ECONSTOR
Make Your Publications Visible.

A Service of

ZBW Leibniz-Informationszentrum Wirtschaft
Leibniz Information Centre for Economics

Loi, Michele

Research Report

People Analytics muss den Menschen zugutekommen: Eine ethische Analyse datengesteuerter algorithmischer Systeme im Personalmanagement

Study der Hans-Böckler-Stiftung, No. 450

Provided in Cooperation with:
The Hans Böckler Foundation

Suggested Citation: Loi, Michele (2021) : People Analytics muss den Menschen zugutekommen: Eine ethische Analyse datengesteuerter algorithmischer Systeme im Personalmanagement, Study der Hans-Böckler-Stiftung, No. 450, ISBN 978-3-86593-366-9, Hans-Böckler-Stiftung, Düsseldorf

This Version is available at:
https://hdl.handle.net/10419/233056

Standard-Nutzungsbedingungen:

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

Terms of use:

Documents in EconStor may be saved and copied for your personal and scholarly purposes.

You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.

If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.

https://creativecommons.org/licenses/by/4.0/de/legalcode

WWW.ECONSTOR.EU

Mitglied der Leibniz-Gemeinschaft

Hans Böckler Stiftung
Mitbestimmung · Forschung · Stipendien

STUDY

Study 450 · April 2021

PEOPLE ANALYTICS MUSS DEN MENSCHEN ZUGUTEKOMMEN

Eine ethische Analyse datengesteuerter algorithmischer Systeme im Personalmanagement

Michele Loi

FORSCHUNGS-
FÖRDERUNG

Dieser Band erscheint als 450. Band der Reihe Study der Hans-Böckler-Stiftung. Die Reihe Study führt mit fortlaufender Zählung die Buchreihe „edition Hans-Böckler-Stiftung“ in elektronischer Form weiter.

Hans Böckler Stiftung
Mitbestimmung · Forschung · Stipendien

STUDY

Study 450 · April 2021

PEOPLE ANALYTICS MUSS DEN MENSCHEN ZUGUTEKOMMEN

Eine ethische Analyse datengesteuerter algorithmischer Systeme im Personalmanagement

Michele Loi

FORSCHUNGS-
FÖRDERUNG

© 2020 by Hans-Böckler-Stiftung
Georg-Glock-Straße 18, 40474 Düsseldorf
www.boeckler.de

„People Analytics muss den Menschen zugutekommen“ von Michele Loi ist lizenziert unter **Creative Commons Attribution 4.0 (BY).**
Diese Lizenz erlaubt unter Voraussetzung der Namensnennung des Urhebers die Bearbeitung, Vervielfältigung und Verbreitung des Materials in jedem Format oder Medium für beliebige Zwecke, auch kommerziell.
(Lizenztext: https://creativecommons.org/licenses/by/4.0/de/legalcode)

Die Bedingungen der Creative-Commons-Lizenz gelten nur für Originalmaterial. Die Wiederverwendung von Material aus anderen Quellen (gekennzeichnet mit Quellenangabe) wie z. B. von Schaubildern, Abbildungen, Fotos und Textauszügen erfordert ggf. weitere Nutzungsgenehmigungen durch den jeweiligen Rechteinhaber.

Satz: DOPPELPUNKT, Stuttgart
Lektorat: Miriam Geoghegan, Königsbach-Stein

ISBN: 978-3-86593-366-1

INHALT

1 Ziel und Umfang 7

2 Methodik 10

3 KI-Ethik-Leitlinien: Auf dem Weg zu einem einheitlichen konzeptionellen Schema? 17

4 Implikationen der KI-Ethik-Leitlinien für HR Analytics 27

4.1. Kenntnisse, Kommunikation, Zueigenmachung und Verbesserung der Datenerhebung, des Datenzugriffs und der Datenspeicherung 27

4.2. Den Algorithmus kennen, kommunizieren, sich zu eigen machen und verbessern 33

4.3. Die Auswirkungen auf die Menschen kennen, kommunizieren, anerkennen und verbessern 69

5 Fazit 84

5.1. DSGVO+: Regeln für die Datenerhebung für HR Analytics sollten über die DSGVO hinausgehen 84

5.2. Die Entwicklung datengetriebener HR-Werkzeuge erfordert ausreichende fachliche Kompetenzen, um Wissen über den Algorithmus zu generieren 87

5.3. Die Auswirkungen der Anwendung des KI-Werkzeugs auf die Mitarbeitenden sollten sorgfältig überwacht werden 90

5.4. HR und Management sollten eine angemessene Transparenz in Bezug auf im HR eingesetzte datengetriebene Werkzeuge gewährleisten 92

6 Literatur 95

Abbildungsverzeichnis

Abbildung 1: Übersicht der verschiedenen Arten von Interessengruppen, die in Tabelle 1 aufgeführt sind 16

Abbildung 2: Potenzielle ethische Vorteile von Erklärbarkeit, Transparenz und Rechenschaftspflicht und deren Beziehung zu inhaltlichen ethischen Werten 57

Tabellenverzeichnis

Tabelle 1: Zwanzig für die Analyse ausgewählte Leitlinien 12

Tabelle 2: Matrix der Empfehlungen in den KI-Leitlinien, basierend auf prozeduralen und inhaltlichen Werten 23

Tabelle 3: Formen der Beteiligung unterschiedlicher Interessengruppen 79

1 ZIEL UND UMFANG

Ziel dieses Berichts ist es, die Implikationen der sich herausbildenden Ethik der künstlichen Intelligenz (KI) für den Einsatz von KI im Beschäftigungskontext zu analysieren. Im Fokus stehen dabei KI-Anwendungen, die sich auf das Wohlbefinden am Arbeitsplatz, die beschäftigungsbezogenen Chancen, die Karriere und die Vergütung der einzelnen Mitarbeitenden auswirken können. Der Einsatz von KI-Technologie im Bereich HR Analytics steckt noch in den Kinderschuhen, auch wenn man eine großzügige Definition der Technologietypen heranzieht, auf die sich KI bezieht.

HR-Analytics-Softwareprodukte beinhalten selten automatisierte Entscheidungen oder gar Empfehlungen, die auf datengetriebenen Vorhersagen basieren. Vielmehr entwickeln und visualisieren sie oft eine Reihe von HR-Kennzahlen und überlassen Bewertungen und Entscheidungen allein den menschlichen Entscheidungsträger*innen. Die Funktion dieser Technologien besteht darin, die analytischen Fähigkeiten der Entscheidungsträger*innen zu verbessern, indem sie die Informationen in einem brauchbareren und aufschlussreicheren Format darstellen und verpacken. Diese Entscheidungsunterstützungssysteme werden oft als deskriptive Analytik (Beantwortung der Frage „Was passiert?“) und diagnostische Analytik (Beantwortung der Frage „Warum ist etwas passiert?“) bezeichnet. Obwohl die deskriptive und diagnostische Analytik sowohl technisch als auch konzeptionell alles andere als trivial sind und eine Vielzahl an ethischen Implikationen haben, sind sie nicht das, was im Rahmen dieses Berichts als KI gilt.

Dieser Bericht befasst sich vielmehr mit den gesellschaftlich umstrittensten (und möglicherweise aus diesem Grund am wenigsten entwickelten) Aspekten der HR Analytics, nämlich der prädiktiven Analytik (Beantwortung der Frage „Was wird als nächstes passieren?) und der präskriptiven Analytik (Beantwortung der Frage „Wie soll man sich verhalten?“). Die Arten von KI-Funktionalitäten, die uns interessieren, sind in erster Linie automatisierte HR-Entscheidungssysteme oder zumindest Empfehlungen für HR-Entscheidungen. Typischerweise umfassen solche Aktivitäten das Profiling und Scoring von Mitarbeitenden, wobei Profiling darin besteht, einzelne Mitarbeitende abstrakten Gruppen zuzuordnen (z. B. produktiv und nicht produktiv, zuverlässig und unzuverlässig), und Scoring darin besteht, einzelne Personen zu feinkörnigeren abstrakten Gruppen der gleichen Art zuzuordnen (z. B. alle Beschäftigte, die auf einer Skala von 0 – nicht zuverlässig – bis 1 – vollkommen zuverlässig – eine Zuverlässigkeit von 0,7 aufweisen).

Die Auswirkungen von KI-Werkzeugen auf Praktiken im Bereich Personalmanagement sind bedeutsam, weil sie das Potenzial haben, das Leben der Menschen am Arbeitsplatz tiefgreifend zu beeinflussen, zu verändern und umzulenken, insbesondere im Hinblick auf Aus- und Weiterbildung (1, 2). Ein Beispiel dafür ist die Versicherungsgesellschaft AXA, die ein Bildungs-/Ausbildungssystem namens „Online-Karriereassistent" einsetzt, um Mitarbeitende dabei zu unterstützen, neue Beschäftigungsmöglichkeiten innerhalb des Unternehmens auf der Grundlage ihrer Fähigkeiten, Bestrebungen und Persönlichkeit zu suchen. Ein solches Matching basiert auf einer automatischen Analyse des Lebenslaufs einer Mitarbeiterin bzw. eines Mitarbeiters, bei der ein Qualifikationsprofil erstellt wird, das dann mit den Qualifikationsprofilen der im Unternehmen verfügbaren Stellen abgeglichen wird (3).

Die vorliegende Analyse befasst sich mit KI, die nicht nur Business-Intelligence-Informationen, wie z. B. Rechnungswesen- und Finanzkennzahlen sowie Produktivitätskennzahlen (4) nutzen kann, sondern potenziell auch Informationen, die nicht primär zur Verfolgung arbeitsbezogener Prozesse generiert werden, wie beispielsweise die Datenspur, die von tragbaren Geräten erzeugt wird, etwa wenn Mitarbeitende zur Teilnahme an einem vom Unternehmen geförderten Fitness-Tracking-Programm eingeladen werden; der Videofluss von tragbaren Kameras; oder die Social-Sensing-Daten, die von sogenannten „sociometric badges" aufgezeichnet werden (5). Die Hoffnung der HR Analytics ist, dass die Analyse von Informationen zu einer effizienteren Personalplanung und einer evidenzbasierten Organisation beitragen kann. Es ist denkbar, dass die prädiktive und die präskriptive Analytik auf dasselbe Spektrum von Entscheidungen angewendet werden können, die das Gebiet der HR Analytics heute kennzeichnen, wie z. B.:

- die Berechnung der optimalen Anzahl von Mitarbeitenden, die an der Rezeption Kunden und Kundinnen betreuen;
- die Auswahl des richtigen Persönlichkeitsprofils für den Umgang mit Kunden und Kundinnen an der Rezeption;
- die Bewertung der Auswirkungen von Programmen der Gesundheits- und Wellnessbranche;
- die Messung der Fähigkeit, Initiative zu ergreifen, und die Nutzung des Messwerts zur Vorhersage der Leistung;
- die Verwendung von Daten bei der Entscheidung über personalisierte Gesundheits- und Fitnessprogramme für Sportler*innen oder bei Vertragsentscheidungen. Im letzteren Fall können Sportmannschaften versuchen, die Risiken vorherzusagen, die mit der Einstellung einer vielversprechenden Athletin bzw. eines vielversprechenden Athleten verbunden

sind, der aufgrund von Unfall oder Krankheit nicht einsatzfähig werden könnte;
- die Analyse des Informationsflusses zwischen den Mitgliedern eines Teams, um die Kommunikation und die Problemlösung durch Gruppen zu verbessern;
- die Analyse der Mitarbeitendenzufriedenheit anhand von Umfragen;
- die Sammlung und Analyse wichtiger Leistungsdaten, um die persönlichen Leistungen der einzelnen Mitarbeitenden und deren Ausrichtung auf die Unternehmensziele zu bewerten;
- die Analyse des Umsatzes und der Geschäftschancen, um Humankapitalengpässe oder -überschüsse vorherzusagen, bevor sie eintreten;
- die Entwicklung von Indikatoren, die die Wahrscheinlichkeit vorhersagen, dass die Mitarbeitenden im Unternehmen verbleiben, indem ermittelt wird, was sie am meisten schätzen;
- die Optimierung des Arbeitsplans eines Ladengeschäfts für den nächsten Tag, basierend auf der vorhergesagten individuellen Verkaufsleistung in Kombination mit anderen Lieferkettenentscheidungen (6).

KI-generierte Vorhersagen und Empfehlungen können für alle Aufgaben verwendet werden, die derzeit dem Bereich der datengetriebenen HR Analytics zugerechnet werden, wie z. B. die Personalisierung von Stellenangeboten und Arbeitsverträgen, die Steuerung der Leistung der einzelnen Mitarbeitenden, die Optimierung von Lern- und Talentförderungsaktivitäten, die Steuerung des Engagements und der Kommunikation der Mitarbeitenden, die Entscheidungen über disziplinarische, gesundheitsrelevante und sicherheitsrelevante Interventionen, die Organisation von Urlaub, Abwesenheit, flexiblen Arbeitszeiten und Mutterschafts-/Vaterschaftsurlaub der Mitarbeitenden sowie die Gewährung von Belohnungen (z. B. Gehalt und Prämien) (4)

2 METHODIK

Im ersten Schritt dieses Projekts wurden Interviews mit Mitgliedern von Gewerkschaften[1] durchgeführt, um wichtige ethische Bedenken im Bereich HR Analytics herauszuarbeiten. Die Erkenntnisse aus diesen Interviews wurden mit Erkenntnissen aus der wissenschaftlichen Literatur zur Datenethik und Algorithmusethik (7) kombiniert, die nicht speziell HR-Anwendungen betreffen. Anschließend wurde ein philosophischer Rahmen für die Analyse des Inhalts von KI-Leitlinien entwickelt, der auf zwei Quellen basiert: a) den Werten, die in einer kürzlich veröffentlichten induktiven Werteanalyse von 84 KI-Leitlinien aufgeführt sind (8); b) der Arbeit des Autors dieses Berichts in seiner Eigenschaft als Ko-Leiter einer Arbeitsgruppe zur Entwicklung einer solchen Leitlinie (9). In der letztgenannten Arbeit wurden ethische Empfehlungen für die Entwicklung datengetriebener Produkte entsprechend ihrer Beziehung zum Datenpipeline-Workflow strukturiert, der mit der Datenerhebung beginnt und mit der Anwendung eines datengetriebenen Modells auf neue Daten, die einzelne Personen betreffen, endet. Dieses Datenpipeline-Konzept wurde als Rahmen für die Analyse anderer Leitlinien benutzt. Die Grundidee bestand darin, ethische Anforderungen und Empfehlungen mit drei verschiedenen Stufen in der Datenpipeline in Verbindung zu bringen, nämlich 1) Datenerhebung und -generierung; 2) Wissensaufbau/Modellbildung; 3) Anwendung des Modells auf konkrete einzelne Personen.

Der letzte Aspekt des konzeptionellen Rahmens basierte auf einer eingehenden Analyse des Inhalts von 20 Leitlinien – eine Teilmenge der 84 Leitlinien, die in der oben erwähnten Übersichtsstudie der globalen Ethik-Leitlinien-Landschaft (8) betrachtet wurden. Die qualitative Analyse dieser 20 Leitlinien durch den Autor führte dazu, durch Induktion andere allgemeine Konzepte zu abstrahieren als die, die in der Übersichtsstudie von Jobin, Ienca und Vayena (8) ermittelt wurden. Während die Analyse von Jobin et al. (8) die allgemeinsten Grundsätze und Wertbegriffe identifizierte, ermittelte unsere ergänzende Analyse die allgemeinsten *Aktivitäts*arten.

1 Die folgenden Gewerkschaftsvertreter*innen wurden befragt: Oliver Suchy, Leiter der Abteilung Digitale Arbeitswelten und Arbeitsweltberichterstattung des Deutschen Gewerkschaftsbundes (DGB); Isabelle Schömann, Bundessekretärin des Europäischen Gewerkschaftsbunds; Marco Bentivogli, Generalsekretär der FIM-CISL (Italienische Föderation der Metallarbeiter im italienischen Gewerkschaftsbund CISL); Michele Carrus, Generalsekretär des Allgemeinen italienischen Gewerkschaftsbunds (CGIL) für die Region Sardinien; sowie Wolfgang Kowalsky, Senior Advisor beim Europäischen Gewerkschaftsbund.

Durch die Kombination von Gewerkschaftsanliegen, verschiedenen ethischen Werten (aus der Analyse von Jobin et al. [8]), verschiedenen Stufen in der Datenpipeline sowie verschiedenen vorgeschriebenen Aktivitätsarten entstand ein neuer konzeptioneller Rahmen zur Analyse des Inhalts von ethischen Leitlinien in der KI.

Alle Leitliniendokumente wurden über die entsprechende Seite auf der AlgorithmWatch-Website abgerufen. Die 20 hier analysierten Leitlinien wurden so ausgewählt, dass sie verschiedene Arten von Interessengruppen so repräsentieren, dass sie eine Stichprobe der vielfältigen Arten von Interessengruppen in den ursprünglich 84 Dokumenten darstellen, die von Jobin et al. (8) analysiert wurden. Die Auswahl wurde auf Quellen beschränkt, die in englischer Sprache verfügbar waren. Es wurden nur europäische, internationale oder supranationale Leitlinien berücksichtigt. Neben den Ethik-Leitlinien der Hochrangigen EU-Expertengruppe für Künstliche Intelligenz, mit denen wir unsere Analyse begannen, wurden weitere 19 der 84 Leitlinien in der Übersichtsstudie von Jobin et al. (8) ausgewählt[2] und analysiert. Tabelle 1 enthält alle analysierten Leitlinien in der Reihenfolge, in der sie abgerufen und zur Aufnahme in die vorliegende Analyse ausgewählt wurden.

2 Wir analysierten die ersten 19 Leitlinien (mit Ausnahme derjenigen, die nicht zugänglich waren oder nicht auf Englisch veröffentlicht wurden), die auf der Website https://aiethics.herokuapp.com/ aufgeführt waren. Dieser spezifische Satz erwies sich als breit gefächert in Bezug auf: a) die Art der Interessengruppe, die die Leitlinie herausgegeben hat; b) die Interessengruppe, an die die Leitlinie gerichtet war; c) den Schwerpunkt (z. B. allgemeiner Anwendungsbereich, Diskriminierung, Fairness, Geschlecht, Schutz der Privatsphäre und Datenschutz, Meinungsfreiheit, schädliche KI, Auswirkungen auf die Arbeit); und d) die Art des Dokuments (z. B. kurze Liste von Grundsätzen, lange Leitlinien, spezialisiertes Weißbuch).

Zwanzig für die Analyse ausgewählte Leitlinien

Titel des Dokuments/ Name der Website	Titel der Leitlinien/ Grundsätze	Herausgeber*in	Land der Herausgeberin/ des Herausgebers	Art der Herausgeberin/ des Herausgebers	Datum der Veröffentlichung	Zielgruppe
Ethics Guidelines for Trustworthy AI [Ethik-Leitlinien für eine vertrauenswürdige KI](29)	Ethical Principles in the Context of AI Systems [Ethische Grundsätze im Kontext von KI-Systemen]	Hochrangige Expertengruppe für Künstliche Intelligenz	EU	IGO/ supranational	8. April 2019	mehrere (alle Interessengruppen sowie internationale politische Entscheidungsträger*innen)
AI Guidelines [KI-Leitlinien] (10)	AI Guidelines [KI-Leitlinien]	Deutsche Telekom	Deutschland	Unternehmen	11. Mai 2018	das Unternehmen selbst
10 Principles of Responsible AI (11)	Summary of our proposed recommendations	Women Leading in AI	k. A.	Denkfabrik	k. A.	der öffentliche Sektor (nationale und internationale politische Entscheidungsträger*innen)
Principles for Accountable Algorithms and a Social Impact Statement for Algorithms (12)	Principles for Accountable Algorithms	Fairness, Accountability and Transparency in Machine Learning (FAT-ML)	k. A.	Zusammenschluss von Forschenden und Praktiker*innen	24. Nov. 2016	mehrere (Entwickler*innen und Produktmanager*innen)
Tenets (13)	Tenets	Partnership on AI	k. A	privatwirtschaftliche Allianz	29. Sept. 2016	die Organisation selbst

Ethically Aligned Design. A Vision for Prioritizing Human Wellbeing with Autonomous and Intelligent Systems, Version II (14)	Ethically Aligned Design. A Vision for Prioritizing Human Wellbeing with Autonomous and Intelligent Systems, Version II	Institute of Electrical and Electronic Engineers (IEEE), The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems	international	Berufsverband	12. Dez. 2017*	k. A.
Universal Guidelines for Artificial Intelligence (15)	Universal Guidelines for Artificial Intelligence	The Public Voice	international	gemischt (Koalition von NGOs, Datenschutzbehörden usw.)	23. Okt. 2018	mehrere (Institutionen, Regierungen)
Declaration on Ethics and Data Protection in Artificial Intelligence [Erklärung zu Ethik und Datenschutz im Bereich der künstlichen Intelligenz] (16)	„… guiding principles …“ [„Leitprinzipien“]	ICDPPC	international	IGO/ supranational	23. Okt. 2018	k. A.
Artificial Intelligence: Open Questions about Gender Inclusion (17)	Proposals	Women 20 (W20)	international	IGO/ supranational	2. Juli 2018	der öffentliche Sektor (Staaten/Länder)
Charlevoix Common Vision for the Future of Artificial Intelligence (18)	Charlevoix Common Vision for the Future of Artificial Intelligence	Die Staats- und Regierungschefs der G7-Länder	International	IGO/ supranational	9. Juni 2018	die G7-Länder selbst (Regierungen)

The Toronto Declaration: Protecting the Right to Equality and Non-Discrimination in Machine Learning Systems (19)	The Toronto Declaration: Protecting the Right to Equality and Non-Discrimination in Machine Learning Systems	Access Now; Amnesty International	international	verschieden (NGO, NPO)	16. Mai 2018	mehrere (Staaten, privatwirtschaftliche Akteur*innen)
Privacy and Freedom of Expression in the Age of Artificial Intelligence (20)	Conclusions and Recommendations	Privacy International und Article (19)	international	NGO	25. April 2018	mehrere (Staaten, Unternehmen, Zivilgesellschaft)
White Paper: How to Prevent Discriminatory Outcomes in Machine Learning (21)	Executive Summary	Weltwirtschaftsforum, Global Future Council on Human Rights 2016–2018	international	NPO	12. März 2018	der private Sektor (Unternehmen)
The Malicious Use of Artificial Intelligence: Forecasting, Prevention and Mitigation (22)	Four High-Level Recommendations	Future of Humanity Institute; University of Oxford; Centre for the Study of Existential Risk; University of Cambridge; Center for a New American Security; Electronic Frontier Foundation; OpenAI	international	Verschieden (gemischt Akademiker*innen, NPO)	20. Feb. 2018	k. A.
For a Meaningful Artificial Intelligence: Towards a French and European Strategy (23)	„Part 5 – What Are the Ethics of AI?; Part 6 – For Inclusive and Diverse Artificial Intelligence“	Mission Villani	Frankreich	staatliche Einrichtung/ Organisation	29. März 2018	der öffentliche Sektor (Regierung/Parlament Frankreichs)

Top 10 Principles for Ethical Artificial Intelligence [Die 10 wichtigsten Grundsätze für ethische künstliche Intelligenz] (24)	Top 10 Principles for Ethical Artificial Intelligence [Die 10 wichtigsten Grundsätze für ethische künstliche Intelligenz]	UNI Global Union	international	globale Gewerkschaftsföderation	17. Dez. 2017	mehrere (Gewerkschaften, Erwerbstätige)
ITI AI Policy Principles (25)	ITI AI Policy Principles	Information Technology Industry Council (ITI)	international	privatwirtschaftlicher Verband	24. Okt. 2017	Mitglieder des Verbands
Ethical Principles for Artificial Intelligence and Data Analytics (26)	Ethical Principles for Artificial Intelligence and Data Analytics	Software & Information Industry Association (SIIA), Public Policy Division	international	privatwirtschaftlicher Verband	15. Sept. 2017	der private Sektor (die Industrie, Organisationen)
Report of COMEST on Robotics Ethics (Berücksichtigt wurde nur der Abschnitt „Recommendations".) (27)	Relevant Ethical Principles and Values	COMEST/UNESCO	International	IGO/ supranational	14. Sept. 2017	k. A.
Artificial Intelligence and Machine Learning: Policy Paper (28)	Artificial Intelligence and Machine Learning: Policy Paper	Internet Society	international	NPO	18. April 2017	mehrere (politische Entscheidungsträger*innen, andere Interessengruppen im breiteren Internet-Ökosystem

Anmerkung: Diese Tabelle bietet eine Teilmenge der in Tabelle S1 in Jobin, Ienca und Vayena (8) aufgeführten Leitlinien, mit zusätzlichen Informationen zur Art der Herausgeberin/des Herausgebers.
* Seltsamerweise ist das Datum der Pressemitteilung von Version II der 12. Dezember 2017, aber das Dokument selbst enthält kein Datum. Version I enthält einen Urheberrechtsvermerk vom 25. März 2019. Daher ist Version II offenbar zwei Jahre vor Version I erschienen.

Abbildung 1

Übersicht der verschiedenen Arten von Interessengruppen, die in Tabelle 1 aufgeführt sind

Herausgeber*in

IGO/supranational	NGO/NPO	Zusammenschluss von Forschenden und Praktiker*innen	
	privatwirtschaftliche Allianz	Denkfabrik	Unternehmen
gemischt	staatliche Einrichtungen/ Organisationen	Berufs- verband/ Gesellschaft gemischt	Gewerk- schafts- föderation

Zielgruppe

mehrere (verschiedene Interessen- träger*innen)	die Herausgeberin/ der Herausgeber selbst	k. A.
	der öffentliche Sektor	der private Sektor

Land des Herausgebers/der Herausgeberin

international	k. A.	
	EU	Frankreich
	Deutschland	

3 KI-ETHIK-LEITLINIEN: AUF DEM WEG ZU EINEM EINHEITLICHEN KONZEPTIONELLEN SCHEMA?

Wir wenden uns nun dem Hauptteil dieses Dokuments zu, der der Anleitung gewidmet ist, die die bestehenden Leitlinien zur ethischen KI bieten.

Werte in KI-Ethik-Leitlinien

Jobin et al. (8) haben die globale Landschaft der KI-Ethik-Leitlinien analysiert, die 84 Dokumente umfasst, die bis zum 23. April 2019 zusammengetragen wurden. Ihrer Analyse zufolge werden die folgenden elf unterschiedlichen Werte im gesamten Korpus erwähnt, wenngleich kein Wert in allen Leitlinien genannt wird:

1. Transparenz
2. Gerechtigkeit und Fairness
3. Schadensverhütung
4. Verantwortung
5. Schutz der Privatsphäre/Datenschutz
6. Benefizienz [Gutes tun]
7. Freiheit und Autonomie
8. Vertrauen
9. Nachhaltigkeit
10. Würde
11. Solidarität

Aber was ist ein Wert? Im weiteren Sinne ist ein Wert das, was eine wertende Behauptung charakterisiert, eine Behauptung, die eine Situation nicht einfach beschreibt, sondern nach ihrer Wünschbarkeit beurteilt. Im engeren Sinne ist ein Wert etwas, das gut ist – eine Form von Güte (31). Typischerweise befasst sich die Wertetheorie im engeren Sinne nur mit intrinsischer und/oder finaler Güte (Güte als Zweck), nicht mit irgendetwas, das als Mittel zur Förderung irgendeiner Form von Güte dienen kann. Viele Punkte in der obigen Liste sind jedoch keine „Werte“, wenn man unter Werten *intrinsische* Formen von Güte versteht. Bei einigen von ihnen – z. B. Transparenz – ist es plausibler, sie als instrumentelle Güter zu betrachten, d. h. als Mittel zur Erreichung anderer Ziele, die wirklich wertvoll sind, z. B. menschliches Wohl-

ergehen oder Gerechtigkeit. Daher werde ich den Begriff „wertegeladene Anliegen“ anstelle von „Werten“ verwenden, um die elf Punkte in dieser Liste zu bezeichnen, und zwar sowohl Güter, die als Zweck wertvoll sind, als auch solche, die als Mittel wertvoll sind, sowie andere (z. B. Menschenrechte), die in diesen Leitlinien erscheinen.

Ähnlicher Inhalt, anderes Gerüst

Es ist zu beachten, dass es keine Eins-zu-eins-Entsprechung gibt zwischen Grundsätzen, Werten oder moralischen Zielen einerseits und den Empfehlungen, die als praktische Wege zur Verwirklichung dieser Werte genannt werden, andererseits. Mit anderen Worten, auch wenn wir eine erhebliche Überschneidung sowohl der hochrangigen Werte als auch der niederrangigen Empfehlungen in den verschiedene Leitlinien feststellen, ist die Zuordnung der praktischen Empfehlungen zu den allgemeinen wertegeladenen Zielen (oder Grundsätzen) in den verschiedenen Dokumenten nicht kohärent. Anders ausgedrückt, die Leitlinien, die dieselben wertegeladenen Wörter erwähnen und ähnliche Empfehlungen geben, haben nicht dasselbe konzeptionelle Schema.

Um die Vielfalt der konzeptionellen Schemata zu veranschaulichen, kann man die Frage der algorithmischen Erklärbarkeit betrachten. Diese wird beschrieben als Teil eines Rechts auf bzw. einer Pflicht zur (z. B.):

- Transparenz (15)
- Kontrolle („Wir legen das Fundament“ (10); „Human-in-Command Approach“ [Ansatz der Gesamtsteuerung durch einen Menschen] (24))
- Verstehen (21)
- Erklärbarkeit (12)
- Rechenschaftspflicht (14)
- Interpretierbarkeit (25)

Was können wir aus der Untersuchung dieser Fallstudie über Erklärbarkeit lernen? Erstens gibt es wenig theoretische Kohärenz in der Art und Weise, wie Wertebegriffe verwendet werden. Diese Begriffe werden nie klar oder erschöpfend definiert, und das logische Schlussfolgern von „top-down“, wertegeladenen Zielen oder Grundsätzen auf Empfehlungen wird nicht von einem kohärenten zugrunde liegenden theoretischen Konstrukt geleitet. Vielmehr funktionieren diese wichtigen wertegeladenen Begriffe als Bezeichnungen, unter denen mehrere praktische Ideen abgelegt werden können, und zwar auf eine Art und Weise, die nicht völlig, aber doch erheblich willkürlich ist.

Infolgedessen sind Zweifel berechtigt, ob eine Struktur, die sich ausschließlich oder hauptsächlich auf wertegeladene Ziele (oder Grundätze) stützt, den an der Umsetzung von Ethik-Leitlinien interessierten Nutzenden hilft, die für ihre Aufgaben relevanten Leitlinien problemlos zu finden.

Verschiedene Interessengruppen

Darüber hinaus unterscheiden sich die Leitlinien in Bezug auf die Interessengruppen, die sie erstellt haben, und an die sie sich richten. Jobin et al. (8) schlüsseln 84 Leitlinien nach Interessengruppen auf. Die meisten wurden von privaten Unternehmen erstellt, gefolgt von staatlichen Einrichtungen. Was die Zielgruppen betrifft, so richten sich die meisten von Jobin et al. (8) untersuchten Leitlinien an mehrere Interessengruppen, gefolgt von Leitlinien, die an die herausgebende Organisation selbst gerichtet sind (d. h. von der Organisation erstellt, um sich selbst oder ihre Mitglieder anzusprechen). Jobin et al. (8) zufolge sind die Interessengruppen in den Zielgruppen am vielfältigsten:

- die Führungskräfte und Mitarbeitenden der herausgebende Organisation, einschließlich Entwickler*innen und Designer*innen
- Entwickler*innen und Designer*innen im Allgemeinen
- Forschende
- der private Sektor im Allgemeinen
- der öffentliche Sektor im Allgemeinen
- „Organisationen“ im Allgemeinen
- alle, die die Entwicklung der KI beeinflussen können (8).

Die begrenztere Stichprobe, auf die sich die vorliegende Dokumentenanalyse stützt, umfasst ebenfalls private Unternehmen, staatliche Einrichtungen, NGOs, Forschende, privatwirtschaftliche Verbände sowie Berufsverbände und -organisationen. Somit weist sie dieselbe Heterogenität an Interessengruppen auf wie die breite Stichprobe, die von Jobin et al. (8) analysiert wurde.

