S. (2021). *On the Dangers of Stochastic Parrots* (S. 610–623) [2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT'21)]. Association for Computing Machinery. <https://doi.org/10.1145/3442188.3445922>
- Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., Sydney, V. A., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Cas-

- tellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. Q., Demszky, D. ... Liang, P. (2021). *On the Opportunities and Risks of Foundation Models*. <https://arxiv.org/abs/2108.07258>
- Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral And Brain Sciences*, 36(3), 181–204. <https://doi.org/10.1017/s0140525x12000477>
- Dehaene, S. & Changeux, J.-P. (2011). Experimental and Theoretical Approaches to Conscious Processing. *Neuron*, 70(2), 200–227. <https://doi.org/10.1016/j.neuron.2011.03.018>
- Frey, C. B. & Osborne, M. (2017). The future of employment: How susceptible are jobs to computerisation? *Technological Forecasting and Social Change*, 114, 254–280. <https://doi.org/10.1016/j.techfore.2016.08.019>
- Frey, C. B. & Osborne, M. (2023). Generative AI and the Future of Work: A Reappraisal. *The Brown Journal of World Affairs*, 30(1), 1–17.
- Garcez, A. D. & Lamb, L. C. (2023). Neurosymbolic AI: the 3rd wave. *Artificial Intelligence Review*, 56(11), 12387–12406. <https://doi.org/10.1007/s10462-023-10448-w>
- Giering, O. & Kirchner, S. (2021). Künstliche Intelligenz am Arbeitsplatz. Forschungsstand, Konzepte und empirische Zusammenhänge zu Autonomie. *Soziale Welt*, 72(4), 551–588. <https://doi.org/10.5771/0038-6073-2021-4-551>
- Hassabis, D., Kumaran, D., Summerfield, C. & Botvinick, M. (2017). Neuroscience-Inspired Artificial Intelligence. *Neuron*, 95(2), 245–258. <https://doi.org/10.1016/j.neuron.2017.06.011>
- Hatch, S. G., Goodman, Z. T., Vowels, L., Hatch, H. D., Brown, A. L., Guttman, S., Le, Y., Bailey, B., Bailey, R. J., Esplin, C. R., Harris, S. M., Holt, D. P., McLaughlin, M., O’Connell, P., Rothman, K., Ritchie, L., Top, D. N. & Braithwaite, S. R. (2025). When ELIZA meets therapists: A Turing test for the heart and mind. *PLOS Mental Health*, 2(2), e0000145. <https://doi.org/10.1371/journal.pmen.0000145>
- Hemmer, P., Westphal, M., Schemmer, M., Vetter, S., Vössing, M. & Satzger, G. (2023). *Human-AI Collaboration: The Effect of AI Delegation on Human Task Performance and Task Satisfaction* (S. 453–463) [Proceedings of the 28th International Conference on Intelligent User Interfaces (IUI’23)]. Association for Computing Machinery. <https://doi.org/10.1145/3581641.3584052>
- Jobin, A., Ienca, M. & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, 1(9), 389–399. <https://doi.org/10.1038/s42256-019-0088-2>
- Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Židek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J. ... Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873), 583–589. <https://doi.org/10.1038/s41586-021-03819-2>
- Koch, C., Massimini, M., Boly, M. & Tononi, G. (2016). Neural correlates of consciousness: progress and problems. *Nature Reviews. Neuroscience*, 17(5), 307–321. <https://doi.org/10.1038/nrn.2016.22>
- Lake, B. M., Ullman, T. D., Tenenbaum, J. B. & Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and Brain Sciences*, 40. <https://doi.org/10.1017/s0140525x16001837>
- LeCun, Y., Bengio, Y. & Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436–444. <https://doi.org/10.1038/nature14539>
- Lillicrap, T. P., Santoro, A., Marris, L., Akerman, C. J. & Hinton, G. (2020). Backpropagation and the brain. *Nature Reviews. Neuroscience*, 21(6), 335–346. <https://doi.org/10.1038/s41583-020-0277-3>
- Nachtwei, J. (2021). KI und Psychologie: ein Weckruf. *Business Psychology To-Go*, 2, 30–33.