Verschiedene Aktivitätsarten

Da die Leitlinien nicht kohärent nach wertegeladenen Anliegen (bei Jobin et al. [8] „Werte“ genannt) organisiert sind, schlägt der vorliegende Bericht ein ergänzendes konzeptionelles Schema zur Analyse ihrer Inhalte vor, das auf *Aktivitätsarten* basiert:

- Wissen und Kontrolle über Ziele, Prozesse und Ergebnisse
- Transparenz in Bezug auf Ziele, Prozesse und Ergebnisse
- Rechenschaftspflicht für Ziele, Prozesse und Ergebnisse
- Ergebnis- und Prozessverbesserungen

Die vier Aktivitätsarten sind nicht unabhängig voneinander, sondern hängen wie folgt zusammen:

(a) Wissen und Kontrolle beziehen sich darauf zu kommentieren, i) was eine Organisation zu tun versucht (Ziel); ii) wie sie tut, was sie zu tun beabsichtigt (Prozesse); und iii) was sich daraus ergibt (Ergebnisse). Der Akt des Dokumentierens produziert die menschlichen Güter des Wissens und der Kontrolle. Diese Güter sind *schon vor – und unabhängig von – der Ermöglichung von Transparenz nach außen sowie Rechenschaftspflicht* wertvoll. Wissen und Kontrolle werden durch Transparenz und Rechenschaftspflicht *vorausgesetzt*, können jedoch von internen Prozessen eigenständig hervorgebracht werden, die weder für Außenstehende transparent noch mit klarer rechtlicher Verantwortung und moralischer Schuldigkeit verbunden sind.[3]

b) Transparenz wird durch die Kombination von Wissen und erfolgreicher Kommunikation nach außen erreicht. Transparenz setzt *Wissen voraus*: Um Transparenz in Bezug auf ein Ziel, einen Prozess oder ein Ergebnis herzustellen, muss man es bzw. ihn zuerst kennen und dokumentieren. Das erklärt, warum viele andere Leitlinien Aktivitäten der Ergebnis- und Prozessbewertung und -dokumentation unter Transparenz aufführen.[4]

c) Aufgaben der Rechenschaftspflicht ergeben sich aus der Kombination des Gutes der Kontrolle mit der Zuweisung i) moralischer oder rechtlicher Verantwortung an Organisationen und ii) organisatorischer Verantwortung an einzelne Personen innerhalb von Organisationen. Ein Mangel an Rechenschaftspflicht kann auf mangelnde Kontrolle zurückzuführen sein (z. B. das Versäumnis zu wissen und zu kontrollieren, was man tut und wie man seine Ergebnisse erzielt hat). Kann, nicht muss, denn mangelnde Rechenschaftspflicht kann auch auf das Fehlen klarer organisatorischer Verantwortlichkeiten für die Qualität der Ziele, Prozesse oder Ergebnisse einer Organisation zurückzuführen sein. Viele Empfehlungen über Rechenschaftspflicht beschreiben soziale, administrative, politische oder

3 Die Aktivitäten der Wissensgenerierung und der Kontrolle erfüllen tendenziell zwei der im Dokument der von der Europäischen Kommission eingesetzten unabhängigen Expertengruppe (29) genannten Anforderungen an eine vertrauenswürdige KI, nämlich „Vorrang menschlichen Handelns und menschliche Aufsicht“ und den Teil der Anforderung „technische Robustheit und Sicherheit“, der sich auf „Robustheit“ bezieht. Während der Vorrang menschlichen Handelns und menschliche Aufsicht (als Anforderungen) sowie die technische Robustheit in Bezug auf verschiedene ethische Werte und Grundsätze von instrumentellem Wert sind, kann die Förderung der Sicherheit als ein Aspekt des Grundsatzes der Schadensverhütung angesehen werden, wie von Jobin und et al. (8) vorgeschlagen.

4 Transparenz wird ebenfalls als eine der sogenannten „Anforderungen“ an eine vertrauenswürdige KI aufgeführt (29).

rechtliche Aufgaben, die es ermöglichen oder erleichtern zu bestimmen, wer moralisch oder rechtlich für die Festlegung von Zielen, die Überwachung von Prozessen, die Kontrolle von Ergebnissen oder deren Verbesserung verantwortlich sein sollte. Andere Empfehlungen betreffen die *technischen* Voraussetzungen der Verantwortlichkeit, d. h. die *wissenschaftlichen Erkenntnisse* und die *Techniken*, die die Kontrolle einer Datenwissenschaftlerin bzw. eines Datenwissenschaftlers über die Datenpipeline verbessern, einschließlich der Daten, des Trainings und des Ergebnisses (d. h. des Algorithmus bzw. der Entscheidungsregel, wie wir es später nennen werden). In manchen Fällen ist es schwierig, sinnvolle Erkenntnisse über den Algorithmus zu erlangen, die für eine bedeutsame Kontrolle und menschliche Verantwortlichkeit erforderlich sind (32). Diese Herausforderung kann der Technologie immanent sein – dies ist die Frage der algorithmischen Undurchsichtigkeit und der Blackboxes, auf die später eingegangen werden soll.

d) Ergebnisverbesserung ist der Vorgang, bei dem Prozesse modifiziert werden, um Ergebnisse in wünschenswerter Weise zu verändern. Die Verbesserung macht sich die generierten Erkenntnisse zunutze und erfordert ein gewisses Maß an Kontrolle über Ziele und Prozesse. Liegt z. B. ein mathematisches Maß für Unfairness vor, kann es verwendet werden, um die Nutzenfunktion eines Algorithmus einzuschränken (Zielkontrolle bei Ansätzen zur Sicherstellung von konzeptuell integrierten Werten [„values by design"]) und so bereits in der Entwurfsphase einige Fairness-relevante Eigenschaften des Algorithmus zu verbessern.
   Verbesserung ist ein vager Begriff und kann in Richtung jedes der oben genannten inhaltlichen Werte erreicht werden. Beispielsweise kann sich die Ergebnisverbesserung auf den Bau einer nutzbringenden KI (Benefizienz), auf den Bau einer sicheren KI (Schadensverhütung), auf den Bau einer nicht-diskriminierenden KI (Gerechtigkeit) und auf den Erhalt von bedeutsamer menschlicher Kontrolle bei Interaktionen mit KI beziehen, und zwar nicht nur als Mittel zum Bau einer sicheren und robusten KI, sondern weil menschliche Kontrolle *um ihrer selbst willen* wertvoll ist (Autonomie).[5]

5 Aktivitäten der Verbesserung entsprechen den folgenden Anforderungen an eine vertrauenswürdige KI (29): „gesellschaftliches und ökologisches Wohlergehen" (in Bezug auf den materiellen Wert Benefizienz), „Vielfalt, Nichtdiskriminierung und Fairness" (in Bezug auf den materiellen Wert Gerechtigkeit) und „Schutz der Privatsphäre und Datenqualitätsmanagement" (in Bezug auf die materiellen Werte Schadensverhütung und Autonomie).

Zusammenfassend ein Beispiel: Das Messen algorithmischer Fairness fördert *Erkenntnisse;* das Kommunizieren der gemessenen algorithmischen Fairness nach außen fördert die *Transparenz;* das Erkennen modifizierbarer Ursachen von Unfairness (z. B. Daten oder die Definition der mathematischen Funktion) schafft *Kontrolle;* die Übernahme rechtlicher Verantwortung für diese Prozesse ist eine Form der *Rechenschaftspflicht;* und die Milderung von Unfairness in den Ergebnissen ist eine *Ergebnisverbesserung.*

Diese makroskopischen Aktivitätsarten können weiter spezifiziert werden in Bezug auf die Aufgaben von Datenwissenschaftler*innen sowie von Personen, die für die Implementierung von Technologien in der HR-Abteilung von Organisationen verantwortlich sind. Man betrachte z. B. die Aktivitätsart der Wissensproduktion und -kontrolle. Nach einer vereinfachten, vierstufigen Version von Standardmodellen der datenwissenschaftlichen Pipeline (9) kann man unterscheiden:

- Wissen und Kontrolle über die Daten (entspricht den beiden Stufen Datenerhebung und Datenmanagement)
- Wissen und Kontrolle über den Algorithmus (entspricht der Stufe Algorithmusentwurf und Testphase)
- Wissen und Kontrolle über die Auswirkungen auf die Menschen (entspricht der Stufe Einsatz des Algorithmus in konkreten Fällen).

Eine Matrix von Werten und Aktivitätsarten

Einige der in der Analyse von Jobin et al. (8) identifizierten Werte sind instrumentelle, prozedurale Werte, während andere inhaltliche und intrinsische Werte sind. Werte wie Transparenz und Rechenschaftspflicht sind prozedural, da sie eine bestimmte Art und Weise beschreiben, Dinge zu tun; sie sind instrumentell, da Transparenz und Rechenschaftspflicht normalerweise geschätzt werden, weil sie zu moralisch besseren Ergebnissen und besseren Handlungen führen.

Andere Werte – zumindest Gerechtigkeit und Fairness, Schadensverhütung, Schutz der Privatsphäre, Benefizienz, Freiheit und Autonomie, sowie Würde und Solidarität – sind in dem Sinne intrinsisch, dass sie allgemein als charakteristisch für Ergebnisse und Handlungen angesehen werden, die intrinsisch besser sind, weil sie diese Werte verwirklichen. (So ist es beispielsweise intrinsisch besser, freier und autonomer zu sein als weniger frei und autonom; eine Gesellschaft kann als einer anderen Gesellschaft intrinsisch überlegen angesehen werden, wenn sie gerechter ist und die Beziehungen zwischen ihren Mitgliedern stärker durch Brüderlichkeit und Solidarität gekennzeichnet sind.)

Matrix der Empfehlungen in den KI-Leitlinien, basierend auf prozeduralen und inhaltlichen Werten

Prozedurale und instrumentelle Grundsätze

Ergebnisverbesserung entlang vier Dimensionen (inhaltlichen ethischen Grundsätzen/Werten)*	Wissen und Kontrolle (Dokumentieren Sie die von Ihnen gewählten Ziele/Vorgehensweisen/ erreichten Ergebnisse in Bezug auf:)			Transparenz (Teilen Sie mit, was Sie dokumentiert haben in Bezug auf:)			Rechenschaftspflicht (Bestimmen Sie, wer moralisch oder rechtlich verantwortlich ist für dokumentierte Ziele, Prozesse und Ergebnisse in Bezug auf:)		
Wert 1: Benefizienz (verbunden mit den Werten Wohlbefinden/Nachhaltigkeit/Vertrauen)	Daten	Algorithmus	betroffene Menschen	Daten	Algorithmus	betroffene Menschen	Daten	Algorithmus	betroffene Menschen
Daten	Algorithmus	betroffene Menschen							
Wert 2: Schadensverhütung (verbunden mit den Werten Sicherheit/ Wohlbefinden/Schutz der Privatsphäre)	Daten	Algorithmus	betroffene Menschen	Daten	Algorithmus	betroffene Menschen	Daten	Algorithmus	betroffene Menschen
Daten	Algorithmus	betroffene Menschen							
Wert 3: Gerechtigkeit/Fairness (verbunden mit den Werten Solidarität/Grundrechte)	Daten	Algorithmus	betroffene Menschen	Daten	Algorithmus	betroffene Menschen	Daten	Algorithmus	betroffene Menschen
Daten	Algorithmus	betroffene Menschen							
Wert 4: Autonomie (verbunden mit den Werten Freiheit/Würde/Grundrechte/Schutz der Privatsphäre)	Daten	Algorithmus	betroffene Menschen	Daten	Algorithmus	betroffene Menschen	Daten	Algorithmus	betroffene Menschen
Daten	Algorithmus	betroffene Menschen							

Anmerkung: * Die Werte Autonomie, Schadensverhütung und Gerechtigkeit entsprechen den ethischen Werten, die in den EU-Ethik-Leitlinien für eine vertrauenswürdige KI genannt werden (29): (i) Achtung der menschlichen Autonomie, (ii) Schadensverhütung und (iii) Fairness. Darüber hinaus enthalten die EU-Leitlinien den Wert Erklärbarkeit. Wir denken, dass der Wert Erklärbarkeit sowohl instrumentell als auch auf einer anderen Abstraktionsebene ist, da er eine der Hauptvoraussetzungen für die Werte Rechenschaftspflicht und Transparenz darstellt und eng verbunden mit den instrumentellen Werten Wissen und Kontrolle ist.

Wir können daher die beiden fraglichen Arten von Werten zusammenfassen und sie mit den unterschiedenen Aktivitätsarten kombinieren, was zu einer dreidimensionalen Matrix führt, die sowohl prozedurale (aufgabenbezogene) als auch inhaltliche (ergebnisbezogene) Werte enthält, die sich auf Daten, Algorithmen und betroffene Menschen beziehen.

Die meisten Empfehlungen in diesen Leitlinien können als eine Aufgabe oder eine Kombination von Aufgaben beschrieben werden, die eine Aktivität zur Erlangung von Kontrolle und/oder Dokumentation und/oder Kommunikation und/oder Zueigenmachung und/oder Verbesserung von Prozessen beinhält, um Verbesserungen in einer oder typischerweise in mehreren Dimensionen, (z. B. Gutes hervorbringen, Schaden minimieren und Ungerechtigkeit mindern) gleichzeitig herbeizuführen. Die vier wichtigsten Wertedimensionen entsprechen den vier Prinzipien der biomedizinischen Ethik (33).[6]

Wir haben diese Liste um die anderen intrinsischen Werte in der Liste von Jobin et al. (8) erweitert, die als enger mit ihnen verbunden angesehen werden können. So ist Benefizienz, die darin besteht, Gutes zu tun, mit der Förderung des Wohlergehens verbunden; Nachhaltigkeit ist typischerweise auch mit der Möglichkeit verbunden, in Zukunft das Wohlergehen zu fördern; Schadensverhütung ist mit den Werten Sicherheit, Wohlbefinden und einigen Aspekten des Schutzes der Privatsphäre verbunden; Gerechtigkeit, Fairness und Solidarität sind verbunden mit (Nicht-)Diskriminierung und allgemeiner mit Gleichheit bei der Verteilung des Nutzens sowie mit der sozialen Teilhabe an im Gegensatz zum Ausschluss von den Vorteilen, die durch KI entstehen; Freiheit, Autonomie und Würde hängen mit einigen Aspekten des Schutzes der Privatsphäre zusammen (z. B. Kontrolle über die eigenen Daten und Informationen); außerdem ist die Idee, dass Menschen die Kontrolle über ihr Leben behalten sollten anstatt von außen manipuliert oder kontrolliert zu werden, etwas Wertvolles um seiner selbst willen.

Unterschiedliche Implementierungsinfrastrukturen

Schließlich lassen sich die Aufgaben, die als Antwort auf wertegeladene Anliegen empfohlen werden, nach der Art der sozialen Rollen und der Organisationsstruktur unterscheiden, die sie voraussetzen, um ausgeführt werden zu können. Es ist fruchtbar, drei Haupttypen von Lösungen zu unterscheiden, die sich in Bezug auf die sozialen Rollen sowie die Aufgaben, die sie voraussetzen, unterscheiden:

6 Das einheitliche Rahmenwerk für die KI-Ethik nach Floridi und Cowls (77) umfasst die vier ethischen Werte unseres Rahmens sowie den Wert Erklärbarkeit.

- **Technische Lösungen:** Die Ausführung dieser Lösungen setzt vor allem technische Hilfsmittel voraus. Die FAT-ML-Leitlinie (35) empfiehlt in Bezug auf Datenpräzision eine Validitätsprüfung durchzuführen, indem man eine Zufallsstichprobe der Daten (z. B. Input- und/oder Trainingsdaten) bildet und deren Richtigkeit manuell kontrolliert. Diese Prüfung sollte zu einem frühen Zeitpunkt im Entwicklungsprozess durchgeführt werden, bevor abgeleitete Informationen verwendet werden. Die FAT-ML-Leitlinie empfiehlt ferner, die Gesamtdatenfehlerquote der Daten in der Zufallsstichprobe zu veröffentlichen. Bei der Validitätsprüfung handelt es sich um eine technische Aufgabe, die lediglich die übliche soziale Rolle der Datenwissenschaftlerin bzw. des Datenwissenschaftlers erfordert sowie die technische Infrastruktur, über die die in einer Organisation tätigen Spezialist*innen für maschinelles Lernen normalerweise verfügen.[7]
- **Organisatorische Lösungen:** Diese Lösungen setzen eine Infrastruktur von Regeln (und regelgesteuerten Verhaltensweisen) voraus, die innerhalb einer einzelnen Organisation aufgestellt werden kann. So fordert beispielsweise die zweite Fassung der IEEE-Leitlinien (14), dass Rahmenbedingungen für Lenkung und Kontrolle („Governance"), einschließlich Normen und Regulierungsstellen, geschaffen werden sollten, erstens um Prozesse zu überwachen, die sicherstellen, dass die Nutzung von autonomen und intelligenten Systemen (AIS) nicht gegen Menschenrechte, Freiheiten, Würde und den Schutz der Privatsphäre verstößt, und zweitens die Rückverfolgbarkeit zu gewährleisten, um zum Aufbau von öffentlichem Vertrauen in AIS beizutragen.[8]
- **Institutionelle Lösungen:** Diese Lösungen setzen eine Infrastruktur von Regeln (und regelgesteuerten Verhaltensweisen) voraus, die innerhalb einer einzelnen Organisation aufgestellt werden kann. Dazu gehören z. B. Empfehlungen für neue Gesetze, politische Ziele und die Förderung neu-

7 Die EU-Ethik-Leitlinien für eine vertrauenswürdige KI (29) enthalten eine ähnliche Unterscheidung zwischen den „Methoden" zur Erreichung einer vertrauenswürdigen KI. Die in diesem Absatz erwähnten Lösungen entsprechen den „Methoden" mit den Bezeichnungen „Architekturen für eine vertrauenswürdige KI", „Konzeptuell integrierte Ethik und Rechtsstaatlichkeit (X-by-design)", „Erklärungsmethoden", und „Erproben und Prüfen".

8 In den „EU-Ethik-Leitlinien für eine vertrauenswürdige KI" (29) umfassen die organisatorischen Lösungen „Dienstqualitätsparameter", „Verhaltenskodizes", „Standardisierung", „Zertifizierung", „Rechenschaftspflicht durch Rahmenbedingungen für die Lenkung und Kontrolle", „Bildung und Bewusstsein zur Förderung einer ethischen Mentalität" (innerhalb der Organisation), „Beteiligung der Interessenträger und sozialer Dialog" sowie „Vielfalt und and inklusive Entwurfsteams".

er zivilgesellschaftlicher Gremien oder Interessengruppen. So stellt beispielsweise der französische Bericht „For a Meaningful Artificial Intelligence: Towards a French and European Strategy“ (23) im Zusammenhang mit der Gleichstellung der Geschlechter fest, dass Bildungsanstrengungen im Bereich Gleichstellung und digitale Technologie natürlich von entscheidender Bedeutung seien, eine Erhöhung der Geschlechterdiversität jedoch auch mit einer Anreizpolitik erreicht werden könne, die darauf abziele, bis zum Jahr 2020 den Anteil der weiblichen Studierenden in digitalen Fächern an Universitäten und Wirtschaftshochschulen sowie in deren Vorbereitungskursen auf 40 Prozent zu erhöhen.[9]

9 In den EU-Ethik-Leitlinien für eine Vertrauenswürdige KI (29) entsprechen diese in erster Linie den „Methoden zur Schaffung einer vertrauenswürdigen KI“, die als „Regulierung“ bezeichnet werden. Dazu gehören auch die Methoden „Rechenschaftspflicht durch Rahmenbedingungen für die Lenkung und Kontrolle“, „Bildung und Bewusstsein zur Förderung einer ethischen Mentalität“ und „Beteiligung der Interessenträger und sozialer Dialog“.

4 IMPLIKATIONEN DER KI-ETHIK-LEITLINIEN FÜR HR ANALYTICS

Die in 20 Leitlinien enthaltenen Empfehlungen wurden im Rahmen der vorliegenden Studie analysiert, um ihre Relevanz für die Ethik des Einsatzes der KI im HR-Bereich zu beurteilen. Die Struktur des vorliegenden Berichts spiegelt die Hauptunterscheidung zwischen drei verschiedenen Themen wider, nämlich den Aktivitäten der Datenerhebung, des Aufbaus eines HR-Werkzeugs (Algorithmus) und der Verwendung dieses HR-Werkzeugs zur Unterstützung der Entscheidungsfindung im HR-Bereich. Für jedes Thema werden nacheinander die drei prozeduralen Werte a) Wissen/Kontrolle, b) Kommunikation/Transparenz, c) Rechenschaftspflicht/Zueigenmachung des Prozesses bzw. des Ergebnisses betrachtet.

4.1. Kenntnisse, Kommunikation, Zueigenmachung und Verbesserung der Datenerhebung, des Datenzugriffs und der Datenspeicherung

Das Unterthema „Wissen und Kontrolle über Daten“ umfasst alle Empfehlungen im Zusammenhang mit den Kenntnissen über und die Dokumentation und Überwachung der Prozesse der Datenerhebung und -generierung, der Datenspeicherung und des Datenzugriffs, insbesondere wenn es sich um Daten von identifizierbaren natürlichen Personen (d. h. personenbezogene Daten) handelt.[10] Die Generierung von Wissen über die Daten und den Prozess der Datenerhebung wird von Empfehlungen zum Thema Transparenz und Rechenschaftspflicht vorausgesetzt. Daten können auf unterschiedliche Weise mit KI zusammenhängen: Es kann sich um die Daten handeln, die zum Trainieren eines statistischen Modells verwendet werden, oder um die Daten, auf deren Grundlage eine KI-Anwendung eine Empfehlung ausgibt oder eine Entscheidung trifft über einen konkreten Einzelfall. Es ist klar, dass die Prozesse der Datenerhebung und des Datengriffs von jeder Organisation, die in dieser Hinsicht transparent und rechenschaftspflichtig sein will, beschrieben und dokumentiert werden müssen. Zum Beispiel erfordert der

10 Dieses Unterthema entspricht der Anforderung „Schutz der Privatsphäre und Datenqualitätsmanagement“ in den EU-Ethik-Leitlinien für eine vertrauenswürdige KI (29).

Transparenzgrundsatz des Institute of Electrical and Electronics Engineers (IEEE; Grundsatz 4 in Version 2 [14]) die Sicherung und Erfassung von Daten von Sensoren (die von der KI verwendet werden) wie denen in einem Flugdatenschreiber. Aber die detaillierten Kenntnisse und die Kontrolle über alle Prozesse der Datenerhebung, des Datenzugriffs und der Datenspeicherung sind wertvoll für den Schutz der Privatsphäre (z. B. vor unberechtigtem Zugriff) und für die Verbesserung der Robustheit und Zuverlässigkeit von KI-Systemen, auch wenn diese Prozesse nicht nach außen kommuniziert werden, und sogar unabhängig von rechtlichen Verantwortlichkeiten. So ist es nicht verwunderlich, dass viele Empfehlungen die Dokumentation der erhobenen und von Algorithmen verwendeten Daten verlangen, auch im Zusammenhang mit anderen Werten, wie z. B. dem Schutz der Privatsphäre.

Mehrere Empfehlungen beinhalten Vorgaben oder Checklistenpunkte, die vorschreiben, dass die Datenerhebung, der Datenzugriff und die Datenspeicherung durch KI-Systeme immer in einer kontrollierten, transparenten und rechenschaftspflichtigen Weise erfolgen müssen:

> „Is the data collected in an authorized manner? If from third parties, can they attest to the authorized collection of the data? Has consumer been informed of data collection, and where appropriate provided permission about its use?"
> [Werden die Daten auf zulässige Weise erhoben? Falls von Dritten bereitgestellt, können diese die zulässige Erhebung der Daten bescheinigen? Wurde der Verbraucher über die Erhebung der Daten informiert, und hat er ggf. seine Zustimmung zur Verwendung der Daten erteilt?] (26)

> „AI systems must be data responsible. They should use only what they need and delete it when it is no longer needed (‚data minimization'). They should encrypt data in transit and at rest, and restrict access to authorized persons (‚access control'). AI systems should only collect, use, share and store data in accordance with privacy and personal data laws and best practices."
> [KI-Systeme müssen verantwortungsvoll mit Daten umgehen. Sie sollten nur die Daten verwenden, die sie benötigen, und diese löschen, wenn sie nicht mehr benötigt werden (‚Datenminimierung'). Sie sollten Daten während der Übertragung und bei der Speicherung verschlüsseln und den Zugriff auf befugte Personen beschränken (‚Zugriffskontrolle'). KI-Systeme sollten Daten nur in Übereinstimmung mit den Gesetzen zum Schutz der Privatsphäre und personenbezogener Daten sowie mit bewährten Praktiken erheben, verwenden, austauschen und speichern.] (28)

> „Wir vereinfachen und bereichern das Leben unserer Kunden. Wenn künstliche Intelligenz und die Nutzung kundenbezogener Daten uns dabei helfen, Lösungen im Sinne unserer Kunden zu entwickeln, begrüßen wird dies als Chance, die Bedarfe und Erwartungen unserer Kunden zu erfüllen." (10)

In den hier analysierten Texten werden diese Aktivitäten mit Transparenz (14), Achtung der Privatsphäre und personenbezogener Daten (16) und Rechenschaftspflicht in Verbindung gebracht. Transparenz sollte jedoch nicht als erreicht gelten, wenn die Prozesse der Datenerhebung, des Datenzugriffs und der Datenspeicherung dokumentiert sind.[11] Transparenz in dem Sinne, wie wir diesen Wert hier verstehen, bedeutet im Wesentlichen, diese Fakten auf einfache und wirksame Weise den interessierten Parteien zu vermitteln. Verpflichtungen zur Datentransparenz finden sich in den „KI-Leitlinien" der Deutschen Telekom (10; Grundsatz 4) sowie in der „Erklärung zu Ethik und Datenschutz im Bereich der künstlichen Intelligenz" der ICDPPC (16; Grundsatz 5), in der das Recht auf Information und das Auskunftsrecht ausdrücklich erwähnt werden.

Der Grundsatz der Rechenschaftspflicht erfordert Maßnahmen, mit denen eine Organisation – oder eine Person oder eine Rolle innerhalb einer Organisation – die Verantwortung für einen Prozess übernimmt, bei dem es um Daten geht bzw. sich diesen Prozess zu eigen macht. Ein wichtiges Konzept ist das der *Rückverfolgbarkeit der Quelle*. Wenn eine KI-Anwendung eine Entscheidung für eine Organisation trifft, sollten die Daten, die für diese Entscheidungsfindung verwendet werden, einer Person – d. h. einer organisationsinternen oder -externen Prüfer*in – bekannt und für sie verwendbar sein, um die Entscheidung des KI-Systems zu erklären. Der Gedanke der Rückverfolgbarkeit der Quelle kommt auf unterschiedliche Weise in verschiedenen Leitlinien zum Ausdruck. Einige Leitlinien enthalten konkrete Vorschläge, wie z. B. die Empfehlung zur sicheren Speicherung von Sensorendaten (14), die auch als „Ethik-Black Box" bezeichnet wird (24):

> „Auf Roboter angewendet würde die Ethik-Black Box alle Entscheidungen, die Grundlagen für die Entscheidungsfindung, Bewegungen und sensorische Daten für ihren Roboter-Host aufzeichnen." (24)

Wie das Datenschutzrecht (z. B. die EU-Datenschutzgrundverordnung, DSGVO) heben auch die Ethik-Leitlinien die Bedeutung der Gewährleistung eines angemessenen Niveaus der *Cybersicherheit* hervor:

11 Die erste Version der IEEE-Leitlinien (74) enthielt auch den Grundsatz der Datenkontrolle („data agency"), der die digitale Souveränität und die Kontrolle über die Identität als Kontrolle über die Daten, auch durch digitale Agenten, interpretiert. Interessanterweise fehlen diese Gedanken in der zweiten Version (14) völlig. Ein solcher Gedanke fand sich in den anderen 19 hier untersuchten Leitlinien nicht.

> „Wir sorgen nicht nur dafür, dass unsere Sicherheitsmaßnahmen dem aktuellsten Entwicklungsstand entsprechen, sondern behalten auch den Überblick darüber, wie kundenbezogene Daten genutzt werden – und wer auf welche Daten zugreifen darf. Wir verarbeiten keine datenschutzrelevanten Daten ohne Rechtsgrundlage. […] Gleichzeitig schützen wir unsere Systeme vor unerlaubtem externen Zugriff, um die Datensicherheit und den Datenschutz sicherzustellen." (10)

4.1.1. Relevanz für HR Analytics

Es ist nicht überraschend, dass das Thema Datenkontrolle, Transparenz und Rechenschaftspflicht nicht zu den Themen gehört, die in den Leitlinien zur KI-Ethik am häufigsten diskutiert werden. Schließlich überschneidet sich dieses ethische Gebiet stark mit den Rechtsvorschriften über den Schutz der Privatsphäre und den Datenschutz. Somit hat es den geringsten Neuheitswert und muss am wenigsten durch neue Ethik-Leitlinien kodifiziert werden. Fragen des Datenschutzes und des Schutzes der Privatsphäre wurden von den für diese Studie interviewten Gewerkschafter*innen als wesentlich erachtet, und zwar als eine Art notwendige Voraussetzung für die Geltung jeder anderen Ethik-Leitlinie. Die Bedeutung dieser Empfehlungen für HR Analytics lässt sich wie folgt erklären:

1. Die Aufzeichnung der Daten von Sensoren, die zur Überwachung von Mitarbeitenden eingesetzt werden, würde wesentlich dazu beitragen, die Ursachen für kontraintuitive HR-Entscheidungen von KI-Modellen zu bewerten, die mit diesen Daten trainiert wurden. (Relevant für Wert 1, Benefizienz/Vertrauen, und Wert 3, Gerechtigkeit, in Tabelle 2.) Wenn solche Daten jedoch nicht anonymisiert werden (was in vielen Fällen nicht durchführbar sein kann), führt dies zu einer weiteren Gefährdung der Privatsphäre der Beschäftigten und erhöht die mit ihrer Überwachung verbundenen Risiken. Daher sind Leitlinien zum Datenschutz so wichtig. (Relevant für Wert 2, Schadensverhütung, in Tabelle 2.)
2. Es kann zumindest in einigen Fällen angebracht sein, die *informierte Zustimmung* der Mitarbeitenden zur Verwendung ihrer Daten vorzuschreiben. (Relevant für Wert 4, Freiheit, in Tabelle 2.)
3. Neben der *informierten Zustimmung,* und auch in Fällen, in denen ein anderer Rechtsgrund für die Erhebung von Arbeitnehmendendaten herangezogen wird, könnte der Grundsatz der Datenminimierung geltend gemacht werden, um zu versuchen, den Umfang der Daten, die nicht für die Beschäftigung relevant sind, zu begrenzen. Dies wird dem Eingriff der

Arbeitgeberin bzw. des Arbeitgebers in die Privatsphäre der Mitarbeitenden gewisse Grenzen setzen. (Relevant für Wert 4, Freiheit/Schutz der Privatsphäre, in Tabelle 2.)

4. Eine effektive Kommunikation darüber, wie Daten innerhalb eines Unternehmens verwendet werden, kann auch dazu dienen, den Beschäftigten einen gewissen Schutz vor Missbräuchen zu gewähren, wie z. B. wenn dokumentierte Datenprozesse den Arbeitnehmervertreter*innen oder externen Prüfer*innen zugänglich gemacht werden. Aus der Sicht der einzelnen Beschäftigten ist Transparenz über die Datennutzung eine Voraussetzung für die informierte Zustimmung zur Datennutzung. (Relevant für Wert 1, Vertrauen; Wert 2, Schadensverhütung; sowie Wert 4, Freiheit, in Tabelle 2.)
5. Transparenz bei der Erstellung von Profilen trägt dazu bei, Beschäftigte vor der Erhebung von Daten zu schützen, die unzuverlässig oder inkorrekt sind oder für unrechtmäßige Formen der Diskriminierung verwendet werden können. (Wert 2, Schadensverhütung, und Wert 4, Freiheit, in Tabelle 2.)
6. Rechenschaftspflicht für Datenprozesse bietet Anreize gegen unethische Verwendungen von Daten. (Relevant für Wert 1, Vertrauen; Wert 2, Schadensverhütung; Wert 3, Gerechtigkeit; sowie Wert 4, Freiheit, in Tabelle 2.)

4.1.2. Offene Herausforderungen

Andererseits ist der Fokus auf Datenkenntnisse, Transparenz, Kontrolle, Rechenschaftspflicht und Ergebnisverbesserung unzureichend, um Mitarbeitende vor Bedrohungen ihrer Privatsphäre und vor moralisch anstößigen Formen der Diskriminierung zu schützen. Es stimmt, dass es Arbeitgebenden (durch einige Gesetze zum Schutz der Privatsphäre) untersagt ist, Zugang zu Social-Media-Daten zu verlangen. Aber je nach Inhalt der Gesetze in dem Land, in dem die Arbeitgeberin bzw. der Arbeitgeber tätig ist, kann der Grundsatz der Zweckbindung damit vereinbar sein, dass Arbeitgebende Daten aus verschiedenen Quellen erheben. So dürfen Arbeitgebende z. B. die öffentlichen Tweets von Mitarbeitenden zwecks weiterer Analyse sammeln.