- Richards, B. A., Lillicrap, T. P., Beaudoin, P., Bengio, Y., Bogacz, R., Christensen, A., Clopath, C., Costa, R. P., De Berker, A., Ganguli, S., Gillon, C. J., Hafner, D., Kepecs, A., Kriegeskorte, N., Latham, P., Lindsay, G. W., Miller, K. D., Naud, R., Pack, C. C. ... Kording, K. P. (2019). A deep learning framework for neuroscience. *Nature Neuroscience*, 22(11), 1761–1770. <https://doi.org/10.1038/s41593-019-0520-2>
- Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5), 206–215. <https://doi.org/10.1038/s42256-019-0048-x>
- Russell, S. & Norvig, P. (2021). *Artificial Intelligence: A Modern Approach* (4. Aufl.). Pearson Education Limited.
- Shettleworth, S. J. (2010). *Cognition, Evolution, and Behavior* (2. Aufl.). Oxford University Press. <https://doi.org/10.1093/oso/9780195319842.001.0001>
- Spitzer, P., Holstein, J., Hemmer, P., Vössing, M., Kühl, N., Martin, D. & Satzger, G. (2024). *Human Delegation Behavior in Human-AI Collaboration: The Effect of Contextual Information*. <https://arxiv.org/abs/2401.04729>
- Stern, W. (1914). *The psychological methods of testing intelligence*. Warwick & York. <https://doi.org/10.5962/bhl.title.31110>
- Strzelecki, A. (2025). ‘As of my last knowledge update’: How is content generated by ChatGPT infiltrating scientific papers published in premier journals? *Learned Publishing*, 38(1), e1650. <https://doi.org/10.1002/leap.1650>
- World Economic Forum. (2023). *Future of Jobs Report 2023*. https://www3.weforum.org/docs/WEF_Future_of_Jobs_2023.pdf
- Yamins, D. L. K. & DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience*, 19(3), 356–365. <https://doi.org/10.1038/nn.4244>
- Zöller, N., Berger, J., Lin, I., Fu, N., Komarneni, J., Barabucci, G., Laskowski, K., Shia, V., Harack, B., Chu, E. A., Trianni, V., Kurvers, R. H. J. M. & Herzog, S. M. (2024). *Human-AI collectives produce the most accurate differential diagnoses*. <https://arxiv.org/abs/2406.14981>

Autorinnen und Autoren

Helene BAUMGARTNER, *Mag.^a*, Betriebswirtin und Expertin im Büro für Digitale Agenden der Arbeiterkammer Wien. Arbeitsschwerpunkte: Digitalisierung aus der Perspektive der Arbeitnehmer:innen mit besonderem Fokus auf die Themen algorithmisches Management und Cybersecurity im Kontext der Arbeitswelt.

Philipp BESINGER, *Dipl.-Ing.*, Senior Projektleiter bei Fraunhofer Austria Research GmbH, Center für nachhaltige Produktion und Logistik. Forschungsschwerpunkte: generative KI in der Produktion, verantwortungsvolle Anwendung von KI (Responsible AI).

Sandra BILEK-STEINDL, *Mag.*, Senior Economist am Österreichischen Institut für Wirtschaftsforschung (WIFO) in der Forschungsgruppe Makroökonomie und öffentliche Finanzen. Arbeitsschwerpunkte: Analyse und Diagnose der österreichischen und internationalen Wirtschaftsentwicklung, Volkswirtschaftliche Gesamtrechnung, Konjunktur und Wachstum.

Julia BOCK-SCHAPPELWEIN, *Mag.*, Senior Economist am Österreichischen Institut für Wirtschaftsforschung (WIFO) in der Forschungsgruppe Arbeitsmarktökonomie, Einkommen und soziale Sicherheit. Arbeitsschwerpunkte: Arbeitsmarkt und Bildung, Arbeitsmarkt und Gender-Fragen, Arbeitsmarkt und Migration, Zukunft der Arbeit, Demographie, Digitalisierung, Green Jobs.

Eva EGDELING, *Dr.*, Center-Leitung „Data Driven Design“ bei Fraunhofer Austria.

Julia EISNER, *MA*, wissenschaftliche Mitarbeiterin an der Fachhochschule Wiener Neustadt und Dissertantin am Institut für Soziologie der Universität Wien. Forschungsschwerpunkt: Verbindung einer soziologischen und kommunikationswissenschaftlichen Perspektive auf kommunikative Künstliche Intelligenz in der Unternehmenskommunikation.

Samuel EKE, *Dr. med. univ. B.Sc.*, wissenschaftlicher Mitarbeiter am Department für Psychosomatik der Medizinischen Universität Innsbruck, davor am Department für Onkologie, wo er Deep-Learning-Modelle zur 3D-Rekonstruktion von Tumorsphäroiden und Analyse der Infiltration von Immunzellen entwickelte. Forschungsschwerpunkt: der Einsatz von Maschinellem Lernen und erklärbarer Künstlicher Intelligenz in der psychosomatischen Medizin und Psychotherapieforschung.