Da Algorithmen des maschinellen Lernens verwendet werden, um neue und überraschende Korrelationen zu entdecken, können Arbeitgebende z. B. behaupten, dass die Nutzung von Social Media-Daten eine legitime Grundlage für die Bewertung und Vorhersage des Reputationsrisikos bieten kann, das

beispielsweise mit der Social Media-Präsenz ihrer Mitarbeitenden verbunden ist. (So kann z. B eine Organisation, die sich mit Migrant*innen beschäftigt, dafür sorgen wollen, dass ihre Mitarbeitenden keine fremdenfeindlichen Ansichten öffentlich in sozialen Medien äußern). Was in den Rechtsvorschriften über den Datenschutz fehlt, ist die Definition dieser Grenze, d. h. welche Art von Daten rechtmäßig von Arbeitgebenden zum Zwecke der Entscheidungsfindung im HR-Bereich gesammelt werden dürfen. Sollte z. B. eine NGO, die mit Migrant*innen arbeitet, ein KI-Werkzeug einführen, das die Wahrscheinlichkeit vorhersagt, dass ihre Mitarbeitenden öffentlich fremdenfeindliche oder diskriminierende Ansichten äußern, und zwar auf der Grundlage früherer Interaktionen in sozialen Medien?

Selbst wenn Grenzen hinsichtlich der Art der Arbeitnehmendendaten gesetzt werden, die Arbeitgebende rechtmäßig erheben und analysieren dürfen, reicht dies noch nicht aus, um Arbeitnehmende vor einem Eingriff in die Privatsphäre zu schützen. Die Herausforderung für einen Ansatz, der sich auf die Art von Daten konzentriert, die Arbeitgebende berücksichtigen oder nicht berücksichtigen sollten, besteht darin, dass er die Privatsphäre der Arbeitnehmenden nicht vor (etwas ungenauen) Rückschlüssen schützt, die aus scheinbar harmlosen Daten gezogen werden können (34). Die Möglichkeit, statistische Methoden (z. B. maschinelles Lernen) einzusetzen, bringt die Möglichkeit mit sich, dass Daten, die anscheinend keine Auskunft über Geschlecht, sexuelle Neigung, soziale Klasse oder Familienstand geben, mit einem bestimmten Grad an Unsicherheit aus anderen Daten, die Arbeitgebende leichter erheben können, abgeleitet werden können (34).

Modelle des maschinellen Lernens können Arbeitgebende in die Lage versetzen, mit einem bestimmten Grad an Unsicherheit die Lebensplanung ihrer Beschäftigten (z. B. ob eine Arbeitnehmerin plant, ein Kind zu bekommen) auf der Grundlage von Daten zu beurteilen, die aus dem Beschäftigungskontext stammen, wie z. B. dem Zeitplan einer Arbeitnehmerin, ihrer Sonderwünsche usw. Der Schutz der Privatsphäre wird nicht dadurch gewährleistet, dass kontrolliert wird, welche Art von Daten ausgetauscht werden. Der Schutz der Privatsphäre – im Sinne des Schutzes vor Einflussnahme der Arbeitgeberin bzw. des Arbeitgebers in einem Bereich persönlicher Entscheidungen – wird von KI auch dann bedroht, wenn keine sensiblen Daten ausgetauscht werden, wenn KI eine Technologie zur Verfügung stellt, die auf der Basis von Daten, die rechtmäßig am Arbeitsplatz erhoben werden, fundierte Vermutungen über sensible Eigenschaften anstellt.

4.2. Den Algorithmus kennen, kommunizieren, sich zu eigen machen und verbessern

Die meisten hier analysierten Leitlinien betreffen die Art der Algorithmen selbst und nicht die Datenerhebung. Im Kontext von HR können die Leitlinien mit dem Begriff „Algorithmus" oder „KI" zwei verschiedene elektronisch umgesetzte Sätze von Regeln meinen: erstens den lernenden Algorithmus, bei dem es sich z. B. um einen Prozess handeln kann, mit dem eine allgemeine Regel auf der Grundlage historischer Daten hergeleitet wird; zweitens die (algorithmische) *Entscheidungsregel*, bei der es sich um eine Regel handelt, die auf konkrete einzelne Personen angewendet wird und nach der Verarbeitung von Daten über diese konkreten einzelnen Personen durch eine Art Modell eine Vorhersage, Empfehlung oder Entscheidung über sie erzeugt.

Die *Entscheidungsregel* könnte in vielen Fällen auch als „gelernter Algorithmus" bezeichnet werden, da heutzutage Entscheidungsregeln selten auf der Grundlage von Domänenwissen von Hand kodiert werden. Vielmehr werden Algorithmen des maschinellen Lernens verwendet, um allgemeine Regeln abzuleiten (zu lernen), die dann Outputs (z. B. Vorhersagen, Klassifikationen, Empfehlungen, Entscheidungen) über bestimmte Fälle erzeugen. Im Falle der *prädiktiven Analytik* gibt die Entscheidungsregel lediglich eine Vorhersage über eine Person aus und lässt den Menschen entscheiden, was mit dieser Vorhersage zu tun ist. Im Falle der *präskriptiven Analytik* könnte die Entscheidungsregel beispielsweise eine Empfehlung sein (um einer Mitarbeiterin oder einem Mitarbeiter eine Gehaltserhöhung zu geben), bei der ein Mensch entscheidet, oder eine automatisierte Entscheidung (z. B. um Beschäftigten effizient Urlaub an bestimmten Tagen zuzuweisen).

Eine gewisse Verwirrung entsteht durch die Verwendung des Begriffs „KI" sowohl für den lernenden Algorithmus als auch für die Entscheidungsregel. In einem Algorithmus für neuronale Netze beispielsweise ist der lernende Algorithmus eine Art mathematische Regel, die bestimmt, wie die Gewichtungen der Knoten des Netzes aufgrund der Trainingsdaten festgelegt werden. Informatiker*innen sind mit diesen mathematischen Regeln (z. B. dem Backpropagation-Algorithmus, der Differential- und Integralrechnung zur Minimierung von Fehlern verwendet) sehr gut vertraut und verstehen sie vollständig. Die Anwendung dieser algorithmischen Regeln auf einen (z. B. statistischen) Lernprozess erzeugt eine Entscheidungsregel, wie z. B. ein neuronales Netz, das Bilder von Katzen erkennt.

Die intrinsische Logik der Entscheidungsregel eines neuronalen Netzes kann ein äußerst komplexes Kriterium sein. Sie kann als die Berechnung be-

schrieben werden, die sich aus der Interaktion einzelner Neuronen ergibt, die auf eine bestimmte Art kombiniert und gewichtet werden, die nicht a priori bestimmt, sondern durch einen lernenden Algorithmus festgelegt wird. Wenn man sagt, dass ein neuronales Netz entscheidet, ob ein Bild eine Katze ist, so meint „neuronales Netz" hier die gelernte Entscheidungsregel (nicht den lernenden Algorithmus!). Die Regel hat eine bestimmte Leistung, wie etwa Präzision. Beispielsweise klassifiziert sie Katzen in 90 Prozent der Fälle richtig. Die Generierung von Wissen über den Algorithmus und dessen Dokumentation kann sich auf den lernenden Algorithmus und auf die daraus resultierende Entscheidungsregel beziehen, deren Leistung im Labor bewertet werden kann, etwa indem Tests mit bekannten Daten durchgeführt werden (d. h. Daten über Fälle, deren Bezeichnungen, z. B. „Katze" oder „Hund", der Datenwissenschaftlerin bzw. dem Datenwissenschaftler bereits bekannt sind).

Einige Empfehlungen verlangen, dass Organisationen, die KI-Anwendungen erstellen oder einsetzen, den Lernprozess dokumentieren, einschließlich des lernenden Algorithmus (z. B. der Art des Algorithmus und seiner vom Menschen definierten Parameter) und der Daten, die für das Training und Testen des Algorithmus verwendet werden. Andere Leitlinien sind am besten als Anforderungen zu verstehen, Wissen über die Entscheidungsregel zu produzieren. Zu diesen Leitlinien gehören beispielsweise solche, die von Herstellern oder Betreibern von KI-Systemen verlangen, die Präzision (12, 29), Zuverlässigkeit und Reproduzierbarkeit (29) der Entscheidungen oder Vorhersagen der KI zu bewerten.

Zum Beispiel empfehlen die IEEE-Leitlinien (14) den Herstellern, Betreibern und Eigentümern von KI-Systemen, die folgenden Parameter zu melden: den vorgesehenen Anwendungszweck; (ggf.) die Trainingsdaten/Trainingsumgebung; Sensoren/Datenquellen aus der realen Welt; Algorithmen; Prozessgrafiken; Modelleigenschaften (auf verschiedenen Ebenen); Benutzerschnittstellen; Aktoren/Outputs; sowie das Optimierungsziel, die Verlustfunktion und die Belohnungsfunktion. In den EU-Ethik-Leitlinien für eine vertrauenswürdige KI wird dies als „Rückverfolgbarkeit" bezeichnet und als ein Element der Transparenz aufgeführt (29). Die Rückverfolgbarkeit ist jedoch eindeutig auch für Robustheit[12] und Sicherheit von wesentlicher Bedeutung. Viele Leitlinien heben die Wichtigkeit der Sicherstellung der Qualität

12 Wenn man beispielsweise bestätigen kann, dass ein Modell zur Unterscheidung von Wölfen und Hunden mit Bildern von Wölfen trainiert wurde, die überwiegend Schnee in der Umgebung enthielten, und Bildern von Hunden, die überwiegend keinen Schnee in der Umgebung enthielten, kann man leichter gültige Hypothesen darüber aufstellen, dass es dem Modell nicht gelungen ist, Verallgemeinerungen in Situationen der realen Welt vorzunehmen.

der Trainings- und Testdaten hervor. So schlägt beispielsweise das Weltwirtschaftsforum den Unternehmen vor:

> „Determine whether certain data sets fit internally agreed upon standards of ‚adequate' and ‚representative' data (looking to both quantitative and qualitative metrics); identify opportunities to expand data collection efforts where contextually appropriate, viable and possible to do so without violating privacy."
> [Stellen sie fest, ob bestimmte Datensätze den intern vereinbarten Standards für „angemessene" und „repräsentative" Daten entsprechen (sowohl im Hinblick auf quantitative als auch auf qualitative Metriken); ermitteln Sie Möglichkeiten zur Ausweitung Ihrer Datenerhebungsbemühungen, wo dies kontextuell angemessen, durchführbar und möglich ist, ohne die Privatsphäre zu verletzen.] (21)

Dieser Gedanke – die Forderung nach „angemessenen" und „repräsentativen" Daten und „Möglichkeiten zur Ausweitung der Datenerhebungsbemühungen" – findet sich in verschiedenen Leitlinien.[13] Eine Leitlinie enthält sogar die Empfehlung, Daten auszuschließen, die für die Vorhersage des Ergebnisses nicht relevant sind (26), wenn unklar ist, ob eine solche „Relevanz" vor dem algorithmischen Lernprozess oder a posteriori bewertet werden sollte.[14]

---

13 Zum Beispiel enthält die Bewertungsliste (Pilotversion) in den EU-Ethik-Leitlinien für eine vertrauenswürdige KI (29) die folgenden Fragen: „Haben Sie Maßnahmen ergriffen, um sicherzustellen, dass die verwendeten Daten umfassend und aktuell sind? Haben Sie Maßnahmen ergriffen, um zu beurteilen, ob zusätzliche Daten erforderlich sind, z. B. um die Präzision zu verbessern oder Verzerrungen zu vermeiden?"

14 Ein Problem des Vorschlags besteht darin, dass viele Algorithmen des maschinellen Lernens als statistische Methoden betrachtet werden können, um die Relevanz von Daten für eine bestimmte Vorhersage-Aufgabe zu bestimmen. Die Relevanz von Informationen kann somit nicht beurteilt werden, bevor die Daten in die Datenpipeline des maschinellen Lernens einbezogen werden. Die Leitlinie erscheint redundant, wenn sie implizieren soll, dass eine (über einen Algorithmus des maschinellen Lernens identifizierte) Entscheidungsregel keine Vorhersagen auf der Grundlage von Daten treffen sollte, von denen a posteriori (d. h. durch maschinelles Lernen) festgestellt wird, dass sie nicht zur Präzision des Modells beitragen.
So verstanden, wird es der Algorithmus des maschinellen Lernens selbst sein, der über eine Entscheidungsregel bestimmt, ob eine Quellinformation es wert ist erhoben und analysiert zu werden. Aber eine solche Bewertung erfolgt a posteriori, was bedeutet, dass F&E-Abteilungen berechtigt sind, jede Art von Informationen zu erheben, um festzustellen, ob ein Modell daraus gelernt werden kann. Alternativ kann der Anspruch als Versuch interpretiert werden, das Vorwissen der Fachgebietsexpert*innen über die Faktoren, die die Präzision von Vorhersagen erhöhen, als Kriterium für die Entscheidung zu nutzen, welche Daten a priori aus der Pipeline für maschinelles Lernen ausgeschlossen werden sollen. Wenn ja, dann ist dies eine sehr konservative Anforderung, da es eines der Ziele des maschinellen Lernens und von Big Data ist, neue und sogar nicht intuitive Korrelationen zu entdecken, die nicht zum gefestigten Wissen von Fachgebietsexpert*innen gehören.

Neben den für das Training des Algorithmus verwendeten Daten ist ein weiteres Element der Datenpipeline, das den Leitlinien zufolge dokumentiert werden sollte, der Trainingsprozess, d. h. der verwendete Algorithmus sowie die Parameter, mit denen er eingerichtet wurde. Die IEEE-Leitlinien empfehlen den Herstellern, Betreibern und Eigentümern von KI-Systemen, die Algorithmen, die zur Erzeugung eines Modells verwendet werden, zu registrieren, und zwar insbesondere die Modelleigenschaften sowie das Optimierungsziel, die Verlustfunktion und die Belohnungsfunktion (14). Ähnlich fordern die Leitlinien des Software & Information Industry Association (SIIA) Software-Entwickler*innen auf, ein Inventar und eine Dokumentation aller Modelle zu erstellen, die für automatisierte Entscheidungen verwendet werden (26), und zu erklären, was das Modell vorhersagen soll (26). Andere Leitlinien verlangen, dass die Modellbildung sensibel sein sollte gegenüber den Normen und Werten spezifischer Gruppen, die von den Outputs von KI-Systemen betroffen seien, (21), und dass Informationen über die Leistung des Algorithmus generiert und veröffentlicht werden sollten.

Das IEEE-Dokument (14) empfiehlt Organisationen, die *Transparenz* der fraglichen KI-Systeme anhand strenger Metriken zu bewerten. Die Pilotversion der Bewertungsliste in den EU-Ethik-Leitlinien für eine vertrauenswürdige KI (29) empfiehlt den Designer*innen und Entwickler*innen von KI-Systemen: die Präzisionsziele zu erklären; die Angemessenheit der Daten präzise zu bestimmen; die Präzision zu verbessern (unter „Präzision" in der Bewertungsliste); Mechanismen einzurichten, mit denen sie die Nutzenden über „die Gründe und Kriterien, die den Ergebnissen des KI-Systems zugrunde liegen, informieren"; klarzustellen, „worin der Zweck des KI-Systems besteht und wer oder was von dem Produkt/der Dienstleistung profitieren kann"; und die „Merkmale, Grenzen und potentielle Mängel des KI-Systems" mitzuteilen (29).

Wie bereits erwähnt, fordern einige Leitlinien, die Hersteller, Betreiber und Eigentümer von KI-Systemen auf, die Ziele hinter den Algorithmen zu definieren und aufzuzeichnen (z. B. was sie vorhersagen sollen) und ihre Leistung zu testen. Es handelt sich um Standardschritte im Arbeitsablauf der Datenwissenschaft, aber sie werden wohl in Kombination mit drei zusätzlichen Anforderungen ethisch bedeutsam:

– Die Anforderung, dass Informationen über diese Schritte zur Verbesserung der Transparenz des Systems genutzt werden, idealerweise in dem Sinne, dass sie unabhängigen Prüfer*innen zur Verfügung gestellt werden können, falls die ethische Vertrauenswürdigkeit des KI-Systems infrage gestellt wird. (Relevant für Wert 1, Benefizienz/Vertrauen, und Wert 2,

Schadensverhütung, in Tabelle 2, da robuste und vertrauenswürdige KI-Systeme wahrscheinlich mehr Vorteile bringen und sicherer sind.) In mehreren Leitlinien, wird die Dokumentation des Algorithmus in der Tat als ein Element von Transparenz und angemessener Kommunikation betrachtet.[15]

- Die Anforderung, dass die Ziele des Algorithmus in einer Weise definiert werden müssen, die sensibel (oder zumindest nicht unsensibel) gegenüber den Werten und Normen der Interessengruppen ist. Idealerweise soll damit die Kluft zwischen konkreter Praxis und Automatisierung vermieden werden, die mit Produkten verbunden ist, die von Ingenieur*innen entwickelt werden, die nur wenig Kenntnis von der realweltlichen, gelebten Existenz der von ihren Produkten betroffenen Menschen haben. (Relevant für Wert 1, Benefizienz/Vertrauen, und Wert 2, Schadensverhütung, in Tabelle 2.)
- Die Anforderung, dass die Leistungsbewertung in einer Art und Weise erfolgen muss, die sowohl wissenschaftlich als auch ethisch vertretbar ist. Einige Leitlinien erwähnen ausdrücklich die Präzision (29, 35) von KI-Systemen oder Modellen des maschinellen Lernens. Die scheinbare Präzision eines Modells kann jedoch die Tatsache verschleiern, dass das Modell hochgradig ungenau ist, wenn es auf eine andere Population angewandt wird als die, aus der die Daten stammen (ein Fall von „Überanpassung"). Oder die Art und Weise, wie die Präzision des KI-Systems gemessen wird, kann eine geringe ökologische Validität haben, d. h. ein schlechter Indikator dafür sein, wie sich das KI-System in Situationen der realen Welt verhält, auch unter veränderten Bedingungen, in denen die Menschen wissen, dass ihr Verhalten gemessen und bewertet wird. Aus diesem Grund erwähnen einige Leitlinien neben der Präzision auch die Zuverlässigkeit und Reproduzierbarkeit von KI-Systemen (29).[16]

15 In der Bewertungsliste in den EU-Ethik-Leitlinien für eine vertrauenswürdige KI z. B. ist die Kommunikation des Ziels, der Kriterien und der Grenzen des Algorithmus unter „Erklärbarkeit" aufgeführt (29).

16 Beispielsweise enthält die Bewertungsliste in den EU-Ethik-Leitlinien für eine vertrauenswürdige KI (29) unter der Überschrift „Zuverlässigkeit und Wiederholbarkeit" die folgenden Fragen: „Haben Sie getestet, ob bestimmte Kontexte oder Bedingungen berücksichtigt werden müssen, damit die Wiederholbarkeit gewährleistet ist? Haben Sie Prüfverfahren oder -methoden zur Messung und Sicherstellung verschiedener Aspekte der Zuverlässigkeit und Wiederholbarkeit eingeführt? Haben Sie Verfahren eingeführt, die beschreiben, wann ein KI-System bei bestimmten Einstellungen ausfällt? Haben Sie diese Prozesse zur Erprobung und Prüfung der Zuverlässigkeit von KI-Systemen eindeutig dokumentiert und operativ umgesetzt? Haben Sie Vorkehrungen getroffen oder ein Kommunikationssystem eingerichtet, mit dem Sie (End-)Nutzern die Zuverlässigkeit des Systems garantieren können?"

– Alternativ kann man auch von Robustheit und Richtigkeit (Fakten- und Realitätstreue) sprechen. Diese Werte können aufgrund einer schlechten Auswahl der Trainingsdaten (daher die Betonung der Datenqualität in vielen Leitlinien) beeinträchtigt werden oder aufgrund der Unfähigkeit der Entwickler*innen, zusätzliche Umgebungsvariablen, die die Leistung des Modells beeinflussen und mit dem Einsatz von KI in der realen Welt verbunden sind, vorherzusehen und zu bewerten. (Relevant für Wert 1, Benefizienz/Vertrauen; Wert 2, Schadensverhütung; und Wert 3, Fairness in Tabelle 2.)

Solche Empfehlungen erscheinen in anderen Leitlinien als Methoden zur Umsetzung der Grundsätze der Achtung der Menschenrechte (14), der Transparenz (14, 19), des wertesensiblen Designs [(27–29)], der Rechenschaftspflicht (14), der aktiven Beteiligung (21), der Kontrolle („Wir legen das Fundament" [10]), der Präzision (12), sowie der Transparenz und Kommunikation (29).

4.2.1. Verbesserung der Fairness

Ein weiteres weit verbreitetes Thema betrifft die Diskriminierung durch einen Algorithmus, was in etwa seiner Fairness entspricht. Genauer gesagt, der Algorithmus, um dessen Fairness oder diskriminierende Eigenschaften es geht, nennen wir die *Entscheidungsregel.*[17] Zwei verschiedene Aspekte dieses Problems werden in den meisten Kodizes hervorgehoben, wobei eine Minderheit von Kodizes beide Aspekte betont. Ein Aspekt ist, dass der Gedanke als Fairness, gleiche Robustheit oder Richtigkeit bezeichnet werden könnte. Die Idee dabei ist, dass man robuste und repräsentative Daten benötigt, um sowohl fair als auch (gleichermaßen) genau zu sein, wenn man Vorhersagen und Entscheidungen über alle Gruppen trifft (25). Angemessene Datensätze werden oft als Voraussetzung für faire und nichtdiskriminierende Algorithmen genannt; insbesondere wird der Gedanke geäußert, dass die Daten repräsentativ für die verschiedenen in der Gesellschaft vorkommenden Gruppen, einschließlich Minderheiten, sein sollten. In anderen Dokumenten wird nur allgemein von der Vermeidung von Voreingenommenheit oder verzerrten Daten gesprochen, ohne explizit die Sprache der

17 Dies entspricht der Forderung „Vielfalt, Nichtdiskriminierung und Fairness" – insbesondere der „Vermeidung unfairer Verzerrungen" – in den EU-Ethik-Leitlinien für eine vertrauenswürdige KI (29).

Diskriminierung (die sich typischerweise auf soziale Gruppen bezieht) zu verwenden.[18]

Die Grenze zwischen der Frage der Präzision und Zuverlässigkeit und der Frage der Fairness ist nicht klar definiert. In einigen Fällen handelt es sich bei einem unfairen oder diskriminierenden Algorithmus um einen Algorithmus, der bei bestimmten Einstellungen ausfällt[19] – d. h. in Situationen, in denen Minderheitengruppen (z. B. schwarze Frauen) betroffen sind. Andere Dokumente sind weitaus konkreter in ihrer Bezugnahme auf Fairness und Gleichbehandlung von Gruppen, wie z. B. Gruppen, die nach Geschlecht oder Rasse definiert sind (36–38), wie es für die meisten Antidiskriminierungsgesetze typisch ist. Zum Beispiel empfiehlt das Dokument von Women 20 den G20 -Regierungen, offen zugängliche, nach Geschlecht aufgeschlüsselte Datensätze zu erstellen, damit die Systeme des maschinellen Lernens ihre Leistung verbessern können, mit einem ausdrücklichen Verweis auf das Geschlecht (17). In Bezug auf die Trainingsdaten enthalten die EU-Ethik-Leitlinien für eine vertrauenswürdige KI die folgenden Bewertungslistenpunkte als Methoden zur „Vermeidung unfairer Verzerrungen" (29):

- „Haben Sie mögliche Einschränkungen, die sich aus der Zusammensetzung der verwendeten Datensätze ergeben könnten, bewertet und zur Kenntnis genommen?"
- „Haben Sie beachtet, dass die Daten die Vielfalt der Nutzer widerspiegeln und repräsentativ sind? Haben Sie das KI-System auf spezifische Anwendergruppen oder problematische Anwendungsfälle getestet?"
- „Haben Sie die verfügbaren technischen Hilfsmittel recherchiert und eingesetzt, um ein besseres Verständnis von Daten, Modell und Leistung zu erzielen?" (29)[20]

Einige Leitlinien erwähnen explizit den Wert Fairness (z. B. „Vermeidung unfairer Verzerrungen" [29] im Gegensatz zu „Vermeidung von Verzerrungen") und betonen, dass die Definition dessen, was Fairness für eine Entschei-

18 Zum Beispiel die Empfehlung, dass KI-Systeme nicht mit Daten trainiert werden sollten, die verzerrt, ungenau, unvollständig oder irreführend seien (28).

19 Eine Frage zum Thema „Zuverlässigkeit und Wiederholbarkeit" in der Bewertungsliste in den EU-Ethik-Leitlinien lautet: „Haben Sie Verfahren eingeführt, die beschreiben, wann ein KI-System bei bestimmten Einstellungen ausfällt?" (29).

20 Analog dazu empfiehlt das ICDPPC-Dokument in Bezug auf unrechtmäßige Voreingenommenheit oder Diskriminierungen „angemessene Schritte zur Sicherzustellung, dass die bei einer automatisierten Entscheidungsfindung genutzten personenbezogenen Daten und Informationen richtig, aktuell und so vollständig wie möglich sind" (16).

dungsregel bedeutet, schwierig und stark kontextabhängig sein kann. Beispielsweise empfehlen die Leitlinien des Weltwirtschaftsforums über die Vermeidung diskriminierender Ergebnisse ausdrücklich, dass

> „Persons involved in conceptualizing, developing and implementing machine learning systems should consider which definition of fairness best applies to their context and application, and prioritize it in the architecture of the machine learning system and its evaluation metrics."
> [Personen, die an der Konzeptualisierung, Entwicklung und Implementierung von Systemen des maschinellen Lernens beteiligt sind, sollten überlegen, welche Definition von Fairness am besten auf ihren Kontext und ihre Anwendung zutrifft, und dieser Definition in der Architektur des Systems und dessen Bewertungsmetriken Priorität einräumen.] (21)

Neben der Einsicht, dass es eine Vielzahl von Fairness-Definitionen gibt, heben diese und andere Leitlinien eine Reihe weiterer wichtiger Überlegungen bezüglich konzeptuell integrierter Fairness („fairness by design") hervor: Fairness sollte mit technischen Methoden als ein Aspekt der algorithmischen Leistung (z. B. zusammen mit Präzision und anderen Maßen) messbar sein und als Ziel im Datenpipeline-Prozess festgesetzt werden.[21] Das angemessene Gleichgewicht zwischen Fairness, Schutz der Privatsphäre und Präzision ist ein Aspekt des algorithmischen Designs, der technisch umgesetzt wird. Formulierungen, die sich auf die technische Umsetzung von Fairness beziehen, sind z. B. „fairness-aware data mining algorithms" (fairnessbewusste Data-Mining-Algorithmen; 12) oder „values which may need to be embedded in the machine" (Werte, die möglicherweise in die Maschine eingebettet werden müssen; 11).[22] Die FAT-ML-Grundsätze (12) enthalten weitere Implementierungsvorschläge, wie z. B. die Berechnung von Fehlerquoten und -typen (z. B. falsch-positive vs. falsch-negative Fehler)

21 Analog dazu enthält die Bewertungsliste in den EU-Ethik-Leitlinien für eine vertrauenswürdige KI (29) die folgenden Fragen: „Haben Sie eine angemessene Arbeitsdefinition von ‚Fairness" festgelegt, die Sie bei der Gestaltung von KI-Systemen anwenden? Wird Ihre Definition üblicherweise verwendet? Haben Sie andere Definitionen in Betracht gezogen, bevor Sie sich für diese Definition entschieden haben? Haben Sie eine quantitative Analyse oder Metriken zur Messung und Prüfung der angewandten Definition von ‚Fairness" vorgesehen?"

22 Analog dazu enthält die Bewertungsliste in den EU-Ethik-Leitlinien für eine vertrauenswürdige KI (29) die folgenden Fragen: „Haben Sie während der Phasen der Entwicklung, Einführung und Nutzung des Systems Prozesse eingerichtet, mit denen Sie das System auf mögliche Verzerrungen untersuchen und überwachen können? […] Haben Sie Mechanismen eingeführt, mit denen Sie gewährleisten können, dass Sie faire KI-Systeme einsetzen?"

für verschiedene Teilpopulationen.[23] Es ist auch explizit von „disparate impact" die Rede, dem im US-Recht verwendeten Ausdruck für indirekte Diskriminierung.

Die überwiegende Mehrheit der Verweise auf Fairness bezieht sich auf die Vermeidung von Diskriminierung, die typischerweise als ungleiche oder unfaire Behandlung in Bezug auf soziale Gruppen definiert wird. Die FAT-ML-Grundsätze erfordern die Entwicklung eines Bewusstseins für Ungleichheiten beim Zugang zu Gütern für Gruppen, die durch Rasse, Geschlecht, Geschlechteridentität, Fähigkeitsstatus, sozio-ökonomischen Status, Bildungsniveau, Religion und Herkunftsland (12) definiert sind. (Dies unterscheidet sich von der Forderung nach der Beseitigung aller Ungleichheiten, die, wie wir später erläutern werden, problematisch wäre.)

Einige Dokumente enthalten das abstraktere Konzept der „groups that may be advantaged or disadvantaged by the algorithm" (Gruppen, die durch den Algorithmus begünstigt oder benachteiligt werden können; 21). In einigen Leitlinien werden die relevanten Gruppen unter Bezugnahme auf frühere und bestehende soziale Ungerechtigkeiten und, im weiteren Sinne, Ungleichheiten identifiziert. Die Aufforderung, Ziele für Systeme des maschinellen Lernens zu vermeiden, die sich selbst erfüllende Erfolgsmarker schaffen und Muster der Ungleichheit verstärken, denn sonst könnten bestehende Muster struktureller Diskriminierung reproduziert und verschärft werden (19), stellt ein Hinweis auf historische Ungleichheit und Ungerechtigkeit dar.

Zusammenfassend lässt sich feststellen, dass Leitlinien, die der Frage der Fairness große Aufmerksamkeit widmen, sinngemäß die folgenden Empfehlungen enthalten:

Dokumentieren Sie Voreingenommenheit, Diskriminierung, Unfairness und die ausgewählten Fairnessziele

1. Erwerben Sie Wissen über Normen, die sich auf den Kontext des Einsatzes eines Algorithmus und auf den kulturellen Kontext im Allgemeinen beziehen, einschließlich sozialer und rechtlicher Normen (14), sowie Wissen über die verschiedenen möglichen Definitionen von Fairness, die in Bezug auf diese Besonderheiten übernommen werden können (21, 29). (Relevant für Wert 1, Vertrauen, und Wert 3, Gerechtigkeit/Fairness, in Tabelle 2.)

23 Dieselbe Forderung findet sich auch in anderen Dokumenten, wie z. B. den Leitlinien des Weltwirtschaftsforums (21).

2. Ziehen Sie Fachgebietsexpert*innen zu Rate und berücksichtigen Sie interdisziplinäre Erkenntnisse, um potenzielle Voreingenommenheiten und Unfairness zu verstehen (21). (Relevant für Wert 1, Wert 2, Wert 3 und Wert 4 in Tabelle 2.)
3. Ermitteln Sie ungleiche Fehlerquoten nach Bevölkerungsgruppen unter Berücksichtigung verschiedener Fehlertypen (12, 21). (Relevant für Wert 3, Fairness, in Tabelle 2.)
4. Ermitteln Sie indirekte Diskriminierungen (wenn Gruppen, in unterschiedlichem Maße von algorithmischen Entscheidungen profitieren; 12, 21, 26), z. B. indem Sie indirekte Diskriminierungen bei Mitgliedern geschützter Personengruppen dokumentieren. Es wird anerkannt, dass indirekte Diskriminierungen nicht vollständig eliminiert werden können. Daher sollte die Vermeidung solcher indirekten Diskriminierungen unter Berücksichtigung der damit verbundenen Kosten verhältnismäßig und notwendig sein (26). Das SIIA-Dokument (26) empfiehlt sogar, sensible Informationen (Rasse, Geschlecht, ethnische Zugehörigkeit und Religion) zu erheben und, soweit dies gesetzlich zulässig sei, in Datenanalyse-Systemen zu verwenden, um indirekte Diskriminierungen zu bewerten.
5. Einige Leitlinien betonen die Bedeutung der Messung von Unfairness nicht nur in Bezug auf die Testdaten (bei denen es sich um historische Daten handelt, die in der Regel aus derselben Quelle stammen wie die, die für das Training verwendet wurden), sondern auch durch die Überwachung der Leistung der Anwendung der algorithmischen *Entscheidungsregel* im jeweiligen Anwendungsfall. Die Toronto-Erklärung (19) erwähnt z. B. die Bedeutung präziser Vorabversuche und der Einrichtung eines *fortlaufenden Evaluierungssystems* während des gesamten Lebenszyklus des Produkts. Die EU-Ethik-Leitlinien für eine vertrauenswürdige KI (29) verlangen von Organisationen, die KI einsetzen, dass sie sich selbst bewerten, indem sie fragen: „Haben Sie während der Phasen der Entwicklung, Einführung und Nutzung Prozesse eingerichtet, mit denen Sie das System auf mögliche Verzerrungen überprüfen und überwachen?“

Verbessern Sie die Fairness der Ergebnisse

1. Erheben Sie mehr bzw. repräsentativere Daten, um Diskriminierung zu verringern, oder erstellen Sie disaggregierte Datensätze, einschließlich Möglichkeiten zur Erweiterung der Datensammlungen (17, 19, 21, 29). (Relevant für Wert 1, Vertrauen, und Wert 3, Fairness, in Tabelle 2.)
2. Verwenden Sie kein algorithmisches System, wenn dasselbe Ziel mit einem Algorithmus erreicht werden kann, der weniger indirekt diskrimi-

nierende Auswirkungen hat (26). Diese Empfehlung scheint aus dem US-Gesetz über indirekte Diskriminierung („disparate impact") abgeleitet zu sein, das jedoch nur für einen begrenzten Bereich von Entscheidungen gilt (Entscheidungen, die Beschäftigung und das Wohnungswesen betreffen). Das SIIA-Dokument postuliert sogar, dass es eine ethische Verpflichtung gibt, ein System neu zu gestalten, damit es weniger indirekt diskriminierende Auswirkungen auf geschützte Personengruppen hat, auch wenn dadurch die organisatorische Effektivität geopfert werde, wenn das Kosten-Nutzen-Verhältnis eines solchen Schrittes vertretbar sei (26). (Relevant für Wert 3, Fairness, in Tabelle 2.)

3. Andere Dokumente, die die Verhütung oder Minderung von Diskriminierung fordern, sind sehr allgemein gehalten, was die Definition von unfairer Diskriminierung betrifft, sowie die Art von Diskriminierung, die beseitigt werden sollte.[24] In einigen Leitlinien werden Unfairness und Diskriminierung in Bezug auf Fehler und Verzerrungen charakterisiert, wenn solche Fehler oder Verzerrungen bestimmte Personengruppen unverhältnismäßig stark beeinträchtigen können (20). Zu beachten ist jedoch, dass die Beseitigung von Fehlern, die sich unterschiedlich auf verschiedene Gruppen auswirken, eine Auffassung von Ungerechtigkeit als „disparate mistreatment" (39), darstellt, d. h. als schlechte Behandlung, von denen manche Gruppen stärker betroffen sind als andere: Die Milderung indirekter Diskriminierungen (siehe Punkt 7) ist ein ganz anderes Ziel und kann sogar dazu führen, die unausgewogene Fehlerverteilung zwischen Gruppen zu verstärken.
4. Die Ungleichheit der Geschlechter und die geschlechtsspezifische Voreingenommenheit werden besonders hervorgehoben. Einer der Grundsätze im Dokument der UNI Global Union lautet „Sicherung einer geschlechterlosen, unvoreingenommenen KI" (24). Eine Leitlinie kann als Vorschlag gelesen werden für (durch algorithmisches Design hervorgebrachte) positive/umgekehrte Diskriminierung in Bereichen, in denen Frauen

24 So z. B. die Erklärung von Toronto (19), die von Unternehmen verlangt, wirksame Maßnahmen zur Verhütung und Minderung von Diskriminierung zu ergreifen; die Universal Guidelines von „The Public Voice" (15), die empfehlen, dass die Institutionen sicherstellen müssten, dass die KI-Systeme keine Voreingenommenheit widerspiegeln oder unzulässige diskriminierende Entscheidungen treffen – wobei das Attribut „unzulässig" suggeriert, dass einige diskriminierende Entscheidungen zulässig sein könnten; und schließlich die ICDPPC-Erklärung (16), die feststellt, dass" unrechtmäßige Voreingenommenheit oder Diskriminierungen […] verringert und gemindert" werden sollten. Es sei darauf hingewiesen, dass Diskriminierung im Sinne des Völkerrechts (wie in der Erklärung von Toronto erwähnt) möglicherweise eine engere Reichweite hat als Fairness.

in der Gesellschaft benachteiligt sind, denn sie empfiehlt algorithmisch gerechte Maßnahmen zur Korrektur von Voreingenommenheiten und Barrieren im realen Leben, die Frauen daran hindern, volle Teilhabe und gleichberechtigte Wahrnehmung von Rechten zu erreichen (17). (Relevant für Wert 3, Gerechtigkeit, in Tabelle 2.)

5. Führen Sie Abhilfemechanismen ein, wie z. B. Dringlichkeitsverfahren zur Korrektur unvorhergesehener Fälle von Unfairness (21). (Relevant für Wert 1, Vertrauen, und Wert 3, Fairness, in Tabelle 2.)

Transparenz in Bezug auf die Beseitigung von Diskriminierung bzw. die Förderung von Fairness

1. Einige Leitlinien fordern die Dokumentation der Formen von Unfairness und Diskriminierung, die entdeckt werden, sowie der Anstrengungen, die unternommen werden, um Diskriminierung in Systemen des maschinellen Lernens zu erkennen, zu verhindern und zu mildern (19, 21). Dazu gehört beispielsweise die Einführung eines Fairnesszertifikats für KI-Systeme (11). (Relevant für Wert 1, Vertrauen, und Wert 3, Fairness, in Tabelle 2.) Interessanterweise schlägt ein Dokument den Unternehmen sogar vor, zu erläutern, wenn Diskriminierung ein erwünschtes Ergebnis eines Modells ist, und die Verantwortlichen zur Rechenschaft zu ziehen (21).
2. Einige Leitlinien empfehlen, es Dritten zu ermöglichen, Formen der algorithmischen Voreingenommenheit und Diskriminierung zu überwachen, zu melden und zu bewerten. So enthält z. B. das FAT-ML-Dokument einen Grundsatz der Nachprüfbarkeit (12), und die Bewertungsliste in den EU-Ethik-Leitlinien für eine vertrauenswürdige KI enthält die Aufforderung, einen Mechanismus vorzusehen, „der es anderen ermöglicht, Probleme im Zusammenhang mit Verzerrungen, Diskriminierung oder schlechter Leistung des KI-Systems zu kennzeichnen“ (29).

Inklusive F&E-Teams

1. Einige Leitlinien betonen die Bedeutung der Vielfalt in F&E-Teams im Bereich KI. Die EU-Ethik-Leitlinien für eine vertrauenswürdige KI (29) bezeichnen die Bildung von vielfältig zusammengesetzten und inklusiven Entwurfsteams sogar als ein nicht-technisches Verfahren zur Sicherung einer vertrauenswürdigen KI. Das Ziel der Inklusion kann auf unterschiedliche Weise spezifiziert werden. Zum Beispiel empfiehlt das Weißbuch des Weltwirtschaftsforums (21) Unternehmen, die für die Entwicklung von KI-Systemen verantwortlich sind, verschiedene Perspek-

tiven zusammenzuführen und Einblicke darüber zu gewähren, ob bestimmte Personengruppen angemessen in die Trainingsdaten einbezogen und vertreten werden. In diesem Sinne verstanden, geht der Vorschlag weit über die Geschlechterinklusion oder die Inklusion von Mitgliedern von Minderheitengruppen in F&E-Teams hinaus, da er die Inklusion von Forschenden oder Fachleuten mit einem anderen (d. h. nicht Informatikrelevanten) disziplinären Profil rechtfertigen kann. (Relevant für Wert 1, Vertrauen, und Wert 3, Fairness, in Tabelle 2.)

2. Ein Thema, das manchmal mit Fragen der Fairness und Diskriminierung zusammenhängt, ist die Bekämpfung negativer Stereotypen, wie z. B. Geschlechterstereotypen bei Spielzeug- und Sex-Robotern (27). (Relevant für Wert 3, Fairness, in Tabelle 2.)
3. In einer Leitlinie wird behauptet, dass Stereotypisierung (wie z. B. in Punkt 2 oben erwähnt) auf die geringe Beteiligung und marginale Inklusion von Frauen in die Kodierung und Gestaltung von KI-Systemen und Technologien des maschinellen Lernens zurückzuführen sei (17). Folglich empfiehlt sie den G20-Regierungen:
   - […] proaktive Schritte zu unternehmen, um mehr Frauen, die KI-Systeme entwerfen, in die Arbeitnehmerschaft zu integrieren […].
   - […] von den Unternehmen zu verlangen, das zahlenmäßige Verhältnis von Frauen und Männern in ihren Designteams proaktiv offen zu legen.
   - […] von den Empfänger*innen von Forschungsförderung zu verlangen, das zahlenmäßige Verhältnis von Frauen und Männern in den antragstellenden Forschungsteams offen zu legen.
   - […] dafür zu sorgen, dass das zahlenmäßige Verhältnis von Frauen und Männern in Entscheidungsprozessen ausreichend ausgewogen ist […]
   - […] von Frauen geführte Technologieunternehmen im KI-Bereich finanziell zu fördern […].
   - […] anderen Unternehmen Anreize für eine vielfältigere Belegschaft auf allen Ebenen zu bieten. (17)

Ähnliche Empfehlungen, die auf ein besseres Geschlechtergleichgewicht in den für KI zuständigen technischen Teams abzielen, finden sich z. B. in anderen Leitlinien, etwa eine Anreizpolitik, die darauf abzielt, bis zum Jahr 2020 den Anteil der weiblichen Studierenden in digitalen Fächern an Universitäten und Wirtschaftshochschulen sowie in deren Vorbereitungskursen zu erhöhen (23). (Relevant für Wert 3, Fairness, in Tabelle 2.)

4.2.2. Verbesserung der Verständlichkeit

Transparenz besteht nicht nur in der Dokumentierung von Zielen, Prozessen und Ergebnissen. Sie besteht auch darin, diese in einer verständlichen Weise zu kommunizieren. Eine der am breitesten und intensivsten diskutierten Fragen ist, inwieweit Algorithmen von verschiedenen Interessengruppen, einschließlich der breiten Öffentlichkeit, verstanden werden können und sollten. Auch hier ist der Gegenstand der Verständlichkeit die *Entscheidungsregel,* d. h. der *gelernte* Algorithmus, nicht der Algorithmus des maschinellen Lernens. Die Entscheidungsregel macht Vorhersagen oder Empfehlungen oder trifft Entscheidungen über einzelne Fälle in konkreten Anwendungen. Ein besonders wichtiger Aspekt der Unterwerfung unter oder Beeinflussung durch Entscheidungen auf der Grundlage von Regeln, die von Algorithmen bestimmt werden, ist die Möglichkeit, den Grund oder die Gründe für eine solche Entscheidung nachzuvollziehen. Der Anspruch, dass Algorithmen (hier: Entscheidungsregeln) verständlich oder erklärbar sein sollten, bezieht sich somit auf einen vorgeblich moralischen oder rechtlichen Anspruch auf Erklärungen:

> „All individuals have the right to know the basis of an AI decision that concerns them. This includes access to the factors, the logic and techniques that produced the outcome."
> [Jeder Einzelne hat das Recht, die Grundlage einer KI-Entscheidung zu erfahren, die ihn betrifft. Dazu gehört der Zugang zu Informationen über die Faktoren, die Logik und die Techniken, die das Ergebnis hervorgebracht haben.] (15)[25]

Ein häufig erwähnter Gedanke ist, dass Verständlichkeit und Erklärbarkeit relative und nicht absolute Eigenschaften sind. Eine aus Sicht von Datenwissenschaftler*innen gute Erklärung kann für einen Laien nicht verständlich sein, und eine Erklärung, die die Neugier eines Laien befriedigen kann, kann aus datenwissenschaftlicher Sicht als obskur angesehen werden. Daher empfehlen einige Leitlinien, diese zielgruppenbezogene Relativität zu berücksichtigen, z. B.:

> „Explainability Guiding Questions: Who are your end-users and stakeholders?"

25 Analog dazu: „Beschäftigte müssen auch ein „Recht auf Erklärung" haben, wenn KI-Systeme im Personalwesen eigesetzt werden, wie etwa bei Einstellungen, Beförderungen oder Entlassungen" (24).

> [Erklärbarkeit – Leitfragen: Wer sind Ihre Endnutzer*innen und Interessengruppen?] (12)[26]

Es gibt grundsätzlich zwei verschiedene Stränge zur Erklärbarkeit.

Der erste Strang begreift Erklärbarkeit ganzheitlich: Um zu verstehen, *warum* eine KI etwas tut, gilt es, gemeinsam oder in kombinierter Form Folgendes zu dokumentieren:

- der Quellcode (24) (obwohl dieser weithin sowohl als unzureichend als auch als unnötig angesehen wird)
- die Datenquellen (d. h. wann und wo Daten erhoben werden) sowohl für Trainingsdaten als auch für Testdaten (11, 12, 29)
- der Prozess der Datenbereinigung oder Datentransformation (11, 12)
- die Features, die verwendet werden, um einen Algorithmus zu trainieren, Entscheidungen zu treffen (11)
- die Gewichtungen dieser Features (falls bekannt) (11)
- der Algorithmustyp sowie das Ausmaß seiner Undurchsichtigkeit (11, 12)
- die verschiedenen Leistungsmetriken (11, 12)
- das Validierungsverfahren (11, 12)
- das allgemeine Ziel und der Zweck des Algorithmus (29) im Sinne des Verwendungszwecks (14)
- die mathematischen Ziele des Algorithmus im Sinne seines Optimierungsziels/seiner Verlustfunktion/seiner Belohnungsfunktion (14)

Allgemeiner und abstrakter ausgedrückt, besteht dieser Ansatz zur Erreichung von Nachvollziehbarkeit darin, die Faktoren, die Logik und die Techniken (15), „die Gründe und Kriterien, die den Ergebnissen des KI-Systems zugrunde liegen" (29), und in einigen Fällen das „Kochbuch" zu vermitteln, d. h. wie diese Logik technisch umgesetzt wurde. Transparenz über die Entwurfsziele, -gründe und -kriterien kann auch dann erreicht werden, wenn die Entscheidungsregel selbst, d. h. „der interne Workflow des Modells" (29), nicht transparent ist, weil sie zu komplex ist, um vom menschlichen Verstand in ihrer Gesamtheit erfasst zu werden (40). Mit anderen Worten: Diese Art der Nachvollziehbarkeit besteht darin, die Ziele und Desiderata der Entscheidungsregel zu erklären; im Fokus stehen nicht einzelne Entscheidungen (41–44).

26 Analog dazu erwähnen die EU-Ethik-Leitlinien für eine vertrauenswürdige KI (29) in der Bewertungsliste unter „Kommunikation", die eine Unterrubrik von „Transparenz" ist, verschiedene Interessengruppen: „Endnutzer", „Personen, die das KI-System in ein Produkt oder eine Dienstleistung einbauen", „Dritte" und „die Öffentlichkeit".

Die Aufgaben, die für die in diesem Sinne aufgefasste Verständlichkeit empfohlen werden, umfassen auch die oben beschriebenen Aufgaben der Dokumentation des Modellentwurfs. Verständlichkeit ist jedoch am besten als ein Aspekt der algorithmischen Transparenz zu verstehen, der neben der (internen) Dokumentation das Element der zielgerichteten Kommunikation (nach außen) beinhaltet. Wenn Verständlichkeit ein Aspekt der Transparenz ist, reicht es möglicherweise nicht aus, dass die Datenwissenschaftler*innen die Logik und die implementierte Methodik in einer Weise dokumentieren, die andere Datenwissenschaftler*innen oder technisch versierte Prüfer*innen verstehen können. Tatsächlich kann Transparenz andere Zielgruppen haben als Fachkolleg*innen und Prüfer*innen. Als ein Aspekt der Transparenz verstanden, ist Verständlichkeit auch die Anforderung, die Logik und die Verallgemeinerungen, die der Regel zugrunde liegen, so zu erklären, dass sie für bestimmte andere Zielgruppen nachvollziehbar sind[27] (siehe Punkt 3 unten).

Der zweite Diskussionsstrang zum Thema Verständlichkeit betrifft die Erklärbarkeit des Ergebnisses der Anwendung einer Entscheidungsregel auf bestimmte Einzelfälle. Sie lässt sich anhand von Empfehlungen veranschaulichen, wie z. B.:

- der Vorschlag, einen „Warum haben Sie das getan?"-Knopf in KI-Anwendungen einzubauen, die mit Menschen interagieren (14)
- die Idee, dass „[d]ie von der Black Box gelieferten Daten Roboter auch dabei unterstützen [könnten], ihr Handeln in einer für Menschen verständlichen Sprache zu erklären" (24)
- die Idee, dass es in manchen Fällen angebracht sein könne, für jede Entscheidung eine automatisierte Erklärung zu entwickeln (12)

Diese Empfehlungen könnten mit der Idee dessen verbunden sein, was in der Literatur als Post-hoc-Erklärungen einzelner Entscheidungen einer KI bezeichnet worden ist, was weiter unten näher untersucht wird (siehe Abschnitt 4.2.4). Analog zu menschlichen Erklärungen wird die entsprechende

27 Das Element der Kommunikation (neben der Dokumentation) ist in der Bewertungsliste in den EU-Ethik-Leitlinien für eine vertrauenswürdige KI (29) ersichtlich, die folgende Fragen enthält: „Haben Sie je nach Anwendungsfall auch den Informationsaustausch und die Transparenz gegenüber anderen Zielgruppen, Dritten und der Öffentlichkeit erwogen? […] Haben Sie eindeutig die Merkmale, Grenzen und potenzielle Mängel des KI-Systems folgenden Personen mitgeteilt? Bei der Entwicklung: denjenigen Personen, die das KI-System in ein Produkt oder eine Dienstleistung einbauen? Bei der Einführung: den Endnutzern oder Verbrauchern?"

Erklärung nicht ganzheitlich sein, sondern vielmehr den wichtigsten Faktor oder eine begrenzte Anzahl wichtiger Faktoren ermitteln und diese Faktoren als Gründe für eine bestimmte Empfehlung oder Entscheidung charakterisieren (44–46).

Einige Verständlichkeitsleitlinien geben keine Präferenz für den ersten (ganzheitlichen) oder den zweiten (Post-hoc-) Ansatz zur Erreichung von Verständlichkeit an. Sie betonen lediglich, wie wichtig es ist, dass der Schwierigkeitsgrad der Erklärungen zielgruppengerecht ist. Sie bleiben jedoch agnostisch darüber, was das ist. Der Hintergrundgedanke besteht darin, anzuerkennen, dass es nicht interpretierbare Blackboxes gibt (47)[28] und dennoch die Logik, die der Annahme dieser Regeln zugrunde liegt, in einer Weise zu dokumentieren, zu erklären und zu kommunizieren, die die Bewertung dieser Logik für alle relevanten pragmatischen Zwecke, wie z. B. rechtlichen Zwecke, bewertbar macht (45, 46, 48). Solche Leitlinien können im Hinblick auf die Art der Verständlichkeit, um die es hier geht, als „agnostisch" bezeichnet werden. Die Mehrzahl der Erklärbarkeitsempfehlungen in den 20 hier analysierten Leitlinien entsprechen diesem Typus. Dazu gehören z. B.:

- die Anforderung, dass eine Organisation Folgendes bewerten sollten:
  - Inwieweit der Algorithmus undurchsichtig (eine Blackbox) ist (11)
  - Wie viel von dem System/Algorithmus den Nutzenden und Interessengruppen erklärt werden kann (12)
- dass die Organisation einen Plan haben sollte, wie Entscheidungen den Nutzenden und den Personen, die von diesen Entscheidungen betroffen sind, erklärt werden sollen (12)
- Es soll geprüft werden, ob ein direkt interpretierbares oder erklärbares Modell verwendet werden kann. (12)[29]
- Die Systeme müssen in der Lage sein, eine Erklärung ihrer Entscheidungsfindung zu liefern, die für die Endnutzenden verständlich ist und von einer kompetenten menschlichen Instanz überprüft werden kann. Wo dies nicht möglich ist und Rechte auf dem Spiel stehen, müssen die für den Entwurf, den Einsatz und die Regulierung der Technologie des maschinellen Lernens Verantwortlichen sich fragen, ob sie zur Anwendung kommen soll oder nicht (21).

28 Manchmal liegt der Grund dafür, dass es sich um Blackboxes handelt, in der Technologie begründet; manchmal entscheidet man sich bewusst dafür, innere Abläufe des Algorithmus nicht zu enthüllen.

29 Analog dazu enthält die Bewertungsliste in den EU-Ethik-Leitlinien für eine vertrauenswürdige KI (29) die folgende Frage an Organisationen: „Haben Sie recherchiert und versucht, das einfachste und am besten interpretierbare Modell für die jeweilige Anwendung zu verwenden?"

Bislang haben wir uns mit Leitlinien befasst, die die Bedeutung *technischer* Lösungen betonen. Einige Leitlinien betonen jedoch *organisatorische* Lösungen zur Förderung der Verständlichkeit, wie z. B. eine menschliche Ansprechstelle für die von den Entscheidungen betroffenen Personen, die über die damit verbundene Logik Bescheid wissen wollen. Zum Beispiel:

> „Die Transparenz und Verständlichkeit der Systeme mit künstlicher Intelligenz sollte mit dem Ziel einer wirksamen Umsetzung, insbesondere durch folgende Maßnahmen verbessert werden: […] c. eine transparentere Gestaltung der Praktiken der Organisationen, insbesondere durch die Förderung der Transparenz von Algorithmen und die Überprüfbarkeit von Systemen, wobei gleichzeitig die Aussagekraft der bereitgestellten Informationen sicherzustellen ist […]." (16)[30]

> „Introduce a new ‚Certificate of Fairness for AI Systems' alongside a ‚kite mark' type scheme to display it. Criteria to be defined at industry level, similarly to food labelling regulations."
> [Führen Sie ein neues „Fairness-Zertifikat für KI-Systeme" ein zusammen mit einem System vom Typ Gütesiegel zur Darstellung dieses Zertifikats. Die Kriterien sind auf Industrie-Ebene zu definieren, ähnlich wie bei den Vorschriften für die Kennzeichnung von Lebensmitteln.] (11)

Andere Leitlinien (in der Regel solche, die sich an Regierungen, Organisationen des öffentlichen Sektors oder mehrere Interessengruppen richten) schlagen auch *institutionelle* Lösungen zur Förderung der Einführung leicht verständlicher und damit transparenter KI-Systeme vor.[31] Dazu gehören z. B.:

> „Establish an AI regulatory function working alongside the Information Commissioner's Office and Centre for Data Ethics – to audit algorithms, investigate complaints by individuals, issue notices and fines […] and ensure algorithms must be fully explained to users and open to public scrutiny."
> [Richten Sie eine KI-Regulierungsstelle ein, die mit dem Information Commissioner's Office (der Datenschutzbehörde) und dem Centre for Data Ethics zusammenarbeitet, um Algorithmen-Audits durchzuführen und Beschwerden von einzelnen Personen zu prüfen, Bescheide zu erlassen und Bußgelder zu verhängen, […] und stellen Sie sicher, dass die Algorithmen den Nutzenden vollständig erklärt werden und für öffentliche Überprüfung offen sein müssen.] (11)

30 Dabei handelt es sich um „nicht-technische Verfahren" im Sinne der EU-Ethik-Leitlinien für eine vertrauenswürdige KI (29), nämlich „Rechenschaftspflicht durch Rahmenbedingungen für die Lenkung und Kontrolle „ und „Zertifizierung".

31 Dies sind ebenfalls Beispiele für ein „nicht-technisches Verfahren für eine vertrauenswürdige KI" (29), nämlich „Regulierung".

„Introduce a ‚reduced liability' incentive for companies that have obtained a Certificate of Fairness to foster innovation and competitiveness."
[Führen Sie einen Anreiz „eingeschränkte Haftung" für Unternehmen ein, die ein Fairness-Zertifikat erhalten haben, um Innovation und Wettbewerbsfähigkeit zu fördern.] (11)

„Introduce a mandatory requirement for public sector organizations using AI for particular purposes to inform citizens that decisions are made by machines, explain how the decision is reached and what would need to change for individuals to get a different outcome."
[Führen Sie eine zwingende Anforderung für Organisationen des öffentlichen Sektors ein, die KI für bestimmte Zwecke einsetzen, wonach sie die Bürgerinnen und Bürger darüber informieren müssen, dass Entscheidungen von Maschinen getroffen werden, und sie erklären müssen, wie die Entscheidung zustande kommt, und was sich für Personen ändern müsste, um ein anderes Ergebnis zu erzielen.] (11)

4.2.3. Relevanz für HR Analytics

Lassen Sie uns nun die Relevanz dieser ethischen Empfehlungen für die KI-Anwendungen erklären, die im Bereich HR Analytics verwendet werden. Betrachten wir zunächst die Vorgabe, dass man wissen, bewerten und dokumentieren sollte, wie eine KI (ein gelernter Algorithmus) funktioniert (d. h. was ihre Leistungsmetriken und -grenzen sind) und warum (d. h. was ihre Quellen, ihr Ziel, ihre Parameter sind und wie sie erstellt wurde). Das Wissen über den Algorithmus ist insofern ethisch wertvoll, als es dazu beiträgt, eine vertrauenswürdige Technologie aufzubauen. In Anlehnung an Ferrario, Loi und Viganò (49) wird hier die Vertrauenswürdigkeit von KI so verstanden, dass die Eigenschaften, über die die KI verfügt, dergestalt sind, dass es für die Nutzenden rational ist, *ihr einfach zu vertrauen*. Mit anderen Worten, es ist für die Nutzenden rational, sich auf die Technologie zu verlassen, selbst wenn sie nur über wenig Wissen und Kontrolle über ihre Funktionsweise verfügen. Das heißt, wenn die Technologie für den Zweck eingesetzt wird, für den sie entwickelt und empfohlen wurde, selbst ohne detaillierte Kenntnisse und Kontrolle seitens der Endnutzenden, ergibt sich für die durchschnittlichen Nutzenden ein Gewinn oder ein Nutzen. Eine vertrauenswürdige Technologie ist bequem und sicher in der Anwendung. Sie kann es nur geben, wenn diejenigen, die die Technologie entwerfen, sowohl die Technologie selbst als auch den Kontext ihrer Anwendung (und ihres Missbrauchs) gut verstehen, so dass sie mögliche Probleme vorhersehen und vermeiden können. Daher

ist die im HR-Bereich eingesetzte vertrauenswürdige KI so konzipiert, dass sie HR-Manager*innen zugute kommt, die sich auf sie verlassen – selbst wenn diese nicht die Fähigkeit haben, die Vertrauenswürdigkeit der KI technisch zu bewerten – solange die Technologie entsprechend ihrem festgelegten Zweck und ihren genau definierten Grenzen eingesetzt wird. Eine vertrauenswürdige KI im HR-Bereich bringt Vorteile für HR-Manager*innen, die nicht vollständig verstehen können, warum die Technologie vertrauenswürdig ist (50). Vertrauenswürdige KI im HR-Bereich ist ein Ziel, und es ist noch nicht klar, ob dieses Ziel erreichbar ist. Die Verbesserung des Prozesses des Erwerbs von Wissen und Kontrolle über den Algorithmus sollte die KI vertrauenswürdiger machen, denn sie beinhaltet, die Grenzen der Technologie zu kennen und klar zu dokumentieren (z. B. welche Personengruppen sie genau und welche sie ungenau klassifiziert) und auch zu wissen und zu dokumentieren, wie die Technologie missbraucht werden kann. Dies sollte zu einem besseren Design führen und durch transparente Kommunikation die Möglichkeit des Missbrauchs mindern.

Im Falle der KI reicht dieses einfache Vertrauen (sich darauf zu verlassen, dass die Hersteller alle technischen Schritte und Bewertungsschritte korrekt durchführen) jedoch wohl nicht aus, um eine vertrauenswürdige Technologie zu liefern. Eine andere Art von Vertrauen, nämlich das *reflexive* Vertrauen (49), sollte vorhanden sein, um zu verhindern, dass eine solche Technologie den Nutzenden schadet. Der Gedanke dabei ist, dass der marktwirtschaftliche Wettbewerb keine ausreichende Gewähr dafür bietet, dass erfolgreiche Unternehmen vertrauenswürdige Technologie produzieren.[32] Die KI im HR-Bereich braucht wohl „*Wächter*innen*" (z. B. Prüfer*innen, Zertifizierungsstellen, NGOs, Enthüllungsjournalist*innen usw.), um die KI, die in HR-Lösungen eingesetzt wird, zu überwachen und zu kritisieren, und zwar als Ergänzung zu den Marktanreizen für Unternehmen, vertrauenswürdige Technologie zu liefern. Die Bewertung durch kompetente und unabhängige „Wächter*innen" kann dazu beitragen, *reflexives* Vertrauen aufzubauen – Vertrauen, das auf der Überzeugung beruht, dass es sich lohnt, sich auf eine Technologie zu verlassen (für den Zweck, für den sie vermarktet wird), was letztlich auf einer vertrauenswürdigen Bewertung ihrer Qualität beruht (49). Reflexives Vertrauen impliziert also *Transparenz* in Bezug auf die Gestaltung

32 Man mag z. B. skeptisch sein, dass Facebook, ein in Bezug auf Marktfähigkeit eindeutig erfolgreiches Unternehmen, vertrauenswürdige Technologie produziert – Technologie, die Nutzenden zugute kommt, die ihr einfach vertrauen (ohne sie zu verstehen). Zum Beispiel hätten Nutzende (zumindest in der Vergangenheit) Facebook-Algorithmen nicht vertrauen sollen, seriöse Nachrichteninhalte anzuzeigen.

der KI für HR-Lösungen, mit „Wächter*innen als Zielgruppe der Kommunikation. Das Wissen der Designer*innen über den Algorithmus sollte diesen Instanzen in angemessener Weise mitgeteilt werden.
Der Grund, warum Transparenz besonders wichtig für prädiktive und präskriptive KI im HR-Bereich ist und weniger für bestehende Produkte, liegt im Unterschied zwischen KI und aktuellen Produkten. Erstens wird eine qualitativ minderwertige KI von KI-Nutzenden möglicherweise nicht sofort als qualitativ minderwertig erkannt. Die Ungenauigkeit von KI-getriebenen Entscheidungen im HR, die Verzerrungen und die unfairen Entscheidungen, die sie hervorrufen können, werden möglicherweise lange Zeit nicht erkannt. Sinnlose, schädliche Entscheidungen können getroffen werden, die den Mitarbeitenden schaden, bevor Fehler in der KI erkannt werden. Managemententscheidungen gelten dann als fair, wenn sie sich auf klare und konsequent angewandte Regeln stützen (51).
Bei KI-getriebener prädiktiver und insbesondere präskriptiver HR sind die Regeln in der HR- Analytics-Lösung eingebettet; sie können undurchsichtig eingebettet sein, z. B. wenn sie mathematisch zu komplex sind. Die Undurchsichtigkeit solcher Regeln kann das Vertrauen in das Management untergraben. Meta-analytische Studien im Bereich „Organizational Behavior“ haben gezeigt, dass die verschiedenen Teildimensionen organisatorischer Gerechtigkeit in einzigartiger und positiver Weise zu Leistung, Vertrauen, Arbeitszufriedenheit und organisatorischem Engagement beitragen und dass sie negativ mit Absichten, den Arbeitsplatz zu wechseln, sowie mit Fehlzeiten korrelieren (52, 53). Mechanismen zur Umsetzung von Transparenz und Rechenschaftspflicht in Bezug auf den Algorithmus können daher als Mittel angesehen werden, um reflexives Vertrauen in KI im HR-Bereich zu erreichen. In Übereinstimmung mit diesem Gedanken heißt es in einer Leitlinie:

> „Die Arbeitnehmer müssen das Recht haben, Transparenz bei den Entscheidungen/Ergebnissen von KI-Systemen und der zugrunde liegenden Algorithmen zu fordern […].“ (24)

Es ist jedoch klar, dass nicht alle Arbeitnehmerinnen und Arbeitnehmer Transparenz über einen Algorithmus in dem Sinne erhalten können, dass ihnen der Algorithmus direkt erklärt wird. Vielmehr kann der Algorithmus realistischerweise für Prüfer*innen und andere „Wächter*innen“ transparent gemacht werden. Prüfung und Zertifizierung können durch neue Rechtsinstrumente unterstützt werden, die Anreize für Unternehmen schaffen, transparent zu sein und Prüfer*innen ihre Arbeit tun zu lassen. Es ist noch eine offene Frage, ob Reputationsbedenken ausreichen, um die notwendigen

Anreize zur Verbesserung der Algorithmen zu schaffen, oder ob auch legislativer Druck erforderlich ist. Gegenwärtig ist zu hoffen, dass Reputationsbedenken einen Markt für Prüfer*innen und Zertifizierer*innen vertrauenswürdiger KI schaffen, und dass Reputationsbedenken die Hersteller von KI im Bereich HR Analytics veranlassen, sich um solche Formen der Transparenz zu bemühen. Transparenz ohne Rechenschaftspflicht führt nicht zu Verbesserungen. Wenn niemand verantwortlich ist, wenn KI-Systeme nicht so funktionieren, wie es vernünftigerweise erwartet wird, wird sich Transparenz in Bezug auf die Probleme der KI nicht in Prozessverbesserungen niederschlagen, es sei denn, die für das Problem und seine Behebung verantwortlichen Rollen sind identifizierbar.