Benedikt FUCHS, *Dipl.-Ing.*, Business Development Manager bei Fraunhofer Austria Research GmbH, Center für nachhaltige Produktion und Logistik. Forschungsschwerpunkte: menschenzentrierte Arbeitsgestaltung in der digitalen Transformation.

Barbara GEYER, *Prof.ⁱⁿ (FH) Mag.^a (FH) PhD*, Studiengangsleiterin des Masterstudiengangs E-Learning und Wissensmanagement und Leiterin der Stabstelle Instructional Design an der Hochschule Burgenland, aktive Wissensvermittlerin über ihren Blog „KI in Lehre und Weiterbildung“ sowie auf LinkedIn. Forschungsschwerpunkte: Künstliche Intelligenz in der Aus- und Weiterbildung, Blended Learning und Wissensmanagement.

Pia-Zoe HAHNE, *M.Sc.*, Researcherin für Digitalen Humanismus an der Fachhochschule des BFI Wien und Doktorandin an der philosophischen Fakultät der Universität Wien. Forschungsschwerpunkte: Technikphilosophie, KI-Ethik, Arbeitsverhältnisse in der KI-Branche, unsichtbare Arbeit.

Nils HOFMANN, *Dr. med. univ. B.A.*, Arzt und Masterstudent in Philosophie am Institut für Politikwissenschaft der Universität Wien. Arbeitsschwerpunkte: neuropsychologische Grundlagenforschung, Bioethik, politische Theorie und dekonstruktive Rechtsphilosophie.

Wolfgang KATZIAN, Präsident des Österreichischen Gewerkschaftsbundes und des Europäischen Gewerkschaftsbundes, war mehr als zehn Jahre lang Abgeordneter zum Nationalrat. Arbeitsschwerpunkte: Arbeit und Soziales, Wirtschaft, Industrie und Digitalisierung sowie Energie.

Wenke KLINGBEIL-DÖRING, *Dr.*, beschäftigt sich als freie Philosophin mit der Bedeutung und dem Wandel von Arbeit und ist als Lektorin tätig.

Sebastian KLOCKER, Leiter der Stabstelle Digitalisierung und Technik des Österreichischen Gewerkschaftsbundes. Arbeitsschwerpunkte: Digitalisierung, Automatisierung, Safety & Security und Künstliche Intelligenz sowie deren Auswirkungen auf die betriebliche Mitbestimmung und den Beschäftigtendatenschutz, Datenschutz- und Technikberatung für Betriebsräte und Gewerkschaften.

Viktoria A. KRENN, *Dr. Ing.*, Projektleiterin bei Fraunhofer Austria sowie Lektorin an den Universitäten Wien und Zürich. Fachschwerpunkt: Anthropologie.

Erich PREM, *DDr. MBA*, Chef-FTI-Stratege und Geschäftsführer von eutema GmbH, Autor zahlreicher wissenschaftliche Arbeiten, Forschungstätigkeit an der TU Wien, der Universität Wien, am MIT und OFAI. Arbeitsschwerpunkte: AI, Technikethik und Innovationsforschung.

Lejla RIZVANOVIK, AI Privacy Expert und Geschäftsführerin der Unternehmensberatung Servus Data Group GmbH, externe Datenschutzbeauftragte (ISO 17024, TÜV Süd), anerkannte Sachverständige für Datenschutz und KI-Regulierung sowie zertifizierte Auditorin. Arbeitsschwerpunkte: Verbindung von Datenschutz, KI-Compliance und Automatisierung, Beratung und Unterstützung von nationalen und internationalen Unternehmen unterschiedlichster Branchen.

Josef SAWETZ, *Dr. Mag.*, Kommunikations- und Marketingpsychologe, Lektor an der Universität Wien, der Universität für Weiterbildung Krems, der Hochschule Burgenland, der Fachhochschule Wien und der New Design University St. Pölten. Forschungs- und Arbeitsschwerpunkte: persuasive Kommunikation, Profiling-Techniken, AI in der Unternehmenskommunikation.

Sebastian SCHLUND, *Univ.-Prof. Dr.-Ing.*, Leiter des Forschungsbereichs Industrial Engineering an der TU Wien, Geschäftsführer der Fraunhofer Austria Research GmbH, Vorstand der Gesellschaft für Arbeitswissenschaft (GfA). Arbeitsschwerpunkt: digital und automatisiert unterstützte Arbeitsgestaltung in der Produktion.