Auf der anderen Seite können algorithmische Entscheidungen den Mitarbeitenden direkt erklärt werden. Diese Erklärung kann sich von der Erklärung des Algorithmus, d. h. seines Ziels und seiner allgemeinen Logik, unterscheiden. Tatsächlich befasst sich ein bedeutender Teil der Forschung über die Transparenz und Erklärbarkeit von Algorithmen mit Methoden, die individuelle von Algorithmen getroffene Entscheidungen leichter interpretierbar machen (45).

Diese Art der Erklärbarkeit ist für HR Analytics von zweifacher Relevanz. Erstens benötigen durchschnittliche Endnutzende von KI möglicherweise Erklärungen von KI-Entscheidungen, die einfacher sind als die, die von ganzheitlichen Transparenzmodellen geliefert werden. Ohne solche Erklärungen können sie möglicherweise kein (reflexives) Vertrauen in das Modell entwickeln (45). HR-Manager*innen, die KI-Empfehlungen umsetzen, könnten es beispielsweise für unverantwortlich halten, Mitarbeitende aufgrund von Empfehlungen auszuwählen, deren Logik sie nicht vollständig verstehen. Dies wird in den Leitlinien der UNI Global Union hervorgehoben, in denen es heißt:

> „Für die Nutzer ist Transparenz wichtig, weil sie Vertrauen in das System und Verständnis des Systems schafft, indem sie einen einfachen Weg für die Nutzer vorgibt, zu verstehen, was das System tut und warum." (24)

Zweitens können Mitarbeitende, die von Entscheidungen eines Algorithmus betroffen sind, die von Endnutzenden im HR verwendet wird, Erklärungen von den HR-Entscheidungsträger*innen verlangen. Wie Endnutzende benötigen betroffene Mitarbeitende Erklärungen, von denen man erwarten kann, dass sie für sie verständlich sind. Es ist bekannt, dass ein Verständnis der Logik von Entscheidungen, insbesondere von Entscheidungen, die mit Ungleichheit verbunden sind, die Wahrnehmung einer Entscheidung als ge-

recht und fair wahrscheinlicher macht (54). Ein weiterer Gedanke ist, dass wenn Menschen, die durch die Entscheidung eines Algorithmus benachteiligt sind, den Grund für die Entscheidung kennen, unternehmen sie möglicherweise angemessene Schritte, die zu einem anderen, für sie günstigeren Ergebnis führen könnten (46, 55).

Ein Ansatz, um dies zu erreichen, besteht darin, Algorithmen, die als Blackboxes gelten, zu vermeiden und stattdessen Modelle zu verwenden, die leicht interpretierbar sind. Ein weiterer Ansatz besteht darin, komplexere Modelle zu verwenden, die gemeinhin als nicht interpretierbar gelten (sogenannte Blackboxes), und diese interpretierbar zu machen.[33] Einige Modelle können so gestaltet werden, dass sie so einfach sind, dass die meisten Nutzenden intuitiv verstehen können, was sie tun. Diese Modelle sind nicht so präzise wie viele Algorithmen, die tendenziell Blackboxes sind. Daher besteht ein modisch gewordener Weg, Verständlichkeit zu erreichen, darin, einen einfachen Algorithmus zu bauen, der das Verhalten einer Blackbox nachbildet, zumindest für eine begrenzte Anzahl von Inputs, die für die Interessengruppen von Interesse sind (40).

Solche Algorithmen können nur eine ungefähre Erklärung dafür liefern, was die Vorhersagen oder Entscheidungen der KI antreibt. Durch eine solche Erklärung wird die tatsächliche innere Logik einer sehr komplexen Entscheidungsregel nicht erkennbar (45). Daher können solche Versuche, KI erklärbar zu machen, nicht die ganze Bandbreite des Verhaltens eines komplexen undurchschaubaren Algorithmus in allen möglichen betrieblichen Szenarien erklären.

Ein Ansatz dieser Art konzentriert sich auf die Quantifizierung des Einflusses verschiedener Inputs auf eine Entscheidung (55). Wenn z. B. einer Person eine Beförderung verweigert wird, kann ihr mitgeteilt werden, welche Faktoren die Entscheidung beeinflusst haben, wie sie sich (positiv oder negativ) ausgewirkt haben und was ihre jeweiligen Gewichtungen waren. Alternativ dazu liefern kontrafaktische Modelle eine Liste der wichtigsten Merkmale, die die Person hätte besitzen müssen, um das gewünschte Ergebnis zu erzielen (46). Eine kontrafaktische Erklärung kann z. B. lauten: „Sie hätten die Stelle erhalten, wenn Sie über bessere Englischkenntnisse und mindestens drei zusätzliche Jahre Erfahrung in Ihrer derzeitigen Funktion verfügt hätten." Die psychologischen Auswirkungen dieser und anderer Arten von

33 Der Unterschied besteht hier nicht darin, dass neuronale Netze Blackboxes sind und alles andere verständlich ist. Auch das Verhalten von linearen Modellen mit einer Vielzahl von interagierenden Faktoren kann sich dem intuitiven menschlichen Verständnis entziehen (40).

Erklärungen sind empirisch getestet worden, auch in einem fiktiven Szenario, in dem es um die Verwendung von Algorithmen zur Entscheidung über eine Beförderung ging (55). Wie in der psychologischen Literatur vorhergesagt, zeigte die Studie, dass wenn man den Menschen Erklärungen gibt, steigert dies im Allgemeinen die Wahrnehmung einer Entscheidung als gerecht. Die Studie wies jedoch nicht nach, dass Erklärungen, die den Einfluss von Inputdaten quantifizieren, oder kontrafaktische Erklärungen[34] erfolgreicher als andere Arten von Erklärungen sind, wenn es darum geht, die Wahrnehmung von Entscheidungen als fair zu steigern.

Plausibel und rational wäre es, einzelne Entscheidungen im Dienste mehrerer inhaltlicher Werte verständlich und transparent zu machen:

- Post-hoc-Erklärungen von algorithmisch getroffenen HR-Entscheidungen gegenüber Betreiber*innen von KI können die Vertrauenswürdigkeit der Technologie in realweltliche Anwendungen erhöhen. Dies wird erreicht, wenn Post-hoc-Erklärungen einzelner Entscheidungen Betreiber*innen von KI im HR-Bereich mit eingeschränkten Kenntnissen über die Technologie in die Lage versetzen, Fehler in der KI zu erkennen oder eindeutige Fälle von KI-Missbrauch zu vermeiden. Diese Informationen von Betreiber*innen von KI im HR-Bereich können es den Entwickler*innen ermöglichen, besser zu verstehen, wie sich die KI verhält und damit auch, wie (und warum und in welchen Szenarien) sie sich falsch verhalten kann. Wenn dies der Fall ist, sind gemäß den Grundsätzen der Benefizienz, Schadensverhütung und Gerechtigkeit Ad-hoc-Erklärungen erforderlich (z. B. wenn der Missbrauch eine unrechtmäßige Diskriminierung beinhaltet).
- Post-hoc-Erklärungen können die Freiheit von einzelnen Personen, die algorithmischen Entscheidungen unterworfen sind, stärken. Im HR-Kontext ist dies dann der Fall, wenn Post-hoc-Erklärungen von KI-getriebenen HR-Empfehlungen gegenüber den Mitarbeitenden es ihnen ermöglichen, ihr Verhalten so zu ändern, dass bessere Ergebnisse für sie erzielt werden. Dies wird durch den Grundsatz der Autonomie unterstützt.

Alle potenziellen ethischen Vorteile von Erklärbarkeit, Transparenz und Rechenschaftspflicht und deren Beziehung zu inhaltlichen ethischen Werten werden in Abbildung 2 hervorgehoben.

34 Von allen psychologischen Zwecken in dieser Studie ähnelt der Begriff „sensitivity-based explanations“ hinreichend dem Begriff „kontrafaktische Erklärungen“.

Abbildung 2

Potenzielle ethische Vorteile von Erklärbarkeit, Transparenz und Rechenschaftspflicht und deren Beziehung zu inhaltlichen ethischen Werten

4.2.4. Offene Herausforderungen für algorithmische Transparenz und Rechenschaftspflicht:

Es gibt eine mystische Vorstellung, dass neuronale Netze Entitäten sind, die sich der Kontrolle ihrer Designer*innen entziehen. Diese Mystik sollte zurückgewiesen werden. Designer*innen können den Algorithmus immer in dem Sinne kontrollieren, dass sie die Verantwortung für seine Ziele übernehmen, für die Art und Weise, wie die Ziele in eine mathematische Funktion übertragen werden, für die Qualität der ihm zugeführten Daten, für die Art und Weise, wie seine Leistung bewertet wird, für die Überwachung der Leistung im Live-Kontext und für die Empfehlung seiner Verwendung in diesem Kontext (41–44). Es gibt jedoch mehrere Herausforderungen, von denen wir drei untersuchen wollen.

Erste Herausforderung für die Transparenz: Manipulation

Viele Forderungen nach algorithmischer Transparenz ignorieren, dass in bestimmten Fällen transparente Algorithmen moralisch problematisch sein können. In einigen Fällen kann Transparenz in Bezug auf den Algorithmus oder seine Entscheidungen sensible private Informationen über einzelne Personen offenlegen. In anderen Fällen kann Transparenz dazu führen, dass der Algorithmus ausgespielt wird. Wenn z. B. der Algorithmus, der verwendet wird, um Steuerhinterziehung zu erkennen, gut bekannt ist, werden Personen, die die Zahlung von Steuern vermeiden wollen, Strategien optimieren, um nicht erwischt zu werden. Sogar im Kontext von HR Analytics könnte dies ein Problem darstellen, d. h. die Transparenz des Algorithmus kann Versuche ermutigen, den Algorithmus auszuspielen: Die Kenntnis der Proxys, die für Leistung verwendet werden, lenkt das Ziel der Aktivität vom ordnungsgemäßen Funktionieren zur Maximierung des Maßes, das als Proxy für Leistung verwendet wird, ab.[35] Nur wenige der hier untersuchten Leitlinien weisen auf dieses Risiko hin; die meisten Leitlinien schlagen algorithmische Transparenz vielmehr als ein uneingeschränktes Gut vor:

> „Can you provide for public auditing (i.e. probing, understanding, reviewing of system behavior) or is there sensitive information that would necessitate auditing by a designated 3rd party?"
> [Können Sie eine öffentliche Prüfung (d. h. Sondieren, Verstehen, Überprüfen des Systemverhaltens) vorsehen, oder gibt es sensible Informationen, die eine Prüfung durch einen dafür benannten Dritten erfordern würden?] (12)

35 Dies ist vergleichbar mit dem Problem der Studierenden, die lernen, um einen Test zu bestehen, oder der Wissenschaftler*innen die sich auf die Maximierung der Zitationen konzentrieren. Testergebnisse und Zitate sind bestenfalls ein unvollkommener Ersatz für das Lernen bzw. den Wert wissenschaftlicher Beiträge. Das Streben nach der Maximierung von Proxys lenkt die Energien der beteiligten Personen ab, um Ziele zu verfolgen, die strategisch wertvoll, aber aus der Sicht des zu fördernden Gutes suboptimal sind, wie z. B. Lernen oder wissenschaftliche Erkenntnisse. Neben dem Ausspielen des Systems (wenn der Indikator zum Ziel an sich wird [56, 57]) sind weitere Effekte im akademischen Bereich Risikovermeidung (hoch innovative oder interdisziplinäre Themen werden vermieden, weil sie nicht gut abschneiden [58]) und Aufgabenreduktion (Lehre und öffentliches Engagement werden vermieden, um sich auf publizierte Forschung zu konzentrieren [59]). Es sei auf den Fall der Wissenschaft in Italien verwiesen, wo der aggregierte Impact-Faktor trotz erheblicher Mittelkürzungen in den letzten zehn Jahren aufgrund von Praktiken, die darauf abzielen, die individuellen Impact-Faktoren der Forschenden zu steigern, kontraintuitiv zugenommen hat. Die Erklärung dafür ist, dass während die Fördermittel für Forschung gekürzt wurden, Maßnahmen wie Zitationsnachweise zu einer gesetzlichen Bedingung für die Genehmigung bestimmter Beförderungen gemacht wurden (60).

> „How will you facilitate public or third-party auditing without opening the system to unwarranted manipulation?“
> [Wie werden Sie die öffentliche Prüfung oder die Prüfung durch Dritte erleichtern, ohne das System für ungerechtfertigte Manipulationen zu öffnen?] (12)

Zweite Herausforderung für die Transparenz (und die Rechenschaftspflicht): Zuverlässigkeit und Wiederholbarkeit

Die zweite Herausforderung besteht darin, dass die ordnungsgemäße Prüfung der ethisch bedeutsamen Eigenschaften eines Algorithmus schwierig oder unmöglich sein kann, wenn sie aus verschiedenen technischen und sozialen Gründen nur unter *Laborbedingungen* durchgeführt werden kann. Es stimmt zwar, dass lernbasierte Modelle mit historischen Daten getestet und validiert werden, es gibt jedoch keine strenge Garantie dafür, dass das Modell sein angestrebtes Ziel erreicht, wenn es auf neue Fälle, in einem neuen Kontext und mit neuen Daten angewendet wird. Deshalb sollte ein System überwacht werden, wenn es in der realen Welt eingesetzt wird. Dies wird z. B. in den EU-Ethik-Leitlinien für eine vertrauenswürdige KI (29) anerkannt:

- „Haben Sie getestet, ob bestimmte Kontexte oder Bedingungen berücksichtigt werden müssen, damit die Wiederholbarkeit gewährleistet ist? […].“
- Haben Sie Verfahren eingeführt, die beschreiben, wann ein KI-System bei bestimmten Einstellungen ausfällt? […].“
- „Haben Sie das KI-System auf bestimmte Anwendergruppen oder problematische Anwendungsfälle getestet?“ (29)

Die methodische Herausforderung besteht darin, dass nur eine begrenzte Anzahl von Szenarien vorhergesagt und im Labor getestet werden kann. Es kann unvorhergesehene, kritische Szenarien geben, die nur in der Realität eintreten. Das bedeutet, dass Situationen entstehen können, in denen die KI in einer Weise reagiert, die nicht erwartet wird. Man kann nicht transparent sein und sicherstellen, dass sich das Modell in solchen Kontexten fair verhält. Es handelt sich um ein ingenieurtechnisches Problem der Identifizierung vernünftiger, sicherer, robuster und dennoch durchführbarer Testverfahren. Bei tiefen neuronalen Netzen sind Praktiken, die mit gewöhnlicher Software ausreichend robust sind, wie z. B. die Ausweitung der Testphase auf einige wenige Iterationen, in denen das Verhalten der Software in der Praxis bewertet wird, möglicherweise nicht ausreichend. Das Problem wird noch verstärkt, wenn sich das neuronale Netz auf der Grundlage der Ergebnisse der Entschei-

dungen, die es während des Betriebs trifft, selbst trainiert. Solche Probleme werden von einigen der Empfehlungen, die wir untersucht haben, nicht ignoriert. Zum Beispiel heißt es in einer Empfehlung des Netzwerks Women Leading in AI:

> „Introduce a regulatory approach governing the deployment of AI which mirrors that used for the pharmaceutical sector."
> [Führen Sie ein Regelwerk für den Einsatz von KI ein, das das Regelwerk für die Arzneimittelbranche widerspiegelt.] (11)[36]

Der Gedanke, der dem zugrunde liegt, lautet plausiblerweise: Wenn es das Ziel ist, zu wissen und zu dokumentieren, wie sich die AI verhält, reicht die im Labor durchgeführte Prüfung mit historischen Daten nicht aus. Im Hinblick auf die Datenpipeline bedeutet dies, die tatsächlichen Ergebnisse der Algorithmen (Entscheidungsregeln) zu überwachen und diese Informationen zur Verbesserung der KI-Design-Stufe zu nutzen. Die am weitesten verbreiteten Methoden für die Datenpipeline beinhalten bereits Iterationen (61). Es kann jedoch erforderlich sein, noch mehr Iterationen und eine kontinuierliche Überwachung der realweltlichen Ergebnisse zu verlangen, wenn bei einem datengetriebenen Modell viel auf dem Spiel steht und die Modelle Blackboxes sind.

Dritte Herausforderung für die Transparenz: Blackbox-Modelle

Die dritte Herausforderung betrifft die Erklärung von Blackbox-Modellen. Die oben genannten Technologien entfalten ihre ethisch vorteilhaften Folgen nur dann, wenn sich einige spekulative psychologische Hypothesen als zutreffend erweisen, nämlich, dass die Erklärung von KI-Entscheidungen das Bewusstsein der KI-Endnutzenden (die keine Datenwissenschaftler*innen sind) für die möglichen Fehler und den möglichen Missbrauch der KI schärft, und dass die Aufklärung der Betroffenen diese in die Lage versetzt, vernünftige Handlungsalternativen in ihrem Interesse zu finden. Darüber hinaus mag die Idee, eine Blackbox mit einem einfacheren Modell oder anhand einer kontrafaktischen Erklärung zu erklären, aufschlussreicher erscheinen als sie tatsächlich ist. Wenn sich ein Blackbox-Modell tatsächlich auf eine große Anzahl von Faktoren stützt, um eine Entscheidung zu treffen, kann die Reduzierung seiner Dimensionalität in der Tat einige willkürliche Entscheidungen

36 Analog dazu lautet eine Frage in der Bewertungsliste in den EU-Ethik-Leitlinien für eine vertrauenswürdige KI: „Haben Sie eine Strategie zur Überwachung und Erprobung eingeführt, die ihnen zu erkennen hilft, ob ihr KI-System die Ziele, Zwecke und vorgesehenen Anwendungen erfüllt?" (29).

erfordern. So kann es z. B. zehn Faktoren geben, die alle etwa die gleiche Gewichtung haben, und das Erklärungsmodell präsentiert nur vier, wobei diesen vier eine Bedeutung zugeschrieben wird, die sie nicht wirklich haben. Ein kontrafaktisches Modell kann eine Handlungsweise vorschlagen, die in der Tat nicht durchführbar ist, wie z. B. die Änderung des bisherigen (d. h. bereits ausgezahlten) Gehalts einer Person oder der von ihr in der Vergangenheit getroffenen Bildungsentscheidungen. Und wenn das Modell nur durchführbare Vorgehensweisen vorschlagen soll, wird es möglicherweise keine Handlungsweisen vorschlagen, die für die von den Entscheidungen betroffene Person wünschenswert sind (44).

4.2.5. Algorithmische Fairness und HR Analytics

Das Problem der Diskriminierung und Unfairness bei prädiktiven und präskriptiven Analysen ist eindeutig eine der wichtigsten Herausforderungen, wenn KI genutzt wird, Personalentscheidungen anzuleiten. Nirgendwo ist das Potenzial für einen Zusammenprall von Visionen stärker, nämlich von der statistischen Rechtfertigung der Entscheidungsfindung einerseits und der Fairness im Sinne der Werte, die die Arbeitsbeziehungen regeln, anderseits. Die Bedeutung dieses Themas für eine ethische KI zeigt sich auch darin, dass das Risiko von Unfairness und/oder Diskriminierung von der Mehrheit der hier untersuchten Ethik-Kodizes ebenfalls erwähnt wird. Wenn dieses Anliegen nicht explizit als Fairness oder Gleichbehandlung (und ihr Gegenteil, Diskriminierung) erwähnt wird, erscheint es manchmal in Form einer Diskussion von Verzerrungen und Fehlern, von denen bestimmte Gruppen überproportional betroffen sind.

Es hat den Anschein, als seien Empfehlungen bezüglich Fairness, Diskriminierung und Voreingenommenheit am unkompliziertesten. So heißt es z. B. in dem von der UNI Global Union vorgeschlagenen Grundsatz „Sicherstellung einer geschlechterneutralen, unvoreingenommenen KI“ (24):

> „Bei der Gestaltung und Wartung von KI ist es wichtig, dass das System im Hinblick auf negative oder schädliche menschliche Voreingenommenheit hin kontrolliert wird und dass jegliche Voreingenommenheit, sei es im Hinblick auf Geschlecht, Rasse, sexuelle Orientierung oder Alter, erkannt und nicht vom System verbreitet wird.“ (24)

Dies ist offensichtlich im Kontext von HR Analytics relevant, da ein Problem mit KI-Werkzeugen darin besteht, dass sie eine implizite Voreingenommenheit in den Entscheidungsfindungsprozess einführen können – d. h. eine Ten-

denz, Personen, die bestimmten Gruppen angehören, aufgrund statistisch signifikanter Ähnlichkeiten zwischen ihnen zu bevorzugen. Dies geschieht z. B. bei KI, die auf (statistischem) maschinellem Lernen basiert, das aus historischen Daten Ähnlichkeiten zwischen einzelnen Personen lernt. Informationen zu Rasse und Geschlecht müssen dem System maschinellen Lernens nicht explizit mitgeteilt werden, damit es z. B. eine Ähnlichkeit wahrnimmt und berücksichtigt, die Mitglieder derselben Rasse oder Geschlechtergruppe (überproportional) zusammenfasst (37). Daher sind Personengruppen, die in der Vergangenheit menschliche und strukturelle Voreingenommenheit erfahren haben – auch geschützte Personengruppen genannt – anfällig für Schäden durch ungenaue Vorhersagen oder Ressourcenzuweisungen, wodurch historische Ungleichheiten verstärkt werden.

Die KI hilft z. B. indem sie offene Stellen dem „richtigen" Typ Mensch in persönlichen Jobbörsen wie ZipRecruiter anbietet. Im Laufe der Zeit lernen die Jobbörsen die Präferenzen von Personalvermittler*innen für bestimmte Stellenbewerber*innen kennen, die zugunsten von Personen eines bestimmten Geschlechts, einer bestimmten Rasse oder einer bestimmten sozialen Schicht voreingenommen sein können (62, 63). Die unterschiedlichen Muttersprachen von Beschäftigten können zu ungerechtfertigten Ungleichheiten bei der Auswahl von Personen für eine bestimmte Bildungs- bzw. Ausbildungsmaßnahme oder eine bestimmte Stelle führen (64).

Die Empfehlungen zur Fairness umfassen die vier in der Einleitung beschriebenen grundlegenden Aktivitätsarten: 1) Unfairness dokumentieren, 2) Unfairness transparent machen, 3) die moralische oder rechtliche Verantwortung für Unfairness zuweisen, 4) Unfairness mildern. All diese Maßnahmen sind für HR relevant.

4.2.6. Offene Fragen für Fairness beim maschinellen Lernen

Leider besteht eine Kluft zwischen dem alltagssprachlichen Begriff der Fairness und den Versuchen einer statistischen Definition von Fairness. Darüber hinaus gibt es eine konzeptionelle Kluft zwischen fairer Vorhersage/Klassifikation und dem Rechtsbegriff der Nichtdiskriminierung. Dies macht es schwierig, Unfairness zu dokumentieren, sie transparent zu machen, sie zu mildern und Schuld oder moralische Verantwortung für Ergebnisse zuzuschreiben, die einigen Interessengruppen aus nachvollziehbaren Gründen unfair erscheinen mögen. All diese Aktivitäten setzen voraus, dass es ein objektives oder zumindest intersubjektives Maß für unfaire Voreingenommen-

heit oder unfaire indirekte Diskriminierung gibt. Das Problem ist außerdem nicht rein technischer Natur. Es gibt tiefe, konzeptionelle Fragen, die noch offen sind. Die wissenschaftliche Literatur hat gerade erst begonnen, die normative Bedeutung dieser statistischen Fragen aus moralischer und rechtlicher Sicht zu interpretieren.

Eine der wenigen Gedanken, bei denen Konsens herrscht, ist dass das, was als (ungerecht oder unrechtmäßig) diskriminierend oder unfair angesehen werden sollte, kontextabhängig ist. Dies spiegelt sich in der umsichtigeren Sprache wider, die sich in einigen bereits erwähnten Leitlinien findet. Die oben erwähnten Leitlinien des Weltwirtschaftsforums (21) erkennen ausdrücklich an, dass Fairness kontextabhängig ist, dass einige Formen der „Diskriminierung" der Aufgabe der KI immanent sind, und dass es verschiedene Formen der Fairness gibt. Sie erkennen auch an, dass je nach Kontext unterschiedliche Definitionen von Fairness angebracht sein können, und dass es zur Feststellung dessen, was in einem bestimmten Kontext fair ist, notwendig ist, Fachgebietsexpert*innen und interdisziplinäre Erkenntnisse einzubeziehen.

Ein relativ einfaches Element besteht darin, die ungleichen Fehlerquoten nach Personengruppen zu ermitteln, wobei verschiedene Arten von Fehlern berücksichtigt werden, insbesondere die Unterscheidung zwischen falsch-positiven und falsch-negativen Fehlern. Die Herausforderung besteht jedoch darin, eine Ungleichheit in solchen Fehlerquoten zu interpretieren, da es, wie weiter unten erläutert, kein Einzelmaß für die Unfairness von Ungleichheiten in Fehlerquoten gibt, und einem Algorithmus Voreingenommenheit zuzuschreiben ist weniger offensichtlich, als es auf den ersten Blick erscheinen mag.

Wir werden nun einige schwierige moralische Probleme der Fairness beim maschinellen Lernen anhand von zwei hypothetischen Beispielen veranschaulichen, die sich beide auf den HR-Kontext beziehen.[37] Angenommen, Sie trainieren ein HR-Analytics-Werkzeug, um Stellen für Programmierer*innen innerhalb Ihres Unternehmens auszuschreiben. Es handelt sich um ein riesiges multinationales Unternehmen mit Niederlassungen in vielen Ländern. Ihr Ziel ist es, das Stellenangebot Mitarbeitenden anzuzeigen, die wahrscheinlich Programmierer*innen sind, und es zu vermeiden, es Mitarbeitenden zu zeigen, die es nicht sind, so dass Letztere nicht durch die Informationen belastet werden und sich auf offene Stellen konzentrieren können, die für sie

37 Ein ähnliches Beispiel findet sich bei Gilbert (65). Der hier vorgestellte Fall ist dem Beispiel „kurzes Haar/langes Haar" in Zachary et al. (48) nachempfunden.

relevanter sind. Sie möchten ein Modell, das vorhersagt, wer sich für die Anzeige interessieren wird, indem es die Browsingverläufe der Mitarbeitenden analysiert, die sich bereit erklärt haben, diese persönlichen Daten ausschließlich zum Zweck einer solchen Verarbeitung an Sie weiterzugeben.

Ihr Modell behauptet, dass die Wahrscheinlichkeit, auf eine solche Stellenanzeige zu klicken, für eine Person, die stackexchange.com besucht hat, höher und für eine Person, die pinterest.com besucht hat, niedriger ist als der Durchschnitt. Der Grund dafür ist, dass in Ihren Trainingsdaten ein sehr großer Anteil der Personen, die stack-exchange.com besuchten, tatsächlich Computerprogrammierer*innen waren, die sich für solche Stellen interessierten und fähig waren, sie anzunehmen. Das ist nicht überraschend, wenn man bedenkt, dass diese Website hauptsächlich von Leuten besucht wird, die wissen, wie man programmiert, oder es gerade lernen. Andererseits ist die negative Gewichtung von pinterest.com, darauf zurückzuführen, dass die meisten Personen im Trainingsdatensatz, die auf diese Website zugriffen, Frauen waren, und die Frauen, die auf dieser Website landeten, mit großer Wahrscheinlichkeit auch keine Computerprogrammiererinnen waren.

Wenn man den Algorithmus verwendet, stellt sich heraus, dass 95 % der Mitarbeitenden, denen die Anzeige für eine offene Stelle als Computerprogrammierer*in angezeigt wurde, männlich waren. Sowohl in den Rechtswissenschaften als auch in der Philosophie ist dies als „indirekte Diskriminierung" bekannt („disparate impact" in der US-amerikanischen Rechtssprache). Indirekte Diskriminierung bedeutet, dass ein faktisch neutrales Kriterium unbeabsichtigt zu unterschiedlichen Ergebnissen führt, wenn es auf verschiedene Personengruppen angewandt wird.

In der Machine-Learning-Community ist man uneins darüber, wie solche Fälle zu behandeln sind. Wenn Fairness darin besteht, dass demographische Parität herrscht (66), ist es unfair, eine Regel anzuwenden, die die Wahrscheinlichkeit erhöht, dass die Anzeige einer Frau angezeigt wird. Denn Männer und Frauen sollten im Durchschnitt mit gleicher Wahrscheinlichkeit die Anzeige angezeigt bekommen, unabhängig von ihren tatsächlichen Fähigkeiten und Ihrem Interesse an solchen Anzeigen. Gemäß einem anderen Fairnesskriterium, das als „equalized odds" (67) oder „equal mistreatment" (39) bezeichnet wird, ist es unfair, eine Regel anzuwenden, die die Wahrscheinlichkeit, dass die Anzeige angezeigt wird, erhöht, *wenn die Person tatsächlich an der Stelle interessiert ist*, wenn es sich bei der Person um eine Frau handelt. Eine mögliche Form des Audits auf der Grundlage von „equalized odds" müsste zunächst herausfinden, ob die Mitarbeitenden tatsächlich Interesse an die betreffende Stelle haben und dafür qualifiziert sind, und

dann feststellen, ob diejenigen, die Interesse haben, mit gleicher Wahrscheinlichkeit die Anzeige angezeigt bekommen, unabhängig von ihrem Geschlecht.

Man kann sich Argumente für beide Fairnesskriterien vorstellen. Zugunsten von „equalized odds" könnte man sagen, dass es nicht unfair ist, wenn Mitarbeitende, die nicht einschlägig qualifiziert und nicht an einer solchen Stelle interessiert sind, die Anzeige nicht angezeigt bekommen. Wenn diese Mitarbeitenden überproportional einem bestimmten Geschlecht angehören, ist eine Abweichung vom Grundsatz der demographischen Parität dergestalt, dass die Anzeige vorwiegend Angehörigen des anderen Geschlechts angezeigt wird, gerechtfertigt. Gegen „equalized odds" und zugunsten der statistischen Parität kann man argumentieren, dass dies eine Möglichkeit ist, den *Status quo* zu reproduzieren. Es ist wichtig, Frauen (auch Frauen, die keine Programmiererinnen sind und sich nicht auf solche Stellen bewerben wollen) zu zeigen, wie viele gute Chancen es in diesem Bereich gibt. Dies kann ein Anreiz für Frauen sein, in ihrer Freizeit Programmieren zu lernen, oder um sicherzustellen, dass ihren Töchtern eine faire Chance geboten wird, Programmieren zu lernen.

Eine weitere Kritik an „equalized odds" (dem Ausgleich der Falsch-positiv- und Falsch-negativ-Raten[38]) folgt aus Argumenten, die für ein Kriterium der *prädiktiven* Parität, den Ausgleich der *Spezifizität*, sprechen. Wie „equalized odds" ist auch die prädiktive Parität damit vereinbar, dass die Wahrscheinlichkeit, eine Stelle angeboten zu bekommen (oder eine Stellenanzeige angezeigt zu bekommen), von Gruppe zu Gruppe unterschiedlich ist, z. B. insofern, als eine Gruppe mehr Mitglieder enthält, die für diese Art von Stelle qualifiziert sind oder sich dafür interessieren (66, 68, 69).

Prädiktive Parität misst den statistischen Tendenzen, die mit der Gruppenzugehörigkeit verbunden sind, jedoch noch mehr Bedeutung bei als „equalized odds". Tatsächlich ist es möglich, dass eine Entscheidungsregel das Kriterium der prädiktiven Parität erfüllt und dennoch zwei Personen, die für eine Stelle gleich gut geeignet sind, unterschiedliche Chancen haben, eine Stellenanzeige angezeigt zu bekommen, wenn die eine zu einer Gruppe gehört, die statistisch gesehen häufiger die erforderlichen Merkmale aufweist, und die andere zu einer Gruppe, die statistisch gesehen weniger häufig die erforderlichen Merkmale aufweist (68, 69).