Alexander SCHMÖLZ, *Dr.*, Professor für digitalen Humanismus an der Fachhochschule des BFI Wien, Chefredakteur der wissenschaftlichen Zeitschrift Digital Culture & Education (DCE) und Researcher-in-Residence bei fit4internet. Forschungsschwerpunkte: Humanismus und Digitalisierung mit besonderem Fokus auf Governance, Inklusion, Ko-Kreativität und digitale Kompetenzen sowie förderliche und hinderliche pädagogische und politische Rahmenbedingungen.

Petra SCHÖN, *Mag. Dr. BA MA MBA*, österreichische Juristin und Fakultätsmanagerin an der Universität Wien. Arbeitsschwerpunkte: Rechtslinguistik und Translationswissenschaft.

Stefanie SPIELBERGER, *Prof. Dr.*, Professorin für Human Resources Management an der Technischen Hochschule Nürnberg Georg Simon Ohm. Arbeitsschwerpunkte: Transformation der Arbeitswelt im Kontext von New Work, insbesondere in Bezug auf flexible Arbeitsmodelle, den Einfluss Künstlicher Intelligenz und die Vereinbarkeit von Beruf und Familie.

Rita STAMPFL, *Ing.ⁱⁿ BSc MSc MA*, Studiengangsleiterin des Bachelorstudiengangs Wirtschaftsinformatik an der Hochschule Burgenland und Lehrende in den Studiengängen Business Process Engineering & Management, Cloud Computing Engineering sowie E-Learning und Wissensmanagement. Forschungsschwerpunkte: Schnittstelle zwischen Wirtschaft und Informatik, Künstliche Intelligenz in der Hochschullehre.

Tarim STÜWE, *Bsc*, wissenschaftlicher Mitarbeiter bei Fraunhofer Austria.

Titus UDREA, *Dr.*, Experte für Technikfolgenabschätzung (TA) im Büro für Digitale Agenden der Arbeiterkammer Wien. Arbeits- und Forschungsschwerpunkte: Technikfolgenabschätzung aus Arbeitnehmer:innen-Perspektive, Monitoring der digitalen Transformation am Arbeitsplatz, Konstruktiver Einsatz von KI in der Arbeitswelt, Governance und Ethik KI-basierter Systeme.

Torsten ULLRICH, *Dr.*, Forschungs koordin ator und stv. Center-Leitung „Data Driven Design“ bei Fraunhofer Austria sowie Lektor und wissenschaftlicher Mitarbeiter der Technischen Universität Graz.

Elena WIEGELMANN, *MI*, wissenschaftliche Mitarbeiterin bei Fraunhofer Austria.

Jan-Ove WIESNER, *BSc.*, hilfswissenschaftlicher Mitarbeiter bei Fraunhofer Austria Research GmbH, Center für nachhaltige Produktion und Logistik, Masterstudent der Psychologie an der Universität Wien. Forschungsschwerpunkte: menschenzentrierte Gestaltung digitaler Assistenzsysteme in Produktion und Instandhaltung.

Herausgeberin und Herausgeber



Agnes Raschauer ist Programm Managerin am Postgraduate Center der Universität Wien. Nach ihrem Studium der Soziologie an der Universität Wien und der University of Washington, Seattle, WA, war sie in der Hochschulforschung (New Design University, St. Pölten und Stiftung Universität Hildesheim) sowie der universitären Weiterbildung tätig. Im Rahmen der Offenen Weiterbildung am Postgraduate Center der Universität ist sie für die Programmentwicklung verantwortlich und koordiniert Veranstaltungen zum Wissenstransfer zwischen Wissenschaft, Unternehmen und Organisationen sowie der Öffentlichkeit.



Nino Tomaschek ist Leiter des Postgraduate Center der Universität Wien. Er ist Privatdozent für Wissenschaftstheorie und habilitierte über Systemisches Transformationsmanagement an der Universität Augsburg, wo er auch die Augsburger Schule des Innovations-Coaching mitbegründete. Er war Visiting Professor und Senior Research Director an der Flensburgschool for Advanced Research Studies, Adjunct Associate Professor an der University of Nikosia und ist Lehrbeauftragter an der EBS-Universität für Wirtschaft und Recht in Oestrich-Winkel. Seine Arbeitsschwerpunkte sind Management Development und Organisationsentwicklung. Nino Tomaschek ist Autor von zahlreichen Fachbüchern und als Consultant und Aufsichtsrat für Unternehmen und Organisationen tätig.