38 Dies impliziert, dass die Empfindlichkeit eines Klassifikators für beide Gruppen gleich ist.

Einige Statistiker*innen behaupten, dass das angemessene Fairness-Kriterium prädiktive Parität und nicht „equalized odds" sei. Prädiktive Parität erfordert, dass, wenn eine Organisation eine HR-Entscheidung trifft, die auf einer Vorhersage beruht (z. B. die Entscheidung, einer Mitarbeiterin oder einem Mitarbeiter eine offene Stelle in der Organisation anzuzeigen, basierend auf der Vorhersage, dass die Person an dieser offenen Stelle interessiert sein und auf die Anzeige klicken wird), derselbe Anteil von positiven Vorhersagen sich als richtig erweisen sollte, unabhängig von der Gruppe, zu der die Person gehört.

So ist z. B. der Algorithmus zur Anzeige von offenen Stellen für Programmierer*innen im Unternehmen auch dann fair, wenn Männer die Anzeige mit größerer Wahrscheinlichkeit sehen als Frauen, solange der Anteil der Frauen (denen die Anzeige angezeigt wird), die auf die Anzeige klicken, (ungefähr) gleich groß ist wie der Anteil der Männer (denen die Anzeige angezeigt wird), die auf die Anzeige klicken. Obwohl dieses Kriterium scheinbar dem Fairnesskriterium „equalized odds" entspricht, ist es in der Tat mathematisch in allen außer seltenen Fällen damit unvereinbar (70).[39]

Ein Argument für die prädiktive Wertparität (und damit in den meisten Fällen gegen das Kriterium der „equalized odds") ist, dass es das ist, was eine Arbeitgeberin bzw. ein Arbeitgeber tun würde, wenn sie bzw. er den Beitrag aller potenziellen Mitarbeitenden zum Unternehmen gleich bewertet unabhängig von den Gruppen, denen sie angehören. Das nachfolgende hypothetische Szenario zeigt, warum dies zu einer Verletzung des Kriteriums „equalized odds" (d. h. zu ungleichen Falsch-positiv- und Falsch-negativ-Raten) führen kann: Angenommen, Sie trainieren einen Algorithmus des maschinellen Lernens, um ungeeignete Bewerbende aus einem riesigen Stapel von Lebensläufen potenzieller Kandidat*innen für eine Stelle als Schulbusfahrer*in auszusondern. Sie wollen für diese Stelle niemanden einstellen, dem

39 Der Unterschied beläuft sich auf Folgendes: Während prädiktive Parität nur dann erreicht wird, wenn Frauen und Männer, denen die Anzeige tatsächlich angezeigt wird, mit derselben Wahrscheinlichkeit auf die Anzeige klicken, setzt das Kriterium „equalized odds" voraus, dass die Frauen und Männer, die (auf Nachfrage) Interesse an solchen Anzeigen zeigen (z. B. durch Anklicken), mit derselben Wahrscheinlichkeit auf die Anzeige klicken (68). Im vorliegenden Beispiel können „equalized odds" und prädiktive Parität nur dann gleichzeitig erreicht werden, wenn mindestens eine der folgenden (unwahrscheinlichen) Bedingungen erfüllt ist: (a) der Anteil der Personen, die sich tatsächlich für die Anzeige interessieren, ist in beiden Gruppen genau gleich (d. h. genau der gleiche Anteil weiblicher und männlicher Mitarbeitende würde auf die Anzeige klicken, wenn sie angezeigt würde) oder (b) es ist möglich, mit vollkommener Genauigkeit (100 % korrekte Vorhersagen) vorherzusagen, ob ein Mitarbeiter bzw. eine Mitarbeiterin auf die Anzeige klicken wird.

man nicht trauen kann, und deshalb wollen Sie diejenigen Fahrer*innen ausschließen, die eher betrunken Auto fahren. Sie haben Zugang zu Daten über Fahrer*innen, gegen die eine Geldstrafe verhängt wurde oder denen der Führerschein entzogen wurde, weil sie wegen Trunkenheit am Steuer erwischt wurden. Sie haben auch Zugriff auf die Browser-Verläufe dieser Internet-Nutzenden. Ihre Trainingsdaten beziehen sich auf verschiedene Personengruppen und enthalten ebenso viele Daten aus Personengruppen, die überwiegend muslimischen Glaubens sind, wie aus Personengruppen, die überwiegend christlichen Glaubens sind. Es stellt sich heraus, dass die höchste Genauigkeit durch einen Algorithmus erreicht wird, der berücksichtigt, ob eine Person eine alkoholbezogene Website (z. B. die Website eines auf alkoholische Getränke spezialisierten Händlers oder eine Website für Wein-Ratings) besucht hat. Wir gehen ferner davon aus, dass es keine Verzerrungen in den Daten aufgrund der Praxis des Anhaltens von Fahrer*innen für Alkoholkontrollen gibt – d. h. Fahrer*innen wurden nicht aufgrund ihrer wahrgenommenen Religion oder anderer Merkmale, die statistisch mit der Religion korreliert sind, bevorzugt angehalten.

Wir können erklären, warum prädiktive Parität den voraussichtlichen Gewinn für die Arbeitgeberin bzw. den Arbeitgeber unabhängig von der Gruppenzugehörigkeit ausgleicht, indem wir untersuchen, wie eine auf Vorhersagen basierende Regel Personen behandelt, deren tatsächliches Label „Trunkenheit am Steuer" lautet und bekannt ist. Prädiktive Parität zwischen Personen christlichen und muslimischen Glaubens wird erreicht, wenn bei Betrachtung der Profile der Kandidat*innen, die die *Entscheidungsregel* ausgeschlossen hat, der Anteil der Fahrer*innen, die betrunken am Steuer erwischt wurden, bei Personen christlichen und muslimischen Glaubens gleich ist, und wenn bei Betrachtung der Profile der Kandidat*innen, die die *Entscheidungsregel* nicht ausgeschlossen hat, der Anteil der Fahrer*innen, die tatsächlich betrunken am Steuer erwischt wurden, für Personen christlichen und muslimischen Glaubens gleich ist.

Nehmen wir an, dass Personen christlichen und muslimischen Glaubens, die alkoholbezogene Websites besuchen, im Durchschnitt mit gleicher Wahrscheinlichkeit letztendlich betrunken fahren, und dass Christ*innen und Muslim*innen, die keine alkoholbezogenen Websites besuchen, im Durchschnitt mit gleicher Wahrscheinlichkeit beim Fahren nüchtern bleiben. Die Entscheidungsregel bringt für das Unternehmen den gleichen Nutzen (den Nutzen der Auswahl einer Fahrerin bzw. eines Fahrers, der nicht ungeeignet ist), unabhängig von der Religion des Bewerbenden: Ungeachtet dessen, ob das Unternehmen aufgrund der Empfehlung des Werkzeugs eine Person

christlichen oder muslimischen Glaubens eingestellt hat, es wird mit gleicher Wahrscheinlichkeit eine Person eingestellt haben, die betrunken fahren wird, unabhängig von der Religion, der diese Person angehört.

Aber während dies erreicht wird, können keine *„equalized odds"* erreicht werden. Das heißt, es wird ungleiche Falsch-positiv- und Falsch-negativ-Raten für Personen christlichen und muslimischen Glaubens geben, wenn es (wie es plausibel erscheint) mehr Christ*innen im Vergleich zu Muslim*innen unter den Besucher*innen alkoholabhängiger Websites gibt. Personen christlichen Glaubens, die nicht betrunken Auto fahren, werden mit größerer Wahrscheinlichkeit ausgeschlossen als Personen muslimischen Glaubens, die nicht betrunken Auto fahren, da sie mit größerer Wahrscheinlichkeit alkoholbezogene Websites besucht haben.[40]

Daher ist es in Situationen wie dieser nicht möglich, sowohl das Kriterium der prädiktiven Parität als auch das Kriterium der „equalized odds" zu erfüllen. Die Wahl zwischen den statistischen Kriterien der prädiktiven Wertparität und der „equalized odds" ist eine *moralische* und keine *technische* Wahl. Da es sich um eine *Wertefrage* handelt, gibt es keine offensichtliche Antwort auf diese Frage von den Expert*innen für Fairness im maschinellen Lernen (d. h. basierend auf ihrer Autorität qua Expert*innen auf diesem Gebiet oder auf Statistiken). Möglicherweise gibt es nicht die eine richtige

40 Diese Ungleichheit ist bei einer Regel möglich, die bei der Beurteilung von Personen muslimischen und Personen christlichen Glaubens gleichermaßen zutreffend ist. Wenn es z. B. nur drei Muslim*innen gibt, die eine solche Website besuchen, und 3.000.000 Christ*innen, wird eine Regel mit 2/3 Genauigkeit (für die positiven) fälschlicherweise Trunkenheit am Steuer für eine Person muslimischen Glaubens und 1.000.000 Personen christlichen Glaubens vorhersagen. Wenn es nur drei Personen christlichen Glaubens gibt, die keine alkoholbezogenen Websites besuchen, und 3.000.000 Personen muslimischen Glaubens, die das nicht tun, wird eine Regel mit einer Genauigkeit von 2/3 (für die negativen) für 1.000.000 Personen muslimischen Glaubens und eine Person christlichen Glaubens fälschlicherweise voraussagen, dass sie nicht betrunken fahren.

Diese Regel kann als voreingenommen zugunsten der Muslime angesehen werden, da sie fälschlicherweise (der Einfachheit halber) 1.000.000 Personen christlichen Glaubens und ein Person muslimischen Glaubens als „alkoholisierte Fahrer*innen" klassifiziert, während sie fälschlicherweise 1.000.000 Muslim*innen und nur eine Person christlichen Glaubens als „nicht alkoholisierte Fahrer*innen" klassifiziert. Diese Ungleichheiten in der Fehlklassifikationsrate (sowohl bei den positiven als auch bei den negativen) können plausibel als vorteilhaft für Personen muslimischen Glaubens und nachteilig für Personen christlichen Glaubens angesehen werden.

Ein konzeptionell ähnlicher Fall von „unequal mistreatment" (in diesem Fall von Weißen vs. Schwarzen) wurde von ProPublica in Bezug auf den COMPAS-Algorithmus zur Vorhersage der Rückfallwahrscheinlichkeit von Straftätern festgestellt (71). Das Unternehmen verteidigte das Werkzeug, indem es zeigte, dass es trotz der sehr deutlichen „unequal odds" für Weiße und Schwarze fast perfekte prädiktive Parität erreicht hat (69).

Wahl, sondern unterschiedliche statistische Bedingungen werden in unterschiedlichen Szenarien angemessen sein. Philosoph*innen und Theoretiker*innen des maschinellen Lernens diskutieren noch darüber, wie solche Entscheidungen nachvollzogen werden können (72, 73).

Die tiefste Grundlage der Diskussion ist letztlich nicht nur technischer Natur (obwohl die Diskussion den technischen Begriff der bedingten Wahrscheinlichkeiten betrifft), sondern auch ethischer Natur: Es geht um die am besten geeignete Interpretation des Gedankens, „Menschen gleich zu behandeln, wenn sie in der relevanten Hinsicht gleich sind". (Zum Beispiel: Sollten Menschen, die dieselben Ergebnisse erzielen, dieselben Chancen erhalten, auch wenn die Wahrscheinlichkeit, dass sie das relevante Ergebnis erreichen, unterschiedlich ist? Sollte unsere Entscheidung über einzelne Personen von den Informationen abhängen, die wir über sie sammeln können, selbst wenn die fraglichen Informationen sie nicht als einzelne Personen betreffen, sondern nur über ihre Ähnlichkeit mit anderen Personen Auskunft geben).

Aus diesem Grund sind – trotz einer letztlich etwas vereinfachten Erklärung, dass ein Algorithmus von der Presse als „diskriminierend" bezeichnet wurde – die meisten Behauptungen über *Voreingenommenheit* und *Diskriminierung* natürlich kontrovers, denn sie beruhen auf moralischen Annahmen darüber, was bei einer von der Wahrscheinlichkeit geleiteten Entscheidung fair ist (und was nicht), die weiter diskutiert werden sollte. Die wissenschaftliche Debatte darüber, wie man von den Merkmalen einer Situation (die moralisch relevant sind) auf die Wahl eines angemessenen statistischen Constraints schließen kann, ist gerade erst im Entstehen begriffen (72).

4.3. Die Auswirkungen auf die Menschen kennen, kommunizieren, anerkennen und verbessern

4.3.1. Ziele und Bestrebungen der Generierung von Wissen über die Auswirkungen von Algorithmen auf die Menschen

Die obige Diskussion (Abschnitt 4.2.4) zeigt, dass nicht alle moralisch relevanten Features von Algorithmen (d. h. Entscheidungsregeln) durch die Bewertung und das Testen von Algorithmen im Labor ermittelt werden können. Ein Algorithmus kann mit gültigen und unvoreingenommenen Daten trainiert worden sein, die für alle Personengruppen gleichermaßen repräsentativ sind; seine Präzision kann gemessen worden sein und angesichts der Aufgabe, für die er verwendet wird, völlig zufriedenstellend sein; alle für die

Fairness relevanten Kennzahlen können gemessen und der Algorithmus kann geändert worden sein, um eine Fehlerverteilung zwischen den verschiedenen Personengruppen zu erreichen, die die meisten Menschen für den fraglichen Kontext als fair ansehen würden.

Und doch beruhen all diese Bewertungen und Messungen auf Daten aus der Vergangenheit. Wenn angemessene Trainings- und Testdaten ausgewählt worden sind, sollte das zukünftige Verhalten der Entscheidungsregel dem im Labor gezeigten Verhalten recht ähnlich sein. Man kann jedoch nicht einfach sicher sein, dass dies der Fall sein wird. Die Bewertung auf der Grundlage von Daten aus der Vergangenheit reicht möglicherweise nicht aus, um die Leistung einer Entscheidungsregel (die von einem Lernalgorithmus erzeugt wurde) zu bewerten, wenn sie auf die neuen Mitarbeitenden in der realen Welt angewendet wird. Dieses Problem wird typischerweise zum Ausdruck gebracht, indem man feststellt, dass die KI nicht *robust* ist – dass sie nicht angemessen *verallgemeinert.*

Diese Ungewissheit in Bezug auf die Zukunft ist nicht nur darauf zurückzuführen, dass das Verhalten von Blackbox-KI-Systemen, wie z. B. neuronalen Netzen, für alle möglichen Inputs im Grunde genommen unbekannt ist. Diese Ungewissheit ist auch nicht auf diejenigen KI-Anwendungen beschränkt, die sich während ihres Einsatzes weiter entwickeln, auch wenn diese tendenziell am wenigsten vorhersagbar sind, da sie ihre Regeln, je nachdem, wie die Realität auf sie reagiert, überarbeiten. Die Unvorhersagbarkeit aller KI-Anwendungen ergibt sich aus einem viel umfassenderen Phänomen, nämlich ihrer Einbettung in einem menschlichen Kontext, der seinem Wesen nach immer etwas unberechenbar ist, und zwar auf eine Weise, die sich in den Trainings- und Testdaten nicht widerspiegelt. Dies ist natürlich kein spezifisches Problem der KI, sondern vielmehr ein potenzielles Problem für alle Technologien. Aber im Falle von KI, und insbesondere von KI am Arbeitsplatz, erscheinen die Risiken, die mit negativen Auswirkungen auf die Menschen verbunden sind, die im Labor nicht vorhersehbar sind, besonders schwerwiegend.

Wenn man die Auswirkungen der KI auf die Gesellschaft betrachtet, kann man mindestens zwei verschiedene Ebenen unterscheiden: die Auswirkungen auf einzelne Personen (z. B.: Wie falsch war die Vorhersage über die Leistung eines einzelnen Mitarbeiters?) und auf Gruppen (z. B.: Erschwert die *Entscheidungsregel* die Einstellung von Schwarzen?). Darüber hinaus kann man zwei Arten von Einzelpersonen und Gruppen unterscheiden: erstens diejenigen Einzelpersonen und Gruppen, die eine vertragliche und wirtschaftliche Beziehung haben bzw. eingehen wollen mit der Organisation, die

die KI im HR-Bereich einsetzt (insbesondere, aber nicht ausschließlich, Bewerbende und Mitarbeitende); zweitens alle anderen Einzelpersonen und Gruppen. In diesem Abschnitt interessieren wir uns für die Auswirkungen *auf einzelne Personen, die vertragliche und wirtschaftliche Beziehungen haben bzw. eingehen wollen mit der Organisation*, die KI für HR Analytics einsetzt.

Mehrere Leitlinien empfehlen, Wissen über die tatsächlichen Auswirkungen von Algorithmen auf die Menschen zu generieren. Das philosophische und rechtlich-normative Konzept, das für eine solche Leitlinie verwendet wird, ist nicht immer in allen Leitlinien gleich. Zum Beispiel diskutieren einige Leitlinien die Auswirkungen auf die Menschen in Bezug auf die *grundlegenden Menschenrechte* (16); andere verwenden die Begriffe *Risikoprävention und -minderung* („risk prevention and mitigation"; 14), *„ethics by design"* bzw. *Ethik in der Konzeptionsphase* (16), *Pflicht zur Wahrung der öffentlichen Sicherheit* („public safety obligation"; 15) oder *verantwortungsvoller Einsatz* („responsible deployment"; 28). In gewisser Weise ist dies der am wenigsten konkrete Aspekt der Leitlinien, der bisher untersucht wurde. Viele Leitlinien betonen, wie wichtig es ist, die Implementierung von KI-Anwendungen in Organisationen zu überwachen, doch bieten sie nur sehr wenige konkrete organisatorische Lösungen zur Umsetzung dieser Anforderung. Es überwiegen Grundsatzerklärungen ohne konkrete Anleitungen. Nachfolgend einige Beispiele:

> „We propose that companies work on concrete ways to enhance company governance, establishing or augmenting existing mechanisms and models for ethical compliance."
> [Wir schlagen vor, dass Unternehmen an konkreten Möglichkeiten zur Verbesserung der Unternehmenssteuerung arbeiten, indem sie bestehende Mechanismen und Modelle für die Einhaltung von ethischen Standards einführen oder erweitern.] (21)

> „Im Rahmen des Gesamtansatzes für ‚ethics by design' (‚Ethik bereits in der Konzeptionsphase") sollten Systeme mit künstlicher Intelligenz […] verantwortungsbewusst konzipiert und entwickelt werden, insbesondere durch folgende Maßnahmen: […] b. die Bewertung und Dokumentation der erwarteten Auswirkungen auf Einzelpersonen und auf die Gesellschaft zu Beginn eines Projekts zur künstlichen Intelligenz und zu relevanten Entwicklungen während seines gesamten Lebenszyklus […]." (16)

> „Institutions must assess the public safety risks that arise from the deployment of AI systems that direct or control physical devices."
> [Institutionen müssen die Risiken für die öffentliche Sicherheit bewerten, die durch den Einsatz von KI-Systemen entstehen, die physische Geräte steuern oder kontrollieren.] (15)

„The capacity of an AI agent to act autonomously and to adapt its behavior over time without human direction calls for […] ongoing monitoring."
[Die Fähigkeit eines KI-Agenten, autonom zu handeln und sein Verhalten im Laufe der Zeit ohne menschliche Anleitung anzupassen, erfordert […] eine laufende Überwachung.] (28)

„Organizations should adopt and maintain policies and procedures reasonably designed to collect information sufficient to conduct assessments that would detect any significant disparate impacts, including, if necessary, collecting sensitive information such as race, gender, ethnicity, and religion or constructing accurate proxies for such sensitive information."
[Organisationen sollten Richtlinien und Verfahren einführen und aufrechterhalten, die vernünftigerweise dazu bestimmt sind, Daten zu sammeln, die ausreichen, um Bewertungen durchzuführen, mit denen bedeutende indirekte Diskriminierungen aufgedeckt werden können, ggf. einschließlich der Sammlung sensibler Daten wie Rasse, Geschlecht, ethnische Zugehörigkeit und Religion oder der Konstruktion genauer Proxy-Variablen für solche sensiblen Informationen.] (26)

„A/IS should prioritize human well-being as an outcome in all system designs, using the best available, and widely accepted well-being metrics as their reference point."
[AIS (autonome intelligente Systeme) sollten bei allen Systementwürfen dem menschlichen Wohlbefinden als Ergebnis Priorität einräumen und die besten verfügbaren und weithin akzeptierten Wohlbefindenskennzahlen als Referenz verwenden.] (14)

4.3.2. Empfehlungen

Abgesehen von allgemeinen und vagen Grundsatzerklärungen (und ebenso allgemeinen Bewertungsforderungen im Zusammenhang mit denselben Grundsätzen) sind einige konkrete Empfehlungen, die in den 20 hier untersuchten Leitlinien zu Tage treten, erwägenswert:

Erstens *Kompetenzangleichung:* Es sollte eine Angleichung von Fähigkeiten und Kompetenzen stattfinden zwischen den Funktionsträger*innen in einer Organisation, die KI-Anwendungen einsetzen, und denjenigen, die für die Entwicklung und Erprobung dieser Anwendungen verantwortlich sind. Die erste Fassung der IEEE-Leitlinien zum ethisch ausgerichteten Design (74) enthielt eine interessante Empfehlung sowohl für die Entwickler*innen als auch für die Nutzenden von KI-Produkten, nämlich: Die Entwickler*innen von KI-Produkten müssten das Niveau des Hintergrundwissens und der Fertigkeiten angeben, die KI-Bediener*innen benötigen, um KI-Produkte sicher

zu bedienen; und die Organisationen sollten sicherstellen, dass die Bediener*innen über erforderliche Kompetenzen verfügen. Im Kontext von HR Analytics bedeutet dies, dass nur die HR-Fachkräfte KI-basierte Instrumente verwenden sollten, die die Logik der Entscheidungsregel und ihre potenziellen Fehler in einem erforderlichen Maß verstehen. Man könnte sich vorstellen, dass KI-Produkte in HR Analytics immer mit einem „Ethik-Handbuch" ausgeliefert werden, in dem erklärt wird: 1) was die Grenzen der Modelle sind, einschließlich der Umstände, die zu Fehlern des Modells führen können; 2) wie die Entscheidungen des Modells erklärt werden sollten, insbesondere wenn sie angefochten werden; 3) was getan wurde, um das Modell fair zu gestalten; 4) welche potenziellen Probleme zu erwarten sind; 5) wie sie überwacht und bewertet werden können; 6) was unternommen werden muss, um die Software an die sozialen Umstände anzupassen (und Voreingenommenheit zu vermeiden); und 7) wie Probleme kommuniziert werden können. Wie die FAT-ML-Empfehlung hervorhebt, besteht ein wichtiger Aspekt dieser Kompetenzangleichung darin, dass Datenwissenschaftler*innen bestimmen sollten, wie die Unsicherheit/Fehlermarge für (KI-getriebene) Entscheidungen zu kommunizieren sei (12).

Die zweite erwägenswerte Empfehlung ist eine konkrete Umsetzungsleitlinie, die das Verfahren der *Bereitstellung eines Feedback-Mechanismus* betrifft, das oft einem Recht gleichkommt, *algorithmische Entscheidungen*, insbesondere solche, die vollständig automatisiert sind, *anzufechten oder zu korrigieren*. Dieser Gedanke findet sich in mehr als einer Leitlinie wieder:

> „5. […] die Gewährleistung des Rechts der Einzelnen, dass keine Entscheidung über sie getroffen wird, die allein auf einer automatisierten Verarbeitung beruht, wenn sie dadurch erheblich beeinträchtigt sein würden, und, wenn das nicht möglich ist, durch die Gewährleistung des Rechts der Einzelnen auf die Anfechtung einer solchen Entscheidung." (16)

> „Develop a process by which people can correct errors in input data, training data, or in output decisions."
> [Entwickeln Sie ein Verfahren, mit dem Menschen Fehler in Inputdaten, Trainingsdaten oder Output-Entscheidungen korrigieren können.] (12)[41]

41 Ähnliche Punkte erscheinen in der Bewertungsliste in den EU-Ethik-Leitlinien für eine vertrauenswürdigen KI (29): „Haben Sie abhängig vom Anwendungsfall einen Mechanismus vorgesehen, der es anderen ermöglicht, Probleme im Zusammenhang mit Verzerrungen, Diskriminierung oder schlechter Leistung des KI-Systems zu kennzeichnen? Haben Sie klare Schritte und Kommunikationswege darüber in Betracht gezogen, wie und an wen solche Themen herangetragen werden sollten? Haben Sie außer den (End-)Nutzern auch weitere, möglicherweise indirekt vom KI-System Betroffene berücksichtigt?"

Drittens wird in einigen Leitlinien empfohlen, dass Organisationen, die eine KI-getriebene Lösung umsetzen, über *Abhilfeverfahren* verfügen sollten, um Fälle zu behandeln, in denen im Laufe der Zeit durch die Anwendung der KI erheblicher Schaden oder Unfairness entsteht:

> „Access to Redress: Leaders, designers and developers of ML systems are responsible for identifying the potential human rights impacts of their systems. They must make visible avenues for redress for those affected by disparate impacts, and establish processes for the timely redress of any discriminatory outputs."
> [Zugang zu Abhilfe: Führungskräfte sowie Designer*innen und Entwickler*innen von ML-Systemen (Systemen des maschinellen Lernens) sind dafür verantwortlich, die möglichen negativen Auswirkungen ihrer Systeme auf die Menschenrechte zu erkennen. Sie müssen für diejenigen, die von indirekten Diskriminierungen betroffen sind, sichtbare Möglichkeiten der Abhilfe aufzeigen und Verfahren für die rechtzeitige Abhilfe bei diskriminierenden Outputs einführen.] (21)

Viele Leitlinien fordern zudem, dass Organisationen, die KI entwickeln oder KI-gesteuerte Lösungen implementieren (oder Organisationen, die beides tun), sich der Auswirkungen von KI und KI-gesteuerten Lösungen auf die Gesellschaft als Ganzes und nicht nur auf ihre Kund*innen bewusst sind. In den hier untersuchten Dokumenten scheint in Bezug auf solche „gesamtgesellschaftlichen" Fragen eine konkrete Anleitung jedoch völlig zu fehlen:

> „KI soll in erster Linie den Menschen und dem Planeten dienen" (24)

> „Teilen der Vorteile von KI-Systemen" (24)

> „Sicherung eines gerechten Übergangs und Unterstützung für grundlegende Freiheiten und Rechte" (24)

> „Verbot eines Wettrüstens mit KI-Waffen" (24)

> „Responsible Design and Deployment: We recognize our responsibility to integrate principles into the design of AI technologies, beyond compliance with existing laws. […] As an industry, it is our responsibility to recognize potentials for use and misuse, the implications of such actions, and the responsibility and opportunity to take steps to avoid the reasonably predictable misuse of this technology by committing to ethics by design."
> [Verantwortungsbewusster Entwurf und Einsatz: Wir erkennen unsere Verantwortung an, über die Einhaltung bestehender Gesetze hinaus Grundsätze in den Entwurf von KI-Technologien zu integrieren. […] Als Industrie ist es unsere Verantwortung, Potenziale für Nutzung und Missbrauch, die Auswir-

> kungen solcher Maßnahmen sowie die Verantwortung und die Möglichkeit zu erkennen, Maßnahmen zu ergreifen, um den vernünftigerweise vorhersehbaren Missbrauch dieser Technologie zu vermeiden, indem wir uns dem Grundsatz der konzeptuell integrierten Ethik verpflichten.] (25)

> „We will seek to ensure that AI technologies benefit and empower as many people as possible."

> [Wir werden versuchen sicherzustellen, dass KI-Technologien so vielen Menschen wie möglich zugute kommen und befähigen.] (13)

> „Societal and Organizational Impact: The AIA [algorithmic impact assessment] needs to highlight the impact on the workforce as well as society/community as a whole. For example, it needs to demonstrate how the system augments human capabilities and how the algorithm does not become policy, thus removing human autonomy in wider decision-making."
> [Gesellschaftliche und organisatorische Auswirkungen: Die Bewertung der Auswirkungen des Algorithmus muss die Auswirkungen auf die Mitarbeitenden sowie auf die Gesellschaft/Gemeinschaft als Ganzes hervorheben. Zum Beispiel muss sie zeigen, wie das System die menschlichen Fähigkeiten steigert und wie der Algorithmus nicht zu Politik wird, wodurch die menschliche Autonomie in der weiteren Entscheidungsfindung aufgehoben würde.] (11)

Möglicherweise werden solche Empfehlungen vor dem Hintergrund spezifischer KI-Anwendungsfälle konzipiert, insbesondere KI-Anwendungen, die die Verbreitung von (Fehl-)Informationen regeln und die politischen Grundrechte, insbesondere die demokratischen Rechte, beeinträchtigen. (Die Internationale Konferenz der Beauftragten für den Datenschutz und den Schutz der Privatsphäre, ICDPPC, erwähnt ausdrücklich „Technologien, die die persönliche Entwicklung oder Meinungen beeinflussen" und „die Achtung der verwandten Schutzrechte einschließlich der freien Meinungsäußerung und der Informationsfreiheit" (16). Außerhalb dieser spezifischen Bereiche sind die weitreichenden gesellschaftlichen Auswirkungen von KI entweder unbekannt oder (angesichts unserer derzeitigen Vorhersagefähigkeiten) nicht vorhersehbar.

Ein weiterer Bereich, in dem breite gesellschaftliche Risiken aufgezeigt wurden, betrifft den Einsatz von KI in der Cybersicherheit und in autonomen Waffen (Brundage et al. [22], UNI Global Union [24]). Außerhalb dieses Bereichs gibt es nur sehr wenige Diskussionen über globale Risiken (mit Ausnahme der Debatte über das Aussterben der Menschheit aufgrund – oder ihre Unterwerfung durch – eine boshafte Super-KI u. ä.). Dies mag darauf zurückzuführen sein, dass Anwendungen in anderen Bereichen keine eindeutigen

breiten gesellschaftlichen Auswirkungen haben, oder dass – mit Ausnahme einiger weniger klarer Fälle, wie der Ausbreitung von Fake-News und Online-Hass – das Bewusstsein für das von diesen Anwendungen ausgehende Risiko noch nicht ausgereift ist (75).

Selbst die Pilotversion der Bewertungsliste in den EU-Ethik-Leitlinien für eine vertrauenswürdige KI (29) scheint keine operativen Fragen zu enthalten, die als konkrete Anleitung dienen könnten. Mehrere in den Fragen in der Bewertungsliste enthaltenen Aussagen über die Notwendigkeit, die Auswirkungen der KI auf die Menschen zu bewerten, sagen Organisationen, was sie *tun sollten*; sie sagen ihnen jedoch nicht, *wie* sie es tun sollten:

> „Haben Sie für die Anwendungsfälle, bei denen die Möglichkeit einer Beeinträchtigung der Grundrechte besteht, eine Folgenabschätzung in Bezug auf die Grundrechte durchgeführt?“ (29)[42]

> „Besteht in diesen Fällen die Gefahr, dass das KI-System die menschliche Autonomie beeinträchtigt, indem es unbeabsichtigt in den Entscheidungsprozess des Endverbrauchers eingreift?“ (29)[43]

> „Verbessert oder erweitert das KI-System die menschlichen Fähigkeiten?“ (29)

> „Welche Art von Erkennungs- und Reaktionsmechanismen haben Sie etabliert, um zu beurteilen, ob etwas schief gehen könnte?“ (29)

42 Die von der internationalen Gemeinschaft anerkannten Menschenrechte sind zahlreich, was sich in den verschiedenen internationalen Übereinkommen, Abkommen und Erklärungen widerspiegelt, die viele Staaten unterzeichnet haben (76). Das Völkerrecht enthält jedoch umfassendere und weniger umfassende Listen von Menschenrechten, und nicht alle Menschenrechte werden von allen Staaten anerkannt. Die Leitlinien geben keine Anleitung dazu, welche Menschenrechte berücksichtigt werden sollten (sollen sämtliche Menschenrechte berücksichtigt werden, dann ist die Folgenabschätzung möglicherweise für keine Organisation durchführbar). Die Leitlinien bieten auch keine Anleitung dazu, welche praktischen Schritte notwendig sind, um die Auswirkungen der KI auf die auf einer solchen Liste aufgeführten Menschenrechte zu bewerten. Dies ist problematisch, da Folgenabschätzungen in Bezug auf die Menschenrechte für die meisten Organisationen keine üblichen Tätigkeiten sind.

43 Da der Begriff der menschlichen Autonomie viele Bedeutungen hat, erscheint diese Frage so vage, dass sie in der Praxis ziemlich nutzlos ist. Beeinträchtigt z. B. gezielte Werbung die menschliche Autonomie, indem sie unbeabsichtigt in den Entscheidungsprozess der (End-)Nutzenden (d. h. in die Entscheidung, für eine bestimmte Kandidatin bzw. einen bestimmten Kandidaten zu stimmen) eingreift? Wohl kaum, denn die Wirkung der Einmischung ist voll und ganz beabsichtigt. Es ist nicht klar, ob die Tatsache, dass die Antwort auf diese Frage in der Bewertungsliste „nein“ lautet, diese Art der Einmischung in die menschliche Autonomie für eine vertrauenswürdige KI ethisch akzeptabel macht.

„Haben Sie das Verhalten Ihres Systems in unerwarteten Situationen und Umgebungen bewertet?“ (29)

„Haben Sie beurteilt, ob die Möglichkeit besteht, dass das KI-System den Anwendern oder Dritten Schaden zufügt? Wenn ja haben Sie die Wahrscheinlichkeit, den potenziellen Schaden, den betroffenen Personenkreis und die Schwere des Schadens beurteilt?“ (29)

„Haben Sie die möglichen Auswirkungen eines Ausfalls Ihres KI-Systems eingeschätzt, und ob dieser zu falschen Ergebnissen, zur Nichtverfügbarkeit Ihres Systems oder zu gesellschaftlich inakzeptablen Ergebnissen (z. B. diskriminierenden Praktiken) führen könnte?“ (29)

„Haben Sie ein gutes Verständnis der sozialen Auswirkungen des KI-Systems sichergestellt? Haben Sie beispielsweise das Risiko drohender Arbeitsplatzverluste oder der Dequalifizierung von Arbeitnehmerinnen und Arbeitnehmern beurteilt? Welche Maßnahmen wurden ergriffen, um diesen Risiken entgegenzuwirken?“ (29)

„Haben Sie die weiter reichenden gesellschaftlichen Auswirkungen der Nutzung des KI-Systems über die einzelnen (End-)Nutzer hinausgehend bewertet und dabei beispielsweise andere, möglicherweise indirekt betroffene Akteure berücksichtigt?“ (29)

4.3.3. Die Bedeutung der Einbeziehung von Interessengruppen

Es scheint ein weit verbreitetes Bewusstsein für das Problem des Mangels an Fachkenntnissen für weitreichende Fragen im Zusammenhang mit KI im Allgemeinen zu bestehen. Dies gilt auch für HR-Anwendungen im Besonderen. Als Reaktion darauf plädieren die meisten Leitlinien für die Einbeziehung von Interessengruppen als praktikable Möglichkeit, die Wissenslücken zu schließen. Die Einbeziehung von Interessengruppen ist eine der Anforderungen an „Vielfalt, Nicht-Diskriminierung und Fairness“ in den EU-Ethik-Leitlinien für eine vertrauenswürdige KI (29). Sie ist bei weitem die stärkste Empfehlung in den von der Partnership on AI herausgegebenen Grundsätzen („Tenets“) (13) und wird in den Grundsätzen 2, 3, 4, 5 und 8 in irgendeiner Form erwähnt. Der Grund für die große Anziehungskraft dieses Begriffs liegt darin, dass die Einbeziehung von Interessengruppen die Variable für alles ist, was benötigt wird, um eine Ethik-, Legitimations- oder Wissenslücke zu schließen. Es wird davon ausgegangen, dass die Einbeziehung von Interessengruppen eine Rolle spielt in Bezug auf:

- Rückmeldungen über den Fokus der ethischen Untersuchung, d. h.: Ermitteln Unternehmen und/oder Prüfer*innen alle relevanten Risiken und Schwachstellen? (10,13)
- Die Notwendigkeit einer offenen, interdisziplinären Forschung (13), insbesondere zur Ermittlung und Zusammenführung verschiedener *Fähigkeiten und Kompetenzen* (11)
- Zusammenführung der interdisziplinären Kompetenzen, die erforderlich sind, um *potenzielle Voreingenommenheiten und Formen der Diskriminierung* zu erkennen (19, 20, 25). Einigen Leitlinien zufolge sollten Interessengruppen einbezogen werden, um die gesamte Bandbreite der Datentypen zu identifizieren, die notwendig sind, um ein System des maschinellen Lernens für einen bestimmten Kontext angemessen zu trainieren, und um besser zu verstehen, wie man die benötigten Daten angemessen beschafft (21).
- Kenntnis der *Normen und Werte* der von KI-getriebenen Entscheidungen betroffenen Personen oder Personengruppen (11, 21)
- Ermittlung der verschiedenen *Interessengruppen*, die von KI-Forschung *betroffen* sind (13)
- Ermittlung *bereichsspezifischer Bedenken* (13, 21, 25)
- Vermeidung von Ängsten und Verwirrungen bezüglich KI (14)
- Förderung neuer Formen der Lenkung und Kontrolle, die verschiedene Interessengruppen, wie z. B. die Zivilgesellschaft, die Regierung, den privaten Sektor, oder die akademische Welt und die technische Community, einbeziehen (13, 27)

Neben der Benennung der verschiedenen *Ziele und Anliegen*, die durch die Einbeziehung von Interessengruppen angegangen werden sollten, zeigen diese Empfehlungen auch verschiedene praktische Formen auf, die die Beteiligung von Interessengruppen annehmen kann. Diese sind in Tabelle 3 zusammengefasst.

Tabelle 3

Formen der Beteiligung unterschiedlicher Interessengruppen

Art der Interessengruppe	Art der Einbeziehung	Leitlinie
Die Privatwirtschaft	„[…] Partnerschaften mit anderen Unternehmen eingehen und unser Know-how […] zur Verfügung stellen, um die anstehenden Herausforderungen gemeinsam zu bewältigen." (10)	(10, 13, 25)
Bürger*innen, die breitere Gemeinschaft	Panels für Bürger*innen/Entscheidungsgruppen in allen Phasen des Entwicklungsprozesses; die Einbeziehung einer Ethik-Politik (11) Die Veranstaltung von öffentlichen Diskussionen über die Auswirkungen der neuen Robotertechnologien auf die verschiedenen Dimensionen der Gesellschaft und des Alltagslebens (27) Die Öffentlichkeit über [die Arbeit der Partnership on AI] informieren und auf ihre Fragen eingehen (13) Bereitstellung eines Mechanismus, mit dem die Zielgruppen der KI sichere Rückmeldungen geben können (21) „Wir wollen uns in der Bildung zu KI und Ethik engagieren." (10)	(10, 11, 13, 14, 21, 25, 27)
Politische Entscheidungsträger*innen, Regierungen, Vollzugsbehörden	„[…] unser Know-how politischen Entscheidungsträgern und Bildungsanbietern zur Verfügung stellen, um die anstehenden Herausforderungen gemeinsam zu bewältigen." (10) Aufklärung von Regierungen, Gesetzgebern und Vollzugsbehörden über diese Themen, damit die Bürger*innen mit ihnen zusammenarbeiten, um Angst oder Verwirrung zu vermeiden (z. B. auf dieselbe Weise, wie Polizeibeamt*innen seit Jahren Vorträge über öffentliche Sicherheit in Schulen halten; in naher Zukunft könnten sie Workshops über sichere autonome intelligente Systeme anbieten). (14)	(10, 14, 25)
NGOs	Aufbau zivilgesellschaftlicher Koalitionen und Expert*innennetzwerke: Es sei wichtig, die Notwendigkeit zu betonen, Wissensaustauschprogramme und bessere Voraussetzungen für gemeinsame Strategieentwicklung zwischen zivilgesellschaftlichen Organisationen zu schaffen. (20)	(19, 20)
Wissenschaftler*innen und Ingenieur*innen; die akademische Welt	Kein spezifischer Mechanismus	(10, 13, 19)
Arbeitnehmende	„Wir wollen uns in der Bildung zu KI und Ethik engagieren." (10) „4 […] Die Menschen sollten das Recht auf den Zugriff, die Verwaltung und die Kontrolle der von KI-Systemen erzeugten Daten haben, da die genannten Systeme die Fähigkeit besitzen, diese Daten zu analysieren und zu verwenden." (24)	Deutsche Telekom, UNI Global Union (10, 24)

4.3.4. Lenkungs- und Kontrollstrukturen und Rechenschaftspflicht

Viele Leitlinien empfehlen, die ethischen Ergebnisse durch verstärkte Rechenschaftspflicht zu verbessern. Leider bleiben die meisten Leitlinien recht allgemein, was die Art der (neuen?) Lenkungs- und Kontrollsysteme betrifft, die zu diesem Zweck eingerichtet werden sollten. Zum Beispiel:

> „Systems for registration and record-keeping should be created so that it is always possible to find out who is legally responsible for a particular A/IS."
> [Es sollten Systeme zur Registrierung und Aufzeichnung geschaffen werden, so dass es immer möglich ist, herauszufinden, wer für ein bestimmtes autonomes intelligentes System (AIS) rechtlich verantwortlich ist.] (14)

> „Organizations should publicly describe the model governance programs they have in place to detect and remedy any possible discriminatory effects of the data and models they use, including the standards they use to determine whether and how to modify algorithms to be fairer."
> [Organisationen sollten die Programme öffentlich beschreiben, die sie zur Lenkung und Kontrolle der Modelle eingerichtet haben, um mögliche diskriminierende Auswirkungen der Daten und Modelle, die sie verwenden, aufzudecken und zu beheben, einschließlich der Standards, die sie verwenden, um zu bestimmen, ob und wie Algorithmen geändert werden müssen, um fairer zu sein.] (26)

> „If accidents occur, the AI will need to be transparent and accountable to an accident investigator, so the internal process that led to the accident can be understood."
> [Wenn sich Unfälle ereignen, muss die KI transparent und einem Unfallermittler gegenüber rechenschaftspflichtig sein, damit der innere Prozess, der zu dem Unfall geführt hat, nachvollzogen werden kann.] (24)

> „Kontinuierliche Aufmerksamkeit und Wachsamkeit sowie Rechenschaftspflicht für die möglichen Auswirkungen und Folgen von Systemen mit künstlicher Intelligenz sollten insbesondere durch […] die Festlegung nachweisbarer Governanceprozesse für alle wichtigen Akteure […]." (16)

Die wenigen konkreten Ideen, die genannt werden, sind: „[Governanceprozesse] auf der Grundlage vertrauenswürdiger Dritter oder der Einsetzung unabhängiger Ethikkommissionen" (16); die Bereitstellung einer Anwendungsprogrammschnittstelle (API), die es Dritten ermöglicht, den Algorithmus abzufragen; die Ermöglichung der Durchführung automatisierter Audits durch die Forschungsgemeinschaft; die Einrichtung eines Plans zur Kommu-

nikation mit Dritten, wie z. B. der Forschungs- und Entwicklungsgemeinschaft, die möglicherweise Interesse hätten, einen Algorithmus zu prüfen (12).

Eine etwas spezifischere Empfehlung über die Art dieser Verantwortung besagt, dass KI nicht benutzt werden darf, um die Verantwortung der für die Entscheidungen verantwortlichen Manager*innen zu verschleiern. Das heißt, dass für die Rechenschaftspflicht immer ein oder mehrere Menschen hinter der KI stehen:

> „For the foreseeable future A/IS should not be granted rights and privileges equal to human rights: A/IS should always be subordinate to human judgment and control."
> [Auf absehbare Zeit sollten autonomen intelligenten Systemen (AIS) keine den Menschenrechten gleichwertigen Rechte und Privilegien gewährt werden: AIS sollten immer dem menschlichen Verstand und der menschlichen Kontrolle untergeordnet sein.] (14)

> „Legal accountability has to be ensured when human agency is replaced by the decisions of AI agents."
> [Die rechtliche Rechenschaftspflicht muss sichergestellt werden, wenn menschliche Handlungsmächtigkeit durch die Entscheidungen von KI-Agenten ersetzt wird.] (28)

> „The true operator of an AI system must be made known to the public."
> [Der wahre Betreiber eines KI-Systems muss der Öffentlichkeit bekannt gemacht werden.] (15)

4.3.5. Relevanz für HR Analytics

Viele der hier analysierten Leitlinien können dahingehend interpretiert werden, dass Organisationen, die KI-Technologie zur Unterstützung ihrer HR-Entscheidungen einsetzen, ein System zur Überwachung der Auswirkungen dieser Technologie auf ihre Mitarbeitenden sowie ein speziell entwickeltes System ethischer Lenkung und Kontrolle (Governance) einführen sollten. Die spezifischeren Empfehlungen zu einer solchen Überwachung und Lenkung und Kontrolle, die sich in den 20 hier analysierten Leitlinien finden, lassen sich wie folgt zusammenfassen:

- Stellen Sie eine Anwendungsprogrammschnittstelle (API) zur Verfügung, die es der Forschungsgemeinschaft ermöglicht, den Algorithmus, der für prädiktive und präskriptive HR-Entscheidungen verwendet wird, abzufragen.

- Führen Sie einen unabhängigen Prozess der ethischen Governance ein.
- Entwickeln Sie ein Verfahren, mit dem Mitarbeitende Fehler in Input-Daten oder Output-Entscheidungen berichtigen können.
- Richten Sie einen Mechanismus ein, der es den Mitarbeitenden ermöglicht, Entscheidungen anzufechten, die ausschließlich auf der automatisierten Verarbeitung ihrer Informationen beruhen.
- Schaffen Sie sichtbare Möglichkeiten der Abhilfe für diejenigen, die von indirekten Diskriminierungen und anderen diskriminierenden Outputs betroffen sind.
- Führen Sie Aufzeichnungen über die Rolle, die algorithmische Empfehlungen und Vorhersagen bei HR-Entscheidungen spielen.

Die erste Empfehlung mag unrealistisch erscheinen. Es ist unwahrscheinlich, dass Organisationen die Durchführung automatisierter Audits ermöglichen, um die Arbeit von Forschenden zu erleichtern, die die von den Organisationen verwendeten Algorithmen öffentlich als diskriminierend kritisieren könnten, was ein ständiges Risiko von Reputationsverlusten mit sich bringen würde. Die zweite Empfehlung könnte in einigen Kontexten, in denen KI eingesetzt wird (z. B. in der Medizin), akzeptabel sein, ist aber im Zusammenhang mit HR recht radikal. Höchstwahrscheinlich würde sie die KI im HR-Bereich zu teuer machen, und es würde sich daher nicht lohnen, in sie zu investieren. Plausibler ist, dass große Unternehmen gut beraten sind, einen internen Ethikausschuss einzurichten, die die Einführung der KI in Industrieprozessen mit erheblichen Auswirkungen auf ihre Mitarbeitenden überprüft.

Natürlich hängt die Durchführbarkeit solcher Ideen selbst für ein großes Unternehmen davon ab, wie solche Ethikausschüsse funktionieren sollen. Sollte der Ethikausschuss alle HR-Entscheidungen überprüfen, die mithilfe der KI getroffen werden? Und wenn ja, wie würde sich eine solche Überprüfung von der gewöhnlichen Arbeit der HR-Fachkräften im Unternehmen unterscheiden? Vielmehr ist die Idee eines Ethik-Ausschusses plausibler, wenn er als ein Gremium konzipiert ist, das zusammentritt, um die Einführung der KI im HR-Bereich zu planen, insbesondere was getan werden muss, um sie fair und verständlich zu gestalten. Der Ethikausschuss sollte auch eindeutig festlegen, welche Schritte zu unternehmen sind, um das Verhalten der KI im Betrieb zu überwachen.

Den meisten hier untersuchten Empfehlungen folgend, sollte ein solches Gremium transdisziplinär sein und Vertreter*innen verschiedener Abteilungen umfassen, z. B. der Unternehmensleitung, der Datenwissenschaftsabtei-

lung sowie der verschiedenen Funktionen, die unmittelbar oder mittelbar von der Innovation betroffen sind. Im Idealfall kann auch eine externe Expertin bzw. ein externer Experte für KI-Ethik hinzugezogen werden. Die Idee eines Prozesses, der es den Mitarbeitenden ermöglicht, Fehler in Daten zu berichtigen, die verwendet werden, um HR-Vorhersagen und Entscheidungen über diese Vorhersagen zu treffen, ist durch das Datenschutzgesetz impliziert.

Der Gedanke, dass Mitarbeitende in der Lage sein sollten, KI-Entscheidungen anzufechten, kann als die Idee verstanden werden, dass Betreiber*innen von KI im HR-Bereich in der Lage sein sollten, Handlungsempfehlungen von KI-Systemen anzufechten. Dies ist nicht nur plausibel, sondern höchstwahrscheinlich Teil der Art und Weise, wie KI in HR Analytics überall umgesetzt wird. Es ist schwierig, sich eine konkrete Anwendung von KI in HR Analytics, insbesondere für präskriptive Analysen, vorzustellen, die zu einer vollständigen Automatisierung von HR-Entscheidungen führen würde. Der Gedanke, dass Mitarbeitende, die KI-Entscheidungen unterliegen, in der Lage sein sollten, HR-Entscheidungen, die sie betreffen, anzufechten, ist nicht nur sinnvoll, sondern wird zumindest in den meisten EU-Ländern vom Arbeitsrecht grundsätzlich vorausgesetzt.

Theoretisch ist in den meisten EU-Ländern ein Abhilfemechanismus für die Missachtung der Rechte von Mitarbeitenden aufgrund der Verwendung von KI in HR Analytics durch arbeitsrechtliche Schutzvorschriften (oder zumindest durch Vorschriften zum Schutz von offiziell angestellten Mitarbeitenden) impliziert. Aber es mag für Mitarbeitende nicht einfach sein, Abhilfe *zu erlangen*, wenn sie ihnen zusteht. So sollten Mitarbeitende z. B. einen Rechtsanspruch haben, eine unfaire Entscheidung anzufechten und aufheben zu lassen, unabhängig von der Rolle, die die KI bei der Begründung der Entscheidung gespielt hat. Darauf kann irgendeine Form der Entschädigung für die betroffene Person folgen. Es ist jedoch unklar, wie effektiv Menschen, die von algorithmischen Entscheidungen betroffen sind, Abhilfe erlangen können, da z. B. Diskriminierung durch Algorithmen schwer nachzuweisen sein kann.

Schließlich scheint die Empfehlung, Aufzeichnungen über die Empfehlungen von Algorithmen zu führen, einschließlich der Mitarbeitendendaten und der Software, die zu diesen Empfehlungen führen, eine realistische Maßnahme zu sein, um die Rechenschaftsplicht der Menschen hinter der KI zu erhöhen. Solche Aufzeichnungen können eine angemessene Bewertung einzelner unfairer Entscheidungen sowie der allgemeinen Mängel der zugrunde liegenden Software ermöglichen.

5 FAZIT

Dieses Fazit basiert auf in den Leitlinien enthaltenen Schlüsselgedanken, die in den vorangegangenen Kapiteln in allgemeiner Form ausgedrückt wurden. Wir wenden die Empfehlungen der 20 hier untersuchten Leitlinien zu KI und Algorithmen auf den Bereich der HR Analytics an und kommen zu vier allgemeinen Empfehlungen:

- DSGVO+: Regeln für die Erhebung von Daten für HR Analytics sollten über die EU-Datenschutz-Grundverordnung (DSGVO) hinausgehen.
- Die Entwicklung KI-getriebener HR-Werkzeuge erfordert ausreichende fachliche Kompetenzen, um Wissen über den Algorithmus zu generieren.
- Die Auswirkungen der Anwendung des KI-Werkzeugs auf die Mitarbeitenden sollten sorgfältig überwacht werden.
- Das angemessene Ausmaß an Transparenz in Bezug auf algorithmische Entscheidungen sollte ermittelt und implementiert werden.

Ich erläutere jede dieser Empfehlungen in drei Schritten:

1. Zunächst werden Schlüsselfragen der Implementierung extrahiert.
2. Anschließend werden praktische Empfehlungen entwickelt, die, wenn sie befolgt werden, den Einsatz von KI im HR-Bereich ethischer machen würden.
3. Im letzten Schritt wird aufzeigt, wie sich diese Empfehlungen mit den wichtigsten (inhaltlichen) ethischen Werten – Benefizienz, Schadensverhütung, Gerechtigkeit und Fairness sowie Autonomie – zusammenhängen.

Die Zusammenfassung folgt der dreiteiligen Unterteilung der Themen in Datenerfassung/Datenspeicherung/Datenzugriff, die Entwicklung algorithmischer Werkzeuge und die Bewertung ihrer Auswirkungen.

5.1. DSGVO+: Regeln für die Datenerhebung für HR Analytics sollten über die DSGVO hinausgehen

Die erste Anforderung an eine ethische KI ist es, das höchste Niveau des Datenschutzes zu erreichen. Dies betrifft die Mitarbeitendendaten, die für das Training der KI verwendet werden, sowie die Daten der einzelnen Mitarbei-

tenden, auf deren Grundlage die KI Empfehlungen ausgibt, die die einzelnen Personen betreffen. Es gibt eine deutliche Überschneidung zwischen den KI-Leitlinien bezüglich der Erhebung von Daten und den rechtlichen Grundsätzen des Datenschutzes. Die EU-Datenschutz-Grundverordnung (DSGVO), die neueren Datums und umfassend ist, könnte für viele Organisationen einen adäquaten Rechtsstandard bieten, der auch die Kompromisse zwischen verschiedenen Werten und Zielen der Organisationen berücksichtigt. Aber die DSGVO reicht nicht aus. Ein DSGVO+-Ansatz könnte als ein ethischer Ansatz betrachtet werden, der die Achtung des Datenschutzes durch die Grundsätze der Einbeziehung von Interessengruppen (4.3.3) und der Lenkungs- und Kontrollstrukturen zur Förderung der Rechenschaftspflicht (4.3.4 und 4.3.5) ergänzt.

Schlüsselfragen der Implementierung

Die wichtigsten ethischen Fragen, die vor der Durchführung der Datenerhebung beantwortet werden müssen, sind:

- Garantieren wir den strengsten Schutz der Privatsphäre der Mitarbeitenden, der mit den Zielen von HR Analytics im Einklang steht?
- Ist der Zweck, für den wir um Daten von Mitarbeitenden bitten, einer, den wir rechtfertigen können, und steht er mit unserem Auftrag als Organisation im Einklang?
- Ist es möglich, die Mitarbeitenden in den Prozess der Lenkung und Kontrolle von datengetriebenen Algorithmen aktiv einzubeziehen, anstatt ihnen die Rolle passiver Empfänger*innen algorithmischer Entscheidungen zuzuweisen?

Wichtige Implementierungsschritte

Praktische Empfehlungen zur Umsetzung ethischer Prozessverbesserungen sind:

- *Richten Sie einen internen oder unabhängigen Ethikausschuss* zur sorgfältigen Bewertung der *Zwecke* der Datenerhebung unter ethischen Gesichtspunkten ein. Ein interner Ethikausschuss sollte alle Schlüsselkompetenzen in der Organisation umfassen (z. B. einschließlich, aber nicht beschränkt auf HR, Datenschutz und Compliance). Ein unabhängiger Ethikausschuss sollte Interessengruppen einbeziehen, die in der Lage sind, verschiedene Sichtweisen zu bieten.
- *Bitten Sie die Mitarbeitenden, ihre Meinungen zu äußern.* Die Stimmen der Mitarbeitenden, die sich über die Zwecke, für die ihre Daten analysiert werden, Gedanken machen, sollten nicht ignoriert werden. Ein Diskus-

sionsforum für Mitarbeitenden, in das auch Gewerkschaftsvertreter*innen als Moderator*innen einbezogen werden können, kann dazu beitragen, Wege zur Nutzung von Daten im HR-Bereich zu finden, die als eher problematisch gelten, sowie von Daten, die als eher wünschenswert angesehen werden.

Wichtige ethische Werte

Aus der Sicht der hier betrachteten inhaltlichen Werte (3.1.5) sind die inhaltlichen ethischen Ziele und Zwänge in dieser Phase der Umsetzung von KI die folgenden:

- *Benefizienz:* Im Idealfall sollte KI in HR Analytics allen Mitarbeitenden zugute kommen, nicht nur ihren Arbeitgeber*innen. Der beste Weg, Daten für die KI zu erheben, besteht darin, den Mitarbeitenden Wege aufzuzeigen, wie die KI ihren Arbeitsplatz *für sie* besser machen kann. HR Analytics sollte auf Daten basieren, die von den Mitarbeitenden zur Verfügung gestellt werden, weil sie sich dafür engagieren und den Zielen, für die die KI eingeführt wird, verpflichtet sind. Um die Mitarbeitenden einzubinden, sollten die Ziele der HR Analytics transparent sein und der Nutzen für die Mitarbeitenden klar umrissen werden. Es sollte ein Mechanismus vorhanden sein, um nicht nur Daten von Mitarbeitenden zu erheben, sondern auch die Meinungen der Mitarbeitenden über legitime und illegitime Möglichkeiten der Nutzung dieser Daten in HR.
- *Schadensverhütung*: HR Analytics sollte nicht dazu benutzt werden, Mitarbeitenden zu schaden. Ein starker Schutz der Privatsphäre sowie starke Cybersicherheit sollten gewährleistet sein, um sicherzustellen, dass 1) die Daten nur für die erklärten Zwecke verwendet werden; 2) Datenmissbrauch und Verletzungen des Schutzes personenbezogener Daten vermieden werden können. HR Analytics sollte nicht dazu verwendet werden, schwerwiegende Strafen und Sanktionen (z. B. Entlassungen, Gehaltskürzungen) auf der Grundlage automatisierter Entscheidungen oder Entscheidungen zu verhängen, bei denen der menschliche Input lediglich „formell" ist, d. h. sich auf die Anerkennung der Empfehlung der Software beschränkt.
- *Gerechtigkeit und Fairness:* Mitarbeitende, die nicht bereit sind, ihre Daten für HR Analytics verarbeiten zu lassen, sollten fair behandelt werden. Auch wenn sie von einigen Vorteilen ausgeschlossen werden können, die mit der Bereitschaft zur Bereitstellung von Daten für HR Analytics verbunden sind (z. B. dem Erhalt von datengesteuerten, personalisierten Rückmeldungen und Vorhersagen), muss jede weitere (d. h. vermeidba-

re) Benachteiligung von Personen vermieden werden, die sich bei der Weitergabe ihrer Daten weniger wohl fühlen.
- *Autonomie*: Es müssen Schritte unternommen werden, um datengestützte Erkenntnisse und Vorhersagen mit den Mitarbeitenden zu teilen. Mitarbeitende sollten immer in die Lage versetzt werden, a) aus datengetriebenen Einsichten und Vorhersagen etwas Nützliches für sich selbst lernen zu können, b) auf eine datengetriebene Vorhersage oder Beurteilung mit einer positiven Verhaltensänderung reagieren zu können. Das übergeordnete ethische Ziel in Bezug auf die Autonomie besteht darin, sicherzustellen, dass die Mitarbeitenden nicht zu „*passiven* Untertanen" der algorithmischen Governance werden, sondern aktiv zur Verbesserung der Leistung der Organisation beitragen können, indem sie die Vorteile datengesteuerter Einsichten nutzen, die KI im HR-Bereich hervorbringen kann.

5.2. Die Entwicklung datengetriebener HR-Werkzeuge erfordert ausreichende fachliche Kompetenzen, um Wissen über den Algorithmus zu generieren

Die Generierung adäquater Modelle, die auf *ethische Weise* aus Daten gelernt wurden, impliziert grundsätzlich, sich mit der Erklärbarkeit und Fairness der Algorithmen auseinanderzusetzen. Die Nachvollziehbarkeit und Fairness eines Modells kann zu einem erheblichen Teil durch technische Methoden verstanden, dokumentiert und verbessert werden. Die erforderliche fachliche Kompetenz sollte durch weniger formalisierte Einsichten ergänzt werden, die von Interessengruppen geliefert werden, die zu diesem Zweck möglicherweise hinzugezogen werden müssen. Dazu gehören beispielsweise Einsichten darüber, was *im jeweiligen Kontext* moralisch angemessen ist.

Eine Herausforderung für Unternehmen besteht darin, ihre Fähigkeit zu verbessern, die Auswirkungen von KI-Entscheidungen auf bestimmte Gruppen, oder die Daten, die indirekt diskriminierend sein können, besser zu verstehen. Es ist möglich, dass ethnisch vielfältigere Forschungsteams mit einem ausgewogeneren Verhältnis von Männern und Frauen eine höhere Sensibilität für diese Themen haben. Noch entscheidender ist, dass es eine kognitive Vielfalt geben sollte. Insbesondere sollte ein Bedürfnis bestehen, dass Expert*innen „anders denken" sollten als die meisten Informatiker*innen (dazu gehören auch Informatiker*innen, die Kompetenzen in Ethik erworben haben). Die Forschung wird hoffentlich Erkenntnisse über die Wirksamkeit spezifischer Einstellungs- und Diversity-Strategien für KI liefern.

Schlüsselfragen der Implementierung

Die wichtigsten ethischen Fragen, die vor dieser Phase der Datenverarbeitung beantwortet werden sollten, sind

- Verfügen die HR-Manager*innen in der Organisation über die fachlichen Kompetenzen, die erforderlich sind, um die Erklärbarkeit und Fairness von KI-Werkzeugen angemessen zu beurteilen?
- Sammeln die Datenwissenschaftler*innen genügend Informationen über das Training von KI-Systemen, um Fehler in den KI-Anwendungen zu orten und beheben, wenn diese nicht angemessen funktionieren? Sind die potenziellen Mängel und Grenzen der Trainingsdatensätze hinreichend bekannt? Sind die technischen Schritte zur Bewertung und Dokumentation von Fairness und Nachvollziehbarkeit bereits vorhanden?
- Gibt es eine ausreichende kognitive und moralisch-politische Vielfalt unter den Bewertenden der KI, um Fragen der Fairness und Transparenz kritisch zu beurteilen?
- Ist die Organisation in der Lage, Inputs und Rückmeldungen von Expert*innen außerhalb der Organisation einzuholen, die für die Beurteilung der Fairness und Verständlichkeit von KI-Werkzeugen relevant sind?

Wichtige Implementierungsschritte

Praktische Empfehlungen zur Umsetzung ethischer Verbesserungen sind:

- *Stellen Sie sicher, dass Sie über angemessene Kompetenzen für den Aufbau und die ethische Umsetzung von KI verfügen.* Organisationen, deren Ziel es ist, KI-Werkzeuge im HR-Wesen zu produzieren und zu implementieren, müssen Expert*innen mit den Kompetenzen (sowie dem informellen Wissen und den kognitiven Stilen) rekrutieren, die erforderlich sind, um die Nachvollziehbarkeit und Fairness eines Algorithmus zu bewerten.
- *Beziehen Sie zivilgesellschaftliche Organisationen und Fachwissen von Hochschulen ein.* Lassen Sie sich extern bestätigen, dass die technischen Methoden, die Sie zum Training Ihrer Modelle einsetzen, wissenschaftlich fundiert und ethisch vertretbar sind. Entwickeln Sie ein Bewusstsein für mögliche ethische Kritik, indem Sie Rückmeldungen von außerhalb der Organisation einholen. Wenn möglich, stellen Sie den Forschenden (anonymisierte) Datensätze und Ihre Algorithmen zur Verfügung, damit sie Ihr Werkzeug einem unabhängigen Audit unterziehen können.
- *Betreiben Sie Wissensvermittlung im Bereich KI und Ethik:*
  - Schulen Sie potenzielle Endnutzende (z. B. HR-Fachkräfte), um sicherzustellen, dass sie über die Kompetenzen verfügen, die für den korrekten Betrieb von KI-Anwendungen im HR-Bereich erforderlich sind.

Endnutzende sollten kein blindes Vertrauen in KI-Werkzeuge haben, sondern ein angemessenes Maß an Vertrauen mit einer kritischen Einstellung verbinden. Beziehen Sie die beabsichtigten Nutzenden Ihres KI-Werkzeugs ein, um sicherzustellen, dass Sie genug über das Werkzeug, seine Grenzen und seinen korrekten Anwendungsbereich wissen. Fördern Sie Bildungsressourcen, die dem Personal in HR-Abteilungen helfen, falsche Vorstellungen über KI im HR-Bereich zu vermeiden.[44]

- Schulen Sie Mitarbeitende, die potenziell von KI betroffen sind, oder ihre Vertreter*innen (z. B. den Betriebsrat), um sicherzustellen, dass sie über das Know-how und die Kompetenzen verfügen, die notwendig sind, um die Anwendung von AI im HR-Bereich richtig zu verstehen. Sie sollten kein blindes Vertrauen in KI-Werkzeuge haben, sondern ein angemessenes Maß an Vertrauen in Verbindung mit einer kritischen Haltung. Binden Sie beabsichtigte Betroffene ihres Werkzeugs oder deren Vertreter*innen ein, um sicherzustellen, dass sie genug über das Werkzeug, seine Grenzen und seinen richtigen Anwendungsbereich wissen. Fördern Sie Bildungsressourcen, die ihnen helfen, Missverständnisse über KI in im HR-Bereich zu vermeiden.

Zentrale (inhaltliche) ethische Werte

Unter dem Gesichtspunkt der hier betrachteten Werte (3.1.5) sind die wesentlichen ethischen Ziele und Zwänge dieser Phase der Umsetzung der KI die folgenden:

– *Benefizienz:* Die produzierten Werkzeuge sollten der Organisation und ihren Mitarbeitenden, die bereit sind, sich auf der Grundlage ihrer persönlichen Daten bewerten zu lassen, einen Nutzen bringen.
– *Schadensverhütung*: Die Risiken, die sich aus Ungenauigkeiten des Modells ergeben (z. B. Entscheidungen, die auf falschen Vorhersagen beruhen), sollten eingehend bewertet werden. Sobald die Risiken sorgfältig verstanden sind, sollten die Risiken und der potenzielle Nutzen gegeneinander abgewogen werden. Insbesondere sollte man auch den Nutzen abwägen, der sich aus dem Einsatz eines Mechanismus ergibt, der möglicherweise (sowohl in Bezug auf Fairness als auch in Bezug auf Effizienz) besser funktioniert als das bereits bestehende Verfahren. Die Mitarbeitenden sollten einbezogen werden, um ihre Ansichten über die Risiken und Vorteile solcher Werkzeuge besser zu verstehen.

44 Beispielsweise können Datenwissenschaftler*innen Analysen und Beispiele für Situationen liefern, in denen das Werkzeug missbraucht wird.

- *Gerechtigkeit und Fairness:* Die Fairness des Werkzeugs sollte unter Anwendung modernster Methoden und einer angemessenen Mischung aus technischen (auf statistischen Berechnungen basierenden) und nichttechnischen (psychologischen oder philosophischen) Ansätzen bewertet werden.
- *Autonomie*: Um einen Beitrag zur menschlichen Autonomie zu leisten, muss die Logik hinter dem KI-Werkzeug verstanden werden, das für Personalbeurteilungen und -empfehlungen verwendet wird. Dabei handelt es sich um eine Mischung aus fundierten wissenschaftlichen Verfahren, die es erlauben, zu rekonstruieren, wie das Werkzeug gelernt hat, Empfehlungen so zu formulieren, wie es das tut. Zum Beispiel sollten im Falle des statistischen Lernens die Daten, die Trainingsmethode und die Spezifikation der Nutzen-/Verlustfunktion des Algorithmus notiert werden. Darüber hinaus sollte es möglich sein, Erläuterungen zu geben, die eine sinnvolle und konstruktive Debatte zwischen Datenwissenschaftler*innen und HR-Expert*innen mit verschiedenen Formen von Domänenwissen ermöglichen. Zum Beispiel kann es nützlich sein, eine Möglichkeit zu haben, einzelne Empfehlungen, die von einer KI gemacht wurden, zu erklären. Solche Erklärungen können weniger präzise sein als die, die Datenwissenschaftler*innen verwenden, aber sie spielen dennoch eine wichtige Rolle.

5.3. Die Auswirkungen der Anwendung des KI-Werkzeugs auf die Mitarbeitenden sollten sorgfältig überwacht werden

Es sollte möglich sein, die tatsächlichen Auswirkungen des Einsatzes von KI zur Unterstützung von HR-Entscheidungen zu überwachen und zu dokumentieren. Dies kann durch die Implementierung technischer Verfahren (z. B. die automatische Erhebung von Daten über Entscheidungen, die auf der Grundlage der Inputs von KI-Anwendungen getroffen werden) und durch die Implementierung sozialer Prozesse erreicht werden, wie z. B. die Möglichkeit für HR-Mitarbeitende, Rückmeldungen zu geben und die Outputs der KI zu diskutieren, auch mit den Designer*innen und Entwickler*innen der Anwendung.

Schlüsselfragen der Implementierung

Die wichtigsten ethischen Fragen, die vor dieser Phase der Datenverarbeitung beantwortet werden sollten, sind

- Verfügen wir über Mechanismen, die sicherstellen, dass die von Algorithmen getroffenen Entscheidungen bewertet und korrigiert werden können, wenn etwas schief geht?
- Sind unsere Verfahren zur Dokumentation der von Algorithmen getroffenen Entscheidungen angemessen? Können wir dafür sorgen, dass sie angemessen sind, ohne die Privatsphäre unserer Mitarbeiter zu gefährden?
- Wie können wir einen Mechanismus für sichere Rückmeldungen einrichten, der das Funktionieren des HR-Werkzeugs nicht beeinträchtigt, sondern es den Datenwissenschaftler*innen erlaubt, es Schritt für Schritt zu verbessern?
- Gibt es einen Mechanismus, um Personen zu entschädigen, die aufgrund einer unzutreffenden Bewertung oder Vorhersage durch einen Algorithmus ungerecht behandelt werden?

Wichtige Schritte zur Umsetzung der Ethik

Den 20 hier analysierten Leitlinien zufolge sind die wichtigsten organisatorischen Schritte, die in dieser Phase der Datenverarbeitung zur Umsetzung der Leitlinien vorhanden sein sollten, die folgenden:

- Entwickeln Sie einen (angemessen datengeschützten) Mechanismus zur Aufzeichnung von Entscheidungen über Mitarbeitende, die mithilfe von algorithmischen Empfehlungen oder Vorhersagen getroffen werden.
- Entwickeln Sie einen (angemessen datengeschützten) Mechanismus, um festzustellen, ob bestimmte Personengruppen (z. B. Nicht-Muttersprachler*innen, Schwangere, Angehörige von Minderheiten usw.) durch algorithmische Entscheidungen negativ oder positiv beeinflusst werden.
- Entwickeln Sie ein Verfahren, mit dem Mitarbeitende Fehler in Input-Daten oder Outputs berichtigen können.
- Entwickeln Sie ein Verfahren, mit dem Mitarbeiterinnen und Mitarbeiter gegebenenfalls Entscheidungen anfechten können, die vollständig automatisiert sind.
- Wenn eine Untergruppe Ihrer Mitarbeitenden (z. B. Nicht-Muttersprachler*innen, Eltern mit Kindern, Angehörige religiöser Gruppen usw.) durch die Einführung algorithmischer Entscheidungen systematisch benachteiligt zu sein scheint, richten Sie ein Verfahren zur Abhilfe ein, oder besser noch, verbessern Sie das Ergebnis für diese Gruppe langfristig.

Zentrale (inhaltliche) ethische Werte

Unter dem Gesichtspunkt der hier betrachteten materiellen Werte (3.1.5) lassen sich die ethischen Ziele dieser Phase wie folgt charakterisieren:

- *Benefizienz:* Es sollte möglich sein, jede Behauptung über die positive Wirkung der KI in HR Analytics am Arbeitsplatz mit Daten zu untermauern. Wenn die Daten keinen Nutzen aus dem Einsatz eines solchen Werkzeugs erkennen lassen, sollte der Einsatz des Werkzeugs überdacht werden. Wenn die für eine solche Bewertung erforderlichen Daten nicht vorhanden sind, sollten sie erstellt werden.
- *Schadensverhütung:* Es sollte ein Verfahren vorhanden sein, um Probleme zu erkennen, die durch algorithmische Entscheidungen verursacht werden. Diese Probleme können auf ungenauen Entscheidungen oder auf (möglicherweise irrationaler) Furcht vor und Ablehnung von solchen Werkzeugen beruhen. Eine Möglichkeit, dies zu erreichen, ist ein Verfahren, das es den Mitarbeitenden ermöglicht, HR-Entscheidungen, die mithilfe von Algorithmen getroffen wurden, anzufechten und zu kritisieren und Erklärungen darüber zu erhalten.
- *Gerechtigkeit und Fairness:* Neben technischen Bewertungen, die routinemäßig beim Training eines Werkzeugs vorgenommen werden, ist die Fairness von KI anhand von realweltlichen Ergebnissen zu bewerten.
  Zu diesem Zweck ist es wichtig, die Gruppen von Beschäftigten, die von einem solchen Werkzeug nachteilig betroffen sein könnten, zu ermitteln. Die Auswirkungen auf die Beschäftigten aus verschiedenen Gruppen sollten gemessen und bewertet werden. Gruppen, die unter Benachteiligungen leiden, oder nicht wie andere Gruppen von der Einführung eines solchen Werkzeugs profitieren, sollten Möglichkeiten zur Verbesserung ihrer Situation erhalten.
- *Autonomie*: Vollautomatisierte Entscheidungen im HR-Bereich sollten in der Regel vermieden werden. Wenn es sie gibt, muss ein Verfahren zur Anfechtung solcher Entscheidungen vorhanden sein.

5.4. HR und Management sollten eine angemessene Transparenz in Bezug auf im HR eingesetzte datengetriebene Werkzeuge gewährleisten

Der Wert Transparenz wird, wie die vorangegangene Analyse zeigt, von allen hier untersuchten KI-Leitlinien berücksichtigt. Allerdings ist Transparenz nur dann von Vorteil, wenn sie richtig ausgestaltet ist, sonst wird sie nach hinten losgehen. Sie wird dann von Vorteil sein, wenn ein besseres Verständnis der Logik des Algorithmus den Mitarbeitenden Anreize bietet, besser zu arbeiten. Dieses Ergebnis ist möglich, wird jedoch nicht sicher deswegen er-

reicht, nur weil der HR-Algorithmus auf statistisch genauen statistischen Modellen basiert. Schlecht ausgestaltete Transparenz kann auf zweierlei Weise nach hinten losgehen:

Erstens können die Mitarbeitenden die Transparenz ausnutzen, um die Kennzahlen, die zur Entscheidungsfindung über sie verwendet werden, für die Erlangung persönlicher Vorteile zu verwenden, und zwar auf eine Art und Weise, die nicht gut für das Unternehmen ist.

Zweitens besteht die Gefahr einer Pervertierung der von den Mitarbeitenden verfolgten Ziele, wenn sie ihre Ergebnisse auf der Grundlage von Proxys für Leistung und Exzellenz maximieren, anstatt eine echte Verbesserung anzustreben, die als Nebeneffekt eine gute Wertung produziert.[45]

Allgemeiner gesagt: Wenn die KI von Modellen des maschinellen Lernens abgeleitet wird, die blind für strategische Überlegungen sind, kann ihre Vorhersagekraft untergraben werden, da die Kenntnis des Algorithmus die Anreize der Arbeitnehmenden beeinflusst und somit zu neuen Verhaltensmustern führt (die sich von denen unterscheiden, die die Daten erzeugen, die beim Training der KI verwendet werden). Es sollte jedoch betont werden, dass Mitarbeitende und Manager*innen versuchen könnten, selbst intransparente KI-Anwendungen auszuspielen, und zwar anhand von Vermutungen oder „Firmenmythen" über die Funktionsweise der undurchschaubaren KI. Daher kann intransparente KI ebenfalls verzerrende Auswirkungen haben und Vorhersagen untergraben, und es ist fraglich, ob eine transparentere KI immer zu schlechteren Versuchen führen wird, das System auszuspielen. Eine wissenschaftliche Möglichkeit, solche Verzerrungen zu vermeiden, besteht darin, eine *strategiesichere* KI zu entwerfen (d. h. Entscheidungsregeln, die im spieltheoretischen Sinne „Dominant-Strategy-Incentive-kompatibel", DSIC, sind). Wenn Mitarbeitende einer strategiesicheren bekannten Entscheidungsregel unterworfen sind, dann wird per definitionem ihrem eigennützigen Interesse am besten gedient, wenn sie die sie betreffenden Angaben wahrheitsgetreu machen. Dies kann fortgeschrittenere Modelle des maschinellen Lernens erfordern, die nicht nur auf den Grundsätzen der Statistik, sondern auch auf den Grundsätzen der Wirtschaftswissenschaften, insbesondere der Spieltheorie, beruhen. (77)

Aufgrund ihrer Komplexität muss eine adäquate Transparenzstrategie in der Tat mit verschiedenen, auf unterschiedliche Kontexte zugeschnittenen Lösungen geplant, entworfen und erreicht werden. Verschiedene Arten von

45 Wie bereits im Abschnitt 4.2.4 („Offene Herausforderungen für algorithmische Transparenz und Rechenschaftspflicht") dargelegt.

Transparenz sollten an die richtigen Interessengruppen (z. B. Arbeitnehmervertreter*innen und HR-Manager*innen im Falle von HR) gerichtet werden, indem die Unternehmensleitung dafür gewonnen wird, parallel zur Einführung von KI eine Transparenzstrategie zu entwerfen. Es ist auch möglich, dass verschiedene Formen der Transparenz gleichzeitig eingesetzt werden sollten. So kann z. B. die Dokumentation des maschinellen Lernprozesses für einige Zwecke und einige Interessengruppen ausreichend, für andere jedoch nutzlos sein; eine detaillierte Kenntnis der Features und ihrer Auswirkungen auf das Ergebnis kann in einigen Kontexten fair, nützlich und möglich sein (z. B. wenn die Algorithmen keine „Blackboxes" sind und die Kenntnis der Algorithmen keine perversen Ziele verwirklicht), aber eine Erklärung der Logik der betreffenden Entscheidungen auf höherer Ebene und die vollständige Kenntnis der Features kann in anderen Fällen angemessener sein (z. B. wenn detailliertes Wissen zum Ausspielen des Systems verwendet werden würde).

6 LITERATUR

1. **Peck D. They're Watching You at Work. The Atlantic. 20. November 2013.** https://www.theatlantic.com/magazine/archive/2013/12/theyre-watching-you-at-work/354681/.

2. **Walker J. Meet the New Boss:** Big Data. Wall Street Journal. 20. September 2012. https://www.wsj.com/articles/SB10000872396390443890304578006252019616768.

3. **Sharp R. Virtual Career Assistants Can Solve Coaching Challenges. Hrmagazine. 13. April 2018.** http://www.hrmagazine.co.uk/article-details/virtual-career-assistants-can-solve-coaching-challenges.

4. **Pape T. Prioritising Data Items for Business Analytics:** Framework and Application to Human Resources. European Journal of Operational Research. 16. Juli 2016; 252(2): 687–698.

5. **Moore P, Robinson A. The Quantified Self:** What Counts in the Neoliberal Workplace. New Media & Society. 1. Dezember 2016; 18(11): 2774–2792.

6. **Davenport TH, Harris J, Shapiro J. Competing on Talent Analytics. Harvard Business Review. 2010; 88(10):** 52–58.

7. **Mittelstadt B, Allo P, Taddeo M, Wachter S, Floridi L. The Ethics of Algorithms:** Mapping the Debate. Big Data & Society. 1. November 2016; 3(2). https://journals.sagepub.com/doi/epub/10.1177/2053951716679679.

8. **Jobin A, Ienca M, Vayena E. Artificial Intelligence:** The Global Landscape of Ethics Guidelines. Nature Machine Intelligence. September 2019; 1(9): 389–399.

9. **Loi M, Heitz C, Ferrario A, Schmid A, Christen M. Towards an Ethical Code for Data-Based Business. In:** 2019 6th Swiss Conference on Data Science (SDS). Bern, Schweiz: IEEE; 2019: 6–12. https://ieeexplore.ieee.org/document/8789855/.

10. **Deutsche Telekom. AI Guidelines. 2018.** https://www.telekom.com/resource/blob/532446/f32ea4f5726ff3ed3902e97dd945fa14/dl-180710-ki-leitlinien-en-data.pdf. Deutsche Originalfassung: Deutsche Telekom. KI-Leitlinien. 2018. https://www.telekom.com/de/konzern/digitale-verantwortung/details/ki-leitlinien-der-telekom-523904.

11. **Women Leading in AI. 10 Principles of Responsible AI.** http://womenleadinginai.org/wp-content/uploads/2019/02/WLiAI-Report-2019.pdf.

12. **Fairness, Accountability, and Transparency in Machine Learning (FAT ML). Principles for Accountable Algorithms and a Social Impact Statement for Algorithms. 2016.** https://www.fatml.org/resources/principles-for-accountable-algorithms.

13. **Partnership on AI. Tenets. 2016.** https://www.partnershiponai.org/tenets/.

14. **Institute of Electrical and Electronics Engineers (IEEE), The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. Ethically Aligned Design. A Vision for Prioritizing Human Wellbeing with Autonomous and Intelligent Systems, Version 2. 2017.** https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead_v2.pdf.

15. **The Public Voice. Universal Guidelines for Artificial Intelligence. 2018.** https://epic.org/international/AIGuidleines DRAFT20180910.pdf.

16. **ICDPPC. Declaration on Ethics and Data Protection in Artificial Intelligence. 2018.** https://edps.europa.eu/sites/edp/files/publication/icdppc-40th_ai-declaration_adopted_en_0.pdf.
Deutsche Übersetzung: ICDPPC. Erklärung zu Ethik und Datenschutz im Bereich der künstlichen Intelligenz. 2018. https://datenschutz.sachsen-anhalt.de/fileadmin/Bibliothek/Landesaemter/LfD/PDF/binary/Konferenzen/Internationale_Konferenz/brussels2018_conference/2018_40.ICDPPC_Bruessel_EntschliessungEthik.pdf.

17. **Women 20 (W20). Artificial Intelligence:** Open Questions About Gender Inclusion. 2018. http://webfoundation.org/docs/2018/06/AI-Gender.pdf.

18. **Leaders of the G7. Charlevoix Common Vision for the Future of Artificial Intelligence. 2018.** https://www.mofa.go.jp/files/000373837.pdf.

19. **Access Now; Amnesty International. The Toronto Declaration:** Protecting the Right to Equality and Nondiscrimination in Machine Learning Systems. 2018. https://www.accessnow.org/cms/assets/uploads/2018/08/The-Toronto-Declaration_ENG_08-2018.pdf.

20. **Privacy International & Article 19. Privacy and Freedom of Expression In the Age of Artificial Intelligence. 2018.** https://www.article19.org/wp-content/uploads/2018/04/Privacy-and-Freedom-of-Expression-In-the-Age-of-Artificial-Intelligence-1.pdf.

21. **WEF, Global Future Council on Human Rights 2016–2018. White Paper:** How to Prevent Discriminatory Outcomes in Machine Learning. 2018. http://www3.weforum.org/docs/WEF_40065_White_Paper_How_to_Prevent_Discriminatory_Outcomes_in_Machine_Learning.pdf.

22. **Brundage M, Avin S, Clark J, Toner H, Eckersley P, Garfinkel B, et al. The Malicious Use of Artificial Intelligence:** Forecasting, Prevention, and Mitigation. arXiv:180207228 [cs]. 20. Feb. 2018. http://arxiv.org/abs/1802.07228.

23. **Villani C. For A Meaningful Artificial Intelligence:** Towards A French And European Strategy. März 2018. https://www.aiforhumanity.fr/pdfs/MissionVillani_Report_ENG-VF.pdf.

24. **UNI Global Union. Top 10 Principles for Ethical Artificial Intelligence. 2017.** http://www.thefutureworldofwork.org/media/35420/uni_ethical_ai.pdf.
Deutsche Übersetzung: UNI Global Union. Die 10 wichtigsten Grundsätze für ethische künstliche Intelligenz. http://www.thefutureworldofwork.org/media/35484/uni-global-union_-kuenstliche-intelligenz.pdf.

25. **Information Technology Industry Council (ITI). ITI AI Policy Principles. 2017.** https://www.itic.org/resources/AI-Policy-Principles-FullReport2.pdf.

26. **Software & Information Industry Association (SIIA), Public Policy Division. Ethical Principles for Artificial Intelligence and Data Analytics. 2017.** http://www.siia.net/Portals/0/pdf/Policy/Ethical%20Principles%20for%20Artificial%20Intelligence%20and%20Data%20Analytics%20SIIA%20Issue%20Brief.pdf?ver=2017-11-06-160346-990.

27. **COMEST/UNESCO. Report of COMEST on Robotics Ethics. 2017.** https://unesdoc.unesco.org/ark:/48223/pf0000253952.

28. **Internet Society. Artificial Intelligence and Machine Learning:** Policy Paper. 2017. https://www.internetsociety.org/wp-content/uploads/2017/08/ISOC-AI-Policy-Paper_2017-04-27_0.pdf.

29. **Independent High-Level Expert Group on Artificial Intelligence Set Up by the European Commission. Ethics Guidelines for Trustworthy AI. European Commission – Digital Single Market; 2019.** https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai. Deutsche Übersetzung: Unabhängige hochrangige Expertengruppe für künstliche Intelligenz eingesetzt von der Europäischen Kommission. Ethik-Leitlinien für eine vertrauenswürdige KI. Europäische Kommission. 2019. https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai.

30. **Jobin A, Ienca M, Vayena E. The Global Landscape of AI Ethics Guidelines. Nature Machine Intelligence. September 2019; 1(9):** 389–399.

31. **Schroeder M. Value Theory. In:** Zalta EN, Hrsg. The Stanford Encyclopedia of Philosophy. Herbst 2016. Metaphysics Research Lab, Stanford University; 2016. https://plato.stanford.edu/archives/fall2016/entries/value-theory/.

32. **Santoni de Sio F, Van den Hoven J. Meaningful Human Control over Autonomous Systems:** A Philosophical Account. Frontiers in Robotics and AI. 28. Februar 2018. https://www.frontiersin.org/articles/10.3389/frobt.2018.00015/full.

33. **Beauchamp TL, Childress JF. Principles of Biomedical Ethics. 6. Auflage. New York:** Oxford University Press; 2008.

34. **Wachter S, Mittelstadt BD. A Right to Reasonable Inferences:** Re-Thinking Data Protection Law in the Age of Big Data and AI. Columbia Business Law Review. 2018.

35. **FAT ML. Principles for Accountable Algorithms and a Social Impact Statement for Algorithms. Ohne Datum.** https://www.fatml.org/resources/principles-for-accountable-algorithms.

36. **Lippert-Rasmussen K. Nothing Personal:** On Statistical Discrimination. Journal of Political Philosophy. 1. Dezember 2007; 15(4): 385–403.

37. **Custers B, Calders T, Schermer B, Zarsky T, Hrsg. Discrimination and Privacy in the Information Society:** Data Mining and Profiling in Large Databases. Ausgabe 2013. New York: Springer; 2012.

38. **Barocas S, Selbst AD. Big Data's Disparate Impact. California Law Review. 2016; 104(671):**671–732.

39. **Zafar MB, Valera I, Rodriguez MG, Gummadi KP. Fairness Beyond Disparate Treatment & Disparate Impact:** Learning Classification Without Disparate Mistreatment. arXiv:161008452 [Stat.ML]. 8. März 2017; 1171–1180. https://arxiv.org/pdf/1610.08452.pdf.

40. **Lipton ZC. The Mythos of Model Interpretability. arXiv:**160603490 [cs.LG]. 10. Juni 2016. https://arxiv.org/pdf/1606.03490.pdf.

41. **Kroll JA. The Fallacy of Inscrutability. Philosophical Transcations of the Royal Society A. 15. Oktober 2018; 376 (2133):** 20180084. https://doi.org/10.1098/rsta.2018.0084.

42. **Kroll JA, Barocas S, Felten EW, Reidenberg JR, Robinson DG, Yu H. Accountable Algorithms. University of Pennsylvania Law Review. 2017; 165:** 633.

43. **Selbst AD, Barocas S. The Intuitive Appeal of Explainable Machines. Fordham Law Review. 2018; 87:** 1085.

44. **Loi M, Ferrario A, Viganò E. Transparency As Design Publicity:** Explaining and Justifying Inscrutable Algorithms. Rochester, NY: Social Science Research Network; Juni 2019. Bericht Nr.: ID 3404040. https://papers.ssrn.com/abstract=3404040.

45. **Ribeiro MT, Singh S, Guestrin C. „Why Should I Trust You?":** Explaining the Predictions of Any Classifier. arXiv:1602.04938 [cs.LG]. 16. Februar 2016. https://arxiv.org/pdf/1602.04938.pdf.

46. **Wachter S, Mittelstadt B, Russell C. Counterfactual Explanations Without Opening the Black Box:** Automated Decisions and the GDPR. 2017. Harvard Journal of Law & Technology. 2017; 31: 841.

47. **Pasquale F. The Black Box Society:** The Secret Algorithms That Control Money and Information. Harvard University Press; 2015.

48. **Lipton ZC, Chouldechova A, McAuley J. Does Mitigating ML's Impact Disparity Require Treatment Disparity? arXiv:**171107076 [stat.ML]. 19. November 2017. https://arxiv.org/pdf/1711.07076.pdf.

49. **Ferrario A, Loi M, Viganò E. In AI We Trust Incrementally:** A Multi-layer Model of Trust to Analyze Human-Artificial Intelligence Interactions. Philosophy & Technology. 23. Oktober 2019. https://doi.org/10.1007/s13347-019-00378-3.

50. **Ferrario A, Loi M, Viganò E. In AI We Trust Incrementally. A Multi-Layer Model of Trust to Analyze Human-Artificial Intelligence Interactions. Philosophy and Technology. Zur Veröffentlichung angenommen.**

51. **Colquitt JA, Zipay KP. Justice, Fairness, and Employee Reactions. Annual Review of Organizational Psychology and Organizational Behavior. 10. April 2015; 2(1):** 75–99.

52. **Colquitt JA, Wesson MJ, Porter COLH, Conlon DE, Ng KY. Justice at the Millennium:** A Meta-Analytic Review of 25 Years of Organizational Justice Research. Journal of Applied Psychology. Januar 2001; 86(3): 425–445.

53. **Colquitt JA, Scott BA, Rodell JB, Long DM, Zapata CP, Conlon DE, et al. Justice at the Millennium, a Decade Later:** A Meta-Analytic Test of Social Exchange and Affect-Based Perspectives. Journal of Applied Psychology. 2013; 98(2): 199–236.

54. **Thibaut JW, Walker L. Procedural Justice:** A Psychological Analysis. L. Erlbaum Associates; 1975.

55. **Binns R, Kleek MV, Veale M, Lyngs U, Zhao J, Shadbolt N. „It's Reducing a Human Being to a Percentage":** Perceptions of Justice in Algorithmic Decisions. SocArXiv. 31. Jan. 2018; https://osf.io/preprints/socarxiv/9wqxr/

56. **Hammarfelt B, de Rijcke S. Accountability in Context:** Effects of Research Evaluation Systems on Publication Practices, Disciplinary Norms, and Individual Working Routines in the Faculty of Arts at Uppsala University. Research Evaluation. 1. Jan. 2015; 24(1): 63–77.

57. **Sousa SB, Brennan JL. The UK Research Excellence Framework and the Transformation of Research Production. In:** Musselin C, Teixeira PN, Hrsg.. Reforming Higher Education: Public Policy Design and Implementation. Dordrecht: Springer Netherlands; 2014: 65–80. (Higher Education Dynamics). https://doi.org/10.1007/978-94-007-7028-7_4.

58. **Müller R, de Rijcke S. Thinking with Indicators. Exploring the Epistemic Impacts of Academic Performance Indicators in the Life Sciences. Research Evaluation. 1. Juli 2017; 26(3):** 157–168.

59. **Dalen HP van, Henkens K. Intended and Unintended Consequences of a Publish-Or-Perish Culture:** A Worldwide Survey. Journal of the American Society for Information Science and Technology. 2012; 63(7): 1282–1293.

60. **Baccini A, Nicolao GD, Petrovich E. Citation Gaming Induced by Bibliometric Evaluation:** A Country-Level Comparative Analysis. PLOS ONE. 11. Sept. 2019; 14(9):e0221212. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0221212.

61. **Azevedo AIRL, Santos MF. KDD, SEMMA and CRISP-DM:** A Parallel Overview. In: Weghorn H, Abraham AP, Hrsg. IADIS Proceedings of Informatics 2008 and Data Mining 2008. Amsterdam: IADIS; 2008: 182–185.

62. **Bogen M. All the Ways Hiring Algorithms Can Introduce Bias. – Google Search. Harvard Business Review Digital Articles:** 2–4.

63. **Bogen M, Rieke A. Help Wanted – An Exploration of Hiring Algorithms, Equity, and Bias. Upturn; 2018.**

64. **Loukina A, Madnani N, Zechner K. The Many Dimensions of Algorithmic Fairness in Educational Applications. In:** Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications. Florenz, Italien: Association for Computational Linguistics; 2019: 1–10. https://www.aclweb.org/anthology/W19-4401.pdf.

65. **Gilbert DE. Luck, Bayesian Tensor Completion, and Fairness [A Dissertation in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy]. Ithaca, Faculty of the Graduate School of Cornell University; 2019.**

66. **Berk R, Heidari H, Jabbari S, Kearns M, Roth A. Fairness in Criminal Justice Risk Assessments:** The State of the Art. Sociological Methods & Research. 2. Juli 2018; doi:10.1177/0049124118782533.

67. **Hardt M, Price E, Srebro N. Equality of Opportunity in Supervised Learning. arXiv:**161002413 [cs.LG]. 7. Okt. 2016. http://arxiv.org/abs/1610.02413.

68. **Chouldechova A. Fair Prediction with Disparate Impact:** A Study of Bias in Recidivism Prediction Instruments. arXiv:161007524 [stat.AP]. 14. Okt. 2016. http://arxiv.org/abs/1610.07524.

69. **Brennan T, Dieterich W, Ehret B. Evaluating the Predictive Validity of the COMPAS Risk and Needs Assessment System. Criminal Justice and Behavior. 2009; 36(1):** 21–40.

70. **Kleinberg J, Mullainathan S, Raghavan M. Inherent Trade-Offs in the Fair Determination of Risk Scores. arXiv:**160905807 [cs.LG]. 19. Sept. 2016. http://arxiv.org/abs/1609.05807.

71. **Angwin J, Larson J. Machine Bias. ProPublica. 2016** https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

72. **Heidari H, Loi M, Gummadi KP, Krause A. A Moral Framework for Understanding Fair ML Through Economic Models of Equality of Opportunity. In:** Proceedings of the Conference on Fairness, Accountability, and Transparency. New York: ACM; 2019: 181–190. http://doi.acm.org/10.1145/3287560.3287584.

73. **Binns R. On the Apparent Conflict Between Individual and Group Fairness. arXiv:**191206883 [cs.LG]. 14. Dez. 2019. https://arxiv.org/abs/1912.06883.

74. **Institute of Electrical and Electronics Engineers (IEEE), The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. Ethically Aligned Design. A Vision for Prioritizing Human Wellbeing with Autonomous and Intelligent Systems, Version 1. 2019.** https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead_v2.pdf.

75. **Jamieson KH. Cyberwar:** How Russian Hackers and Trolls Helped Elect a President: What We Don't, Can't, and Do Know. New York, NY: Oxford University Press; 2018.

76. **Beitz CR. The Idea of Human Rights. Oxford; New York:** Oxford University Press; 2009.

77. **Floridi L, Cowls JA. A Unified Framework of Five Principles for AI in Society. Harvard Data Science Review. 2. Juli 2019; 1(1).** https://hdsr.mitpress.mit.edu/pub/l0jsh9d1/release/6.

DANKSAGUNG

Ich möchte den folgenden Personen meinen tief empfundenen Dank aussprechen für den Beitrag, den sie zu diesem Projekt geleistet haben, insbesondere indem sie sich für die Interviews zur Verfügung stellten, die mir bei der Formulierung der Forschungsfragen geholfen haben:

Oliver Suchy, Leiter der Abteilung Digitale Arbeitswelten und Arbeitsweltberichterstattung des Deutschen Gewerkschaftsbundes (DGB); Isabelle Schömann, Bundessekretärin des Europäischen Gewerkschaftsbunds; Marco Bentivogli, Generalsekretär der FIM-CISL (Italienische Föderation der Metallarbeiter im Italienischen Arbeitergewerkschaftsbund CISL); Michele Carrus, Generalsekretär des Allgemeinen italienischen Gewerkschaftsbunds (CGIL) für die Region Sardinien; sowie Wolfgang Kowalsky, Senior Advisor beim Europäischen Gewerkschaftsbund.

Für Unzulänglichkeiten der vorliegenden Arbeit bin ich allein verantwortlich.

Der Einsatz von Künstlicher Intelligenz im Personalmanagement steckt noch in den Kinderschuhen. Gerade deshalb stellt sich die Frage, wie sie überhaupt zu besseren und faireren Personalentscheidungen beitragen kann. In seinem Gutachten entwickelt Michele Loi von der Universität Zürich Empfehlungen dafür, wie Systeme zum automatisierten Personalmanagement ethisch angemessen eingesetzt werden können. Entscheidend sind guter Datenschutz, fachliche Kompetenz, transparente Verfahren und eine sorgfältige Überprüfung der Auswirkungen automatisierter Entscheidungen.

WWW.BOECKLER.DE

ISBN 978-3-86593-366-1